Personalized Federated Learning with Spurious Features: An Adversarial Approach

Anonymous authors
Paper under double-blind review

Abstract

One of the common approaches for personalizing federated learning is fine-tuning the global model for each local client. While this addresses some issues of statistical heterogeneity, we find that such personalization methods are vulnerable to spurious features at local agents, leading to reduced generalization performance. This work considers a setup where spurious features correlate with the label in each client's training environment, and the mixture of multiple training environments (i.e., the global environment) diminishes the spurious correlations. While the global federated learning model trained over the global environment suffers less from spurious features, the local fine-tuning step may lead to personalized models vulnerable to spurious correlations. In light of this practical and pressing challenge, we propose a novel strategy to mitigate the effect of spurious features during personalization by maintaining the adversarial transferability between the global and personalized models. Empirical results on object and action recognition tasks show that our proposed approach bounds personalized models from further exploiting spurious features while preserving the benefit of enhanced accuracy from fine-tuning.

1 Introduction

Federated learning (FL) is a leading framework for clients to collaboratively train a shared global machine learning (ML) model without releasing their local private datasets (McMahan et al., 2017; Kairouz et al., 2019). The jointly trained global model can be further fine-tuned on each client's local dataset to produce personalized models (Fallah et al., 2020; T. Dinh et al., 2020; Li et al., 2021). While existing theoretical and empirical results highlight how personalized models improve accuracy on local data, few works consider what features the personalized models learn from the local dataset. Our motivating hypothesis is that *, not all local features are beneficial*.

Spurious features, which correlate with labels in some environments but do not generalize across all environments, are ubiquitous (Geirhos et al., 2020; Singla & Feizi, 2022). For example, for activity recognition, a rocky cliff can correlate with climbing activity (Nam et al., 2020); for classification, flowers can be a spurious feature in butterfly images (Singla & Feizi, 2022); for recognition, hair color may confound the gender of face images (Liu et al., 2021). These examples are illustrated with images in Appendix A. In these examples, the predictive performance of a model that entangles spurious features can degrade in environments where the spurious correlation no longer holds (e.g., for the activity detection example, consider climbing on icy walls). Such non-uniform performance across environments due to the entanglement of spurious features has led to issues in robustness and fairness, among others (He et al., 2019; Nam et al., 2020; Liu et al., 2021).

This paper considers a setup where each user holds training samples from their own environment but may move to other environments with a deployed model. Such a setup applies to federated learning systems with diverse and heterogeneous clients, who tend to collect training samples around their current environment and may later travel with their mobile devices. Examples include mobile augmented reality applications (Cao et al., 2022), where client-side distributions change temporally and spatially. In such cases, although spurious features correlate with the label on each client's *training environment*, aggregating the clients over different environments (i.e., the *global environment*) contributes to mitigating the spurious correlations.

Table 1: Requirement comparison. Our approach limits personalization from increasing the entanglement level to spurious features without requiring access to multiple environments or their indexes.

Method	Environments	Environment Indexes
Invariant risk minimization (Arjovsky et al., 2019)	Must-have	Must-have
Distributionally robust optimization (Sagawa et al., 2020)	Must-have	Must-have
Just-train-twice (JTT) (Liu et al., 2021)	Must-have	Need-not
Ours	Need-not	Need-not

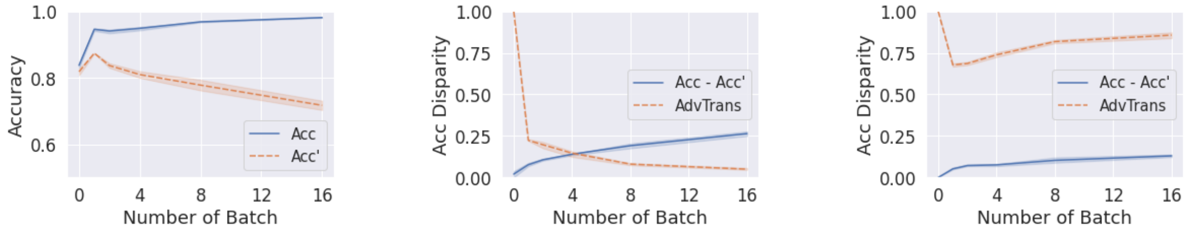
In FL, even if the spurious correlations in the global environment can be diminished, naive local fine-tuning surely increases the entanglement level of personalized models to spurious features of their local training environments. Thus, learning a personalized model for each client without risking entanglement to spurious features is non-trivial. Common methods that aim to disentangle spurious features from models (Wang et al., 2019; Sagawa et al., 2020; Liu et al., 2021; Wang et al., 2022) require full access to multiple environments and are, therefore, not applicable to our federated setting. On the other hand, although using the off-the-shelf global model in local agents without fine-tuning avoids further spurious correlations, the performance on each local client's dataset can be sub-optimal.

To this end, we propose a novel method to limit personalization from increasing the entanglement level to spurious features without requiring access to multiple environments (Table 1) and subsequently improve the generalization performance of personalized models measured by the *accuracy disparity* (Zhao & Gordon, 2022)–the difference between a model's accuracies across environments. We call the difference between the entanglement level of the global and personalized models *entanglement deviation*. Then, our method uses the *adversarial transferability* between the global and personalized models as a proxy to bound the entanglement deviation. The adversarial transferability is measured by the percentage of adversarial examples generated by the global model that also "fool" personalized models. The intuition is that if two models entangle features in the same way, the adversarial examples are more likely to transfer from one to another and vice versa. Based on this intuition, we propose the following hypothesis:

> *If personalized models increase their entanglement to spurious features, fewer adversarial examples generated by the global model transfer to personalized models.*

Note that our approach does not require entangling spurious features to be an exclusive cause of decreasing the adversarial transferability and remains effective as long as such a connection exists. Empirically, we observe that the adversarial transferability between the global and personalized models often decreases when personalized models entangle spurious features. In such cases, the entanglement deviation will increase (Section 3), validating our hypothesis. However, naively including adversarial examples with flipped labels (i.e., mispredictions) to fine-tune and maintain the adversarial transferability is sub-optimal in bounding the entanglement deviation (Section 3.3). Therefore, we theoretically analyze the connection between the adversarial transferability and the entanglement deviation of the global and personalized models. In addition, we show that the disparity can be bounded by the entanglement and present conditions under which the disparity upper bound of personalized models can be close to that of the global model (Section 4). Based on the theoretical results, we develop an improved method to bound the entanglement deviation and the accuracy disparity of personalized models (Section 5), thus improving the robustness of personalized models to local spurious features. Our main contributions are:

- We empirically evaluate the vulnerability of personalized models to spurious features in a federated learning setting, highlighting a critical risk of existing personalization methods.
- We theoretically connect the adversarial transferability and the entanglement levels of the global and personalized models to spurious features.
- We develop a method to bound the increased entanglement of personalized models to spurious features by maintaining the adversarial transferability between the global and personalized models.



(a) Accuracy disparity.

(b) Accuracy disparity correlates with adversarial transferability.

(c) Naively maintaining high adversarial transferability does not help.

Figure 1: An empirical study to evaluate the transferability hypotheses with spurious features. (a) The accuracy of personalized models in the training environments (Acc) and the other environments that are not accessible during training (Acc′) with increasing fine-tuning batches. The personalized models entangle the spurious feature and increase accuracy disparities between environments. (b) As the personalized models entangle the spurious feature and increase their accuracy disparity (Acc - Acc′), adversarial transferability (AdvTrans) decreases. (c) Naively maintaining high adversarial transferability (AdvTrans) helps mitigate the accuracy disparity (Acc - Acc′), but the accuracy disparity still increases from 0.06 to 0.12 along with the increasing adversarial transferability.

We conduct extensive experiments to validate the effectiveness of the proposed methods under FL settings. Our experiments on MNIST (Deng, 2012), Coil20 (Nene et al., 1996), CelebA (Liu et al., 2015; Caldas et al., 2018), and biased action recognition (BAR) (Nam et al., 2020) datasets show that the proposed approach reduces the accuracy disparity of personalized models from 18.38% to 3.42%. Our method also preserves the benefit of the enhanced average accuracy from fine-tuning, resulting in 4.48% accuracy improvement in the global environment.

2 Related Work

Personalized Federated Learning. Fine-tuning is typical for personalizing FL. The meta-learning-based method first trains a global model and fine-tunes the global model locally (Fallah et al., 2020). Other methods using multi-task learning (Li et al., 2021) or Moreau envelopes (T. Dinh et al., 2020) have an interpretation as fine-tuning the local model along with training the global model. Fine-tuning is also compatible with clustering-based methods (Ghosh et al., 2020). This work considers homogeneous model-based personalization. Other approaches using heterogeneous model architectures (Shamsian et al., 2021; Marfoq et al., 2022) are out of our scope but interesting for further work.

Debiasing Machine Learning Models. Debiasing is a way to disentangle spurious features from the model. A few prior works (Li & Vasconcelos, 2019; Sagawa et al., 2020; Wang et al., 2022) utilize group labels, which might require human annotation, to debias ML models. Residual learning-based methods (He et al., 2019; Nam et al., 2020; Liu et al., 2021) train a biased ML model and up-weight the residual, which mainly contains samples from under-represented environments that the biased ML model mis-predicts. Chi et al. (2021) aims to mitigate the accuracy disparity in regression problems via learning the appropriate representations. However, all these methods rely on explicit access to multiple environments (Table 1).

Adversarial Transferability. Prior works (Tramèr et al., 2017; Charles et al., 2019) attempt to understand adversarial transferability. A recent work connects adversarial transferability to knowledge transferability in a transfer learning setting (Liang et al., 2021). Our paper builds on this work to show how to solve a new problem–bounding the personalized models' entanglement to spurious features–by maintaining adversarial transferability between the global and personalized models.

3 An Empirical Study with Spurious Features

We first perform an empirical study on the entanglement deviation between the global and personalized models in an FL setting to gain some insights into the problem. This section focuses on the MNIST dataset due to the limited space, and Appendix E further provides results on the other datasets. In this study, the personalization method is fine-tuning. Our results highlight the risk of existing fine-tuning-based personalization methods and the difficulty of mitigating the risk. We also highlight the correlation between adversarial transferability and entanglement deviation. We provide additional theoretical analysis in Section 4 for more insights into the observed correlation. The client setup and evaluation protocol in this empirical study are as follows.

Client Setup. There are two environments whose combination is the global environment. Each client gets a different set of grayscale MNIST digits, then creates a spurious correlation by coloring the grayscale objects according to their labels differently for each environment (e.g., the red color correlates with digit 0 in the first environment and with digit 1 in the second environment). Then, each client picks a random environment for training, and all clients collaboratively train a global model. After the global model converges, clients further fine-tune the global model in their training environments.

Evaluation Protocol. We evaluate the accuracy of personalized models in their training environments (Acc) and the other environments that are not accessible during training (Acc′) with an increasing number of fine-tuning steps (i.e., batches). The accuracy disparity indicates entanglement to spurious features, defined as the accuracy difference between the training environment and the environment that is not accessible during training (i.e., Acc - Acc′).

3.1 Personalization May Exacerbate Entanglement

As can be observed in Figure 1a, the accuracy disparity of personalized models increases within a few fine-tuning batches, deviating from that of the global model. This observation indicates that personalized models gradually increase the entanglement to the spurious feature. Although, in principle, one may resort to early stopping, this is not feasible using a single training environment alone.

3.2 Transferability Correlates with Entanglement

Because only the training environment is accessible during personalization, it is infeasible to directly measure personalized models' entanglement or accuracy disparity. To this end, in this section, we focus on methods that implicitly measure and determine to which level personalized models can entangle the spurious feature. Following our hypothesis in Section 1, we consider using the adversarial transferability between the global and personalized models as a proxy for measuring their entanglement deviation.

Adversarial Attack An adversarial attack aims to add a calibrated adversarial perturbation to the input to mislead a model's prediction. A perturbed input is an adversarial example. A common way to calibrate the adversarial perturbation is leveraging the model's first-order gradient w.r.t. the input (Madry et al., 2018; Miyato et al., 2018; Liang et al., 2021). Such a first-order gradient suggests a direction along which the prediction changes significantly. In our setting, the first-order gradient shall direct more perturbations to robust features to mislead the global model, which entangles spurious features less. In contrast, if personalized models entangle the spurious features more, they are more likely to resist the attack that targets the global model by using the spurious features with fewer perturbations. Following this intuition, we employ the projected gradient descent (PGD) attack (Madry et al., 2018) that iteratively calibrates the adversarial perturbation using the first-order gradient. Specifically, at a point $\boldsymbol{x}$ with label y, for the global model f_g with loss function ℓ, at iteration $t+1$, the adversarial example is:

$$\boldsymbol{x}_{\text{adv}}^{t+1} = \text{Proj}_{\|\boldsymbol{x}_{\text{adv}}-\boldsymbol{x}\|\leq\epsilon}(\boldsymbol{x}_{\text{adv}}^{t} + \alpha \cdot \text{sign}(\nabla_{\boldsymbol{x}_{\text{adv}}^{t}} \ell(f_g(\boldsymbol{x}_{\text{adv}}^{t}), y))),$$

where Proj is a projection operator, $\boldsymbol{x}_{\text{adv}}^{0} = \boldsymbol{x}$, ϵ is the attack budget, and α is the attack step size.

Adversarial Transferability We generate adversarial examples using the global model and collect the ones that cause the global model to mispredict. Then, personalized models make predictions on the collected adversarial examples. The adversarial transferability is defined as the percentage of the collected adversarial examples that also cause personalized models to mispredict. Recalling that the collected adversarial examples embed more perturbation on robust features, personalized models may resist the adversarial examples if they use spurious features to predict.

Figure 1b plots the accuracy disparity and the adversarial transferability during fine-tuning. As personalized models increase their entanglement to spurious features, the accuracy disparity of the personalized model increases and deviates from that of the global model. Then, the adversarial transferability between the global and personalized models decreases. This result empirically validates our hypothesis.

3.3 Can Maintaining Transferability Mitigate Entanglement Deviation?

Following the empirical observation of the adversarial transferability and the entanglement deviation, we add the collected adversarial examples with flipped labels (i.e., mispredictions) to the training set during personalization, aiming to maintain the adversarial transferability. However, Figure 1c shows that the accuracy disparity still doubles (0.06 to 0.12), indicating that the entanglement deviation increases, even if the adversarial transferability increases. The following section presents a theoretical analysis, outlining conditions under which maintaining the adversarial transferability helps bound the entanglement deviation.

4 Theoretical Insights

In this section, we present a theoretical analysis that supports our hypothesis in Section 1, the experimental results in Section 3.2, and analyze the failure in Section 3.3. Before we proceed to the detailed analysis, some additional definitions and notations are needed for the presentation (we provide a table summarizing all the notations used in Appendix B to ease the reading).

4.1 Setting

Data Model. A data sample $\boldsymbol{x} = \mathrm{Cat}(\boldsymbol{x}_r, \boldsymbol{x}_s)$ is a concatenation of robust features $\boldsymbol{x}_r$ (i.e., $\dim(\boldsymbol{x}_r) \geq 2$) and a spurious feature $\boldsymbol{x}_s$. The label is binary, $y \in \{0, 1\}$.

Environment. We consider a setting with two environments.[1] On the i^{th} client, $\mathcal{D}_i$ denotes the joint distribution of $\boldsymbol{x}$ and y in the training environment and $\mathcal{D}'_i$ denotes the future environment of the i-th client that is not accessible during training. In addition, we use $\mathcal{D}_{i,r}$ to denote the joint distribution of $\boldsymbol{x}_r$ and y and use $\mathcal{D}_{i,s}$ to denote the joint distribution of $\boldsymbol{x}_s$ and y. Our analysis applies to each client's $\mathcal{D}$ and $\mathcal{D}'$, and we shall omit the subscripts i when it is clear from the context.

Distribution Shift. Our analysis focuses on conditional distribution shift caused by spurious correlations, which imply $\mathcal{P}_{\mathcal{D}}(\boldsymbol{x}_s \mid y) \neq \mathcal{P}_{\mathcal{D}'}(\boldsymbol{x}_s \mid y)$ but $\mathcal{P}_{\mathcal{D}}(\boldsymbol{x}_r \mid y) = \mathcal{P}_{\mathcal{D}'}(\boldsymbol{x}_r \mid y)$ and $\mathcal{P}_{\mathcal{D}}(y) = \mathcal{P}_{\mathcal{D}'}(y)$.

Data Generating Process. We consider a data-generating process that produces a pair of samples with four steps: (1) draw a label y according to $\mathcal{P}_{\mathcal{D}}(y)$, (2) draw a robust feature $\boldsymbol{x}_r$ under $\mathcal{P}_{\mathcal{D}}(\boldsymbol{x}_r \mid y)$, (3) draw a spurious feature according to $\mathcal{P}_{\mathcal{D}}(\boldsymbol{x}_s \mid y)$ and concatenate features $\boldsymbol{x} = \mathrm{Cat}(\boldsymbol{x}_r, \boldsymbol{x}_s)$, and (4) draw a spurious feature with $\mathcal{P}_{\mathcal{D}'}(\boldsymbol{x}_s \mid y)$ and concatenate features $\boldsymbol{x}' = \mathrm{Cat}(\boldsymbol{x}_r, \boldsymbol{x}'_s)$. By the linearity of expectation, the four-step data-generating process does not affect the loss disparity:

$$|\mathbb{E}_{\mathcal{D}}[\ell(f(\boldsymbol{x}), y)] - \mathbb{E}_{\mathcal{D}'}[\ell(f(\boldsymbol{x}), y)]| = |\mathbb{E}_{\mathcal{D},\mathcal{D}'}[\ell(f(\boldsymbol{x}), y) - \ell(f(\boldsymbol{x}'), y)]|. \tag{1}$$

---

[1] Our analysis could be generalized to more than two environments in a straightforward way.

Hypothesis Class. We consider a logistic regression model $f(\boldsymbol{x}) = \sigma(\boldsymbol{w}^\top \boldsymbol{x})$ with a Sigmoid activation function σ, and a logistic loss ℓ. Similar to the notations in the data model, we let $\boldsymbol{w} = \text{Cat}(\boldsymbol{w}_r, \boldsymbol{w}_s)$, where $\boldsymbol{w}_r$ includes the weights on robust features $\boldsymbol{x}_r$ and $\boldsymbol{w}_s$ denotes the weight of spurious feature $\boldsymbol{x}_s$. f_g denotes the global model with weight $\boldsymbol{w}_g$ and f_l denotes the personalized model with weight $\boldsymbol{w}_l$. $\boldsymbol{w}_{g,s}$ denotes the global model's weight of the spurious feature.

4.2 Preliminaries

Entanglement. For a given model f with weight $\boldsymbol{w}$, its entanglement to the spurious feature $\|\boldsymbol{x}_s\|$ is defined as $\boldsymbol{w}_s$. The entanglement $\|\boldsymbol{w}_s\|$ can be further decomposed into a norm term and an angle term: $\|\boldsymbol{w}_s\| = \|\boldsymbol{w}\| \cos\theta$, where θ is the angle between $\boldsymbol{w}$ and $\text{Cat}(\boldsymbol{0}, \boldsymbol{w}_s)$. Such a definition of θ implies that $\theta \in [0, \frac{\pi}{2}]$ and $\cos\theta \in [0, 1]$.

Angles Let $\theta' = \arccos \frac{\boldsymbol{w}_g \cdot \boldsymbol{w}_l}{\|\boldsymbol{w}_g\| \|\boldsymbol{w}_l\|}$ be the angle between a global model's weight $\boldsymbol{w}_g$ and a personalized model's weight $\boldsymbol{w}_l$, θ_g be the angle between $\boldsymbol{w}_g$ and $\text{Cat}(\boldsymbol{0}, \boldsymbol{w}_{g,s})$, and θ_l be the angle between $\boldsymbol{w}_l$ and $\text{Cat}(\boldsymbol{0}, \boldsymbol{w}_{l,s})$.

Loss Disparity The loss disparity of a model f is defined as: $\mathbb{E}_{\mathcal{D}'}[\ell(f(\boldsymbol{x}), y)] - \mathbb{E}_{\mathcal{D}}[\ell(f(\boldsymbol{x}), y)]$, quantifying the generalization performance of a model across environments.

Adversarial Perturbation. The adversarial perturbation $\boldsymbol{\delta}_{f,\epsilon} = \boldsymbol{x}_{\text{adv}} - \boldsymbol{x}$ is generated using a model f with budget ϵ. A common way to generate adversarial examples is solving a maximization problem: $\boldsymbol{x}_{\text{adv}} = \arg\max_{\|\boldsymbol{x}^* - \boldsymbol{x}\| \leq \epsilon} \ell(f(\boldsymbol{x}^*), y)$, which returns a data sample that maximizes the model's loss and likely causes a misprediction. Plugging the definition of $\boldsymbol{\delta}_{f,\epsilon}$ into $\boldsymbol{x}_{\text{adv}} = \arg\max_{\|\boldsymbol{x}^* - \boldsymbol{x}\| \leq \epsilon} \ell(f(\boldsymbol{x}^*), y)$, we have $\boldsymbol{\delta}_{f,\epsilon} = \arg\max_{\|\boldsymbol{\delta}\| \leq \epsilon} \ell(f(\boldsymbol{x} + \boldsymbol{\delta}), y)$. With a small budget ϵ, we can approximate the loss function ℓ using the first-order gradient (Miyato et al., 2018; Liang et al., 2021):

$$\boldsymbol{\delta}_{f,\epsilon} = \arg\max_{\|\boldsymbol{\delta}\| \leq \epsilon} \nabla_{\boldsymbol{x}} \ell(f(\boldsymbol{x}), y)^\top \boldsymbol{\delta} = \epsilon \cdot \frac{\nabla_{\boldsymbol{x}} \ell(f(\boldsymbol{x}), y)}{\|\nabla_{\boldsymbol{x}} \ell(f(\boldsymbol{x}), y)\|}. \tag{2}$$

Adversarial Transferability. With the adversarial perturbation, we define the adversarial transferability loss:

$$\ell_{g \to l}(f_g, f_l, \boldsymbol{x}, y) = \Big(\ell(f_g(\boldsymbol{x} + \boldsymbol{\delta}_{f_g,\epsilon}), y) - \ell(f_g(\boldsymbol{x}), y)\Big) - \Big(\ell(f_l(\boldsymbol{x} + \boldsymbol{\delta}_{f_g,\epsilon}), y) - \ell(f_l(\boldsymbol{x}), y)\Big), \tag{3}$$

measuring how effective the adversarial perturbation $\boldsymbol{\delta}_{f_g,\epsilon}$ generated by the global model f_g is after applying to the personalized model f_l. For a logistic regression model, its first-order gradient at $\boldsymbol{x}$ is $\Big(\sigma(\boldsymbol{w}^\top \boldsymbol{x}) - y\Big)\boldsymbol{w}$. Here, we can see that the gradient direction depends on the weight $\boldsymbol{w}$ at a given data sample. This allows us to connect the following adversarial transferability to entanglement using the angles between weights.

4.3 An Upper Bound of the Loss Disparity

Disparity Upper Bound. To show the necessity of reducing the entanglement to the spurious feature, we connect the entanglement to the spurious feature and the loss disparity, an empirical metric that quantifies a model's generalization performance across environments. The following theorem suggests that the loss disparity of a model is upper bounded by its entanglement to spurious features, quantified by the angle θ between $\boldsymbol{w}$ and $\text{Cat}(\boldsymbol{0}, \boldsymbol{w}_s)$.

Theorem 1. *Under the setting in Section 4.1, for a model f parameterized by $\boldsymbol{w}$, assume the composition $\ell \circ \sigma$ of the activation function σ and the loss function ℓ is ρ-Lipschitz, let θ be the angle between $\boldsymbol{w}$ and $\text{Cat}(\boldsymbol{0}, \boldsymbol{w}_s)$, we have:*

$$|\mathbb{E}_{\mathcal{D}}[\ell(f(\boldsymbol{x}), y)] - \mathbb{E}_{\mathcal{D}'}[\ell(f(\boldsymbol{x}), y)]| \leq \rho \|\boldsymbol{w}_s\|_2 \cdot W_1(\mathcal{D}_s, \mathcal{D}'_s) = \rho \|\boldsymbol{w}\|_2 \cos\theta \cdot W_1(\mathcal{D}_s, \mathcal{D}'_s). \tag{4}$$

Note that $\cos\theta \in [0,1]$ by its definition. The bound in Equation 4 is intuitive. With the setting in Section 4.1, the distribution shift between $\mathcal{D}$ and $\mathcal{D}'$ applies to the spurious feature $\boldsymbol{x}_s$. For a given distribution shift that can be quantified by $W_1(\mathcal{D}_s, \mathcal{D}'_s)$, its impact on the loss disparity can be further amplified by up to the product of the model weight $\|\boldsymbol{w}_s\| = \|\boldsymbol{w}\|\cos\theta$ and the ρ-Lipschitz function composition $\ell \circ \sigma$.

Disparity Deviation Upper Bound Following the result in Theorem 1, we further present an upper bound on the disparity deviation between the global and local models. Our upper bound includes the three angles that are defined in Section 4.2: (1) θ_g, (2) θ_l, and (3) θ', where the first two angles measure how much the global and personalized models entangle the spurious feature and the third one relates to the entanglement upper bound deviation between the two models.

Corollary 1. *Under the setting in Section 4.1 and the definitions in Section 4.2, with Theorem 4, we have:*

$$\begin{aligned}&\Big| |\mathbb{E}_{\mathcal{D}}[\ell(f_g(\boldsymbol{x}), y)] - \mathbb{E}_{\mathcal{D}'}[\ell(f_g(\boldsymbol{x}), y)]| - |\mathbb{E}_{\mathcal{D}}[\ell(f_l(\boldsymbol{x}), y)] - \mathbb{E}_{\mathcal{D}'}[\ell(f_l(\boldsymbol{x}), y)]| \Big| \\ &\quad \leq \rho \cdot W_1(\mathcal{D}_s, \mathcal{D}'_s)\Big(\|\boldsymbol{w}_g\|\cos\theta_g + \|\boldsymbol{w}_l\|\cos\theta_l\Big)\end{aligned} \tag{5}$$

In addition, if $\theta' \leq \theta_g$, we have

$$\begin{aligned}&\Big| |\mathbb{E}_{\mathcal{D}}[\ell(f_g(\boldsymbol{x}), y)] - \mathbb{E}_{\mathcal{D}'}[\ell(f_g(\boldsymbol{x}), y)]| - |\mathbb{E}_{\mathcal{D}}[\ell(f_l(\boldsymbol{x}), y)] - \mathbb{E}_{\mathcal{D}'}[\ell(f_l(\boldsymbol{x}), y)]| \Big| \\ &\quad \leq \rho \cdot W_1(\mathcal{D}_s, \mathcal{D}'_s)\Big(\|\boldsymbol{w}_g\|\cos\theta_g + \|\boldsymbol{w}_l\|\cos(\theta_g - \theta')\Big)\end{aligned} \tag{6}$$

Otherwise, if $\theta' > \theta_g$, we have:

$$\begin{aligned}&\Big| |\mathbb{E}_{\mathcal{D}}[\ell(f_g(\boldsymbol{x}), y)] - \mathbb{E}_{\mathcal{D}'}[\ell(f_g(\boldsymbol{x}), y)]| - |\mathbb{E}_{\mathcal{D}}[\ell(f_l(\boldsymbol{x}), y)] - \mathbb{E}_{\mathcal{D}'}[\ell(f_l(\boldsymbol{x}), y)]| \Big| \\ &\quad \leq \rho \cdot W_1(\mathcal{D}_s, \mathcal{D}'_s)\Big(\|\boldsymbol{w}_g\|\cos\theta_g + \|\boldsymbol{w}_l\|\Big)\end{aligned} \tag{7}$$

Corollary 1 shows the benefit of reducing θ' for decreasing the disparity deviation. Later, we will show that maintaining adversarial transferability helps reduce θ'. The intuition behind Equations equation 6 and equation 7 is straightforward: in a higher dimensional setting (e.g., $\dim(\boldsymbol{x}_r) \geq 2$ in Section 4.1), the angle θ_l that quantifies the entanglement of the local model f_l to the spurious feature $\boldsymbol{x}_s$ falls within an interval $[\theta_g - \theta', \theta_g + \theta']$. If $\theta' > \theta_g$, the angle θ_l is possibly 0 all the time, entangling the local model f_l to the spurious feature. Furthermore, we present a directed analysis of the loss disparity of a local model, showing that reducing θ' helps align the loss disparity upper bound of a local model f_l to the global model f_g.

Corollary 2. *Under the setting in Section 4.1 and the definitions in Section 4.2, with Theorem 1, if $\theta' \leq \theta_g$, we have*

$$\begin{aligned}|\mathbb{E}_{\mathcal{D}}[\ell(f_l(\boldsymbol{x}), y)] - \mathbb{E}_{\mathcal{D}'}[\ell(f_l(\boldsymbol{x}), y)]| &\leq \rho\|\boldsymbol{w}_{l,s}\| \cdot W_1(\mathcal{D}_s, \mathcal{D}'_s) \\ &= \rho\|\boldsymbol{w}_l\|\cos\theta_l \cdot W_1(\mathcal{D}_s, \mathcal{D}'_s) \\ &\leq \rho\|\boldsymbol{w}_l\|\cos(\theta_g - \theta') \cdot W_1(\mathcal{D}_s, \mathcal{D}'_s).\end{aligned} \tag{8}$$

Otherwise, if $\theta' > \theta_g$, we have:

$$|\mathbb{E}_{\mathcal{D}}[\ell(f_l(\boldsymbol{x}), y)] - \mathbb{E}_{\mathcal{D}'}[\ell(f_l(\boldsymbol{x}), y)]| \leq \rho\|\boldsymbol{w}_l\| \cdot W_1(\mathcal{D}_s, \mathcal{D}'_s). \tag{9}$$

Adversarial Transferability. In the following theorem, we show that the adversarial transferability between the global and personalized models can help estimate an upper bound of the angle θ' between $\boldsymbol{w}_g$ and $\boldsymbol{w}_l$. Then, maximizing the adversarial transferability can reduce an upper bound of the angle θ' and the disparity deviation upper bound (Theorem 1). In addition to the adversarial transferability loss $\ell_{g \to l}$ defined in Section 4.2, the upper bound also includes a gradient norm term $\bigcirc$ from the personalized model, a constant $\Box$, and a second-order Taylor remainder term $\diamond$.

Theorem 2. *Under the setting in Section 4.1, let ϵ be the attack budget, and θ' be the angle between $\boldsymbol{w}_g$ and $\boldsymbol{w}_l$ we have:*

$$\theta' \leq \arccos\Big(\frac{1}{\epsilon \cdot \bigcirc} + \Box + \diamond + \mathbb{E}_{\mathcal{D}}[\ell_{g\to l}(f_g, f_l, \boldsymbol{x}, y)]\Big), \tag{10}$$

where $\bigcirc = \mathbb{E}_{\mathcal{D}}[\|\nabla_{\boldsymbol{x}}\ell(f_l(\boldsymbol{x}), y)\|]$, $\Box = \epsilon\mathbb{E}_{\mathcal{D}}[\|\nabla_{\boldsymbol{x}}\ell(f_g(\boldsymbol{x}), y)\|]$, and $\diamond = \mathbb{E}_{\mathcal{D}}[R_{g,2}(\boldsymbol{x}_{\text{adv}}) - R_{l,2}(\boldsymbol{x}_{\text{adv}})]$ is the expected difference of second-order Taylor remainders between f_g and f_l.

Necessary Condition. Since the connection between the adversarial transferability loss and the entanglement deviation is angle-based, various norm terms may corrupt such a connection and cause the sub-optimal result in Section 3.3. For example, increasing the gradient norm term $\bigcirc$ of the personalized model f_l may increase θ' with a fixed adversarial transferability. Also, directly increasing the weight norm $\|\boldsymbol{w}_l\|$ can increase the entanglement deviation even if $\theta' = 0$. Therefore, we need to stabilize the gradient norm term $\bigcirc$ and the weight norm $\|\boldsymbol{w}_l\|$ throughout the personalization step to make the angle-based connection helpful. The following section lists two practices for maintaining adversarial transferability while stabilizing the aforementioned norm terms.

5 Methods

With the theoretical results in Section 4, we present an improved approach to maintaining the adversarial transferability while stabilizing the norm terms. To this end, we first introduce a regularizer to encourage the global and personalized models to make consistent predictions on adversarial examples. In addition, we add another L_2 regularization term to align the weight of the global and personalized models because similar weights lead to similar norm terms. Appending the two regularizers to the loss function during personalization mitigates the increasing accuracy disparity.

5.1 Maintaining Adversarial Transferability

Generating Adversarial Examples. We employ the same projected gradient descent (PGD) attack (Madry et al., 2018) as Section 3.2, which iteratively perturbs the input using the model's first-order gradient. At iteration $t+1$, the adversarial example under budget ϵ is: $\boldsymbol{x}_{\text{adv}}^{t+1} = \text{Proj}_{\|\boldsymbol{x}_{\text{adv}} - \boldsymbol{x}\| \leq \epsilon}(\boldsymbol{x}_{\text{adv}}^{t} + \alpha \text{sign}(\nabla_{\boldsymbol{x}_{\text{adv}}^{t}} \ell(f_g(\boldsymbol{x}_{\text{adv}}^{t}), y)))$, where Proj is a projection operator and α is the attack step size.

Enforcing Consistent Predictions. Both the global model f_g and the personalized model f_l take the adversarial example $\boldsymbol{x}_{\text{adv}}$ as input and output $\boldsymbol{z}_g$ and $\boldsymbol{z}_l$ from their last layers, respectively. We increase the adversarial transferability by adding the following regularization term, which minimizes the cross-entropy between $\boldsymbol{z}_g$ and $\boldsymbol{z}_l$ after softmax normalization. Since the global model f_g is fixed as a reference in the personalization step, and its low accuracy disparity is desirable, we use $\boldsymbol{z}_g$ as the ground truth. With K classes, we have:

$$R_{\text{adv}}(\boldsymbol{z}_g, \boldsymbol{z}_l) = \sum_{i=1}^{K} \frac{e^{\boldsymbol{z}_{g,i}}}{\sum_{i=1}^{K} e^{\boldsymbol{z}_{g,i}}} \cdot \log\left(\frac{e^{\boldsymbol{z}_{l,i}}}{\sum_{i=1}^{K} e^{\boldsymbol{z}_{l,i}}}\right).$$

The local model has access to the global model, so there is no additional communication overhead for implementing this regularization. The adversarial examples are computed using the global model once and for all. The computation only needs a few back-propagation, much less than training the global model.

5.2 Aligning Weights

In addition, we employ a simple and effective strategy by adding an L_2 regularization term to the loss function:

$$R_{L_2}(\boldsymbol{w}_g, \boldsymbol{w}_l) = \|\boldsymbol{w}_g - \boldsymbol{w}_l\|. \tag{11}$$

The motivation behind the L_2 term is straightforward: if two models have similar weights, they have similar weight norms and gradient norms, as is shown in the following proposition.

Table 2: Worst-case accuracy $\text{Acc}_{\text{worst}}$ and accuracy disparity Acc_{disp} across environments of personalized models. Our proposed method improves the worst-environment accuracy and mitigates the accuracy disparity during personalization.

Method	MNIST		Coil20		CelebA		BAR	
	$\text{Acc}_{\text{worst}}$	Acc_{disp}	$\text{Acc}_{\text{worst}}$	Acc_{disp}	$\text{Acc}_{\text{worst}}$	Acc_{disp}	$\text{Acc}_{\text{worst}}$	Acc_{disp}
Global	.847 ± 6e-4	.005 ± 6e-4	.888 ± 8e-4	.001 ± 1e-3	.906 ± 6e-5	.002 ± 5e-5	.720 ± 4e-5	.002 ± 1e-4
FT	.704 ± 3e-4	.285 ± 3e-4	.804 ± 4e-3	.156 ± 4e-3	.849 ± 1e-3	.114 ± 1e-3	.620 ± 8e-5	.180 ± 2e-4
Ditto	.724 ± 1e-3	.258 ± 1e-3	.952 ± 2e-4	.031 ± 2e-4	.884 ± 2e-4	.082 ± 2e-4	.644 ± 2e-6	.151 ± 2e-6
UW	.823 ± 7e-4	.145 ± 2e-5	.894 ± 4e-4	**.021 ± 4e-4**	.895 ± 3e-4	.041 ± 4e-4	N/A	N/A
JTT	.707 ± 6e-5	.278 ± 6e-5	.836 ± 1e-3	.060 ± 1e-3	.836 ± 2e-4	.120 ± 2e-4	N/A	N/A
Ours	**.870 ± 8e-4**	**.081 ± 2e-5**	**.963 ± 5e-5**	.023 ± 1e-5	**.925 ± 2e-5**	**.002 ± 8e-5**	**.730 ± 2e-4**	**.031 ± 2e-4**

Each experiment is repeated 9 times with 3 random seeds for the federated learning step and 3 random seeds for the personalization step.

Proposition 1. *Under the settings in Section 4.1, assume the composition* $\ell \circ \sigma$ *of the activation function* σ *and the loss function* ℓ *is* ρ*-Lipschitz, we have:*

$$\left| \|\boldsymbol{w}_l\| - \|\boldsymbol{w}_g\| \right| \leq \|\boldsymbol{w}_l - \boldsymbol{w}_g\| \tag{12}$$

and

$$\left| \mathbb{E}_{\mathcal{D}}[\|\nabla_{\boldsymbol{w}_l} \ell(f_l(\boldsymbol{x}), y)\|] - \mathbb{E}_{\mathcal{D}}[\|\nabla_{\boldsymbol{w}_g} \ell(f_g(\boldsymbol{x}), y)\|] \right| \leq \rho \mathbb{E}_{\mathcal{D}}[\|\boldsymbol{x}\|] \|\boldsymbol{w}_l - \boldsymbol{w}_g\|. \tag{13}$$

Although prior works (Li et al., 2020; T. Dinh et al., 2020; Li et al., 2021) have explored similar regularization methods, we develop the regularization term from a different perspective.

6 Experiments

This section presents our experimental results under a federated learning setup with four datasets: MNIST (Deng, 2012), Coil20 (Nene et al., 1996), CelebA (Liu et al., 2015; Caldas et al., 2018), and BAR (Nam et al., 2020), demonstrating that our method mitigates the accuracy disparity. We also show that the benefit of enhanced average accuracy from fine-tuning is preserved. Appendix E further presents a simulation with Gaussian data.

6.1 Setup

Spurious Features The MNIST and Coil20 datasets employ a synthetic spurious feature where the object color correlates with the label. Such a synthetic spurious feature allows precise intervention on the spurious attribute of each data sample without interfering with the robust features, matching the setting in Section 4.1. The spurious feature in the BAR dataset is the action background (e.g., rocky cliff). The hair color is a spurious feature that correlates with the gender label in the CelebA dataset.

Environments We consider a two-environment setup. Each user gets a different set of samples (Caldas et al., 2018), allocates the samples to the two environments, and randomly picks one environment as the training environment. Combining the training environments eliminates the synthetic spurious correlation in MNIST and Coil20 datasets. The realistic spurious correlation in the BAR and CelebA datasets can diminish but does not necessarily disappear in the global environment.

Metrics We consider three metrics in evaluating our approaches: (1) the accuracy disparity Acc_{disp} across environments, (2) the worst-case accuracy $\text{Acc}_{\text{worst}}$ among all environments, and (3) the average accuracy in the mixture of all environments (i.e., the global environment). Metrics (1) and (2) measure the generalization

Table 3: Average accuracy personalized models over different environments. Our proposed method preserves the enhanced accuracy of personalization.

Method	MNIST	Coil20	CelebA	BAR
Global	.849 ± 6e-4	.888 ± 1e-3	.919 ± 5e-5	.721 ± 1e-4
FT	.846 ± 3e-4	.882 ± 4e-3	.906 ± 1e-3	.710 ± 2e-4
Ditto	.853 ± 1e-3	.968 ± 2e-4	.925 ± 2e-4	.719 ± 2e-6
Ours	**.911 ± 2e-5**	**.974 ± 1e-5**	**.926 ± 8e-5**	**.745 ± 2e-4**

performance of personalized models under environment shifts after deployment. Metric (3) quantifies the overall utility of personalized models.

We provide a detailed experimental setup in Appendix D.

6.2 Result and Comparison

There are two naive personalization methods: using the global model (Global) and vanilla fine-tuning (FT). In addition, we employ three recent works that are applicable to the federated learning setting (Section 3). Up-weighting (UW) (Sagawa et al., 2020) method is implemented via sampling biased samples and counterexamples with equal probability. Just-train-twice (JTT) (Sagawa et al., 2020) leverages models that have already entangled spurious features (e.g., personalized models with naive fine-tuning) to identify and up-weight counter-examples without explicit annotations. Ditto (Li et al., 2021) adds a regularizer $\lambda \|\boldsymbol{w}_l - \boldsymbol{w}_g\|$ during personalization to limit the deviation of personalized models from the global model. We select models using the validation accuracy minus the adversarial transferability degradation for ours and using the validation accuracy for others.

6.2.1 Summary of Results

Tables 2 and 3 show the main results with the two metrics on generalization performance in Section 6.1. Our method reduces the accuracy disparity of personalized models from 18.38% to 3.42% compared to vanilla fine-tuning, outperforming other competitors (Table 2). On average, our method also preserves the enhanced average accuracy from fine-tuning, resulting in 4.48% accuracy improvement in global environments compared to the global model (Table 3).

6.2.2 Analysis of Results

One limitation of up-weighting-based methods (i.e., UW and JTT) relates to the diversity of counter-examples, which do not agree with spurious correlations. Counter-examples are few in datasets, leading to low diversity.

To explore the impact of counter-examples' diversity on disentangling spurious features from a model, we vary the diversity of counter-examples and adjust the up-weighting factors accordingly. Specifically, we sample a factor of 0.02, 0.025, 0.033, 0.05, and 0.1 biased samples from the MNIST dataset and re-color them to become counter-examples. The factor in the sampling step is called the "sampling factor". Then, we up-weight the counter-examples by a factor of 50, 40, 30, 20, and 10, respectively, keeping the total number of counter-examples consistent. Here, the counter-examples have less diversity if generated by a small number of biased data samples with a large up-weighting factor. Experimental results in Figure 2 show that, as the diversity reduces, the accuracy disparity of personalized models increases, supporting our hypothesis on the sample diversity. Therefore, our method is applicable in the scarcity of counter-examples while up-weighting-based methods fail.

The issue with Ditto is that greedily maximizing the accuracy in training environments often leads to small λ, giving personalized models spaces to entangle spurious features.

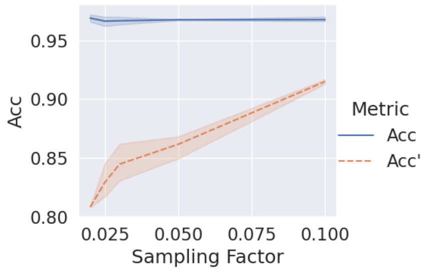
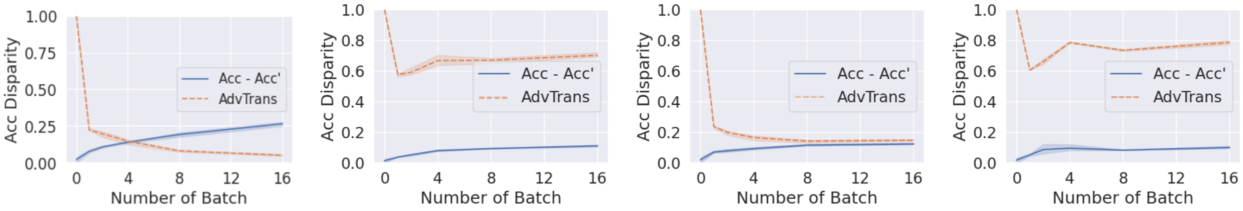


Figure 2: The up-weighting method is less effective with few counter-examples, resulting in a large accuracy disparity of personalized model in the training environment (Acc) and the global environment (Acc′), if the counter-examples are generated by a small number of biased data samples using a small sampling factor and a large up-weighting factor. The up-weighting factor is set to be the reciprocal of the sampling factor.

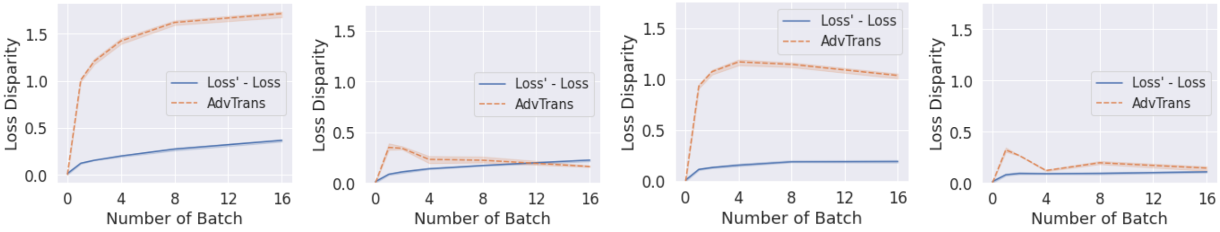
6.3 Ablation Study

We conduct an ablation study on the MNIST dataset and plot the accuracy and loss curves, which are shown to have similar behaviors. Figures 3b and 3f demonstrate that only enforcing consistent predictions on adversarial examples is insufficient. Figure 3c and 3g further show that applying the L_2 regularization term alone does not address the accuracy or loss disparity. Aligning weights of the global and personalized models by applying the L_2 regularization term while maintaining the adversarial transferability address the accuracy and loss disparity as Figures 3d and 3h show, respectively. Both components in our methods are helpful, and a combination of them achieves the best results.

We also explored enforcing the global and personalized models to make consistent predictions on benign samples. However, the accuracy disparity of personalized models increases by 9.66% (19.25% relative increase) with fine-tuning 16 batches. We hypothesize that personalized models could use a combination of robust and spurious features to make the same prediction as the global model on benign samples, which only uses robust features.



(a) Naive Fine-tuning (b) Consistent Prediction (c) Aligning Weights (d) Ours



(e) Naive Fine-tuning (f) Consistent Prediction (g) Aligning Weights (h) Ours

Figure 3: The accuracy disparity (Acc - Acc′) between the training and the global environments, the adversarial transferability accuracy (AdvTrans), the loss disparity (Loss′ - Loss), and adversarial transferability loss (AdvTrans) with different methods. Accuracy and loss disparities have similar behaviors. Combining the two practices in our approach addresses the accuracy and loss disparities.

6.4 Visualizing Gradient Magnitudes

We plot the distribution of the ratio $\frac{\|\nabla_{\boldsymbol{x}}\ell(f_l(\boldsymbol{x}),y)\|}{\|\nabla_{\boldsymbol{x}}\ell(f_g(\boldsymbol{x}),y)\|}$ across data samples with 16 batches of fine-tuning on the MNIST dataset, using the global model as a reference. Figure 4 suggests enforcing consistent predictions using logits (our approach) achieves lower gradient magnitude deviation than maintaining the adversarial transferability using flipped labels (naive approach). The L_2 regularizer also stabilizes gradient magnitudes.

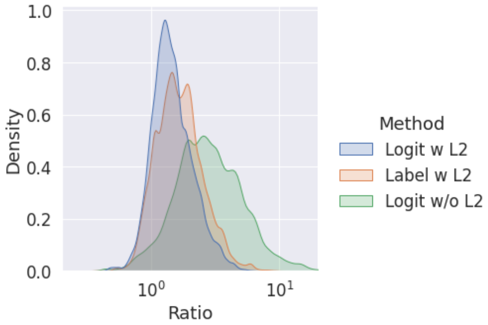


Figure 4: Distribution of the ratio $\frac{\|\nabla_{\boldsymbol{x}}\ell(f_l(\boldsymbol{x}),y)\|}{\|\nabla_{\boldsymbol{x}}\ell(f_g(\boldsymbol{x}),y)\|}$. Using our approach and enforcing consistent predictions with logits achieves lower gradient magnitude deviation than adding adversarial examples with flipped labels to personalization. The L_2 regularizer also stabilizes gradient magnitudes.

6.5 Additional Experiments

Personalization with Imperfect Global Model Our approach still bounds the disparity deviation of personalized models when the global model already entangles spurious features. Specifically, for two global models with accuracy disparities of 0.5% and 2.1% on MNIST, the accuracy disparity of personalized models is 8.1% and 12.4%, respectively. In contrast, fine-tuning increases both disparities to 28.5%.

Robustness to Adversarial Attack Although we add adversarial examples with flipped labels during personalization, the robustness of personalized models from our approach does not degrade significantly. We evaluate our approach against transferable adversarial examples from the global model and direct adversarial attacks on personalized models with attack configurations in Appendix D. Our approach only increases the success rate of transferable attacks from 56.1% to 68.2% and that of direct attacks from 47.8% to 50.1% on MNIST. Both transferable attacks use the global model to generate adversarial examples. The success rate of transferable attacks is not 100% because personalized models are trained with different benign and adversarial samples locally. Transferable attacks can have higher success rates than direct attacks since only the adversarial examples that "fool" global models are transferred.

7 Conclusion and Future Work

In this work, we show a risk of prior federated learning personalization methods with spurious features, which lead to high accuracy disparity between environments. Then, we develop a strategy to reduce the accuracy disparity by maintaining the adversarial transferability between the global and personalized models. Both empirical and theoretical results show that our strategy is effective. Extensions of this work include incorporating text data, which often requires different methods for generating adversarial examples.

References

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Sebastian Caldas, Sai Meher Karthik Duddu, Peter Wu, Tian Li, Jakub Konečnỳ, H Brendan McMahan, Virginia Smith, and Ameet Talwalkar. Leaf: A benchmark for federated settings. *arXiv preprint arXiv:1812.01097*, 2018.

Jacky Cao, Kit-Yung Lam, Lik-Hang Lee, Xiaoli Liu, Pan Hui, and Xiang Su. Mobile augmented reality: User interfaces, frameworks, and intelligence. *ACM Comput. Surv.*, aug 2022. ISSN 0360-0300. doi: 10.1145/3557999. URL `https://doi.org/10.1145/3557999`.

Zachary B. Charles, Harrison Rosenberg, and Dimitris Papailiopoulos. A geometric perspective on the transferability of adversarial directions. In *AISTATS*, 2019.

Jianfeng Chi, Yuan Tian, Geoffrey J Gordon, and Han Zhao. Understanding and mitigating accuracy disparity in regression. In *International Conference on Machine Learning (ICML)*, 2021.

Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach. *Advances in Neural Information Processing Systems*, 33, 2020.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11): 665–673, 2020.

Avishek Ghosh, Jichan Chung, Dong Yin, and Kannan Ramchandran. An efficient framework for clustered federated learning. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 19586–19597. Curran Associates, Inc., 2020. URL `https://proceedings.neurips.cc/paper/2020/file/e32cc80bf07915058ce90722ee17bb71-Paper.pdf`.

He He, Sheng Zha, and Haohan Wang. Unlearn dataset bias in natural language inference by fitting the residual. *arXiv preprint arXiv:1908.10763*, 2019.

Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Keith Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *arXiv preprint arXiv:1912.04977*, 2019.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL `http://arxiv.org/abs/1412.6980`.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. In I. Dhillon, D. Papailiopoulos, and V. Sze (eds.), *Proceedings of Machine Learning and Systems*, volume 2, pp. 429–450, 2020. URL `https://proceedings.mlsys.org/paper/2020/file/38af86134b65d0f10fe33d30dd76442e-Paper.pdf`.

Tian Li, Shengyuan Hu, Ahmad Beirami, and Virginia Smith. Ditto: Fair and robust federated learning through personalization. In *International Conference on Machine Learning*, pp. 6357–6368. PMLR, 2021.

Yi Li and Nuno Vasconcelos. Repair: Removing representation bias by dataset resampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9572–9581, 2019.

Kaizhao Liang, Jacky Y Zhang, Boxin Wang, Zhuolin Yang, Sanmi Koyejo, and Bo Li. Uncovering the connections between adversarial transferability and knowledge transferability. In *International Conference on Machine Learning*, pp. 6577–6587. PMLR, 2021.

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 *International Conference on Machine Learning*, pp. 6781–6792. PMLR, 2021.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. URL `https://openreview.net/forum?id=rJzIBfZAb`.

Othmane Marfoq, Giovanni Neglia, Richard Vidal, and Laetitia Kameni. Personalized federated learning through local memorization. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 15070–15092. PMLR, 17–23 Jul 2022. URL `https://proceedings.mlr.press/v162/marfoq22a.html`.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics*, pp. 1273–1282. PMLR, 2017.

Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1979–1993, 2018.

Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from failure: Training debiased classifier from biased classifier. In *Advances in Neural Information Processing Systems*, 2020.

Sameer A. Nene, Shree K. Nayar, and Hiroshi Murase. Columbia object image library (coil-20. Technical report, 1996.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=ryxGuJrFvS`.

Aviv Shamsian, Aviv Navon, Ethan Fetaya, and Gal Chechik. Personalized federated learning using hypernetworks. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 9489–9502. PMLR, 18–24 Jul 2021. URL `https://proceedings.mlr.press/v139/shamsian21a.html`.

Sahil Singla and Soheil Feizi. Salient imagenet: How to discover spurious features in deep learning? In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=XVPqLyNxSyh`.

Canh T. Dinh, Nguyen Tran, and Josh Nguyen. Personalized federated learning with moreau envelopes. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 21394–21405. Curran Associates, Inc., 2020.

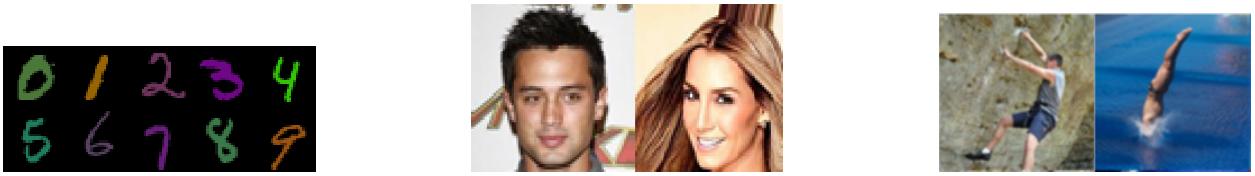
Florian Tramèr, Nicolas Papernot, Ian J. Goodfellow, Dan Boneh, and Patrick Mcdaniel. The space of transferable adversarial examples. *ArXiv*, abs/1704.03453, 2017.

Haohan Wang, Zexue He, Zachary L. Lipton, and Eric P. Xing. Learning robust representations by projecting superficial statistics out. In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=rJEjjoR9K7`.

Haoxiang Wang, Haozhe Si, Bo Li, and Han Zhao. Provable domain generalization via invariant-feature subspace recovery. In *International Conference on Machine Learning*. PMLR, 2022.

Han Zhao and Geoffrey J Gordon. Inherent tradeoffs in learning fair representations. *Journal of Machine Learning Research*, 23:1–26, 2022.

A Spurious Feature Examples



(a) MNIST (b) CelebA (c) BAR

Figure 5: Datasets with spurious features. The object color spuriously correlates with the digit in MNIST (a). The hair color spuriously correlates with the gender in the CelebA dataset (b). The background spuriously correlates with the action in the BAR dataset (c).

B Notation

Table 4: Table of Notation

Symbol	**Description**
$\boldsymbol{x}, y$	A pair of data sample and label
$\mathrm{Cat}(,)$	A concatenation of two vectors
$\boldsymbol{x}_r, \boldsymbol{x}_s$	The robust features and spurious features in $\boldsymbol{x} = \mathrm{Cat}(\boldsymbol{x}_r, \boldsymbol{x}_s)$, respectively
f_g	The global model
f_l	The personalized local model
$\boldsymbol{\delta}_{f_g,\epsilon}$	An adversarial perturbation generated by the global model f_g with attack budget ϵ
$\mathcal{D}$	A distribution in the training environment
$\mathcal{D}'$	A distribution in the other environment that is not accessible during training

C Proofs

Theorem 3. *Under the setting in Section 4.1, for a model f parameterized by $\boldsymbol{w}$, assume the composition $\ell \circ \sigma$ of the activation function σ and the loss function ℓ is ρ-Lipschitz, let θ be the angle between $\boldsymbol{w}$ and* $\mathrm{Cat}(\mathbf{0}, \boldsymbol{w}_s)$*, we have:*

$$|\mathbb{E}_{\mathcal{D}}[\ell(f(\boldsymbol{x}), y)] - \mathbb{E}_{\mathcal{D}'}[\ell(f(\boldsymbol{x}), y)]| \le \rho\|\boldsymbol{w}_s\|_2 \cdot W_1(\mathcal{D}_s, \mathcal{D}'_s) = \rho\|\boldsymbol{w}\|_2 \cos\theta \cdot W_1(\mathcal{D}_s, \mathcal{D}'_s). \tag{14}$$

Proof. With the data generating process in Section 4.1, we have:

$$|\mathbb{E}_{\mathcal{D}}[\ell(f(\boldsymbol{x}), y)] - \mathbb{E}_{\mathcal{D}'}[\ell(f(\boldsymbol{x}), y)]| = |\mathbb{E}_{\mathcal{D},\mathcal{D}'}[\ell(f(\boldsymbol{x}), y) - \ell(f(\boldsymbol{x}'), y)]|. \tag{15}$$

Since the loss function is ρ-Lipschitz, we further have:

$$|\mathbb{E}_{\mathcal{D},\mathcal{D}'}[\ell(f(\boldsymbol{x}), y) - \ell(f(\boldsymbol{x}'), y)]| \le \rho|\mathbb{E}_{\mathcal{D},\mathcal{D}'}[f(\boldsymbol{x}) - f(\boldsymbol{x}')]|. \tag{16}$$

Under the given data generating process in Section 4.2, we know that $\boldsymbol{x} - \boldsymbol{x}' = \mathrm{Cat}(\mathbf{0}, \boldsymbol{x}_s - \boldsymbol{x}'_s)$, yielding:

$$|\mathbb{E}_{\mathcal{D},\mathcal{D}'}[f(\boldsymbol{x}) - f(\boldsymbol{x}')]| \le \|\boldsymbol{w}_s\|_2|\mathbb{E}_{\mathcal{D}_s,\mathcal{D}'_s}[\boldsymbol{x}_s - \boldsymbol{x}'_s]| = \|\boldsymbol{w}\|_2 \cos\theta|\mathbb{E}_{\mathcal{D}_s,\mathcal{D}'_s}[\boldsymbol{x}_s - \boldsymbol{x}'_s]|. \tag{17}$$

By the definition of Wasserstein-1 distance, we have:

$$|\mathbb{E}_{\mathcal{D}_s,\mathcal{D}'_s}[\boldsymbol{x}_s - \boldsymbol{x}'_s]| \le \mathbb{E}_{\mathcal{D}_s,\mathcal{D}'_s}[|\boldsymbol{x}_s - \boldsymbol{x}'_s|] \le W_1(\mathcal{D}_s, \mathcal{D}'_s). \tag{18}$$

□

Theorem 4. *Under the setting in Section 4.1, let* ϵ *be the attack budget, and* θ' *be the angle between* $\boldsymbol{w}_g$ *and* $\boldsymbol{w}_l$ *we have:*

$$\theta' \leq \arccos\Big(\frac{1}{\epsilon \cdot \bigcirc} + \Box + \diamond + \mathbb{E}_{\mathcal{D}}[\ell_{g\to l}(f_g, f_l, \boldsymbol{x}, y)]\Big), \tag{19}$$

where $\bigcirc = \mathbb{E}_{\mathcal{D}}[\|\nabla_{\boldsymbol{x}}\ell(f_l(\boldsymbol{x}), y)\|]$, $\Box = \epsilon\mathbb{E}_{\mathcal{D}}[\|\nabla_{\boldsymbol{x}}\ell(f_g(\boldsymbol{x}), y)\|]$, *and* $\diamond = \mathbb{E}_{\mathcal{D}}[R_{g,2}(\boldsymbol{x}_{\mathrm{adv}}) - R_{l,2}(\boldsymbol{x}_{\mathrm{adv}})]$ *is the expected difference of second-order Taylor remainders between* f_g *and* f_l.

Proof. Expanding $\ell_{g\to l}(f_g, f_l, \boldsymbol{x}, y)$ at $(\boldsymbol{x}, y)$, we have:

$$\begin{aligned}
&\ell_{g\to l}(f_g, f_l, \boldsymbol{x}, y) \\
&= \Big(\ell(f_g(\boldsymbol{x} + \boldsymbol{\delta}_{f_g,\epsilon}), y) - \ell(f_g(\boldsymbol{x}), y)\Big) - \Big(\ell(f_l(\boldsymbol{x} + \boldsymbol{\delta}_{f_g,\epsilon}), y) - \ell(f_l(\boldsymbol{x}), y)\Big) \\
&= \nabla_{\boldsymbol{x}}\ell(f_g(\boldsymbol{x}), y)^\top \boldsymbol{\delta}_{f_g,\epsilon} + R_{g,2}(\boldsymbol{x}_{\mathrm{adv}}) - \nabla_{\boldsymbol{x}}\ell(f_l(\boldsymbol{x}), y)^\top \boldsymbol{\delta}_{f_g,\epsilon} - R_{l,2}(\boldsymbol{x}_{\mathrm{adv}}) \\
&= \epsilon\|\nabla_{\boldsymbol{x}}\ell(f_g(\boldsymbol{x}), y)\| - \nabla_{\boldsymbol{x}}\ell(f_l(\boldsymbol{x}), y)^\top \epsilon \cdot \frac{\nabla_{\boldsymbol{x}}\ell(f_g(\boldsymbol{x}), y)}{\|\nabla_{\boldsymbol{x}}\ell(f_g(\boldsymbol{x}), y)\|} + R_{g,2}(\boldsymbol{x}_{\mathrm{adv}}) - R_{l,2}(\boldsymbol{x}_{\mathrm{adv}})
\end{aligned} \tag{20}$$

For logistic regression models f_g and f_l, we have:

$$\nabla_{\boldsymbol{x}}\ell(f_g(\boldsymbol{x}), y) = \Big(\sigma(\boldsymbol{w}_g^\top \boldsymbol{x}) - y\Big)\boldsymbol{w}_g \text{ and } \nabla_{\boldsymbol{x}}\ell(f_l(\boldsymbol{x}), y) = \Big(\sigma(\boldsymbol{w}_l^\top \boldsymbol{x}) - y\Big)\boldsymbol{w}_l. \tag{21}$$

Since the label $y \in \{0, 1\}$ of each data sample $\boldsymbol{x}$ is fixed, and the logits are in $(0, 1)$, the product of $\sigma(\boldsymbol{w}_g^\top \boldsymbol{x}) - y$ and $\sigma(\boldsymbol{w}_l^\top \boldsymbol{x}) - y$ is always positive, meaning that the angle between gradients equals to the angle θ' between weight vectors $\boldsymbol{w}_g$ and $\boldsymbol{w}_l$. The angle between weight vectors is consistent across data samples. Therefore, we further have:

$$\nabla_{\boldsymbol{x}}\ell(f_l(\boldsymbol{x}), y)^\top \epsilon \cdot \frac{\nabla_{\boldsymbol{x}}\ell(f_g(\boldsymbol{x}), y)}{\|\nabla_{\boldsymbol{x}}\ell(f_g(\boldsymbol{x}), y)\|} = \epsilon\|\nabla_{\boldsymbol{x}}\ell(f_l(\boldsymbol{x}), y)\|\cos\theta'. \tag{22}$$

Plugging Equation 22 into Equation 20 and taking the expectation over $\mathcal{D}$, we have:

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}}[\ell_{g\to l}(f_g, f_l, \boldsymbol{x}, y)] \\
&= \epsilon\mathbb{E}_{\mathcal{D}}[\|\nabla_{\boldsymbol{x}}\ell(f_g(\boldsymbol{x}), y)\|] - \epsilon\mathbb{E}_{\mathcal{D}}[\|\nabla_{\boldsymbol{x}}\ell(f_l(\boldsymbol{x}), y)\|]\cos\theta' + \mathbb{E}_{\mathcal{D}}[R_{g,2}(\boldsymbol{x}_{\mathrm{adv}}) - R_{l,2}(\boldsymbol{x}_{\mathrm{adv}})]
\end{aligned} \tag{23}$$

Moving $\cos\theta$ to the left-hand-side and taking the arc cos operator completes the proof. ☐

Proposition 2. *Under the settings in Section 4.1, assume the composition* $\ell \circ \sigma$ *of the activation function* σ *and the loss function* ℓ *is* ρ*-Lipschitz, we have:*

$$|\|\boldsymbol{w}_l\| - \|\boldsymbol{w}_g\|| \leq \|\boldsymbol{w}_l - \boldsymbol{w}_g\| \tag{24}$$

and

$$|\mathbb{E}_{\mathcal{D}}[\|\nabla_{\boldsymbol{w}_l}\ell(f_l(\boldsymbol{x}), y)\|] - \mathbb{E}_{\mathcal{D}}[\|\nabla_{\boldsymbol{w}_g}\ell(f_g(\boldsymbol{x}), y)\|]| \leq \rho\mathbb{E}_{\mathcal{D}}[\|\boldsymbol{x}\|]\|\boldsymbol{w}_l - \boldsymbol{w}_g\|. \tag{25}$$

Proof. With the triangle inequality, we have:

$$\|\boldsymbol{w}_l\| = \|\boldsymbol{w}_l + \boldsymbol{w}_g - \boldsymbol{w}_g\| \leq \|\boldsymbol{w}_g\| + \|\boldsymbol{w}_l - \boldsymbol{w}_g\| \tag{26}$$

and

$$\|\boldsymbol{w}_g\| = \|\boldsymbol{w}_g + \boldsymbol{w}_l - \boldsymbol{w}_l\| \leq \|\boldsymbol{w}_l\| + \|\boldsymbol{w}_g - \boldsymbol{w}_l\|. \tag{27}$$

Since $\|\boldsymbol{w}_g - \boldsymbol{w}_l\| = \|\boldsymbol{w}_l - \boldsymbol{w}_g\|$, moving $\|\boldsymbol{w}_g\|$ and $\|\boldsymbol{w}_l\|$ to the left-hand-side in Equations 26 and 27, respectively, we have $|\|\boldsymbol{w}_l\| - \|\boldsymbol{w}_g\|| \leq \|\boldsymbol{w}_l - \boldsymbol{w}_g\|$.

Similarly, we have:

$$\begin{aligned}\|\nabla_{\boldsymbol{w}_l}\ell(f_l(\boldsymbol{x}),y)\| &= \|\nabla_{\boldsymbol{w}_l}\ell(f_l(\boldsymbol{x}),y)+\nabla_{\boldsymbol{w}_g}\ell(f_g(\boldsymbol{x}),y)-\nabla_{\boldsymbol{w}_g}\ell(f_g(\boldsymbol{x}),y)\|\\ &\leq \|\nabla_{\boldsymbol{w}_g}\ell(f_g(\boldsymbol{x}),y)\|+\|\nabla_{\boldsymbol{w}_l}\ell(f_l(\boldsymbol{x}),y)-\nabla_{\boldsymbol{w}_g}\ell(f_g(\boldsymbol{x}),y)\|\\ &\leq \|\nabla_{\boldsymbol{w}_g}\ell(f_g(\boldsymbol{x}),y)\|+\rho\|\boldsymbol{x}\|\|\boldsymbol{w}_l-\boldsymbol{w}_g\|\end{aligned} \tag{28}$$

and

$$\|\nabla_{\boldsymbol{w}_g}\ell(f_g(\boldsymbol{x}),y)\| \leq \|\nabla_{\boldsymbol{w}_g}\ell(f_l(\boldsymbol{x}),y)\|+\rho\|\boldsymbol{x}\|\|\boldsymbol{w}_g-\boldsymbol{w}_l\|. \tag{29}$$

Recalling that $\|\boldsymbol{w}_g-\boldsymbol{w}_l\|=\|\boldsymbol{w}_l-\boldsymbol{w}_g\|$, combining Equations 28 and 29 results in:

$$\|\nabla_{\boldsymbol{w}_g}\ell(f_g(\boldsymbol{x}),y)\|-\|\nabla_{\boldsymbol{w}_g}\ell(f_l(\boldsymbol{x}),y)\| \leq |\|\nabla_{\boldsymbol{w}_g}\ell(f_g(\boldsymbol{x}),y)\|-\|\nabla_{\boldsymbol{w}_g}\ell(f_l(\boldsymbol{x}),y)\|| \leq \rho\|\boldsymbol{x}\|\|\boldsymbol{w}_g-\boldsymbol{w}_l\|. \tag{30}$$

Taking the expectation and absolute value over both sides of Equation 30 completes the proof. ☐

D Detailed Experimental Setting

D.1 Data Partition

We distribute the MNIST, Coil20, and BAR datasets across 50 clients. Each client has data samples from 5 different classes. These data samples are further grouped into two environments according to their spurious correlations. Each client randomly picks one environment for training. Local datasets are further partitioned to train/validation/test set with a ratio of 72:8:20, following prior work (Li et al., 2021). Combining test sets from the two environments gets test sets in global environments.

For the CelebA dataset, we let each client represent 20 celebrities. One celebrity only appears on one client. The blond hair correlates with the female in the training environment and correlates with the male in the other environment that is not used for training. Note that non-blond hair colors do not correlate with any gender. In the personalization step, we select the clients with more than 5 blond female data samples in the training environment and more than 5 blond male data samples in the other environment that is not used for training. We select these clients because they provide enough samples to create the spurious correlation and evaluate entanglements to the spurious correlation.

D.2 Hyper-parameters

We use Adam optimizer (Kingma & Ba, 2015) throughout our experiments with a learning rate of 1e-4 for MNIST, CelebA, and BAR and 2e-4 for Coil20. Although stochastic gradient descent (SGD) optimizer is more common in vision-related tasks, the Adam optimizer always leads to lower accuracy disparity. We train the global model for 500 rounds. 5 clients are selected per round, each performing 5 epochs of local updates. We tune the coefficients of the adversarial transferability and L_2 regularization terms from $\{0.01, 0.1, 1.0, 10.0\}$ and select the largest value that does not decrease the validation accuracy during penalization. We start the attack budget at 0.031 (i.e., $\frac{8}{255}$) and gradually decrease it such that $30\% - 50\%$ of the attack succeeds. A large budget will make the attack too strong and push the adversarial examples far across the decision boundary, making the regularization method less effective. We configure ϵ to 0.031/0.01/0.031 for MNIST/CelebA/Coil20, respectively. We fine-tune the global model for 5 epochs on MNIST/BAR and 10 epochs on Coil20/CelebA, which are sufficient for the personalized models to converge. The clients with the most data samples fine-tune the penalized models for a total of 80/40/30 batches on MNIST/CelebA/Coil20 datasets. Note that we may not select the personalized model with the most fine-tuning batches for reporting. In the up-weighting (UW) and just-train-twice (JTT) method, we up-sample the residual by a factor of 50. In Ditto, we tune its λ from $\{0.1, 1.0\}$.

D.3 Neural Network Architecture

We use CNN 28x28 for the MNIST dataset and CNN 64x64 for the CelebA and the Coil20 datasets, as are listed in Table 5. We use a pre-trained ResNet-18 (He et al., 2016) for the BAR dataset.

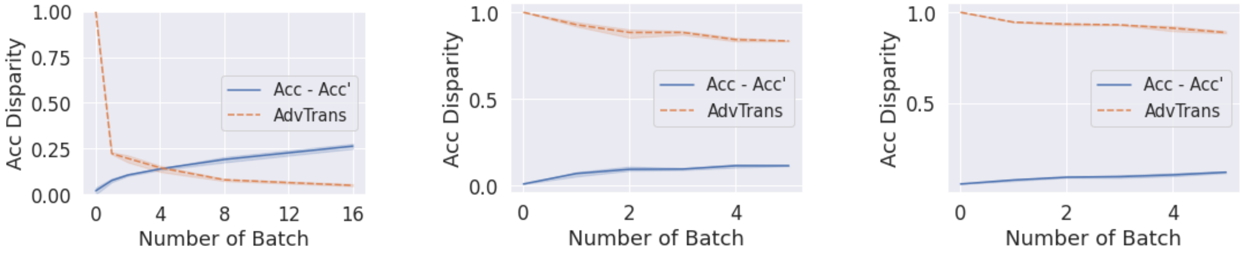
Table 5: Neural Network Architecture

CNN 28 × 28	CNN 64 × 64
Input: $\mathbb{R}^{3\times28\times28}$ 4×4 conv, 64 BN LReLU, stride 2 4×4 conv, 128 BN LReLU, stride 2 FC 4096 ReLU FC 10	Input: $\mathbb{R}^{3\times64\times64}$ 4×4 conv, 64 BN LReLU, stride 2 4×4 conv, 64 BN LReLU, stride 2 FC 4096 ReLU FC 10

E More Experimental Results

E.1 Transferability Correlates with Disparity

The additional empirical studies on Coil20 and CelebA datasets are shown in Figure 6.



(a) MNIST (b) Coil20 (c) CelebA

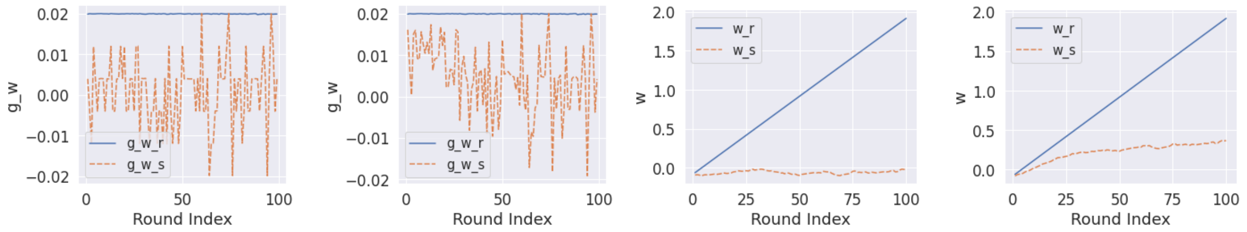
Figure 6: As the personalized models entangle the spurious feature and increase their accuracy disparity (Acc - Acc′), the adversarial transferability (AdvTrans) decreases. Here, the accuracy disparity deviation also increases as the accuracy disparity of personalized models deviates from that of the global model, which is fixed as a reference.

E.2 Simulations

We construct a synthetic dataset where $y \in [0, 1]$, $\boldsymbol{x} = [\boldsymbol{x}_r, \boldsymbol{x}_s]$ and $\boldsymbol{x}_r \sim \mathcal{N}(2y-1, 1)$ and $\boldsymbol{x}_s \sim \mathcal{N}\Big(c(2y-1), 1\Big)$ where the c is 1 if the client index mods 2 is 0. Otherwise, the c is -1 (Figures 7a and 7c). In the experiments shown in Figures 7b and 7d, the c is 1 if the client index mods 2 is 0; otherwise, the c is 0. We employ a linear model $y = \boldsymbol{w}^\top \boldsymbol{x}$ where $\boldsymbol{w} = \mathrm{Cat}(\boldsymbol{w}_r, \boldsymbol{w}_s)$.

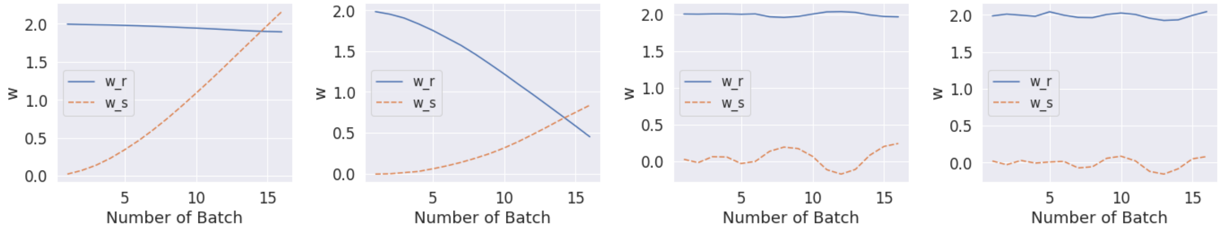
Our simulation starts with the global model. We use the federated averaging algorithm with 50 clients, each with 128 data samples, and select 5 clients per round. Figure 7 shows that the gradient $\boldsymbol{g}_{\boldsymbol{w}_s}$ of $\boldsymbol{w}_s$ is noisier and closer to 0 than the gradient $\boldsymbol{g}_{\boldsymbol{w}_r}$ of $\boldsymbol{w}_r$. As a result, $\boldsymbol{w}_r$ increases during the training while $\boldsymbol{w}_s$ does not increase or increases much slower, as is shown in Figure 7. These results suggest that when the spurious correlation varies across clients, the model does not entangle the spurious feature or only slightly entangle the spurious feature.

Then, we personalize the global model whose $\boldsymbol{w} = [2.0, 0.0]$ on clients with shifting spurious correlations. Results in Figure 8 suggest that our approach limits $\boldsymbol{w}_s$ around 0 with minimal fluctuation.



(a) Spurious Correlation Shifts Across Users (b) 50% Users Have a Stationary Spurious Correlation (c) Spurious Correlation Shifts Across Users (d) 50% Users Have a Stationary Spurious Correlation

Figure 7: The gradients and weights of a linear model during training. The gradient $g_{\boldsymbol{w}_s}$ of $\boldsymbol{w}_s$ is more noisy and closer to 0 than the gradient $g_{\boldsymbol{w}_r}$ of $\boldsymbol{w}_r$. Therefore, $\boldsymbol{w}_s$ stays around 0.



(a) Naive Fine-tuning (b) Consistent Prediction (c) Aligning Weights (d) Ours

Figure 8: Weights of a linear model during fine-tuning. Our approach reduces the co-efficiency $\boldsymbol{w}_s$ of the spurious feature $\boldsymbol{x}_s$ most effectively. Aligning weights also limits $\boldsymbol{w}_s$ but has a more significant fluctuation than ours.