Appendix

A Proof of Theorem 1

In this section, we first propose Corollary 1, then adopt the corollary to prove Theorem 1.

Corollary 1. *For $p(y), f(y), g(y) > 0$, we have*

$$\int_{\mathcal{Y}} p(y)f(y)\mathrm{d}y \leq \int_{\mathcal{Y}} q(y)f(y)\mathrm{d}y + \int_{\mathcal{Y}} |p(y)-q(y)|f(y)\mathrm{d}y$$
$$\int_{\mathcal{Y}} p(y)f(y)\mathrm{d}y \leq \int_{\mathcal{Y}} p(y)g(y)\mathrm{d}y + \int_{\mathcal{Y}} p(y)|f(y)-g(y)|\mathrm{d}y$$

Proof.

$$\begin{aligned}\int_{\mathcal{Y}} p(y)f(y)\mathrm{d}y &= \int_{\mathcal{Y}} q(y)f(y)\mathrm{d}y + \int_{\mathcal{Y}} [p(y)-q(y)]f(y)\mathrm{d}y \\ &\leq \int_{\mathcal{Y}} q(y)f(y)\mathrm{d}y + \int_{\mathcal{Y}} |p(y)-q(y)|f(y)\mathrm{d}y \\ \int_{\mathcal{Y}} p(y)f(y)\mathrm{d}y &= \int_{\mathcal{Y}} p(y)g(y)\mathrm{d}y + \int_{\mathcal{Y}} p(y)[f(y)-g(y)]\mathrm{d}y \\ &\leq \int_{\mathcal{Y}} p(y)g(y)\mathrm{d}y + \int_{\mathcal{Y}} p(y)|f(y)-g(y)|\mathrm{d}y\end{aligned}$$

After proving Corollary 1, we return to prove the theorem.

Theorem 1. *Assume the loss value on the training set satisfies $\frac{1}{|\mathcal{S}|}\sum_{(\boldsymbol{x}_i,y_i,a_i)\in\mathcal{S}} l(\boldsymbol{h}_i,y_i;\theta_h) \leq \epsilon$[10], and $l(\boldsymbol{h},y;\theta_h)$ and f_h satisfy K_l- and K_h-Lipschitz continuity*[11]*, respectively. The generalization loss difference between unprivileged group and privileged group has the following upper bound with probability $1-\gamma$,*

$$\left|\int_{\mathcal{X}_0}\int_{\mathcal{Y}} p(\boldsymbol{x},y)l(\boldsymbol{h},y;\theta_h)\mathrm{d}\boldsymbol{x}\mathrm{d}y - \int_{\mathcal{X}_1}\int_{\mathcal{Y}} p(\boldsymbol{x},y)l(\boldsymbol{h},y;\theta_h)\mathrm{d}\boldsymbol{x}\mathrm{d}y\right| \leq \epsilon + \min\left\{\sqrt{-L^2\log\gamma(2N_{\tilde{a}})^{-1}}, (K_l+K_hL)\delta_{\tilde{a}}\right\}, \quad (14)$$

where $\tilde{a} = \arg\max_{a\in\mathcal{A}} \int_{\mathcal{X}_a}\int_{\mathcal{Y}} p(\boldsymbol{x},y)l(\boldsymbol{h},y;\theta_h)\mathrm{d}\boldsymbol{x}\mathrm{d}y$; $\mathcal{X}_a = \{\boldsymbol{x}_i \in \mathcal{D} | a_i = a\}$; $\delta_{\tilde{a}} = \max_{\boldsymbol{x}_i\in\mathcal{X}_{\tilde{a}}} \min_{(\boldsymbol{x}_j,y_j,a_j)\in\mathcal{S}} ||\boldsymbol{h}_i - \boldsymbol{h}_j||_2$; $N_{\tilde{a}} = |\{(\boldsymbol{x}_i,y_i,a_i) | a_i = \tilde{a}, (\boldsymbol{x}_i,y_i,a_i)\in\mathcal{S}\}|$; $L = \max_{(\boldsymbol{x}_i,y_i)\in\mathcal{U}} l(\boldsymbol{h}_i,y_i;\theta_h)$; and $\boldsymbol{h}_i = f_b(\boldsymbol{x}_i|\theta_b)$.

[10] ϵ is small if the classifier head f_h has been well-trained on the annotated dataset $\mathcal{S}$.

[11] $l(\boldsymbol{h},y;\theta_h)$ and f_h satisfy $|l(\boldsymbol{h}_i,y;\theta_h) - l(\boldsymbol{h}_j,y;\theta_h)| \leq K_l||\boldsymbol{h}_i - \boldsymbol{h}_j||_2$ and $|p(y|\boldsymbol{x}_i) - p(y|\boldsymbol{x}_j)| \leq K_h||\boldsymbol{h}_i - \boldsymbol{h}_j||_2$, respectively, where the likelihood function $p(y \mid \boldsymbol{x}_i) = \mathrm{softmax}(f_h(\boldsymbol{h}_i|\theta_h))$.

Proof. According to the upper bound of generalization error, the generalization error for group $\boldsymbol{x} \in \mathcal{X}_a$ for $\forall a \in \mathcal{A}$ is bounded with probability $1-\gamma$,

$$g_a = \int_{\mathcal{X}_a}\int_{\mathcal{Y}} p(\boldsymbol{x},y) l(\boldsymbol{h},y;\theta_h)\mathrm{d}\boldsymbol{x}\mathrm{d}y \leq \epsilon + \sqrt{-L^2\log\gamma(2N_a)^{-1}},$$

where $L = \max_{(\boldsymbol{x}_i,y_i)\in\mathcal{U}} l(\boldsymbol{h}_i,y_i;\theta_h)$. Moreover, we consider the upper bound of absolute gap $|g_0 - g_1| \leq \max_{a\in\mathcal{A}} g_a$. The generalization error difference between the two groups is bounded with probability $1-\gamma$ as follow,

$$\begin{aligned}\left|\int_{\mathcal{X}_0}\int_{\mathcal{Y}} p(\boldsymbol{x},y) l(\boldsymbol{h},y;\theta_h)\mathrm{d}\boldsymbol{x}\mathrm{d}y - \int_{\mathcal{X}_1}\int_{\mathcal{Y}} p(\boldsymbol{x},y) l(\boldsymbol{h},y;\theta_h)\mathrm{d}\boldsymbol{x}\mathrm{d}y\right| \\ \leq \epsilon + \sqrt{-L^2\log\gamma(2N_{\tilde{a}})^{-1}},\end{aligned} \tag{15}$$

where $\tilde{a} = \arg\max_{a\in\mathcal{A}} \int_{\mathcal{X}_a}\int_{\mathcal{Y}} p(\boldsymbol{x},y) l(\boldsymbol{h},y;\theta_h)\mathrm{d}\boldsymbol{x}\mathrm{d}y$.

To prove the second bound of the generalization error difference, let $\mathcal{N}(\boldsymbol{x}_i)$ denote the nearest neighbour of $\boldsymbol{x}_i \in \mathcal{X}$ which belongs to the annotated dataset, i.e. $\mathcal{N}(\boldsymbol{x}_i) = \arg\min_{(\boldsymbol{x}_j,y_j,a_j)\in\mathcal{S}} ||\boldsymbol{h}_j - \boldsymbol{h}_i||_2$; let $\boldsymbol{h}_i^{\mathcal{N}}$ denote the embedding of $\mathcal{N}(\boldsymbol{x}_i)$; and let $d_{\boldsymbol{x}_i}$ denote the distance between $\boldsymbol{x}_i$ and $\mathcal{N}(\boldsymbol{x}_i)$ in the embedding space, i.e. $d_{\boldsymbol{x}_i} = \min_{(\boldsymbol{x}_j,y_j,a_j)\in\mathcal{S}} ||\boldsymbol{h}_i - \boldsymbol{h}_j||_2$. Accoding to Corollary 1, with $p(y) = p(y|\boldsymbol{x}_i)$, $q(y) = p(y|\mathcal{N}(\boldsymbol{x}_i))$ and $f(y) = l(\boldsymbol{h}_i,y;\theta_h)$, the generalization error can be bounded by

$$\begin{aligned}&\int_{\mathcal{Y}} p(y|\boldsymbol{x}_i) l(\boldsymbol{h}_i,y;\theta_h)\mathrm{d}y \leq \\ &\int_{\mathcal{Y}} p(y|\mathcal{N}(\boldsymbol{x}_i)) l(\boldsymbol{h}_i,y;\theta_h)\mathrm{d}y + \int_{\mathcal{Y}} \big|p(y|\boldsymbol{x}_i) - p(y|\mathcal{N}(\boldsymbol{x}_i))\big| l(\boldsymbol{h}_i,y;\theta_h)\mathrm{d}y.\end{aligned} \tag{16}$$

Note that the classifier head f_h satisfies K_h-Lipschitz continuity $|p(y|\boldsymbol{x}_i) - p(y|\mathcal{N}(\boldsymbol{x}_i))| \leq K_h|||\boldsymbol{h}_i - \boldsymbol{h}_i^{\mathcal{N}}||_2 = K_h d_{\boldsymbol{x}_i}$ and $l(\boldsymbol{h},y;\theta_h) \leq L$, the second term in the right-side of Equation (16) is bounded by

$$\int_{\mathcal{Y}} \big|p(y|\boldsymbol{x}_i) - p(y|\mathcal{N}(\boldsymbol{x}_i))\big| l(\boldsymbol{h}_i,y;\theta_h)\mathrm{d}y \leq K_h L d_{\boldsymbol{x}_i}. \tag{17}$$

Furthermore, taking $p(y) = p(y|\mathcal{N}(\boldsymbol{x}_i))$, $f(y) = l(\boldsymbol{h}_i,y;\theta_h)$ and $g(y) = l(\boldsymbol{h}_i^{\mathcal{N}},y;\theta_h)$ into Corollary 1, we have the first term in the right-side of Equation (16) can be bounded by

$$\begin{aligned}&\int_{\mathcal{Y}} p(y|\mathcal{N}(\boldsymbol{x}_i)) l(\boldsymbol{h},y;\theta_h)\mathrm{d}y \leq \\ &\int_{\mathcal{Y}} p(y|\mathcal{N}(\boldsymbol{x}_i)) l(\boldsymbol{h}_i^{\mathcal{N}},y;\theta_h)\mathrm{d}y + \int_{\mathcal{Y}} p(y|\mathcal{N}(\boldsymbol{x}_i)) |l(\boldsymbol{h},y;\theta_h) - l(\boldsymbol{h}_i^{\mathcal{N}},y;\theta_h)|\mathrm{d}y \\ &\leq \epsilon + K_l d_{\boldsymbol{x}_i},\end{aligned} \tag{18}$$

where we have $\int_{\mathcal{Y}} p(y|\mathcal{N}(\boldsymbol{x}_i)) l(\boldsymbol{h}_i^{\mathcal{N}},y;\theta_h)\mathrm{d}y \leq \epsilon$ due to the upper bound of training error; and we have

$$\int_{\mathcal{Y}} p(y|\mathcal{N}(\boldsymbol{x}_i)) |l(\boldsymbol{h}_i,y;\theta_h) - l(\boldsymbol{h}_i^{\mathcal{N}},y;\theta_h)|\mathrm{d}y \leq K_l d_{\boldsymbol{x}_i}, \tag{19}$$

due to the K_l-Lipschitz continuity of the loss function.

Taking Equations (17) and (18) into Equation (16), the generalization error on group $\boldsymbol{x} \in \mathcal{X}_a$ can be bounded by

$$\int_{\mathcal{X}_a}\int_{\mathcal{Y}} p(\boldsymbol{x},y)l(\boldsymbol{h}_i,y;\theta_h)\mathrm{d}y\mathrm{d}\boldsymbol{x} \leq \epsilon + (K_l + K_hL)\delta_a, \tag{20}$$

where $\delta_a = \max_{\boldsymbol{x}_i\in\mathcal{X}_a} d_{\boldsymbol{x}_i} = \max_{\boldsymbol{x}_i\in\mathcal{X}_a}\min_{\boldsymbol{x}_j\in\mathcal{S}}||\boldsymbol{h}_i - \boldsymbol{h}_j||_2$ denotes the max-min distance between the unannotated and annotated instances in the embedding space. Note that $a \in \mathcal{A}$, we take $\tilde{a} = \arg\max_{a\in\mathcal{A}}\int_{\mathcal{X}_a}\int_{\mathcal{Y}} p(\boldsymbol{x},y)l(\boldsymbol{h},y;\theta_h)\mathrm{d}y\mathrm{d}\boldsymbol{x}$. The generalization error difference between the two groups can be bounded by

$$\left|\int_{\mathcal{X}_0}\int_{\mathcal{Y}} p(\boldsymbol{x},y)l(\boldsymbol{h},y;\theta_h)\mathrm{d}\boldsymbol{x}\mathrm{d}y - \int_{\mathcal{X}_1}\int_{\mathcal{Y}} p(\boldsymbol{x},y)l(\boldsymbol{h},y;\theta_h)\mathrm{d}\boldsymbol{x}\mathrm{d}y\right| \leq \epsilon + (K_l + K_hL)\delta_{\tilde{a}}. \tag{21}$$

Combine Equation (21) with (15), we have the generalization error gap between group $x \in \mathcal{X}_0$ and group $x \in \mathcal{X}_1$ bounded as follow with probability $1-\gamma$,

$$\left|\int_{\mathcal{X}_0}\int_{\mathcal{Y}} p(\boldsymbol{x},y)l(\boldsymbol{h},y;\theta_h)\mathrm{d}\boldsymbol{x}\mathrm{d}y - \int_{\mathcal{X}_1}\int_{\mathcal{Y}} p(\boldsymbol{x},y)l(\boldsymbol{h},y;\theta_h)\mathrm{d}\boldsymbol{x}\mathrm{d}y\right| \leq \epsilon + \min\left\{\sqrt{-L^2\log\gamma(2N_{\tilde{a}})^{-1}}, (K_l + K_hL)\delta_{\tilde{a}}\right\}.$$

B Details about the Datasets

The experiments are conducted on the MEPS[12], Loan default[13], German credit[14], Adult[15] and CelebA[16] datasets to demonstrate the proposed framework is effective to mitigate the socially influential bias such as the gender, race or age bias. The statistics of the datasets is given in Table 1. The details about the datasets including the size and spliting of the datasets, the predicted and sensitive attributes, and the annotation budget are described as follows.

- **MEPS**: The task on this dataset is to predict whether a person would have a *high* or *low* utilization based on other features (*region, marriage, etc.*). The *Race* of each person is the sensitive attribute, where the two sensitive groups are *white* and *non-white*. The vanilla model shows discrimination towards the non-white group. The annotation budget is 8‰ [17].
- **Loan default**: The task is to predict whether a person would default the payment of loan based on personal information (*Bill amount, education, etc.*), where the sensitive attribute is *age*, and the two sensitive groups are people *above 35* and those *below 35*. The vanilla trained model shows discrimination towards the younger group. The annotation budget is 4‰.

[12] https://github.com/Trusted-AI/AIF360/tree/master/aif360/data/raw/meps
[13] https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients
[14] https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)
[15] http://archive.ics.uci.edu/ml/datasets/Adult
[16] http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html
[17] The cost of annotating 8‰ of training instances is affordable.

Table 1: Details about the datasets.

	Adult	MEPS	Loan default	German credit	CelebA
Domain	Social	Medical	Financial	Financial	Social
Data formate	Tabular	Tabular	Tabular	Tabular	Image
Predicted attribute	Salary	Utilization	Defaulting	Credit	Wavy hair Young
Sensitive attribute	Gender	Race	Age	Age	Gender
Number of instance	30162	15830	30000	4521	5000
Number of attribute	13	41	8	16	160×160
Train, Validate, Test splitting	0.25, 0.25 0.5	0.25, 0.25 0.5	0.25, 0.25 0.5	0.25, 0.25 0.5	0.25, 0.25 0.5
Annotation budget	4‰	8‰	4‰	2%	3%

Table 2: Detailed hyper-parameter setting.

	Adult	MEPS	Loan default	German Credit	CelebA-hair	CelebA-young
Classifier body f_b	Perceptron	Perceptron	Perceptron	Perceptron	ResNet-18	ResNet-18
Classifier head f_h	2-layer MLP	3-layer MLP	2-layer MLP	3-layer MLP	3-layer MLP	3-layer MLP
Classifier head f_a	Perceptron	Perceptron	Perceptron	Perceptron	Perceptron	Perceptron
Embedding dim M	64	32	64	32	256	256
Hidden-layer dim	32	32	32	32	64	128

- **German credit**: The goal of this dataset is to predict whether a person has *good* or *bad* credit risks based on other features (*balance, job, education, etc.*). *Age* is the sensitive attribute, where the two sensitive groups are people *older than 35* and those *not older than 35*. The vanilla trained model shows discrimination towards the younger group. The annotation budget is 2%.
- **Adult**: The task for this dataset is to predict whether a person has *high* (more than 50K/yr) or *low* (less than 50K/yr) income based on other features (*education, occupation, working hours, etc.*). *Gender* is considered as the sensitive attribute for this dataset. Thus, we have two sensitive groups *male* and *female*. The vanilla trained classification model shows discrimination towards the female group. The annotation budget is 4‰.
- **CelebA**: This is a large-scale image dataset of human faces. We consider two tasks for this dataset: i) identifying whether a person has wavy hair; ii) identifying whether a person is young. *Gender* is the sensitive feature, where the two sensitive groups are *male* and *female*. The vanilla trained model shows discrimination towards male in task i) and female in tasks ii), respectively. The annotation budget is 3%.

C Implementation Details

The experiment on each dataset follows the pipeline of **pre-training**, **debiasing**, and **head-selection**. Each step is shown as follows.

Pre-training: We pre-train $f_b(\bullet \mid \theta_b)$ to minimize the contrastive loss on the whole training set without any annotations for 50 epochs; and pre-train $f_h(\boldsymbol{h} \mid \theta_h)$ for 10 epochs to minimize the cross-entropy $\text{CE}(\hat{y}, y)$; then pre-train $f_b(\boldsymbol{h} \mid \theta_b)$ for 10 epochs to minimize the cross-entropy $\text{CE}(\hat{a}, a)$, where the initial sensitive annotations are very few (less than 10), randomly selected from each group. θ_b, θ_h and θ_a provide initial solutions for the bias mitigation.
Debiasing: We adopt APOD to debias the classifier head $f_h(\bullet \mid \theta_h)$ for several iterations. Specifically, the number of iterations equals the available annotation number, where APOD selects one instance for annotation, debiases $f_h(\bullet \mid \theta_h)$ and retrains $f_a(\bullet \mid \theta_a)$ for 10 epochs in each iteration, and back up the checkpoint of θ_h and θ_a in the last epoch of each iteration. In the Pre-training and Debiasing stages, the parameters θ_b, θ_h and θ_a are updated using the Adam optimizer with a learning rate of 10^{-3}, mini-batch size 256 and a dropout probability of 0.5. The DNN architectures and detailed hyper-parameter settings on different datasets are given in Appendix D.
Head-selection: We use the trained $f_a(f_b(\bullet \mid \theta_b) \mid \theta_a)$ to generate the proxy sensitive annotations for the validation dataset so that the fairness metrics can be estimated on the validation dataset. The optimal debiased classifier head f_h is selected to maximize the summation of accuracy and fairness score on the validation dataset. We merge the selected f_h with the pre-trained f_b and test the classifier $f_h(f_b(\bullet \mid \theta_b) \mid \theta_h)$ on the test dataset. This pipeline is executed five times to reduce the effect of randomness, and the average testing performance and the standard deviation are reported in the remaining sections.

D Detailed Hyper-parameter Setting

The detailed hyper-parameter setting is given in Table 2.

E Details about the Baseline Methods

We introduce details on the baseline methods in this section.

- **Group DRO**: Group DRO maintains a distribution $\boldsymbol{q} = [q_0, q_1]$ over the sensitive groups $a \in \mathcal{A}$, and updates the classifier $f(\bullet \mid \theta_f)$ via the min-max optimization given by

$$\theta_f = \arg\min_{\theta} \max_{\boldsymbol{q}} \sum_{a \in \mathcal{A}} \frac{q_a}{N_a} \sum_{(\boldsymbol{x}_i, y_i) \sim \mathcal{D}_a} l(\boldsymbol{x}_i, y_i; \theta), \tag{22}$$

  where $\mathcal{D}_a = \{\boldsymbol{x}_i, y_i \mid a_i = a\}$ depends on fully-annotated training set to generate the sensitive groups.
- **FAL**: Original FAL depends on the annotation of sensitive attribute to have active instance selection. Hence, we consider an improved version of the original framework to adapt to the problem in this work. Specifically, our improved

FAL updates the classifier to minimize the cross-entropy loss on the annotated dataset. The annotated instances are selected by

$$(\boldsymbol{x}^*, y^*) = \arg\max_{(x,y)\in\mathcal{U}} \alpha \mathrm{ACC}(f_t) + (1-\alpha)[\mathcal{F}(f_t) - \mathcal{F}(f_{t-1})], \tag{23}$$

where f_t denotes the classifier learned on the annotated dataset $\mathcal{S}$; $\mathcal{F}(f_t)$ denotes the fairness score of classifier f_t, which is the value of Equalized Odds in our experiment; α controls the trade-off between accuracy and fairness; and we have α in the range of $[0.5, 1]$ in our experiments.

- **LfF**: LfF adopts generalized cross entropy loss to learn the biased model f_B to provide proxy annotation, and simultaneously learn the debiased model f_D towards minimizing the cross entropy re-weighted by the proxy annotation. f_B and f_D are updated by

$$\begin{aligned} \theta_B^* &= \min_{\theta_B} \sum_{i=1}^{N} \frac{1 - p(\boldsymbol{x}_i; \theta_B)^q}{q}, \\ \theta_D^* &= \arg\min_{\theta_D} \sum_{i=1}^{N} \frac{l(\boldsymbol{x}_i, \hat{y}; \theta_B) l(\boldsymbol{x}_i, \hat{y}; \theta_D)}{l(\boldsymbol{x}_i, \hat{y}; \theta_B) + l(\boldsymbol{x}_i, \hat{y}; \theta_D)}, \end{aligned} \tag{24}$$

  where we control the hyper-parameter q in the range of $[2.5, 3]$ in our experiments.
- **SSBM**: This method initially randomly select a subset for annoatation, then adopts POD for the bias mitigation.
- **POD+RS**: Different from SSBM, this method randomly selects an annotated instance and adopts POD for bias mitigation in each iteration. The random instance selection and POD executes iteratively. This method is designed for studying the effect of annotation ratio to the mitigation performance.
- **POD+AL**: This method adopts POD for bias mitigation. Different from APOD, the annotated instances are selected by uncertainty sampling. Specifically, we calculate the Shannon entropy of the model prediction for each instance in the unannotated dataset. For $\boldsymbol{x}_i \in \mathcal{U}$, we have the entropy given by

$$\mathcal{H}(\boldsymbol{x}_i) = -p_{\hat{y}_i=1} \log_2 p_{\hat{y}_i=1} - p_{\hat{y}_i=0} \log_2 p_{\hat{y}_i=0}. \tag{25}$$

  where $[p_{\hat{y}_i=1}, p_{\hat{y}_i=0}] = \mathrm{softmax}[f(\boldsymbol{h}_i|\theta_h)]$; and $f(\boldsymbol{h}_i|\theta_h) \in \mathcal{Y}$. The instance for annotation is selected by

$$(\boldsymbol{x}^*, y^*) = \arg\max_{(\boldsymbol{x}_i, y_i)\in\mathcal{U}} \mathcal{H}(\boldsymbol{x}_i). \tag{26}$$

- **POD+CA**: This method adopts POD for bias mitigation. Different from APOD, POD+CA selects the instance for annotation following the max-min rule given by

$$(\boldsymbol{x}^*, y^*) = \arg\max_{\boldsymbol{x}_i\in\mathcal{U}} \min_{\boldsymbol{x}_j\in\mathcal{S}} ||\boldsymbol{h}_i - \boldsymbol{h}_j||_2, \tag{27}$$

  where $\mathcal{S}$ and $\mathcal{U}$ denote the annotated and unannotated datasets, respectively.