Optimal Transport Reweighting for Robust Learning under Spurious Correlations and Label Noise

Sung Ho Jo **Seonghwi Kim** **Wonsang Yun** **Minwoo Chae***
Pohang University of Science and Technology (POSTECH)
{tjdgh1813, kshwi, wonsang, mchae}@postech.ac.kr

Abstract

Machine learning models often suffer performance degradation under subpopulation shift, particularly when spurious correlations cause models to rely on shortcut features that fail to generalize across subgroups. A recent line of work mitigates this issue by using loss-based signals to identify informative samples, but these signals can become severely distorted under label noise: mislabeled samples may also incur large losses and contaminate subsequent reweighting or retraining. Despite its practical importance, this intersection remains largely underexplored. We propose POTER, a reweighting framework based on optimal transport that derives sample importance from the transport geometry between the training distribution and a reference distribution constructed from limited validation group annotations. By measuring alignment at the individual-sample level rather than relying on loss, POTER downweights mislabeled or strongly bias-aligned samples while assigning higher importance to samples better aligned with the reference distribution. In addition, POTER requires only a single ERM training stage, moving beyond the retraining paradigm common in recent work. Across standard benchmarks and noisy-label settings, POTER achieves state-of-the-art worst-group accuracy, including cases where label corruption is concentrated within minority subgroups.

1 Introduction

Empirical risk minimization (ERM) assumes that training and test data are drawn from the same distribution, and has achieved strong average performance when this assumption holds. In practice, however, models are often deployed under distribution shifts. A particularly challenging instance is subpopulation shift, in which the proportions of subpopulations differ between training and test distributions. This becomes especially problematic when certain attributes are spuriously correlated with the target label in the training data, causing models to rely on shortcut features that fail to generalize across subpopulations [35]. For example, in the Waterbirds dataset, bird labels are strongly correlated with backgrounds, leading models to exploit background cues rather than bird-specific features. Performance consequently deteriorates on minority subgroups where such correlations no longer hold—a vulnerability with serious implications in high-stakes domains such as healthcare and fairness-sensitive decision-making [30, 4, 13].

To mitigate spurious correlations, many existing methods rely on group annotations to identify underperforming subpopulations and promote subgroup robustness during training [35, 20, 14]. In realistic settings, however, group annotations for the full training set are often unavailable, with such information available only for a limited validation set. To address this limitation, a prominent line of work relies on sample-wise proxy signals—such as per-sample loss or misclassification indicators—from a preliminary model to identify minority or bias-conflicting samples, which are then upweighted or used in last-layer retraining pipelines [26, 34, 37, 24]. Bias-conflicting samples refer

*Corresponding author.

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

to those that do not conform to the dominant spurious correlation in the training data (e.g., waterbirds on land backgrounds in the Waterbirds dataset). The central weakness of this line of work is that it hinges on high-loss samples being informative. That premise is fragile even in clean settings, since loss-based signals depend on the learned model and optimization trajectory, making the selected samples inconsistent across training procedures.

Under label noise, this fragility becomes especially severe. Mislabeled samples also tend to incur large training losses, so high-loss-based criteria systematically select them along with genuinely informative bias-conflicting examples. As a result, the selected subset becomes contaminated, and the subsequent reweighting or retraining step can no longer reliably emphasize bias-conflicting samples. Despite its practical importance, the intersection of spurious correlations and label noise remains largely overlooked. One recent approach is [36], which proposes label propagation prior to last-layer retraining as a remedy and attains strong subpopulation robustness under symmetric label noise, where each training label is independently flipped to the other class with the same probability. However, this robustness is established when the model used in the subsequent reweighting or retraining stage is trained on clean data, and the gains become substantially smaller once label noise affects the full training pipeline. Moreover, it degrades further when label noise is concentrated within minority subgroups—a setting we term *subgroup-concentrated label noise*—exposing a previously overlooked failure mode.

To address the fragility of proxy-based sample selection, we introduce POTER (Potentials from Optimal Transport for Empirical Reweighting), an OT-based framework that constructs a reweighted empirical training distribution using only limited group annotations from a validation set. Rather than identifying informative samples using a preliminary model and then upweighting or retraining on the selected subset, our method derives sample importance directly from the OT geometry between the training distribution and a constructed reference distribution. This removes optimization-dependent sample selection entirely and does not rely on a clean-data-trained preliminary model. Instead, sample importance is determined by alignment with the reference distribution, so examples that are poorly aligned with the target distribution—including those affected by label noise—are assigned lower importance. To our knowledge, this is the first OT-based reweighting framework for spurious correlation mitigation. More fundamentally, the proposed framework moves beyond the retraining-based paradigm that has largely dominated this field: unlike existing approaches that rely on full-model or last-layer retraining, it requires only a single standard ERM training stage, with OT dual potentials computed within seconds even on the largest benchmark we consider. Despite this simplicity, POTER achieves state-of-the-art performance on standard spurious correlation benchmarks and substantially outperforms prior methods under both symmetric and subgroup-concentrated label noise, without sacrificing clean-setting performance.

Our main contributions are as follows:

- We propose POTER, a novel OT-based reweighting framework for spurious correlation mitigation, deriving sample importance from OT dual potentials using only limited validation-set group annotations, without proxy-based sample selection or preliminary downstream-task training.
- We advance beyond recently established last-layer retraining pipelines, requiring only a single standard ERM training stage with OT computation adding negligible overhead.
- We demonstrate state-of-the-art worst-group accuracy on standard benchmarks and substantially outperform existing methods under symmetric label noise across a wide range of noise levels.
- We identify subgroup-concentrated label noise as a previously overlooked failure mode, and show that our approach remains robust in this regime, where even a method explicitly designed to address label noise under spurious correlations degrades substantially.

2 Related Work

Spurious Correlation Mitigation A large body of work has studied robust learning under spurious correlations. Early approaches typically assume access to group annotations for the full training set and use them to guide robust learning [1, 35, 22, 40, 8]. To reduce reliance on full training-set group labels, later methods use limited supervision to infer pseudo groups or spurious attributes and combine them with existing robust learning methods like Group DRO, but this requires an

additional group-inference stage and incurs extra computational cost [29, 10, 12]. More recently, retraining-based pipelines—especially those based on last-layer retraining—have become a dominant design choice, using proxy signals from a preliminary model to select or reweight samples before retraining or adaptation [26, 34, 37, 24, 9, 11]. Although effective, these methods still depend on additional retraining and on the quality of the preliminary model used to generate such signals. Our work pushes this progression one step further by directly constructing a reweighted empirical training distribution, without requiring any retraining stage. Beyond these established directions, recent work has also begun to examine new failure modes within the spurious correlation framework, including settings with multiple spurious features [32, 18], weak spurious signals [28], distributional shifts within minority subgroups [14], and settings where some subgroups are entirely absent from the training distribution [19]. This broader trend underscores the need to study additional sources of failure beyond standard benchmark assumptions.

Spurious Correlation under Label Noise While recent work has increasingly examined diverse failure modes within the spurious correlation framework, label noise—a classical challenge in machine learning—has received comparatively limited attention. More broadly, prior studies [31, 33, 36] suggest that many existing spurious correlation methods—especially those relying on loss- or error-based signals—can suffer severe performance degradation under label noise, regardless of whether group information is available. Among the few works that directly address this intersection, END [31] uses predictive uncertainty to identify and oversample samples that are likely to be free of spurious cues, but requires two rounds of full-model training and tends to yield weaker clean-setting performance. More recently, a kNN-based label-correction method [36] improves last-layer retraining under symmetric label noise. However, its reliance on local neighborhood structure makes it sensitive to subgroup-concentrated label noise. We show that even in noise regimes where this method appears to perform well, worst-group accuracy can deteriorate substantially when label noise is concentrated within minority subgroups, even if the overall noise level remains moderate.

Optimal Transport and Data Valuation Optimal transport provides a principled framework for comparing and aligning probability distributions, and has found broad application in machine learning, particularly in robust learning under distribution shift [23, 5, 3]. More recently, OT dual potentials have been leveraged as theoretically grounded signals for data valuation [15], with subsequent work extending this framework to federated learning settings [25] and scaling it to larger datasets [17]. Our work builds on this perspective, using OT dual potentials not only for data valuation, but also to derive sample importance weights for model learning. To our knowledge, this is the first work to introduce OT-based reweighting for spurious correlation mitigation, enabling us to construct a reweighted empirical training distribution using only limited group annotations from a validation set—without proxy-based sample selection or any retraining stage.

3 Preliminaries

Problem Setup. We consider a supervised classification problem with input $X \in \mathcal{X}$, clean label $Y \in \mathcal{Y}$, and subgroup (or subpopulation) index $G \in \mathcal{G}$, where $K := |\mathcal{Y}|$ denotes the number of classes. Let P_{train} and P_{test} denote the training and test distributions over (X, Y, G), respectively. We assume a subpopulation shift where the marginal distributions of G under P_{train} and P_{test} differ, while the conditional distribution of (X, Y) given G remains invariant. Under subgroup-dependent label noise, we observe a corrupted label $\tilde{Y}$, instead of Y, governed by the conditional probabilities

$$P_{\mathrm{train}}(\tilde{Y} = \tilde{y} \mid Y = y,\, G = g) = (1 - \eta_g)\mathbf{1}[\tilde{y} = y] + \frac{\eta_g}{K-1}\mathbf{1}[\tilde{y} \neq y], \qquad g \in \mathcal{G},$$

where $\eta_g \in [0, 1)$ denotes the noise rate for subgroup g. Symmetric label noise corresponds to the special case $\eta_g = \eta$ for all $g \in \mathcal{G}$. We define subgroup-concentrated label noise as the scenario in which corruption is restricted to a specific subset $S \subseteq \mathcal{G}$ of subgroups. In our experiments, we specifically consider the case where $\eta_g = \eta$ for $g \in S$ and $\eta_g = 0$ otherwise.

We are given a training set $D_t = \{(x_i, \tilde{y}_i)\}_{i=1}^n$ consisting of independent samples from the marginal distribution of $(X, \tilde{Y})$. Crucially, the subgroup indices G are latent and remain unobserved throughout training. We are also given a limited validation set $D_v = \{(x'_j, y'_j, g'_j)\}_{j=1}^m$ sampled from P_{test}, with clean labels and subgroup annotations.

Our objective is to learn a classifier that remains robust under subpopulation shift even when the training labels are noisy. We evaluate performance on the test distribution using worst-group accuracy, i.e., the minimum accuracy over all subgroups, which effectively captures robustness to subgroup-proportion shifts, including extreme cases where test mass concentrates on a single subgroup. For the distributional formulation used later, we represent the observed training data and validation data by their empirical distributions, $\mu_t = n^{-1} \sum_{i=1}^{n} \delta_{(x_i, \tilde{y}_i)}$ and $\mu_v = m^{-1} \sum_{j=1}^{m} \delta_{(x'_j, y'_j)}$, respectively, where δ_z denotes the Dirac measure at z.

Optimal Transport. Let $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. Given empirical distributions μ_t and μ_v and a cost function $c : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}_+$, we define the optimal transport value between μ_t and μ_v as the optimal value of the Kantorovich problem [16]:

$$\mathcal{T}(\mu_t, \mu_v) := \inf_{\pi \in \Pi(\mu_t, \mu_v)} \int_{\mathcal{Z} \times \mathcal{Z}} c(z, z') \, d\pi(z, z'). \tag{1}$$

Here, $\Pi(\mu_t, \mu_v)$ denotes the set of couplings with marginals μ_t and μ_v. To enable efficient computation, we consider the entropy-regularized optimal transport problem and denote its optimal value by

$$\mathcal{T}_\varepsilon(\mu_t, \mu_v) = \inf_{\pi \in \Pi(\mu_t, \mu_v)} \int_{\mathcal{Z} \times \mathcal{Z}} c(z, z') \, d\pi(z, z') + \varepsilon \, \mathrm{KL}(\pi \,\|\, \mu_t \otimes \mu_v) \,,$$

where $\varepsilon > 0$ is a regularization parameter. The corresponding optimization problem can be solved efficiently by the iterative Sinkhorn algorithm [7].

Moreover, under mild assumptions, Kantorovich duality implies that the optimal value of the primal OT problem in (1) admits the following dual representation [38, Theorem 5.10]:

$$\mathcal{T}(\mu_t, \mu_v) = \sup_{\substack{f \in L^1(\mu_t),\, g \in L^1(\mu_v) \\ f(z) + g(z') \le c(z, z') \;\forall z, z' \in \mathcal{Z}}} \left\{ \int_{\mathcal{Z}} f(z) \, d\mu_t(z) + \int_{\mathcal{Z}} g(z) \, d\mu_v(z) \right\}.$$

Here, f and g are dual potentials, and $L^1(\mu)$ denotes the space of functions integrable with respect to a measure μ. For the empirical measures μ_t and μ_v considered here, with Euclidean cost, the above dual representation holds and the dual optimum is attained. Since both μ_t and μ_v are probability measures, the dual potentials are unique only up to an additive constant: if (f, g) is optimal, then $(f + a, g - a)$ is also optimal for any $a \in \mathbb{R}$. This non-uniqueness up to an additive constant does not affect the resulting sample importance weights, since the reweighting step is invariant to adding a constant to f. For notational convenience, we fix one optimal dual pair and denote it by (f^*, g^*).

Dual Potentials as Sensitivity Signals. The optimal dual potentials admit a sensitivity interpretation with respect to perturbations of the probability mass. Specifically, for empirical measures with supports $\{z_i\}_{i=1}^n$ and $\{z'_j\}_{j=1}^m$, we write

$$f_i^* = f^*(z_i) \quad \text{and} \quad g_j^* = g^*(z'_j)$$

for the values of the optimal dual potentials on the support points of μ_t and μ_v, respectively. For empirical measures with fixed support, the Kantorovich dual objective is linear in the support masses for a fixed dual pair, and the optimal dual potentials appear as the corresponding coefficients. In particular, a positive value of f_i^* corresponds to an increase in the transport value $\mathcal{T}(\mu_t, \mu_v)$ when the mass at z_i is increased, whereas a negative value corresponds to a decrease in the transport value for the same perturbation. Accordingly, f_i^* can be interpreted as a sample-level sensitivity signal that captures how z_i contributes to the discrepancy between μ_t and μ_v. Prior work has leveraged this interpretation for training data valuation [15, 25, 17]; we show that it can be developed into a principled reweighting framework for spurious correlation mitigation under label noise.

4 Proposed Method

We propose POTER, which constructs a reweighted empirical training distribution by deriving sample importance weights from OT dual potentials. The key insight is that, by solving OT between the empirical training distribution and an explicitly constructed reference distribution, the training-side

dual potential f_i^* provides a principled measure of how well each training sample aligns with the desired group composition encoded by the reference distribution. We derive sample importance from these dual potentials through three steps: (i) constructing a reference distribution from limited group supervision, (ii) solving OT in a class-conditioned manner to obtain dual potentials that are not confounded by spurious attribute similarity, and (iii) converting the resulting sensitivity signals into sample weights suitable for robust reweighted ERM training. The following subsections describe each step in turn.

4.1 Reference Distribution Construction

We define the reference distribution μ_{ref} as a mixture of group-conditional empirical validation distributions with specified mixture weights ρ. For each group $g \in \mathcal{G}$, let

$$D_v^g := \{(x'_j, y'_j) \in D_v : g'_j = g\}$$

denote the corresponding validation subset. Given mixture weights $\rho = (\rho_g)_{g\in\mathcal{G}}$ with $\rho_g \geq 0$ and $\sum_{g\in\mathcal{G}} \rho_g = 1$, we define

$$\mu_{\mathrm{ref}} := \sum_{g\in\mathcal{G}} \rho_g \left(\frac{1}{|D_v^g|} \sum_{(x'_j, y'_j)\in D_v^g} \delta_{(x'_j, y'_j)} \right).$$

Since bias-conflicting minority groups are typically underrepresented in the training distribution but crucial for achieving worst-group robustness, ρ assigns larger mass to these groups and smaller, potentially zero, mass to bias-aligned majority groups.

4.2 Class-Conditioned Optimal Transport

Dual Potentials under the Reference Distribution. Consider the optimal dual pair (f^*, g^*) associated with the OT problem between μ_t and μ_{ref}. As discussed above, the quantity f_i^* can be viewed as a sample-level sensitivity of the OT value with respect to perturbations of the mass at the training sample z_i. Samples that are overrepresented or distributionally misaligned with μ_{ref} are expected to induce larger values of f_i^*, since increasing their mass makes the training distribution harder to transport to the reference distribution. In contrast, samples that align well with μ_{ref} are expected to have smaller values of f_i^*. Because μ_{ref} is constructed to encode the target group composition for worst-group generalization, we use f_i^* as a sample-importance signal, with smaller dual potentials indicating closer alignment with the desired distribution.

Class-Conditioned Cost. However, when an attribute is spuriously correlated with the class label, bias-aligned majority samples from one class may lie close in feature space to bias-conflicting minority reference samples from another class that share the same attribute. For example, in the Waterbirds dataset, landbirds on water backgrounds and waterbirds on water backgrounds share the same background, and may therefore be close in feature space despite belonging to different classes. Solving OT jointly over all samples can therefore induce cross-class transport between such samples, causing majority samples that would otherwise be expected to receive large dual potentials to instead be matched to minority reference samples and assigned inappropriately small values. To prevent this effect, we solve OT in a class-conditioned manner using the cost

$$C_{ij} = \begin{cases} \|\phi(x_i) - \phi(x'_j)\|_2^2, & \tilde{y}_i = y'_j, \\ +\infty, & \tilde{y}_i \neq y'_j, \end{cases}$$

where $\phi(\cdot)$ denotes the (pretrained) feature embedding used to define the transport cost. The resulting OT problem admits an exact classwise decomposition, which we leverage in practice by solving OT separately within each class and concatenating the resulting dual potentials. The corresponding formal statement and proof are deferred to Appendix A. Consequently, the resulting dual potentials $\{f_i^*\}_{i=1}^n$ provide a sample-importance signal that reflects alignment with the class-conditional reference distribution and is not confounded by spurious cross-class similarity arising from shared spurious attributes.

4.3 Sample Reweighting from Dual Potentials

Having obtained the class-conditioned dual potentials $\{f_i^*\}_{i=1}^n$, we now convert them into sample importance weights to construct a reweighted empirical training distribution for robust ERM training.

To this end, let

$$\mathcal{P}_n := \left\{ \mu = \sum_{i=1}^n w_i \, \delta_{z_i} \ : \ w_i \geq 0, \ \sum_{i=1}^n w_i = 1 \right\},$$

denote the set of probability measures supported on the training samples $\{z_i\}_{i=1}^n$, and let $\mu_0 = \sum_{i=1}^n w_i^0 \, \delta_{z_i}$ be a base distribution on the same support. For $\tau > 0$, we consider the KL-regularized optimization problem

$$\mu^* \in \operatorname*{argmin}_{\mu \in \mathcal{P}_n} \left\{ \sum_{i=1}^n f_i^* \, w_i + \tau \, \mathrm{KL}(\mu \| \mu_0) \right\}. \tag{2}$$

Here, smaller values of f_i^* indicate training samples that are better aligned with the reference distribution μ_{ref}. Accordingly, minimizing the linear term in (2) assigns larger mass to such samples, thereby reversing the ordering of the dual-potential scores into an importance weighting for reweighted ERM. The KL regularization prevents the degenerate solution that would otherwise concentrate all mass on the lowest-score samples, while also yielding a more general framework that allows prior information on the training distribution to be incorporated through the base measure μ_0.

Proposition 4.1 (KL-regularized reweighting)**.** *Assume that $w_i^0 > 0$ for all $i \in [n]$. Then the optimization problem in* (2) *admits the unique solution*

$$w_i^* = \frac{w_i^0 \exp(-f_i^*/\tau)}{\sum_{j=1}^n w_j^0 \exp(-f_j^*/\tau)}, \qquad i \in [n].$$

Proposition 4.1 yields a unique closed-form solution, giving the exponential reweighting rule. The parameter τ controls the sharpness of the reweighting: smaller values place more mass on samples with lower dual-potential scores, whereas larger values produce a flatter weighting closer to the base distribution μ_0. When additional prior information on the training distribution is available—such as known group information—it can be incorporated through the choice of μ_0. In our setting, however, we assume no such prior information and therefore take $\mu_0 = \mu_t$, the empirical training distribution. For the standard empirical training distribution, where $w_i^0 = 1/n$, this reduces to

$$w_i^* = \frac{\exp(-f_i^*/\tau)}{\sum_{j=1}^n \exp(-f_j^*/\tau)}, \qquad i \in [n].$$

Finally, letting h_θ denote the model parameterized by θ and ℓ a standard classification loss, we train the model by minimizing the reweighted empirical risk

$$\inf_{\theta \in \Theta} \sum_{i=1}^n w_i^* \, \ell\big(h_\theta(x_i), \tilde{y}_i\big).$$

The proof of Proposition 4.1 is deferred to Appendix B.

5 Experiments

5.1 Datasets

We conduct experiments on four widely used benchmarks for spurious correlation: CMNIST, Waterbirds, CelebA, and CivilComments.

CMNIST [1] is a colored variant of MNIST in which the digit label is the target and the color serves as the spurious attribute. Following prior work, digits 0–4 are assigned to label 0 and digits 5–9 to label 1, yielding four groups defined by label and color.

Waterbirds [35] is constructed from CUB bird images [39] and Places backgrounds [42], with groups defined by bird type and background. The target label is bird type, while background is spuriously

Table 1: Worst-group and average accuracy on CMNIST, Waterbirds, CelebA, and CivilComments under the standard evaluation setting. *Group Labels* indicates whether the method uses subgroup annotations from the training data, the validation data, or both. *No Extra Training* indicates whether the method operates without additional retraining. Among methods that use only validation group labels, boldface indicates the best performance and underlining denotes the second-best.

Algorithm	Group Labels (Tr/Val)	No Extra Training	CMNIST		Waterbirds		CelebA		CivilComments	
			Worst Acc	Avg Acc	Worst Acc	Avg Acc	Worst Acc	Avg Acc	Worst Acc	Avg Acc
Group DRO	Tr/Val	✓	$73.1_{\pm 0.3}$	$74.8_{\pm 0.2}$	$90.6_{\pm 0.2}$	$92.7_{\pm 0.1}$	$89.3_{\pm 1.3}$	$92.6_{\pm 0.3}$	$69.0_{\pm 0.9}$	$89.9_{\pm 0.2}$
LISA	Tr/Val	✓	$73.3_{\pm 0.2}$	$74.0_{\pm 0.1}$	$89.2_{\pm 0.6}$	$91.8_{\pm 0.3}$	$89.3_{\pm 1.1}$	$92.4_{\pm 0.4}$	$72.6_{\pm 0.1}$	$89.2_{\pm 0.9}$
DFR^{Tr}	Tr/Val	-	$59.8_{\pm 0.4}$	$62.1_{\pm 0.2}$	$90.2_{\pm 0.8}$	$97.0_{\pm 0.3}$	$80.7_{\pm 2.4}$	$90.6_{\pm 0.7}$	$58.0_{\pm 1.3}$	$92.0_{\pm 0.1}$
PDE	Tr/Val	✓	$72.6_{\pm 0.7}$	$73.0_{\pm 0.4}$	$90.3_{\pm 0.3}$	$92.4_{\pm 0.8}$	$91.0_{\pm 0.4}$	$92.0_{\pm 0.6}$	$71.5_{\pm 0.5}$	$86.3_{\pm 1.7}$
ERM	-	✓	$3.4_{\pm 0.9}$	$12.9_{\pm 0.8}$	$62.6_{\pm 0.3}$	$97.3_{\pm 1.0}$	$47.7_{\pm 2.1}$	$94.9_{\pm 0.3}$	$58.6_{\pm 1.7}$	$92.1_{\pm 0.4}$
JTT	Val	-	$67.3_{\pm 5.1}$	$76.4_{\pm 3.3}$	$83.8_{\pm 1.2}$	$89.3_{\pm 0.7}$	$81.5_{\pm 1.7}$	$88.1_{\pm 0.3}$	$69.3_{\text{N/A}}$	$91.1_{\text{N/A}}$
CnC	Val	-	–	–	$88.5_{\pm 0.3}$	$90.9_{\pm 0.1}$	$88.8_{\pm 0.9}$	$89.9_{\pm 0.5}$	$68.9_{\pm 2.1}$	$81.7_{\pm 0.5}$
SSA	Val	-	$71.1_{\pm 0.4}$	$75.0_{\pm 0.3}$	$89.0_{\pm 0.6}$	$92.2_{\pm 0.9}$	$\underline{89.8}_{\pm 1.3}$	$92.8_{\pm 0.1}$	$69.9_{\pm 2.0}$	$88.2_{\pm 2.0}$
AFR	Val	-	$53.5_{\pm 0.0}$	$60.6_{\pm 0.0}$	$90.4_{\pm 1.1}$	$94.2_{\pm 1.2}$	$82.0_{\pm 0.5}$	$91.3_{\pm 0.3}$	$68.7_{\pm 0.6}$	$89.8_{\pm 0.6}$
GIC	Val	-	$\underline{72.2}_{\pm 0.5}$	$73.2_{\pm 0.2}$	$86.3_{\pm 0.1}$	$89.6_{\pm 1.3}$	$89.4_{\pm 0.2}$	$91.9_{\pm 0.1}$	$\underline{72.5}_{\pm 0.3}$	$90.0_{\pm 0.3}$
LEIA	Val	-	$56.0_{\pm 4.8}$	$63.2_{\pm 2.0}$	$\underline{90.7}_{\pm 0.2}$	$93.3_{\pm 0.7}$	$85.0_{\pm 0.9}$	$95.2_{\pm 0.1}$	$\mathbf{72.9}_{\pm 0.2}$	$90.9_{\pm 0.8}$
POTER (Ours)	Val	✓	$\mathbf{72.5}_{\pm 0.7}$	$73.6_{\pm 0.4}$	$\mathbf{90.9}_{\pm 0.3}$	$92.7_{\pm 0.1}$	$\mathbf{90.4}_{\pm 0.5}$	$91.2_{\pm 0.3}$	$71.0_{\pm 0.6}$	$89.9_{\pm 0.3}$

correlated with the label in the training set. The minority groups are *landbird on water background* and *waterbird on land background*.

CelebA [27] is used for blond-hair classification, where gender serves as the spurious attribute. Groups are defined by the combination of hair color and gender, and the minority group is *blond male*.

CivilComments [6] is a text toxicity classification benchmark. We use the WILDS version of the dataset [21], in which identity-related text can act as spurious attributes. Following prior work [26, 41, 12, 11], we consider eight identity attributes: *male*, *female*, *LGBTQ*, *Christian*, *Muslim*, *other religion*, *black*, and *white*, defining 16 subgroups.

Further details on dataset construction and experimental settings are provided in Appendix C.

5.2 Experimental Settings

We consider three experimental settings: the standard benchmark setting, the symmetric label noise setting, and a subgroup-concentrated label noise setting that we introduce.

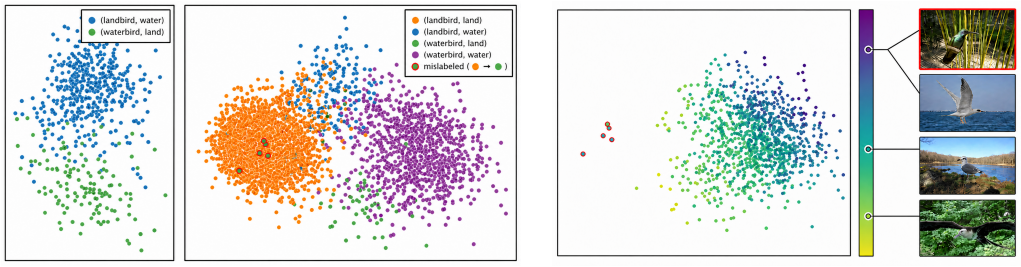
Standard Benchmark Setting. The standard benchmark setting evaluates worst-group robustness under spurious correlations using the original benchmarks without additional label corruption. We note that label noise is already present in standard benchmarks: CMNIST contains 25% label noise by construction, and the Waterbirds benchmark contains mislabeled examples [2], suggesting that label noise is a naturally occurring phenomenon in spurious correlation benchmarks.

Symmetric Label Noise. Following prior work [36], we inject symmetric label noise into the training set by independently flipping each training label with probability η. We vary η across a range of noise levels to evaluate robustness under different degrees of label corruption.

Subgroup-Concentrated Label Noise. We introduce a noise setting in which corruption is concentrated within minority subgroups at rates up to 60%, leaving the remaining groups uncorrupted. While symmetric label noise corrupts all groups uniformly, this setting probes whether methods remain robust to locally concentrated label noise—a distinct and previously overlooked challenge.

5.3 Baselines

We compare our method against representative baselines under two settings: one in which training group labels are available and one in which only validation group labels are available. For methods with training group labels, we consider Group DRO [35], LISA [40], DFR [20], and PDE [8]. For methods with only validation group labels, we consider JTT [26], CnC [41], SSA [29], AFR [34], GIC



(a) Reference set (left) and noisy training set (right).

(b) f^* values for training samples labeled "waterbird".

Figure 1: Analysis of reference-based sample importance on Waterbirds. (a) PCA projection of feature embeddings for the reference set and noisy training set. The reference set consists of validation samples from the two bias-conflicting minority groups, while the noisy training set contains all four groups with mislabeled samples marked by red outlines. (b) Training samples with observed label "waterbird" are colored by the class-conditioned OT dual potential f^* computed against the waterbird-on-land reference samples. Brighter colors indicate smaller f^* values and therefore larger sample importance. The visual examples, from top to bottom, show a mislabeled sample, a strongly bias-aligned sample, a bias-aligned sample containing both land and water background regions, and a bias-conflicting sample.

[12], and LEIA [11]. In the label-noise setting, we additionally focus on Group DRO, a representative loss-based robust optimization method; AFR, which performs loss-based sample upweighting for last-layer retraining; and KNN-RAD [36], the most closely related method that directly addresses label noise under spurious correlations. Detailed descriptions are provided in Appendix D.

5.4 Results

5.4.1 Standard Spurious Correlation Benchmarks

Table 1 reports results on standard spurious correlation benchmarks. Among methods that use only validation group labels, POTER achieves superior worst-group accuracy on CMNIST, Waterbirds, and CelebA, and remains competitive on CivilComments, without relying on additional training stages. Notably, this performance matches or exceeds several methods with access to full training group labels.

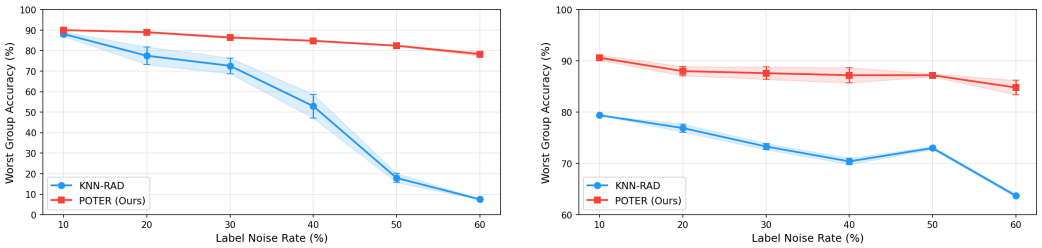
Figure 1 offers insight into the mechanism underlying POTER's strong performance. As shown in Figure 1(b), bias-conflicting samples receive high importance, whereas strongly bias-aligned samples receive low importance. Beyond this group-level distinction, samples near the boundary between land and water backgrounds retain relatively high importance even within the bias-aligned group. Unlike group-level reweighting approaches such as Group DRO, POTER assigns fine-grained weights according to alignment with the reference distribution, allowing informative samples to be emphasized regardless of their group membership. This sample-level flexibility is consistent with the strong worst-group accuracy observed in Table 1.

5.4.2 Robustness under Label Noise

Interpretation of Robustness under Label Noise. The label-noise robustness of POTER follows from using class-conditioned reference alignment rather than loss-based sample selection. For samples with observed label "waterbird", POTER compares them to the waterbird-on-land reference distribution. As illustrated by the red-outlined mislabeled samples in Figure 1, a landbird-on-land sample mislabeled as "waterbird" is evaluated within the observed waterbird class, but its feature embedding remains far from the waterbird-on-land reference. It therefore receives a large f^* value and hence low importance, as do strongly bias-aligned samples. Thus, the alignment-based weighting rule effectively downweights both mislabeled and strongly bias-aligned samples through a single unified mechanism.

Table 2: Worst-group accuracy on Waterbirds and CelebA under different symmetric label noise rates.

Dataset	Method	Group Labels (Tr/Val)	No Extra Training	Label Noise (%)			
				0	10	20	30
Waterbirds	Group DRO	Tr/Val	✓	$90.6_{\pm 0.2}$	$72.9_{\pm 1.6}$	$54.3_{\pm 1.0}$	$52.2_{\pm 3.6}$
	AFR	Val	-	$88.3_{\pm 0.9}$	$58.7_{\pm 0.4}$	$61.2_{\pm 5.9}$	$52.9_{\pm 0.0}$
	KNN-RAD	Val	-	$\mathbf{91.0}_{\pm 0.1}$	$82.4_{\pm 0.7}$	$74.7_{\pm 1.1}$	$68.9_{\pm 2.1}$
	POTER (Ours)	Val	✓	$90.9_{\pm 0.3}$	$\mathbf{89.4}_{\pm 0.4}$	$\mathbf{87.9}_{\pm 1.6}$	$\mathbf{85.3}_{\pm 1.4}$
CelebA	Group DRO	Tr/Val	✓	$89.3_{\pm 1.3}$	$66.9_{\pm 0.3}$	$59.5_{\pm 2.6}$	$54.8_{\pm 1.8}$
	AFR	Val	-	$80.9_{\pm 2.3}$	$69.4_{\pm 4.8}$	$67.7_{\pm 2.4}$	$58.8_{\pm 6.3}$
	KNN-RAD	Val	-	$83.9_{\pm 0.0}$	$80.0_{\pm 0.9}$	$78.1_{\pm 1.3}$	$70.2_{\pm 2.6}$
	POTER (Ours)	Val	✓	$\mathbf{90.4}_{\pm 0.5}$	$\mathbf{87.6}_{\pm 0.3}$	$\mathbf{82.8}_{\pm 1.0}$	$\mathbf{78.2}_{\pm 0.4}$



(a) Waterbirds. (b) CelebA.

Figure 2: Worst-group accuracy under subgroup-concentrated label noise on Waterbirds and CelebA. Label noise is concentrated on the minority subgroups: (waterbird, land background) and (landbird, water background) for Waterbirds, and (blond, male) for CelebA. The y-axis range is adjusted for each dataset to highlight within-dataset performance differences.

Symmetric Label Noise. Table 2 reports worst-group accuracy on Waterbirds and CelebA under increasing symmetric label noise. Loss-based methods degrade substantially as noise increases: on Waterbirds, Group DRO drops from 90.6% to 52.2% and AFR from 88.3% to 52.9% at 30% noise, consistent with the fragility of loss-based signals under label corruption. Since methods such as SSA and GIC incorporate Group DRO as part of their training pipeline, the degradation of Group DRO is indicative of a vulnerability shared by this broader class of approaches. KNN-RAD, which explicitly addresses label noise, is more stable than loss-based baselines but is consistently outperformed by POTER once label noise is introduced. At 30% noise, POTER achieves 85.3% on Waterbirds and 78.2% on CelebA, outperforming KNN-RAD by 16.4 and 8.0 percentage points, respectively.

Subgroup-Concentrated Label Noise. Figure 2 reports worst-group accuracy under subgroup-concentrated label noise, where corruption is restricted to minority subgroups. KNN-RAD degrades sharply over the evaluated range of subgroup noise rates, falling from 88.1% to 7.5% on Waterbirds and from 79.4% to 63.7% on CelebA between 10% and 60% subgroup noise. This suggests that local-neighborhood-based label correction can become unreliable when label corruption is concentrated within the subgroups most critical for worst-group generalization. In contrast, POTER remains stable across the same range, achieving 78.3% on Waterbirds and 84.8% on CelebA at 60% noise, demonstrating robustness to locally concentrated label corruption.

5.4.3 Analysis of the Weighting Signal

We examine the weighting signal before reweighted ERM. Figure 3 shows the group-wise distributions of the class-conditioned OT dual potentials f^* on Waterbirds and CelebA, and Table 3 reports the resulting mean sample weight per group on Waterbirds, CelebA, and CMNIST, normalized so that uniform weighting equals 1 (group definitions follow Appendix C).

Two properties stand out. First, within each class, the bias-conflicting groups receive lower f^* than their bias-aligned counterparts, and hence larger weights: 6.32 and 10.19 for the two Waterbirds minority groups, and 2.22 and 1.91 for the two bias-conflicting CMNIST groups. In CelebA, the

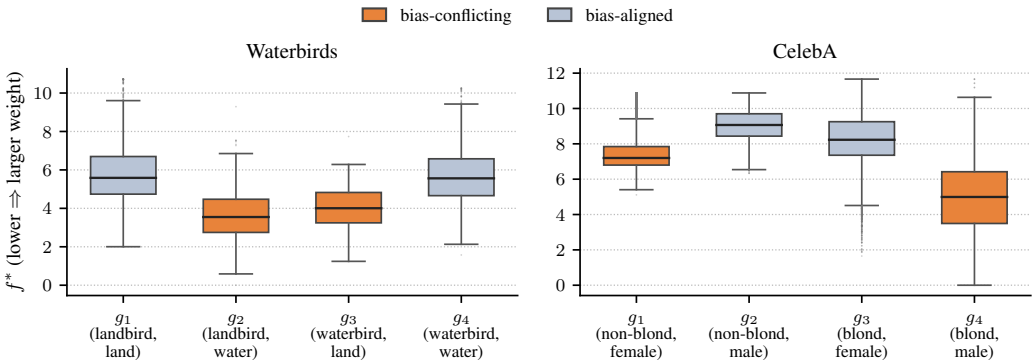


Figure 3: Distribution of class-standardized OT dual potentials f^* by group on Waterbirds and CelebA in the standard setting. Within each class, lower potentials favor larger sample weights. Bias-conflicting groups (orange) have lower median potentials than their bias-aligned counterparts (light blue).

Table 3: Mean sample weight per group in the standard setting, normalized so that uniform weighting equals 1. Boldface indicates the bias-conflicting groups.

Dataset	g_1	g_2	g_3	g_4
Waterbirds	0.31	**6.32**	**10.19**	1.87
CelebA	**0.93**	0.35	2.41	**12.68**
CMNIST	**2.22**	0.78	0.68	**1.91**

within-class mean weights are 0.93 for g_1 versus 0.35 for g_2, and 12.68 for g_4 versus 2.41 for g_3. Class balancing assigns equal total mass to each class, giving the much larger non-blond class a lower per-sample weight scale. Thus, g_1 remains relatively upweighted within its class despite its mean weight being below one.

Second, the weights vary substantially within groups rather than reducing to group-level coefficients. On Waterbirds, the interquartile range of the weights spans 0.34–7.79 for landbirds on water and 0.91–16.07 for waterbirds on land. As illustrated in Figure 1(b), this variation is informative even inside bias-aligned groups: a waterbird-on-water image with a mixed land–water background receives a larger weight than a prototypical open-water image, consistent with weaker reliance on the background shortcut. POTER thus emphasizes underrepresented groups while retaining sample-level variation within them.

6 Conclusion

We presented POTER, a principled reweighting framework based on optimal transport that derives sample importance from distributional alignment. Our experiments reveal a key vulnerability of existing methods: the premise that high-loss samples reliably identify bias-conflicting examples can break down under label noise, with further degradation when corruption is concentrated within minority subgroups. While prior work has diagnosed the label-noise sensitivity of existing mitigation methods [33], targeted remedies remain limited: they can sacrifice clean-setting performance [31] or remain tied to retraining-based pipelines [36]. POTER addresses this issue through a single unified weighting mechanism that handles spurious correlation and label noise simultaneously, requiring no retraining and adding only minimal overhead for computing OT dual potentials. We hope this work motivates broader consideration of overlooked failure modes in robust learning under subpopulation shift, particularly those induced by realistic data imperfections such as label noise. One limitation is that POTER still depends on a representative, class-labeled validation set. Appendix G.1 examines reduced and noisy references, while robustness to entirely missing validation subgroups remains an open challenge.

Acknowledgments and Disclosure of Funding

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT; RS-2026-25468697, RS-2026-25607210), and by the Institute of Information & Communications Technology Planning & Evaluation (IITP) Global Data-X Leader HRD program grant funded by the Korea government (MSIT; IITP-2024-RS-2024-00441244).

References

[1] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

[2] Saeid Asgari, Aliasghar Khani, Fereshte Khani, Ali Gholami, Linh Tran, Ali Mahdavi Amiri, and Ghassan Hamarneh. Masktune: Mitigating spurious correlations by forcing to explore. In *Proc. Advances in Neural Information Processing Systems*, 2022.

[3] Xingjian Bai, Guangyi He, Yifan Jiang, and Jan Obloj. Wasserstein distributional robustness of neural networks. In *Proc. Advances in Neural Information Processing Systems*, 2023.

[4] Imon Banerjee, Kamanasish Bhattacharjee, John L Burns, Hari Trivedi, Saptarshi Purkayastha, Laleh Seyyed-Kalantari, Bhavik N Patel, Rakesh Shiradkar, and Judy Gichoya. "shortcuts" causing bias in radiology artificial intelligence: causes, evaluation, and mitigation. *Journal of the American College of Radiology*, 20(9):842–851, 2023.

[5] Jose Blanchet, Lin Chen, and Xun Yu Zhou. Distributionally robust mean-variance portfolio selection with wasserstein distances. *Management Science*, 68(9):6382–6410, 2022.

[6] Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion proceedings of the 2019 world wide web conference*, pages 491–500, 2019.

[7] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

[8] Yihe Deng, Yu Yang, Baharan Mirzasoleiman, and Quanquan Gu. Robust learning with progressive data expansion against spurious correlation. In *Proc. Advances in Neural Information Processing Systems*, 2024.

[9] Mahdi Ghaznavi, Hesam Asadollahzadeh, Fahimeh Hosseini Noohdani, Soroush Vafaie Tabar, Hosein Hasani, Taha Akbari Alvanagh, Mohammad Hossein Rohban, and Mahdieh Soleymani Baghshah. Exploiting what trained models learn for making them robust to spurious correlations without group annotations. In *Workshop on Spurious Correlation and Shortcut Learning: Foundations and Solutions*, 2025.

[10] Soumya Suvra Ghosal and Yixuan Li. Distributionally robust optimization with probabilistic group. In *Proc. AAAI Conference on Artificial Intelligence*, pages 11809–11817, 2023.

[11] Abinitha Gourabathina, Hyewon Jeong, Teya Bergamaschi, Marzyeh Ghassemi, and Collin Stultz. Robustness beyond known groups with low-rank adaptation. *arXiv preprint arXiv:2602.06924*, 2026.

[12] Yujin Han and Difan Zou. Improving group robustness on spurious correlation requires preciser group inference. In *Proc. International Conference on Machine Learning*, pages 17480–17504, 2024.

[13] Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. In *Proc. International Conference on Machine Learning*, pages 1929–1938, 2018.

[14] Sung Ho Jo, Seonghwi Kim, and Minwoo Chae. Mitigating spurious correlation via distributionally robust learning with hierarchical ambiguity sets. *Proc. International Conference on Learning Representations*, 2026.

[15] Hoang Anh Just, Feiyang Kang, Jiachen T Wang, Yi Zeng, Myeongseob Ko, Ming Jin, and Ruoxi Jia. Lava: Data valuation without pre-specified learning algorithms. *Proc. International Conference on Learning Representations*, 2023.

[16] Leonid V Kantorovich. On the translocation of masses. In *Dokl. Akad. Nauk. USSR (NS)*, volume 37, pages 199–201, 1942.

[17] Samuel Kessler, Tam Le, and Vu Nguyen. Sava: Scalable learning-agnostic data valuation. *Proc. International Conference on Learning Representations*, 2025.

[18] Nayeong Kim, Juwon Kang, Sungsoo Ahn, Jungseul Ok, and Suha Kwak. Improving robustness to multiple spurious correlations by multi-objective optimization. In *Proc. International Conference on Machine Learning*, pages 24040–24058, 2024.

[19] Seonghwi Kim, Sung Ho Jo, Wooseok Ha, and Minwoo Chae. Distributionally robust classification for multi-source unsupervised domain adaptation. In *Proc. International Conference on Learning Representations*, 2026.

[20] Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer re-training is sufficient for robustness to spurious correlations. In *Proc. International Conference on Learning Representations*, 2023.

[21] P. W. Koh, S. Sagawa, H. Marklund, S. M. Xie, M. Zhang, A. Balsubramani, W. Hu, M. Yasunaga, R. L. Phillips, I. Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *Proc. International Conference on Machine Learning*, pages 5637–5664, 2021.

[22] David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). In *Proc. International Conference on Machine Learning*, pages 5815–5826, 2021.

[23] Daniel Kuhn, Peyman Mohajerin Esfahani, Viet Anh Nguyen, and Soroosh Shafieezadeh-Abadeh. Wasserstein distributionally robust optimization: Theory and applications in machine learning. In *Operations Research & Management Science in the Age of Analytics*, pages 130–166. Informs, 2019.

[24] Tyler LaBonte, Vidya Muthukumar, and Abhishek Kumar. Towards last-layer retraining for group robustness with fewer annotations. In *Proc. Advances in Neural Information Processing Systems*, 2024.

[25] Wenqian Li, Shuran Fu, Fengrui Zhang, and Yan Pang. Data valuation and detections in federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12027–12036, 2024.

[26] Evan Z Liu, Behzad Haghgoo, Annie S Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just train twice: Improving group robustness without training group information. In *Proc. International Conference on Machine Learning*, pages 6781–6792, 2021.

[27] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proc. International Conference on Computer Vision*, pages 3730–3738, 2015.

[28] Varun Mulchandani and Jung-Eun Kim. Severing spurious correlations with data pruning. In *Proc. International Conference on Learning Representations*, 2025.

[29] Junhyun Nam, Jaehyung Kim, Jaeho Lee, and Jinwoo Shin. Spread spurious attribute: Improving worst-group accuracy with spurious attribute estimation. In *Proc. International Conference on Learning Representations*, 2022.

[30] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.

[31] Dongpin Oh, Dae Lee, Jeunghyun Byun, and Bonggun Shin. Improving group robustness under noisy labels using predictive uncertainty. *arXiv preprint arXiv:2212.07026*, 2022.

[32] Bhargavi Paranjape, Pradeep Dasigi, Vivek Srikumar, Luke Zettlemoyer, and Hannaneh Hajishirzi. AGRO: Adversarial discovery of error-prone groups for robust optimization. In *Proc. International Conference on Learning Representations*, 2023.

[33] Rui Qiao and Bryan Kian Hsiang Low. Understanding domain generalization: A noise robustness perspective. In *Proc. International Conference on Learning Representations*, 2024.

[34] Shikai Qiu, Andres Potapczynski, Pavel Izmailov, and Andrew Gordon Wilson. Simple and fast group robustness by automatic feature reweighting. In *Proc. International Conference on Machine Learning*, pages 28448–28467, 2023.

[35] S. Sagawa, P. W. Koh, T. B. Hashimoto, and P. Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *Proc. International Conference on Learning Representations*, 2020.

[36] Nathan Stromberg, Rohan Ayyagari, Sanmi Koyejo, Richard Nock, and Lalitha Sankar. Enhancing robustness of last layer two-stage fair model corrections. In *Proc. Advances in Neural Information Processing Systems*, 2024.

[37] Nathan Stromberg, Rohan Ayyagari, Monica Welfert, Sanmi Koyejo, Richard Nock, and Lalitha Sankar. Robustness to subpopulation shift with domain label noise via regularized annotation of domains. *Transactions on Machine Learning Research*, 2024.

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

[39] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The caltech-ucsd birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

[40] Huaxiu Yao, Yu Wang, Sai Li, Linjun Zhang, Weixin Liang, James Zou, and Chelsea Finn. Improving out-of-distribution robustness via selective augmentation. In *Proc. International Conference on Machine Learning*, pages 25407–25437, 2022.

[41] Michael Zhang, Nimit S Sohoni, Hongyang R Zhang, Chelsea Finn, and Christopher Re. Correct-N-Contrast: A contrastive approach for improving robustness to spurious correlations. In *Proc. International Conference on Machine Learning*, pages 26484–26516, 2022.

[42] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1452–1464, 2017.

A Class-Conditioned OT Decomposition

Let

$$\mu_t = \sum_{i=1}^{n} w_i^t \, \delta_{z_i}, \qquad \mu_{\mathrm{ref}} = \sum_{j=1}^{m} w_j^{\mathrm{ref}} \, \delta_{z'_j}.$$

For each class $k \in [K]$, define

$$I_k := \{i : \tilde{y}_i = k\}, \qquad J_k := \{j : y'_j = k\},$$

and assume that the classwise masses match:

$$\sum_{i \in I_k} w_i^t = \sum_{j \in J_k} w_j^{\mathrm{ref}}, \qquad \forall k \in [K].$$

This condition requires equal training and reference mass within each class, while allowing the total mass to differ across classes.

For notational simplicity, we continue to write x_i and x'_j for the feature embeddings on which the transport cost is computed. We use the class-conditioned cost

$$C_{ij} = \begin{cases} \|x_i - x'_j\|_2^2, & \tilde{y}_i = y'_j, \\ +\infty, & \tilde{y}_i \neq y'_j, \end{cases}$$

and define the corresponding classwise empirical measures by

$$\mu_t^{(k)} := \sum_{i \in I_k} w_i^t \, \delta_{z_i}, \qquad \mu_{\mathrm{ref}}^{(k)} := \sum_{j \in J_k} w_j^{\mathrm{ref}} \, \delta_{z'_j}.$$

Accordingly, $\mathcal{T}(\mu_t, \mu_{\mathrm{ref}})$ denotes the OT value under the class-conditioned cost C, and $\mathcal{T}^{(k)}(\mu_t^{(k)}, \mu_{\mathrm{ref}}^{(k)})$ denotes its restriction to class k.

Under this cost, every finite-cost coupling is supported on same-class pairs. The mass-matching condition ensures that a feasible coupling exists within each class, leading to the following decomposition.

Proposition A.1 (Class-conditioned OT decomposition)**.** *Under the class-conditioned cost C, the OT value decomposes across classes as*

$$\mathcal{T}(\mu_t, \mu_{\mathrm{ref}}) = \sum_{k=1}^{K} \mathcal{T}^{(k)}\big(\mu_t^{(k)}, \mu_{\mathrm{ref}}^{(k)}\big).$$

Moreover, if $(f^{(k)*}, g^{(k)*})$ is an optimal dual pair for the k-th classwise OT problem for each $k \in [K]$, then the global dual vectors $f^* \in \mathbb{R}^n$ and $g^* \in \mathbb{R}^m$, defined by

$$f_i^* := f_i^{(k)*} \quad \textit{for } i \in I_k, \qquad g_j^* := g_j^{(k)*} \quad \textit{for } j \in J_k,$$

form an optimal dual pair for $\mathcal{T}(\mu_t, \mu_{\mathrm{ref}})$.

Proof. Under the class-conditioned cost

$$C_{ij} = \begin{cases} \|x_i - x'_j\|_2^2, & \tilde{y}_i = y'_j, \\ +\infty, & \tilde{y}_i \neq y'_j, \end{cases}$$

any feasible coupling π satisfying $\langle C, \pi \rangle < \infty$ must satisfy

$$\pi_{ij} = 0 \qquad \text{whenever } \tilde{y}_i \neq y'_j.$$

Hence every finite-cost feasible plan is supported on

$$\bigcup_{k=1}^{K} (I_k \times J_k).$$

For each $k \in [K]$, let $\pi^{(k)}$ denote the restriction of π to $I_k \times J_k$. Then the marginal constraints imply

$$\pi^{(k)} \in \Pi\big(\mu_t^{(k)}, \mu_{\mathrm{ref}}^{(k)}\big).$$

Conversely, by the classwise mass-matching assumption, any collection of classwise couplings

$$\pi^{(k)} \in \Pi\big(\mu_t^{(k)}, \mu_{\mathrm{ref}}^{(k)}\big), \qquad k \in [K],$$

induces a feasible global coupling π by setting

$$\pi_{ij} := \begin{cases} \pi_{ij}^{(k)}, & (i,j) \in I_k \times J_k \text{ for some } k, \\ 0, & \text{otherwise.} \end{cases}$$

Therefore, the feasible set of finite-cost couplings factorizes across classes.

On this feasible set,

$$\langle C, \pi \rangle = \sum_{k=1}^{K} \langle C^{(k)}, \pi^{(k)} \rangle,$$

where $C^{(k)}$ denotes the restriction of C to $I_k \times J_k$. It follows that

$$\mathcal{T}(\mu_t, \mu_{\mathrm{ref}}) = \inf_{\pi \in \Pi(\mu_t, \mu_{\mathrm{ref}})} \langle C, \pi \rangle = \sum_{k=1}^{K} \inf_{\pi^{(k)} \in \Pi(\mu_t^{(k)}, \mu_{\mathrm{ref}}^{(k)})} \langle C^{(k)}, \pi^{(k)} \rangle = \sum_{k=1}^{K} \mathcal{T}^{(k)}\big(\mu_t^{(k)}, \mu_{\mathrm{ref}}^{(k)}\big).$$

Now let $(f^{(k)*}, g^{(k)*})$ be an optimal dual pair for the k-th classwise OT problem. Define global vectors $f^* \in \mathbb{R}^n$ and $g^* \in \mathbb{R}^m$ by

$$f_i^* := f_i^{(k)*} \quad \text{for } i \in I_k, \qquad g_j^* := g_j^{(k)*} \quad \text{for } j \in J_k.$$

Since $C_{ij} = +\infty$ whenever $\tilde{y}_i \neq y'_j$, the constraint

$$f_i^* + g_j^* \leq C_{ij}$$

is automatically satisfied whenever $i \in I_k$, $j \in J_\ell$, and $k \neq \ell$, and on each set $I_k \times J_k$ it reduces to the classwise dual feasibility constraint. Thus (f^*, g^*) is feasible for the global dual problem. Its dual objective value is

$$\sum_{i=1}^{n} f_i^* \, w_i^t + \sum_{j=1}^{m} g_j^* \, w_j^{\mathrm{ref}} = \sum_{k=1}^{K} \left(\sum_{i \in I_k} f_i^{(k)*} w_i^t + \sum_{j \in J_k} g_j^{(k)*} w_j^{\mathrm{ref}} \right),$$

which, by optimality of each classwise dual pair, equals

$$\sum_{k=1}^{K} \mathcal{T}^{(k)}\big(\mu_t^{(k)}, \mu_{\mathrm{ref}}^{(k)}\big) = \mathcal{T}(\mu_t, \mu_{\mathrm{ref}}).$$

Hence (f^*, g^*) is an optimal dual pair for $\mathcal{T}(\mu_t, \mu_{\mathrm{ref}})$. □

The decomposition is determined by the observed class partition and matched marginal masses and therefore does not require any assumption on the label-corruption mechanism. In our implementation, the two empirical measures are normalized within each class to satisfy the mass-matching condition, as described in Appendix E.3.

Computational Implication. Proposition A.1 also yields a practical computational benefit in a classwise implementation. Let

$$n_k := |I_k|, \qquad m_k := |J_k|.$$

Rather than operating on a dense $n \times m$ transport matrix, the class-conditioned formulation can be implemented as K independent OT subproblems of sizes $n_k \times m_k$. For Sinkhorn-style solvers, this reduces the total per-iteration cost from $O(nm)$ for a dense full-matrix implementation to

$$O\left(\sum_{k=1}^{K} n_k m_k \right).$$

When class sizes are roughly balanced, i.e.,

$$n_k \approx n/K, \qquad m_k \approx m/K,$$

this becomes

$$O(nm/K).$$

In addition, the classwise subproblems are independent and can therefore be parallelized straightforwardly.

B Proof of Proposition 4.1

Proof. The optimization problem in (2) is equivalent to

$$\min_{w \in \Delta_n} \left\{ \sum_{i=1}^{n} f_i^* w_i + \tau \sum_{i=1}^{n} w_i \log \frac{w_i}{w_i^0} \right\},$$

where

$$\Delta_n := \left\{ w \in \mathbb{R}_+^n : \sum_{i=1}^{n} w_i = 1 \right\}.$$

Since $w_i^0 > 0$ for all $i \in [n]$ and $\tau > 0$, the objective is strictly convex on Δ_n, so the minimizer is unique. Moreover, the optimum is attained in the interior of Δ_n. Therefore, the first-order optimality conditions are necessary and sufficient for optimality.

Consider the Lagrangian

$$\mathcal{L}(w, \lambda) = \sum_{i=1}^{n} f_i^* w_i + \tau \sum_{i=1}^{n} w_i \log \frac{w_i}{w_i^0} + \lambda \left(\sum_{i=1}^{n} w_i - 1 \right).$$

For each $i \in [n]$, the first-order condition is

$$\frac{\partial \mathcal{L}}{\partial w_i} = f_i^* + \tau \left(\log \frac{w_i}{w_i^0} + 1 \right) + \lambda = 0.$$

Hence

$$w_i = w_i^0 \exp\left(-\frac{f_i^*}{\tau} \right) \exp\left(-\frac{\lambda}{\tau} - 1 \right).$$

Summing over $i \in [n]$, the constraint $\sum_{i=1}^{n} w_i = 1$ implies

$$\exp\left(-\frac{\lambda}{\tau} - 1 \right) = \left(\sum_{j=1}^{n} w_j^0 \exp\left(-\frac{f_j^*}{\tau} \right) \right)^{-1}.$$

Therefore,

$$w_i = \frac{w_i^0 \exp(-f_i^*/\tau)}{\sum_{j=1}^{n} w_j^0 \exp(-f_j^*/\tau)}, \qquad i \in [n].$$

This proves that the unique minimizer $\mu^* = \sum_{i=1}^{n} w_i^* \delta_{z_i}$ satisfies

$$w_i^* = \frac{w_i^0 \exp(-f_i^*/\tau)}{\sum_{j=1}^{n} w_j^0 \exp(-f_j^*/\tau)}, \qquad i \in [n].$$

□

Scope of the theoretical results. Propositions A.1 and 4.1 characterize POTER's weight construction: they establish the classwise OT decomposition under the stated mass-matching condition and the unique KL-regularized reweighting solution for given dual potentials. These results do not provide a finite-sample guarantee on the target risk of a classifier trained using the resulting weights. Under label shift, where class proportions change while class-conditional distributions remain fixed, establishing such a guarantee would additionally require relating the population reweighted objective to the target risk, controlling estimation error in the OT potentials and sample weights, and analyzing the generalization of weighted ERM with these data-dependent weights. We leave such an analysis to future work.

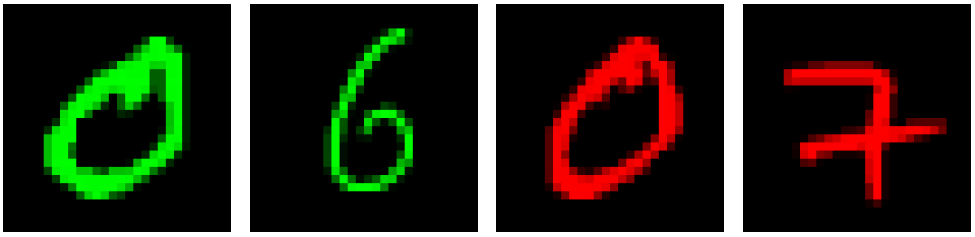
Known target class proportions can be incorporated into the class-mass normalization step. When these proportions are unknown, our default class-balancing scheme is motivated by worst-group robustness. For a fixed classifier and fixed group-conditional distributions, the population worst-group accuracy lower-bounds the accuracy under any mixture of the groups. Thus, this bound continues to hold under changes in group composition, including those induced by changes in class proportions when the groups jointly encode class and spurious attributes. This population-level relation motivates the use of worst-group accuracy when the test-time group composition is unknown.

C Dataset Details

Colored MNIST (CMNIST) [1, 40]. The CMNIST dataset is a binary classification benchmark derived from MNIST. Digits 0–4 are assigned to class 0, and digits 5–9 are assigned to class 1. Digit color is used as the spurious attribute, with two possible colors: red and green. This yields four groups defined by the pair of class label and color:

$$g_1 = \{0, \text{green}\}, \quad g_2 = \{1, \text{green}\}, \quad g_3 = \{0, \text{red}\}, \quad g_4 = \{1, \text{red}\}.$$

The training, validation, and test sets contain 30,000, 10,000, and 20,000 examples, respectively. In the training set, color is spuriously correlated with the label: the red-to-green ratio is $8:2$ for class 0 and $2:8$ for class 1. In the validation set, the color distribution is balanced within each class, with a red-to-green ratio of $1:1$. In the test set, the label-color correlation is reversed: the red-to-green ratio is $1:9$ for class 0 and $9:1$ for class 1. Following [1], labels are randomly flipped with probability 0.25.



(a) 0, green (b) 1, green (c) 0, red (d) 1, red

Figure 4: Example images from the CMNIST dataset. The groups are $g_1 = \{0, \text{green}\}$, $g_2 = \{1, \text{green}\}$, $g_3 = \{0, \text{red}\}$, and $g_4 = \{1, \text{red}\}$.

Waterbirds [35]. The Waterbirds dataset is a binary image classification benchmark constructed by placing bird images from the Caltech-UCSD Birds 200-2011 (CUB) dataset [39] onto backgrounds from the Places dataset [42]. The target label is bird type, and the spurious attribute is background type. Bird species in CUB are partitioned into waterbirds and landbirds; the waterbird class includes Albatross, Auklet, Cormorant, Frigatebird, Fulmar, Gull, Jaeger, Kittiwake, Pelican, Puffin, Tern, Gadwall, Grebe, Mallard, Merganser, Guillemot, and Pacific Loon, while all remaining species are treated as landbirds. The four groups are defined by the pair of bird type and background:

$$g_1 = \{\text{landbird}, \text{land}\}, \quad g_2 = \{\text{landbird}, \text{water}\},$$
$$g_3 = \{\text{waterbird}, \text{land}\}, \quad g_4 = \{\text{waterbird}, \text{water}\}.$$

The standard split contains 4,795 training examples, 1,199 validation examples, and 5,794 test examples. In the training set, bird type is strongly correlated with background: 3,498 examples are landbirds on land backgrounds, 184 are landbirds on water backgrounds, 56 are waterbirds on land backgrounds, and 1,057 are waterbirds on water backgrounds. Thus, the bias-conflicting minority groups are g_2 and g_3. The validation split contains 467, 466, 133, and 133 examples from g_1 through g_4, respectively.



(a) landbird, land (b) landbird, water (c) waterbird, land (d) waterbird, water

Figure 5: Example images from the Waterbirds dataset. The groups are $g_1 = \{\text{landbird, land}\}$, $g_2 = \{\text{landbird, water}\}$, $g_3 = \{\text{waterbird, land}\}$, and $g_4 = \{\text{waterbird, water}\}$.

CelebA [27]. CelebA is a large-scale facial attribute dataset. We use the standard blond-hair classification task, where the target label is hair color and the spurious attribute is gender. The target label is binary, distinguishing blond hair from non-blond hair. The four groups are defined by the pair of hair color and gender:

$$g_1 = \{\text{non-blond hair}, \text{female}\}, \quad g_2 = \{\text{non-blond hair}, \text{male}\},$$
$$g_3 = \{\text{blond hair}, \text{female}\}, \quad g_4 = \{\text{blond hair}, \text{male}\}.$$

The training and validation splits contain 162,770 and 19,867 examples, respectively. In the training set, the group counts for g_1 through g_4 are 71,629, 66,874, 22,880, and 1,387, respectively. The validation split contains 8,535, 8,276, 2,874, and 182 examples from g_1 through g_4, respectively. Thus, the blond male group g_4 is the minority group and is substantially underrepresented compared with the other groups.



(a) non-blond, female (b) non-blond, male (c) blond, female (d) blond, male

Figure 6: Example images from the CelebA dataset. The groups are $g_1 = \{\text{non-blond hair, female}\}$, $g_2 = \{\text{non-blond hair, male}\}$, $g_3 = \{\text{blond hair, female}\}$, and $g_4 = \{\text{blond hair, male}\}$.

CivilComments [6]. CivilComments is a text classification benchmark for online toxicity detection. We use the WILDS [21] version of the dataset, where the target label is binary toxicity: $Y = 0$ denotes a non-toxic comment and $Y = 1$ denotes a toxic comment. The WILDS dataset defines four coarse groups by the combination of toxicity and whether a comment mentions any identity attribute. The training, validation, and test splits contain 269,038, 45,180, and 133,782 comments, respectively.

Following prior work, we use the 16-group evaluation protocol based on eight demographic identity attributes:

$$\mathcal{A} = \{\text{male}, \text{female}, \text{LGBTQ}, \text{Christian}, \text{Muslim}, \text{other religion}, \text{black}, \text{white}\}.$$

Let $A_j(x) \in \{0, 1\}$ indicate whether comment x mentions the j-th identity attribute. For each identity attribute $j \in \{1, \ldots, 8\}$ and toxicity label $y \in \{0, 1\}$, we define

$$g_{j,y} = \{(x, y') : A_j(x) = 1,\ y' = y\}.$$

Thus, the evaluation groups correspond to toxic and non-toxic comments mentioning each of the eight identity attributes. Unlike CMNIST, Waterbirds, and CelebA, these 16 groups are overlapping rather than a partition of the dataset, since a single comment may mention multiple identities. Per-group sample counts across the train and validation splits are reported in Table 4.

D Baseline Details

We provide brief descriptions of the baselines used in our comparisons.

- **ERM**: Empirical risk minimization trains the model by minimizing the average training loss, without using group information or an explicit robustness objective. It serves as the standard non-robust reference point.
- **Group DRO** [35]: Group DRO assumes access to group annotations during training and optimizes the worst-group loss rather than the average loss. It adaptively emphasizes groups with larger losses, encouraging the model to improve performance on underrepresented or bias-conflicting groups.

Table 4: Per-group sample counts in CivilComments-WILDS for each of the eight identity attributes. Counts are obtained by thresholding both the toxicity score and the identity-attribute annotation at 0.5. Groups are overlapping: a comment that mentions multiple identities is counted in multiple rows.

Identity	Train		Validation	
	$Y=0$	$Y=1$	$Y=0$	$Y=1$
male	25,373	4,437	4,050	715
female	31,282	4,962	5,120	771
LGBTQ	6,155	2,265	1,099	358
Christian	24,292	2,446	4,166	384
Muslim	10,829	3,125	1,598	512
other religion	5,541	1,003	824	162
black	6,785	3,111	1,119	533
white	12,016	4,682	2,015	852

- **LISA** [40]: LISA mitigates reliance on spurious features through a selective interpolation strategy. It applies Mixup-style augmentation between example pairs that either share the same label but differ in the spurious attribute, or share the same spurious attribute but differ in the label. This selective pairing encourages the learned representation to be less dependent on shortcut correlations.
- **DFR** [20]: DFR first trains a standard ERM model and then retrains only the last layer using a group-balanced subset. In our comparison, we follow the training-set variant used in prior work and construct the balanced retraining set from training examples with group annotations, denoted by DFR^{Tr}.
- **PDE** [8]: PDE begins training from a group-balanced subset of the training data and progressively expands this subset to include additional training examples. This staged expansion is designed to delay the model's reliance on dominant spurious correlations while still allowing it to use more data as training proceeds.
- **JTT** [26]: JTT is a two-stage approach that first trains an ERM model and uses misclassified examples as proxies for hard or bias-conflicting examples. It then constructs an upsampled training set in which these examples are repeated, and trains a second model on this dataset, improving robustness without requiring group annotations for the full training set.
- **CnC** [41]: CnC is a two-stage approach that first trains a standard ERM model and uses its predictions to identify pairs of examples that share the same class but are likely to differ in spurious attributes. It then trains a second model with a contrastive objective that pulls together same-class examples with different spurious attributes, reducing dependence on shortcut cues.
- **SSA** [29]: SSA estimates spurious attributes for training examples using limited group supervision and forms pseudo groups from the predicted attributes and class labels. Group DRO is then applied to these inferred groups.
- **AFR** [34]: AFR first trains a standard ERM model and then retrains the last layer with a weighted loss on top of fixed representations. The weighting rule emphasizes examples on which the ERM model performs poorly, increasing the influence of minority or bias-conflicting samples without relying on training-set group annotations.
- **GIC** [12]: GIC infers group labels by training a spurious-attribute classifier using comparison data with a different group distribution from the training set. The predicted spurious attributes are combined with class labels to form inferred groups, which can then be used with downstream robust learning methods such as Group DRO.
- **LEIA** [11]: LEIA proposes a two-stage adaptation method that identifies a low-dimensional subspace in the representation where model errors concentrate. It then applies a low-rank adjustment to the classifier logits, improving group robustness in a parameter-efficient manner without modifying the backbone or requiring full training-set group annotations.
- **KNN-RAD** [36]: KNN-RAD addresses label noise under spurious correlations by first performing kNN-based label correction in representation space and then applying a robust last-layer retraining procedure. It is the most closely related baseline in our label-noise experiments because it explicitly targets the joint setting of label noise and spurious correlations.

E Implementation Details

All experiments were conducted using a single NVIDIA GeForce RTX 3090 GPU. We first describe the implementation details of POTER and then provide details on baseline implementation. The code for POTER is available at `https://github.com/Sung-Ho-Jo/POTER`.

E.1 Feature Embeddings and Training Configurations

For each dataset, we compute OT dual potentials in a fixed pretrained feature space $\phi(\cdot)$. For CMNIST, Waterbirds, and CelebA, we use features from an ImageNet-pretrained ResNet-50. For CivilComments, we use features from BERT-base-uncased. Importantly, these feature extractors are not trained on the downstream task before computing OT dual potentials, distinguishing POTER from methods that rely on representations learned by a preliminary task-specific model. Features are extracted once before the OT step and kept fixed for OT computation.

For the final reweighted ERM training stage, we use the same architecture as the feature extractor used for OT: ResNet-50 for image datasets and BERT-base-uncased for CivilComments. We use SGD for CMNIST, Waterbirds, and CelebA, and AdamW for CivilComments. The training configurations are summarized in Table 5. These settings are based on prior protocols for image datasets [14] and CivilComments [26], with the exact values reported in the table. We do not use a learning rate scheduler. For Waterbirds and CelebA, we use standard data augmentations. Specifically, we apply random resized cropping, horizontal flipping, and color jittering for Waterbirds, and random resized cropping, horizontal flipping, and random rotation for CelebA. Model selection is performed using worst-group accuracy on the validation set, and all reported results are averaged over three random seeds.

Table 5: Training configurations for the final reweighted ERM stage.

Dataset	Optimizer	Learning Rate	Weight Decay	Batch Size	Epochs
CMNIST	SGD	10^{-2}	10^{-2}	128	300
Waterbirds	SGD	10^{-5}	1.0	128	300
CelebA	SGD	10^{-5}	10^{-2}	128	50
CivilComments	AdamW	10^{-5}	10^{-2}	32	3

E.2 Reference Distribution Construction

For the standard benchmark setting and subgroup-concentrated label noise setting, we construct the reference distribution from validation examples in the bias-conflicting groups: landbird-on-water and waterbird-on-land for Waterbirds, non-blond female and blond male for CelebA, and the two label-color-conflicting groups for CMNIST. For CivilComments, following the supervision setting used in prior work [26, 12, 11], we assume access only to the binary identity-mention indicator rather than the full set of identity-specific group annotations used for 16-group evaluation. Accordingly, the reference distribution is constructed using the four groups defined by toxicity and whether any identity attribute is mentioned; specifically, we use the non-toxic identity-mentioning group and the toxic non-identity-mentioning group.

For the symmetric label noise setting, we instead construct the reference distribution by including both bias-aligned and bias-conflicting examples within each class. Specifically, within each class, we subsample the bias-aligned and bias-conflicting examples to have matched counts, so that the reference distribution is balanced with respect to the spurious attribute conditional on the class label. Within the selected reference set, we use uniform empirical weights.

E.3 Optimal Transport Computation

We compute OT dual potentials using class-conditioned optimal transport. For each observed class, we solve an entropy-regularized OT problem between the training examples and reference examples belonging to that class. We normalize the training and reference empirical measures separately within each class. For a class k with n_k training examples and m_k reference examples, we assign uniform masses $1/n_k$ and $1/m_k$ to the training and reference examples, respectively. Both marginals therefore

have total mass one, satisfying the classwise mass-matching condition in Appendix A by construction, even when the training and reference sample counts differ.

In the pretrained feature space defined by $\phi(\cdot)$, we use the squared Euclidean distance as the ground cost:

$$C_{ij} = \|\phi(x_i) - \phi(x'_j)\|_2^2,$$

where x_i is a training example and x'_j is a reference example.

We solve each class-conditioned entropy-regularized OT problem using the Sinkhorn algorithm [7]. For each class c, the entropy regularization strength is set to

$$\varepsilon_c = 0.03\,\bar{C}_c,$$

where $\bar{C}_c$ denotes the average cost within class c. Sinkhorn is run for at most 300 iterations with convergence threshold 10^{-6}. After convergence, we obtain the training-side dual potentials f^* for each class-conditioned OT problem. We then standardize the dual potentials within each class using median-MAD normalization to align scales across classes, and concatenate them into a single vector over the training set.

E.4 Sample Reweighting

Given the class-conditioned dual potentials f_i^*, we instantiate the closed-form reweighting rule in Proposition 4.1. In our implementation, we set

$$\tau = \alpha \cdot \mathrm{median}(f^*),$$

where α controls the sharpness of the weighting. As noted in Section 3, for fixed τ, the reweighting rule is invariant to the additive-constant ambiguity of OT dual potentials: for any $a \in \mathbb{R}$,

$$\frac{\exp(-(f_i^* + a)/\tau)}{\sum_{j=1}^n \exp(-(f_j^* + a)/\tau)} = \frac{\exp(-f_i^*/\tau)}{\sum_{j=1}^n \exp(-f_j^*/\tau)}.$$

Therefore, dual potentials that differ only by an additive constant lead to the same normalized sample weights.

We then apply additional normalization steps to the resulting weights. To prevent class imbalance from dominating the reweighted training distribution, we normalize the weights within each class and then rescale them so that each class has equal total mass. The resulting weights are used for weighted sampling during training. As a practical safeguard, we clip the maximum weight to prevent the sampling distribution from concentrating too much mass on a small number of examples.

We tune α and the clipping threshold using validation worst-group accuracy. For Waterbirds and CelebA, we search over $\alpha \in \{0.1, 0.2\}$ and clipping thresholds in $\{3, 5, 10, 20\}$. For CMNIST and CivilComments, we search over $\alpha \in \{0.4, 0.5, 0.6, 0.7\}$ and clipping thresholds in $\{2, 3, 5, 10\}$.

Choice of class masses. Class-mass normalization controls the relative contribution of each observed class to the reweighted training objective while leaving the classifier's hypothesis class unchanged. More generally, this step can assign a prescribed fraction q_k of the total weight to class k before clipping, where $q_k \geq 0$ and $\sum_{k=1}^K q_k = 1$. Our default is $q_k = 1/K$. Alternatives include known target class proportions or the observed training class proportions, $q_k = n_k/n$, where n_k denotes the number of training examples with observed label k. These alternatives can be incorporated by changing the classwise rescaling targets.

The reference mixture weights ρ specify the composition of μ_{ref}, whereas q controls the class masses in the subsequent weight-normalization step. In our experiments, the equal-class default is motivated by worst-group robustness when the test-time group composition is unknown, as discussed in Appendix B.

E.5 Baseline Implementation

For the standard benchmark results in Table 1, we report the performance from the original papers whenever available, so that each baseline is evaluated under its intended experimental protocol. When a result is not reported in the original paper, we use the reported results from [12] or [14]. If the result

is unavailable in these sources as well, we reproduce the method ourselves following the experimental setting described in the corresponding paper.

For the label-noise experiments in Table 2, we directly run all baselines under the same label-noise protocol used for POTER, following the implementation details and hyperparameter settings described in the original papers. All results are averaged over three random seeds.

F Additional Benchmark Results

We report detailed label-noise results and extended baseline comparisons for the benchmarks studied in the main paper. We also evaluate POTER under stronger spurious correlation on CMNIST.

F.1 Full Results under Subgroup-Concentrated Label Noise

Table 6 reports the full numerical results corresponding to Figure 2. Label corruption is restricted to the landbird-on-water and waterbird-on-land groups in Waterbirds and the blond-male group in CelebA. The reported noise rate is the fraction of labels corrupted within each affected subgroup. Across both datasets, POTER remains substantially more robust than KNN-RAD as the corruption rate within these minority subgroups increases.

Table 6: Worst-group accuracy on Waterbirds and CelebA under different subgroup-concentrated label noise rates.

Dataset	Method	Label Noise (%)					
		10	20	30	40	50	60
Waterbirds	KNN-RAD	$88.1_{\pm 1.1}$	$77.5_{\pm 4.3}$	$72.6_{\pm 3.8}$	$52.9_{\pm 5.8}$	$17.9_{\pm 2.2}$	$7.5_{\pm 0.2}$
	POTER (Ours)	$\mathbf{90.0}_{\pm 0.1}$	$\mathbf{89.0}_{\pm 0.4}$	$\mathbf{86.4}_{\pm 0.4}$	$\mathbf{84.8}_{\pm 0.3}$	$\mathbf{82.4}_{\pm 0.3}$	$\mathbf{78.3}_{\pm 0.9}$
CelebA	KNN-RAD	$79.4_{\pm 0.0}$	$76.9_{\pm 0.8}$	$73.3_{\pm 0.6}$	$70.4_{\pm 0.6}$	$73.0_{\pm 0.3}$	$63.7_{\pm 0.3}$
	POTER (Ours)	$\mathbf{90.6}_{\pm 0.5}$	$\mathbf{88.0}_{\pm 0.9}$	$\mathbf{87.6}_{\pm 1.2}$	$\mathbf{87.2}_{\pm 1.5}$	$\mathbf{87.2}_{\pm 1.5}$	$\mathbf{84.8}_{\pm 1.4}$

F.2 Results under Symmetric Label Noise on CivilComments

Table 7 reports additional symmetric label noise results on CivilComments. CivilComments exhibits a different pattern from Waterbirds and CelebA, as the 16-group evaluation structure is less dominated by within-class spurious-attribute imbalance. In Waterbirds and CelebA, each class contains both bias-aligned and bias-conflicting groups with highly imbalanced frequencies, so robustness requires identifying rare bias-conflicting examples within the same class. In contrast, CivilComments groups are defined by toxicity and identity mentions, and for every identity attribute, non-toxic examples are substantially more frequent than toxic examples (Table 4). Thus, the dominant imbalance in this evaluation is closer to class imbalance between toxic and non-toxic comments than to a within-class spurious correlation structure.

This helps explain why AFR remains relatively robust on CivilComments. AFR upweights examples according to a loss-based weighting parameter selected on the validation set. Under label noise, aggressively upweighting high-loss examples can hurt validation performance because mislabeled samples also tend to incur large losses. Consequently, validation selection favors a near-uniform weighting regime, making the second-stage training behave similarly to ERM rather than strongly emphasizing high-loss examples. At the same time, AFR applies class mass matching in the last-layer retraining stage, which balances the toxic and non-toxic classes. This additional class balancing is particularly effective on CivilComments, where the dominant imbalance across the 16 evaluation groups is largely driven by the scarcity of toxic examples. This contrasts with Waterbirds and CelebA, where class balancing alone cannot recover the rare bias-conflicting groups within each class.

Overall, POTER, KNN-RAD, and AFR all remain relatively robust on CivilComments under symmetric label noise. POTER achieves the best performance at 0% and 10% noise and remains competitive at higher noise rates, while the smaller gap among these methods reflects the different nature of the CivilComments evaluation compared with the image benchmarks.

Table 7: Worst-group accuracy on CivilComments under different symmetric label noise rates.

Dataset	Method	Group Labels (Tr/Val)	No Extra Training	Label Noise (%)			
				0	10	20	30
CivilComments	Group DRO	Tr/Val	✓	$69.0_{\pm 0.9}$	$57.6_{\pm 3.0}$	$57.2_{\pm 3.6}$	$50.1_{\pm 6.8}$
	AFR	Val	-	$66.2_{\pm 1.8}$	$64.6_{\pm 2.1}$	$65.4_{\pm 1.4}$	$63.3_{\pm 0.4}$
	KNN-RAD	Val	-	$69.6_{\pm 0.0}$	$65.0_{\pm 1.0}$	$62.7_{\pm 1.4}$	$62.9_{\pm 2.8}$
	POTER (Ours)	Val	✓	$71.0_{\pm 0.6}$	$68.4_{\pm 1.4}$	$64.2_{\pm 0.9}$	$63.1_{\pm 1.7}$

F.3 Extended Baseline Comparison under Symmetric Label Noise

Table 8 provides an extended comparison under symmetric label noise on Waterbirds and CelebA, additionally including JTT and END [31]. END is one of the few methods that directly addresses spurious correlations under label noise. As the original implementation of END is not publicly available, we report END and JTT results as given in [31].

The results reveal distinct robustness profiles across methods. JTT degrades severely under label noise, consistent with the fragility of error-based sample selection. END maintains relatively stable worst-group accuracy as the noise rate increases, but incurs a substantial clean-setting cost: its clean-setting worst-group accuracy is 82.8% on Waterbirds and 82.6% on CelebA, compared with 90.9% and 90.4% for POTER, respectively. END also requires full-model retraining. KNN-RAD outperforms loss-based baselines but exhibits noticeable degradation at higher noise rates. POTER achieves the best or near-best worst-group accuracy across all noise levels on both datasets, without requiring any additional retraining stage.

Table 8: Extended comparison of worst-group accuracy on Waterbirds and CelebA under symmetric label noise. *Extra Training* indicates the additional retraining required after base model training: *Last* for last-layer retraining, *Full* for full-model retraining, and *Not required* for no additional retraining.

Dataset	Method	Group Labels (Tr/Val)	Extra Training	Label Noise (%)			
				0	10	20	30
Waterbirds	Group DRO	Tr/Val	Not required	$90.6_{\pm 0.2}$	$72.9_{\pm 1.6}$	$54.3_{\pm 1.0}$	$52.2_{\pm 3.6}$
	JTT	Val	Full	$84.6_{\pm 3.0}$	$56.5_{\pm 8.0}$	$6.0_{\pm 3.0}$	$2.7_{\pm 1.0}$
	AFR	Val	Last	$88.3_{\pm 0.9}$	$58.7_{\pm 0.4}$	$61.2_{\pm 5.9}$	$52.9_{\pm 0.0}$
	END	Val	Full	$82.8_{\pm 1.0}$	$84.2_{\pm 1.0}$	$83.2_{\pm 1.0}$	$81.8_{\pm 1.0}$
	KNN-RAD	Val	Last	$\mathbf{91.0}_{\pm 0.1}$	$82.4_{\pm 0.7}$	$74.7_{\pm 1.1}$	$68.9_{\pm 2.1}$
	POTER (Ours)	Val	Not required	$90.9_{\pm 0.3}$	$\mathbf{89.4}_{\pm 0.4}$	$\mathbf{87.9}_{\pm 1.6}$	$\mathbf{85.3}_{\pm 1.4}$
CelebA	Group DRO	Tr/Val	Not required	$89.3_{\pm 1.3}$	$66.9_{\pm 0.3}$	$59.5_{\pm 2.6}$	$54.8_{\pm 1.8}$
	JTT	Val	Full	$82.2_{\pm 2.0}$	$74.8_{\pm 2.0}$	$24.5_{\pm 36.0}$	$15.1_{\pm 16.0}$
	AFR	Val	Last	$80.9_{\pm 2.3}$	$69.4_{\pm 4.8}$	$67.7_{\pm 2.4}$	$58.8_{\pm 6.3}$
	END	Val	Full	$82.6_{\pm 2.0}$	$79.7_{\pm 1.0}$	$81.1_{\pm 2.0}$	$77.8_{\pm 3.0}$
	KNN-RAD	Val	Last	$83.9_{\pm 0.0}$	$80.0_{\pm 0.9}$	$78.1_{\pm 1.3}$	$70.2_{\pm 2.6}$
	POTER (Ours)	Val	Not required	$\mathbf{90.4}_{\pm 0.5}$	$\mathbf{87.6}_{\pm 0.3}$	$\mathbf{82.8}_{\pm 1.0}$	$\mathbf{78.2}_{\pm 0.4}$

F.4 Results under Stronger Spurious Correlation on CMNIST

We evaluate POTER as bias-conflicting training examples become increasingly scarce, reducing their proportion from 20% to 10%, 5%, and 1%. To isolate the effect of spurious-correlation strength, we remove label noise throughout the dataset and keep the remaining experimental settings fixed. Table 9 highlights POTER's robustness to increasingly strong spurious correlations: it maintains 95.1% worst-group accuracy even when bias-conflicting examples constitute only 1% of the training data.

Table 9: Worst-group accuracy (%) on CMNIST under stronger spurious correlation. Label noise is removed throughout the dataset for this experiment.

Bias-conflicting training ratio	20%	10%	5%	1%
POTER	98.4 ± 0.2	98.0 ± 0.2	97.3 ± 0.3	95.1 ± 0.2

G Empirical Analysis of POTER

We examine how POTER's performance depends on validation supervision, the feature extractor, and the choice of weighting method. We also measure the computational overhead of weight construction.

G.1 Sensitivity to Validation Supervision

POTER uses a group-annotated validation set for reference construction and model selection. The validation-supervised baselines in our comparison also rely on such supervision. Nevertheless, because the reference distribution is central to POTER's sample weighting, we assess its robustness when validation supervision is limited or imperfect. Specifically, we vary the reference size, reference-label noise, and availability of group annotations on Waterbirds.

Reduced reference size. We subsample the standard reference set of 599 examples to 25% and 10% of its original size. As shown in Table 10, POTER achieves $89.8 \pm 0.1\%$ worst-group accuracy with roughly 60 reference examples, about 1% of the combined training and validation data. This indicates that a small group-annotated reference can provide a useful weighting signal.

Table 10: Worst-group accuracy (%) on Waterbirds when subsampling the reference set. The full reference contains 599 examples; all other settings are unchanged.

Reference size	100%	25%	10%
POTER	90.9 ± 0.3	91.1 ± 0.1	89.8 ± 0.1

Noisy reference labels. We inject symmetric noise into the reference class labels at nominal rates of 10%, 20%, and 30%. Table 11 shows a gradual decrease in worst-group accuracy, from $90.9 \pm 0.3\%$ with clean reference labels to $82.5 \pm 1.9\%$ at 30% noise. Because the reference is class-imbalanced, approximately 60% of the reference examples labeled as waterbirds are actually landbirds at 30% nominal noise. This indicates that POTER can retain strong worst-group performance even when most reference examples assigned to one class are mislabeled.

Table 11: Worst-group accuracy (%) on Waterbirds under symmetric corruption of reference class labels. At 30% nominal noise, the observed-waterbird reference has approximately 60% contamination due to class imbalance.

Reference-label noise rate	0%	10%	20%	30%
POTER	90.9 ± 0.3	87.9 ± 1.2	85.7 ± 0.3	82.5 ± 1.9

Without group annotations. We use the entire class-labeled validation set as the reference, without group-based selection or balancing, and select models and hyperparameters using mean validation accuracy. Group annotations are therefore unused in reference construction, training, and model selection; test group labels are used only to evaluate worst-group accuracy. Table 12 reports $89.3 \pm 0.5\%$ worst-group accuracy for this variant, compared with $90.9 \pm 0.3\%$ for standard POTER. This result likely benefits from the relatively balanced coverage of groups in Waterbirds' validation split. This suggests that POTER can achieve strong worst-group performance without validation group annotations when the validation distribution is representative of the test distribution.

G.2 Sensitivity to the Feature Extractor

We vary the feature extractor used to compute the sample weights on Waterbirds, keeping the remaining pipeline fixed and performing no extractor-specific retuning. In addition to the default

Table 12: POTER with and without validation group annotations on Waterbirds.

Variant	Reference	Selection metric	Worst-group (%)
Standard POTER	Group-selected subset	Worst-group accuracy	90.9 ± 0.3
Without group annotations	Entire validation set	Mean accuracy	89.3 ± 0.5

ImageNet-pretrained ResNet-50, we consider a ResNet-50 trained by ERM on Waterbirds for 300 epochs and SigLIP 2 ViT-B/16, which uses a different architecture and pretraining paradigm.

Table 13 reports mean sample weights within each group and classification accuracy. Across all three extractors, each bias-conflicting group receives a larger mean weight than its bias-aligned counterpart within the same class: g_2 versus g_1, and g_3 versus g_4. ERM-trained features yield worst-group accuracy comparable to the default features (91.0% versus 90.9%). SigLIP 2 also achieves strong worst-group accuracy (87.9%) and the highest average accuracy (95.2%). These results suggest that POTER is robust to the choice of feature extractor, with consistent weighting behavior and strong performance across the tested representations without extractor-specific retuning.

Table 13: Feature-extractor sensitivity on Waterbirds without extractor-specific retuning. Columns g_1–g_4 report the mean sample weight for each group defined in Appendix C; g_2 and g_3 are bias-conflicting groups.

Extractor	g_1	g_2	g_3	g_4	Worst-group (%)	Average (%)
Pretrained ResNet-50	0.31	6.32	10.19	1.87	90.9 ± 0.3	92.7 ± 0.1
ERM-trained ResNet-50	0.37	7.79	12.43	1.30	91.0 ± 0.1	91.3 ± 0.0
SigLIP 2 ViT-B/16	0.25	6.89	11.24	1.92	87.9 ± 0.1	95.2 ± 0.2

G.3 Comparison with Distance-Based Weighting

To isolate the contribution of OT, we replace its dual potentials with class-conditioned kNN distances to the same reference set on Waterbirds. We keep the feature representation, reference set, score normalization, downstream training procedure, and validation-based tuning protocol fixed. For the distance-based alternative, we jointly search $k \in \{1, 5, 10, 20\}$ and a weight-transformation grid that includes and extends the POTER grid. The selected configuration uses $k = 5$.

Table 14 shows that both methods assign higher mean weights to the minority group within each class, suggesting that the shared feature representation contains relevant subgroup information. However, OT produces substantially larger differences in mean weights between minority and majority groups, indicating stronger relative emphasis on bias-conflicting samples. POTER achieves $90.9 \pm 0.3\%$ worst-group accuracy versus $76.7 \pm 0.6\%$ for kNN, a difference of 14.2 percentage points. This controlled comparison suggests that reference proximity alone is insufficient to adequately mitigate spurious correlations in this setting. This limitation arises because kNN computes reference distances independently for each training sample, allowing many samples to receive high weights from the same small reference neighborhood without accounting for local overrepresentation. For example, a densely populated training region can remain close to a few reference points even when its share of training mass greatly exceeds the reference mass in that region. In contrast, OT's fixed marginal constraints make these samples compete for limited reference mass, so its dual potentials reflect regional excess or deficit of training mass.

Table 14: Comparison of OT-based weighting and class-conditioned kNN distance weighting on Waterbirds.

Weighting method	g_1	g_2	g_3	g_4	Worst-group (%)
Class-conditioned kNN weighting	0.66	1.24	3.84	1.94	76.7 ± 0.6
OT-based weighting (POTER)	0.31	6.32	10.19	1.87	90.9 ± 0.3

G.4 Runtime Analysis and Comparison with SELF

Runtime for computing OT dual potentials. Table 15 reports the wall-clock time required to compute the OT dual potentials f^* after feature embeddings have been extracted. Feature extraction and model training are excluded from these measurements. The OT computation takes at most 4.52 seconds on average across the evaluated datasets.

In our implementation, the downstream reweighting objective depends only on the training-side dual potentials, not on the full transport plan. We therefore do not store the full transport plan and retain only the Sinkhorn scaling variables needed to compute f^*. This reduces memory usage and makes the OT step practical even for larger datasets.

Table 15: Runtime for computing OT dual potentials f^* after feature embeddings have been extracted. Results are averaged over five runs.

Dataset	n_{train}	n_{ref}	Dimension	Time (s)
CMNIST	30,000	5,022	2,048	0.41 ± 0.00
Waterbirds	4,795	599	2,048	0.07 ± 0.00
CelebA	162,770	8,717	2,048	4.52 ± 0.03
CivilComments	269,038	7,111	768	4.31 ± 0.00

Comparison with SELF. To assess POTER's computational overhead relative to last-layer retraining, we compare it with SELF [24] on Waterbirds. POTER requires a single model-training stage: reference construction, OT computation, and weight transformation involve no additional model training. Both pipelines include feature extraction and full-model training, so we compare their additional method-specific operations. Table 16 reports the runtime of POTER's complete weight construction and SELF's last-layer retraining. POTER's complete weight construction takes 0.075 s, while SELF's last-layer retraining takes 5.657 s. Although both overheads are small relative to full-model training, POTER requires less additional computation and retains $87.9 \pm 1.6\%$ worst-group accuracy under 20% symmetric label noise, compared with $14.7 \pm 0.2\%$ for SELF. This contrast highlights the fragility of loss-based sample selection under label noise and the robustness of POTER's loss-independent reweighting.

Table 16: Comparison of POTER and SELF on Waterbirds: additional computation and robustness to label noise. Worst-group accuracy is measured under 20% symmetric label noise.

Method	Timed operation	Time (s)	Worst-group (%)
SELF	Last-layer retraining	5.657	14.7 ± 0.2
POTER	Complete weight construction	0.075	87.9 ± 1.6

H Relationship to Prior Work

Relationship to OT-based data valuation. POTER builds on the sensitivity interpretation of OT dual potentials used in prior work such as LAVA [15]: the training-side potential describes how the transport discrepancy changes with the mass assigned to an individual training sample. LAVA uses scores derived from these potentials to value training samples relative to a validation distribution, with applications including the detection of noisy, mislabeled, and poisoned samples. POTER adopts this interpretation and constructs a reference distribution from group-annotated validation data to encode a desired group composition, with greater emphasis on bias-conflicting groups. Under fixed transport marginals, samples in an overrepresented training region compete for limited reference mass. The class-conditioned potentials can therefore be interpreted as reflecting local excess or deficit of training mass relative to this deliberately shifted reference. In this view, even a correctly labeled sample can receive a lower weight when its region is overrepresented relative to the reference. We convert these potentials into sample weights through the KL-regularized reweighting rule in Proposition 4.1 and use the resulting weights for model training. This formulation allows the same weighting rule to emphasize bias-conflicting samples and downweight mislabeled samples that are misaligned with the reference for their observed class, as illustrated in Figure 1. The central contribution is to develop this

view of sample importance into a loss-independent reweighting framework for supervised learning: class-conditioned OT dual potentials provide sample-level signals of relative mass concentration and determine each sample's contribution to the training objective. This directly connects the allocation of training mass to robust learning under spurious correlations and label noise.

Relationship to reweighting methods for robust learning. Group DRO [35] optimizes group losses using a shared weighting coefficient for samples within each group. Methods such as SSA [29] and GIC [12] first infer group labels and then use these labels in a group-robust training procedure. When combined with Group DRO, their weighting likewise operates at the group level and does not distinguish individual samples within a group. POTER retains sample-level variation within each group, as illustrated in Figure 1 and the accompanying analysis of the sample weights. A different family of methods uses a preliminary model to determine which samples to emphasize: AFR [34] and LEIA [11] assign greater importance to examples on which the model performs poorly, while SELF [24] selects examples using misclassification or disagreement between models. Criteria based on high loss or misclassification may confound bias-conflicting examples with mislabeled ones. The performance degradation under label noise in Tables 2, 8, and 16 is consistent with this limitation. POTER instead derives individual sample weights from class-conditioned reference alignment, without requiring training-set group annotations or losses from a preliminary classifier. It thus combines sample-level differentiation with loss-independent weighting for robust learning.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract and introduction state that POTER addresses spurious correlations and label noise via OT-based reweighting, and the claims are supported by the theoretical formulation and experiments in the main text and appendix.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: The conclusion discusses the dependence of POTER on the quality of the validation set used to construct the reference distribution and identifies noisy validation supervision as a future direction.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

Answer: [Yes]

Justification: The assumptions and proofs for the class-conditioned OT decomposition and KL-regularized reweighting rule are provided in the appendix.

Guidelines:

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

4. **Experimental result reproducibility**

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

Answer: [Yes]

Justification: The appendix provides dataset details, model architectures, feature extraction, reference distribution construction, OT computation, reweighting hyperparameters, and baseline implementation details.

Guidelines:

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

some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

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

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

   Answer: [Yes]

   Justification: The submitted code includes instructions for dataset preparation, feature extraction, running POTER, ERM baselines, and label-noise experiments.

   Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: The appendix reports dataset splits, optimizers, learning rates, weight decay, batch sizes, epochs, data augmentations, model selection criteria, and hyperparameter search ranges.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [Yes]

   Justification: Experimental results are reported with mean and standard deviation over three random seeds, and runtime results are averaged over five runs.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: The appendix states that experiments were run on a single NVIDIA GeForce RTX 3090 GPU.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: We have reviewed the NeurIPS Code of Ethics and found no ethical concerns in the data usage, experimental design, or reporting of this work.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [Yes]

Justification: The work has potential positive societal impact by improving robustness for underrepresented groups under subpopulation shift, and we do not identify any direct negative societal impacts specific to this work.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

Answer: [N/A]

Justification: The paper does not release new high-risk datasets, generative models, or pretrained models with substantial misuse potential.

Guidelines:

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

12. **Licenses for existing assets**

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

Answer: [Yes]

Justification: The paper cites the original sources for all datasets, pretrained models, and baseline methods used in the experiments.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.

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

13. **New assets**

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

    Answer: [Yes]

    Justification: The released code is documented with a README describing dependencies, dataset preparation, feature extraction, training commands, label-noise experiments, and file structure.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing experiments or new research with human subjects.

    Guidelines:

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

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

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

Answer: [N/A]

Justification: The paper does not involve new human-subject data collection or interventions.

Guidelines:

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

16. **Declaration of LLM usage**

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

Answer: [N/A]

Justification: LLMs are not used as an important, original, or non-standard component of the proposed method or experiments.

Guidelines:

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