Fairness-Aware Low-Rank Representation Fine-Tuning

Parameswaran Kamalaruban [1] **Mark Anderson** [1] **Stuart Burrell** [1] **Maeve Madigan** [1] **Piotr Skalski** [1] **David Sutton** [1]

Abstract

Pre-trained foundation models can be efficiently adapted for specific tasks using Low-Rank Adaptation (LoRA), but the fairness properties of these adapted classifiers remain underexplored. Existing fairness-aware fine-tuning methods assume that sensitive attribute labels are available alongside downstream task labels, which often fails in practice due to user consent limitations or privacy constraints. To address this gap, we investigate fairness-aware LoRA fine-tuning using separate datasets for downstream tasks and sensitive attributes. We introduce four fairness-aware LoRA strategies: sensitive unlearning, adversarial debiasing, orthogonality-based disentanglement, and entropy maximization. Through comprehensive experiments on standard algorithmic fairness datasets using an ImageNet pre-trained ViT-Base model, we evaluate these methods across multiple utility and fairness metrics. Our orthogonality-based disentanglement and entropy maximization approaches consistently outperform standard fine-tuning in both overall utility and fairness, while adversarial debiasing shows less consistent improvements and sensitive unlearning proves ineffective for classification tasks. However, fairness-aware methods underperform on certain metrics like subgroup-wise false-positive rate ratios, highlighting fundamental incompatibilities between fairness objectives. These findings demonstrate the potential of fairness-aware LoRA fine-tuning while revealing inherent challenges of simultaneously optimizing multiple fairness criteria in parameter-efficient adaptation.

[1]Risk and Security AI Lab, Visa Inc. Correspondence to: Parameswaran Kamalaruban <kaparame@visa.com>.

Trustworthy AI for Good (AI4GOOD) Workshop at the 43rd International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

1. Introduction

Pre-trained foundation models have catalyzed remarkable advances across domains such as computer vision and natural language processing (Awais et al., 2025; Zhao et al., 2023; Zhou et al., 2024). Their ability to transfer learned representations to diverse downstream tasks has led to widespread adoption. However, these models also inherit biases present in their training data, potentially resulting in unfair or discriminatory downstream predictions (Bommasani et al., 2021; Ali et al., 2023). As models grow larger and are fine-tuned on specialized datasets to achieve peak performance, addressing these inherent biases becomes critical, particularly for sensitive applications in law, healthcare, or finance. Recent developments in parameter-efficient fine-tuning (PEFT) techniques (Han et al., 2024), such as Low-Rank Adaptation (LoRA) (Hu et al., 2022), enable efficient adaptation of massive pre-trained models by updating only a small subset of parameters. Despite its computational benefits and modular design, the fairness implications of LoRA and related PEFT methods remain insufficiently understood.

In this work, we explore LoRA adapters for bias mitigation in pre-trained foundation models. We consider a practically relevant setting where we lack a joint dataset $\mathcal{D} = \{(x, y, g)\}$ containing input features x, target labels y, and sensitive attributes g. Instead, we assume access to two separate datasets: a downstream task dataset $\mathcal{D}^{(\text{task})} = \{(x, y)\}$ and a sensitive attribute dataset $\mathcal{D}^{(\text{sen})} = \{(x', g)\}$, where x' is drawn from a distribution similar to x. This setting reflects practical constraints where sensitive attribute labels may be available only for a subset of data due to user consent limitations, or where privacy regulations prohibit maintaining joint datasets with both task and sensitive labels. Our objective is to fine-tune the frozen foundation model to maximize task performance while enforcing fairness by making predictions invariant to sensitive attributes.

Bias mitigation strategies are typically categorized as pre-processing (Feldman et al., 2015; Calmon et al., 2017; Kamiran & Calders, 2012), in-processing (Zhang et al., 2018; Kamishima et al., 2012; Agarwal et al., 2018; 2019), or post-processing (Pleiss et al., 2017; Hardt et al., 2016; Kamiran et al., 2012) approaches. We focus on in-processing methods that integrate fairness constraints directly into

model fine-tuning. Traditional fairness regularization techniques (Kamishima et al., 2012; Agarwal et al., 2019) typically require joint datasets $\mathcal{D} = \{(x, y, g)\}$ to impose fairness penalties and are thus inapplicable to our setting. Instead, we investigate fair representation learning approaches that decouple sensitive attribute influences from learned representations. While fairness regularization methods optimize specific group fairness metrics (e.g., equalized odds, demographic parity) that are often mutually incompatible (Kim et al., 2020; Berk et al., 2021), fair representation learning aims to remove sensitive information from representations themselves, naturally supporting our separate dataset scenario.

Despite growing interest in LoRA's fairness implications, most studies focus on fairness-unaware fine-tuning. Recent work (Ding et al., 2024) evaluates subgroup fairness properties of LoRA but limits evaluation to standard downstream fine-tuning. Similarly, approaches like FairLoRA (Sukumaran et al., 2024) and FairTune (Dutt et al., 2024) integrate fairness objectives but require joint datasets $\mathcal{D} = \{(x, y, g)\}$. In contrast, we leverage separate datasets (x, y) and (x', g) to develop four fairness-aware LoRA strategies: (i) sensitive unlearning (UNL), (ii) adversarial debiasing via gradient reversal (ADV), (iii) orthogonality-based disentanglement that decorrelates task and sensitive subspaces (DIS), and (iv) entropy maximization that encourages maximal uncertainty in sensitive predictions (ENT).

Contributions. Our contributions are:

1. We introduce four fairness-aware LoRA fine-tuning methods applicable to our separate dataset setting.
2. We comprehensively benchmark these methods against a fairness-unaware baseline (ERM) across multiple utility and fairness metrics.
3. Through extensive experiments on standard algorithmic fairness datasets using an ImageNet pre-trained 86M ViT-Base model, we demonstrate that DIS and ENT consistently outperform ERM in both utility and fairness, while ADV shows improvements but less consistently. UNL proves ineffective for classification, often underperforming ERM.
4. We empirically highlight incompatibilities between fairness metrics, observing that while fairness-aware methods generally improve performance, they underperform ERM specifically on subgroup-wise false-positive rate ratios, illustrating the challenges of optimizing multiple fairness objectives simultaneously.

2. Related Work

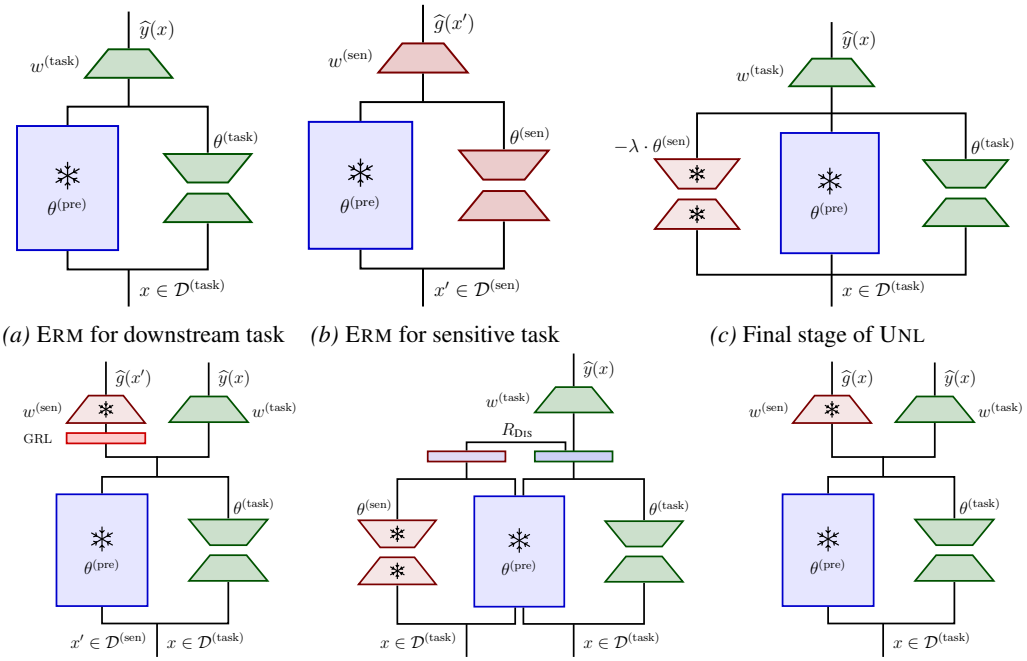
Algorithmic Fairness in Machine Learning. Algorithmic fairness is a rapidly evolving field focused on ensuring equitable outcomes in AI decision-making systems. Various approaches have been proposed to measure and mitigate biases within algorithms (Verma & Rubin, 2018; Bellamy et al., 2019; Weerts et al., 2023; Barocas et al., 2023). Bias mitigation strategies are generally categorized into pre-processing, in-processing, and post-processing methods. Pre-processing methods modify the training data to reduce bias before model training (Feldman et al., 2015; Calmon et al., 2017; Kamiran & Calders, 2012), while post-processing techniques adjust model predictions to improve fairness after training (Pleiss et al., 2017; Hardt et al., 2016; Kamiran et al., 2012). In contrast, in-processing approaches incorporate fairness constraints directly into the learning objective and have been extensively studied in classical machine learning models (Zhang et al., 2018; Kamishima et al., 2012; Agarwal et al., 2018; 2019). For a more comprehensive review of fairness and bias mitigation strategies, we refer interested readers to recent surveys (Mehrabi et al., 2021; Caton & Haas, 2024; Hort et al., 2024; Wan et al., 2023; Ashurst & Weller, 2023).

Fairness-Aware Fine-Tuning. Recent studies have explored fairness-aware fine-tuning in deep learning. For instance, recent work (Mao et al., 2023) investigates last-layer fairness fine-tuning, where fairness constraints are introduced as regularization terms during the fine-tuning phase to mitigate bias in pre-trained deep neural networks. Methods such as FairLoRA (Sukumaran et al., 2024) and FairTune (Dutt et al., 2024) have been proposed to integrate fairness considerations directly into the fine-tuning process. FairLoRA augments the downstream classification loss with a fairness regularization term based on specific fairness metrics, whereas FairTune adopts a bi-level optimization framework where an inner loop performs standard downstream fine-tuning with masked LoRA modules and an outer loop adjusts these masks to optimize a fairness objective on a validation set. However, these approaches typically require joint datasets $\mathcal{D} = \{(x, y, g)\}$.

3. Problem Setup and Background

Problem Setup. We consider a multi-class classification problem with a sensitive attribute. The downstream task dataset is denoted as $\mathcal{D}^{(\text{task})} = \{(x, y)\}$, where $x \in \mathcal{X}$ represents the input features and $y \in \mathcal{Y}$ is the target label. The sensitive task dataset is given by $\mathcal{D}^{(\text{sen})} = \{(x', g)\}$, with $g \in \mathcal{G}$ representing the sensitive attribute and the inputs drawn from a distribution similar to that of $\mathcal{D}^{(\text{task})}$. Given a pre-trained, frozen foundation model, our objective is to fine-tune it so that the resulting downstream predictor achieves high task performance while promoting fairness by making its predictions invariant to the sensitive attribute g.

Utility and Fairness Metrics. We assess model performance using a comprehensive set of utility and group fair-



(a) ERM for downstream task *(b)* ERM for sensitive task *(c)* Final stage of UNL

(d) Final stage of ADV *(e)* Final stage of DIS *(f)* Final stage of ENT

Figure 1. Empirical risk minimization baseline and fairness-aware fine-tuning methods.

ness metrics (Weerts et al., 2023; Hardt et al., 2016; Chouldechova, 2017; Dwork et al., 2012). Utility is measured through standard metrics such as accuracy (ACC), precision (PPV), recall (TPR), false positive rate (FPR), area under the receiver operating characteristic curve (AUROC), and area under the precision-recall curve (AUPRC). In addition to these, we evaluate group fairness with respect to the sensitive attribute g (e.g., gender, race) by considering both difference and ratio metrics. For a utility metric U, let U_g denote its value on subgroup g. The difference metric is $\Delta U = \max_{g,g' \in \mathcal{G}} |U_g - U_{g'}|$ and the ratio metric is U-ratio $= \min_{g,g' \in \mathcal{G}} \frac{U_g}{U_{g'}}$. Smaller ΔU and larger U-ratio indicate improved group fairness.

Low-Rank Adaptation. LoRA is a widely used PEFT method introduced by Hu et al. (Hu et al., 2022) that adapts pre-trained models by learning low-dimensional updates to the weight matrices. In LoRA, given a pre-trained weight matrix $W_0 \in \mathbb{R}^{d \times k}$, the model update is parameterized as a low-rank decomposition: $W = W_0 + \Delta W = W_0 + BA^\top$, where $A \in \mathbb{R}^{d \times r}$ and $B \in \mathbb{R}^{r \times k}$, with rank $r \ll \min(d, k)$. The matrices are initialized such that A is sampled from a Gaussian distribution, $A \sim \mathcal{N}(0, \sigma^2)$ for a small σ, and B is set to the zero matrix, ensuring that the initial update ΔW is zero and the pre-trained model's behavior is preserved. During training, W_0 remains frozen, and only A and B are updated. For transformer-based architectures, LoRA adapters are typically applied to the query and value matrices of the self-attention layers, while an additional task-specific head is attached to the last layer for supervised learning. This approach significantly reduces the number of trainable parameters while maintaining competitive performance on downstream tasks. Recently, (Zhang et al., 2025) observed that freezing the randomly initialized LoRA-A matrices during training reduces parameter interference when merging multiple task-specific adapters; we adopt this design choice across all our fine-tuning methods.

4. Fair Representation Fine-Tuning

In this section, we present a set of LoRA-based fine-tuning strategies for learning fair representations in classification tasks. Our goal is to adapt a pre-trained model such that the learned representations are *predictive* of the target attribute while being *invariant* to a given sensitive attribute (e.g., gender, race), thereby mitigating the influence of sensitive information on downstream predictions.

We begin with a standard empirical risk minimization (ERM) baseline, which fine-tunes the model solely for target prediction without any fairness constraint. While ERM often yields strong task accuracy, it does not address representation bias and may inadvertently preserve or even amplify correlations between sensitive and target attributes.

To address this, we introduce four fairness-aware LoRA fine-tuning methods designed to reduce predictability of the sensitive attribute:

1. Unlearning-based debiasing (UNL), which adapts an unlearning approach from language model detoxification (Zhang et al., 2023a) to classification, explicitly removing sensitive attribute information via a learned sensitive adapter;
2. Adversarial debiasing (ADV), which builds on adversarial representation learning (Ganin et al., 2016; Adel et al., 2019) to encourage target-relevant but sensitive-invariant features via a gradient reversal layer, with stabilization enhancements to mitigate typical adversarial training instabilities;
3. Orthogonality-based disentanglement (DIS), a novel regularization (inspired by disentangled representation learning (Creager et al., 2019; Locatello et al., 2019)) that enforces orthogonality between target-task and sensitive-attribute subspaces in the representation, thereby decorrelating the two; and
4. Entropy maximization regularization (ENT), a stable alternative to adversarial objectives that directly maximizes the entropy of sensitive attribute predictions, pushing them toward maximal uncertainty (Roy & Boddeti, 2019).

Empirical risk minimization (ERM). This baseline adapts a frozen pre-trained model $\theta^{(\text{pre})}$ to a downstream task by learning a task-specific LoRA adapter $\theta^{(\text{task})}_{\text{ERM}}$ and classification head $w^{(\text{task})}_{\text{ERM}}$. Training minimizes the cross-entropy loss on the downstream dataset $\mathcal{D}^{(\text{task})}$ with an additional norm regularization (Zhang et al., 2023b) to stabilize LoRA parameters. The regularizer is $R_{\text{NORM}}(\theta) = \sum_i \|(A_i)^\top A_i - I\|_F^2 + \|(B_i)^\top B_i - I\|_F^2$, where $\theta = \{(A_i, B_i)\}$ are LoRA matrices. The training objective is:

$$(\theta^{(\text{task})}_{\text{ERM}}, w^{(\text{task})}_{\text{ERM}}) \leftarrow \underset{\theta^{(\text{task})}, w^{(\text{task})}}{\arg\min} \big[\lambda_{\text{NORM}} \cdot R_{\text{NORM}}(\theta^{(\text{task})}) + \ell_{\text{CE}}(\theta^{(\text{pre})} \oplus \theta^{(\text{task})}, w^{(\text{task})}; \mathcal{D}^{(\text{task})})\big], \quad (1)$$

where $\oplus$ denotes adding the LoRA adapter to the frozen backbone. The hyperparameter λ_{NORM} controls the strength of the regularization.

Unlearning-based debiasing (UNL). This method removes sensitive attribute information from the frozen backbone before downstream training. We first learn a sensitive attribute adapter $\theta^{(\text{sen})}_{\text{ERM}}$ and classification head on the sensitive dataset $\mathcal{D}^{(\text{sen})}$ by minimizing the following objective:

$$(\theta^{(\text{sen})}_{\text{ERM}}, w^{(\text{sen})}_{\text{ERM}}) \leftarrow \underset{\theta^{(\text{sen})}, w^{(\text{sen})}}{\arg\min} \big[\lambda_{\text{NORM}} \cdot R_{\text{NORM}}(\theta^{(\text{sen})}) + \ell_{\text{CE}}(\theta^{(\text{pre})} \oplus \theta^{(\text{sen})}, w^{(\text{sen})}; \mathcal{D}^{(\text{sen})})\big], \quad (2)$$

with the pre-trained model $\theta^{(\text{pre})}$ remaining frozen. Once $\theta^{(\text{sen})}_{\text{ERM}}$ is trained, we *unlearn* sensitive information by subtracting its scaled contribution from the frozen pre-trained weights: $\theta^{(\text{pre})} \ominus \lambda_{\text{UNL}} \cdot \theta^{(\text{sen})}_{\text{ERM}}$, where $\ominus$ denotes negating either the LoRA-(A) or LoRA-(B) weights, and λ_{UNL} controls the removal strength. Finally, we train a downstream task adapter $\theta^{(\text{task})}_{\text{UNL}}$ and its classification head $w^{(\text{task})}_{\text{UNL}}$ on $\mathcal{D}^{(\text{task})}$:

$$(\theta^{(\text{task})}_{\text{UNL}}, w^{(\text{task})}_{\text{UNL}}) \leftarrow \underset{\theta^{(\text{task})}, w^{(\text{task})}}{\arg\min} \big[\lambda_{\text{NORM}} \cdot R_{\text{NORM}}(\theta^{(\text{task})}) + \ell_{\text{CE}}(\theta^{(\text{pre})} \ominus \lambda_{\text{UNL}} \cdot \theta^{(\text{sen})}_{\text{ERM}} \oplus \theta^{(\text{task})}, w^{(\text{task})}; \mathcal{D}^{(\text{task})})\big], \quad (3)$$

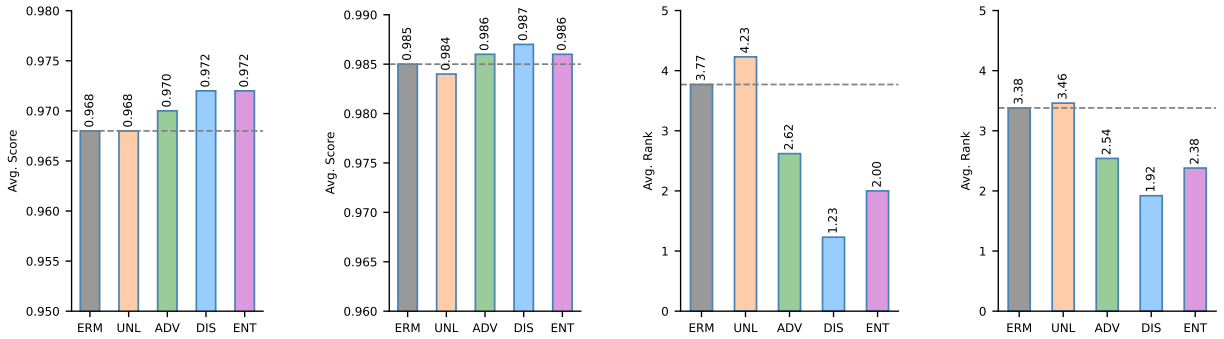
while keeping both $\theta^{(\text{pre})}$ and $\theta^{(\text{sen})}_{\text{ERM}}$ frozen.

Adversarial debiasing (ADV). This method learns task-relevant representations while suppressing sensitive attribute information via adversarial training. The process consists of three stages. We first pre-train a downstream task adapter $\theta^{(\text{task})}_{\text{ERM}}$ and its classification head $w^{(\text{task})}_{\text{ERM}}$ on $\mathcal{D}^{(\text{task})}$ via Eq. (1). Next, using the frozen $\theta^{(\text{task})}_{\text{ERM}}$, we train a sensitive attribute classification head $w^{(\text{sen})}_{\text{ERM}}$ on $\mathcal{D}^{(\text{sen})}$ by minimizing $w^{(\text{sen})}_{\text{ERM}} \leftarrow \arg\min_{w^{(\text{sen})}} \ell_{\text{CE}}(\theta^{(\text{pre})} \oplus \theta^{(\text{task})}_{\text{ERM}}, w^{(\text{sen})}; \mathcal{D}^{(\text{sen})})$, while keeping both $\theta^{(\text{pre})}$ and $\theta^{(\text{task})}_{\text{ERM}}$ frozen. Finally, we adversarially fine-tune the downstream task adapter and the classification head—initialized from the first stage—by simultaneously minimizing the downstream task loss and maximizing the sensitive attribute prediction loss (via a gradient reversal layer (Ganin et al., 2016) applied before $w^{(\text{sen})}_{\text{ERM}}$). The resulting objective is

$$(\theta^{(\text{task})}_{\text{ADV}}, w^{(\text{task})}_{\text{ADV}}) \leftarrow \underset{\theta^{(\text{task})}, w^{(\text{task})}}{\arg\min} \big[\lambda_{\text{NORM}} \cdot R_{\text{NORM}}(\theta^{(\text{task})}) - \lambda_{\text{ADV}} \cdot \ell_{\text{CE}}(\theta^{(\text{pre})} \oplus \theta^{(\text{task})}, w^{(\text{sen})}_{\text{ERM}}; \mathcal{D}^{(\text{sen})}) + \ell_{\text{CE}}(\theta^{(\text{pre})} \oplus \theta^{(\text{task})}, w^{(\text{task})}; \mathcal{D}^{(\text{task})})\big], \quad (4)$$

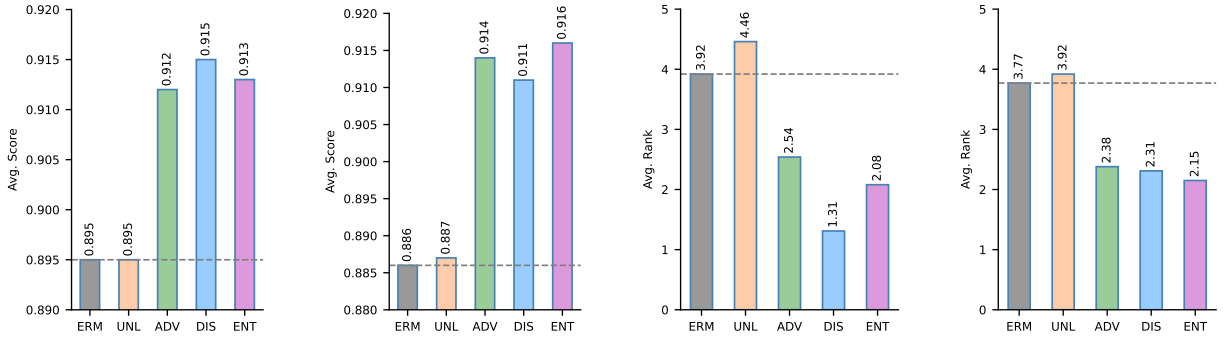
where $\theta^{(\text{task})}$ and $w^{(\text{task})}$ are initialized from stage one, and $\theta^{(\text{pre})}$ and $w^{(\text{task})}_{\text{ERM}}$ remain frozen.

Orthogonality-based disentanglement (DIS). This method aims to prevent sensitive attribute information from being entangled with downstream task representations by explicitly enforcing orthogonality between them. We first train a sensitive adapter $\theta^{(\text{sen})}_{\text{ERM}}$ and its classification head $w^{(\text{sen})}_{\text{ERM}}$ using Eq. (2). Keeping both $\theta^{(\text{pre})}$ and $\theta^{(\text{sen})}_{\text{ERM}}$ frozen, we then train a downstream task adapter $\theta^{(\text{task})}_{\text{DIS}}$ and its classification head $w^{(\text{task})}_{\text{DIS}}$ on $\mathcal{D}^{(\text{task})}$ using a composite objective that combines the standard cross-entropy loss, and an orthogonality regularizer that penalizes cosine similarity between task and sensitive representations. For an input x, let $h_{\text{task}}(x) = f(x; \theta^{(\text{pre})} \oplus \theta^{(\text{task})})$ and $h_{\text{sen}}(x) = f(x; \theta^{(\text{pre})} \oplus \theta^{(\text{sen})}_{\text{ERM}})$ denote the task and sensitive representations, respectively, where $f(\cdot)$ is the frozen backbone plus adapter up to the representation layer. The orthogonality regularizer is defined as $R_{\text{DIS}}(\theta^{(\text{pre})} \oplus \theta^{(\text{task})}, \theta^{(\text{pre})} \oplus$



(a) Overall AUROC Score (↑) *(b)* AUROC Ratio Score (↑) *(c)* Overall AUROC Rank (↓) *(d)* AUROC Ratio Rank (↓)

Figure 2. Average overall AUROC and group-wise AUROC ratio (ratio of the best- to worst-performing subgroup) scores and ranks for each method, aggregated over all experiments.



(a) Overall AUPRC Score (↑) *(b)* AUPRC Ratio Score (↑) *(c)* Overall AUPRC Rank (↓) *(d)* AUPRC Ratio Rank (↓)

Figure 3. Average overall AUPRC and group-wise AUPRC ratio (ratio of the best- to worst-performing subgroup) scores and ranks for each method, aggregated over all experiments.

$\theta^{(\text{sen})}_{\text{ERM}}; \mathcal{D}^{(\text{task})}) = \frac{1}{|\mathcal{D}^{(\text{task})}|} \sum_{x \in \mathcal{D}^{(\text{task})}} \cos(h_{\text{task}}(x), h_{\text{sen}}(x))$, where $\cos(u, v) = \frac{u^\top v}{\|u\| \cdot \|v\|}$ is the cosine similarity. The complete training objective is

$$(\theta^{(\text{task})}_{\text{DIS}}, w^{(\text{task})}_{\text{DIS}}) \leftarrow \underset{\theta^{(\text{task})}, w^{(\text{task})}}{\arg\min} \big[\lambda_{\text{NORM}} \cdot R_{\text{NORM}}(\theta^{(\text{task})}) + \lambda_{\text{DIS}} \cdot R_{\text{DIS}}(\theta^{(\text{pre})} \oplus \theta^{(\text{task})}, \theta^{(\text{pre})} \oplus \theta^{(\text{sen})}_{\text{ERM}}; \mathcal{D}^{(\text{task})}) + \ell_{\text{CE}}(\theta^{(\text{pre})} \oplus \theta^{(\text{task})}, w^{(\text{task})}; \mathcal{D}^{(\text{task})})\big], \tag{5}$$

where λ_{DIS} controls the strength of the orthogonality constraint, encouraging the downstream adapter to capture information complementary to (and thus disentangled from) the sensitive attribute.

Entropy maximization regularization (ENT). This method follows the first two stages of ADV but replaces adversarial training with entropy maximization to encourage uncertainty in sensitive attribute predictions. After pre-training $(\theta^{(\text{task})}_{\text{ERM}}, w^{(\text{task})}_{\text{ERM}})$ on $\mathcal{D}^{(\text{task})}$ and training $w^{(\text{task})}_{\text{SEN}}$ on $\mathcal{D}^{(\text{sen})}$, we fine-tune $(\theta^{(\text{task})}, w^{(\text{task})})$—initialized from stage one—using $R_{\text{ENT}}(\theta^{(\text{pre})} \oplus \theta^{(\text{task})}, w^{(\text{sen})}_{\text{ERM}}; \mathcal{D}^{(\text{task})}) = -\frac{1}{|\mathcal{D}^{(\text{task})}|} \sum_{x \in \mathcal{D}^{(\text{task})}} H(p_{\text{sen}}(x))$, where $p_{\text{sen}}(x) = \text{softmax}(w^{(\text{sen})}_{\text{ERM}} \circ h_{\text{task}}(x))$ is the sensitive attribute probability distribution predicted by the frozen sensitive head, $h_{\text{task}}(x) = f(x; \theta^{(\text{pre})} \oplus \theta^{(\text{task})})$ is the task representation, and $H(p) = -\sum_c p_c \log p_c$ is the Shannon entropy. Maximizing entropy pushes sensitive predictions toward uniformity, reducing leakage. The final training objective is

$$(\theta^{(\text{task})}_{\text{ENT}}, w^{(\text{task})}_{\text{ENT}}) \leftarrow \underset{\theta^{(\text{task})}, w^{(\text{task})}}{\arg\min} \big[\lambda_{\text{NORM}} \cdot R_{\text{NORM}}(\theta^{(\text{task})}) + \lambda_{\text{ENT}} \cdot R_{\text{ENT}}(\theta^{(\text{pre})} \oplus \theta^{(\text{task})}, w^{(\text{sen})}_{\text{ERM}}; \mathcal{D}^{(\text{task})}) + \ell_{\text{CE}}(\theta^{(\text{pre})} \oplus \theta^{(\text{task})}, w^{(\text{task})}; \mathcal{D}^{(\text{task})})\big], \tag{6}$$

where $(\theta^{(\text{task})}, w^{(\text{task})})$ are initialized from stage one, and $\theta^{(\text{pre})}$ and $w^{(\text{task})}_{\text{ERM}}$ remain frozen.

5. Experiments

In this section, we conduct a comprehensive evaluation of the fine-tuning strategies introduced in Section 4, assessing both utility and fairness performance.

Metric	Method	UTKFace		FairFace		WaterBird
		Age × Race	Gender × Race	Age × Race	Gender × Race	Type × Background
ROC (↑)	ERM	0.942 ± 0.001	0.985 ± 0.001	0.915 ± 0.000	0.984 ± 0.000	0.981 ± 0.001
	UNL	0.945 ± 0.000	0.985 ± 0.001	0.913 ± 0.001	0.981 ± 0.000	0.982 ± 0.003
	ADV	0.950 ± 0.000	0.985 ± 0.001	0.916 ± 0.001	0.986 ± 0.000	0.977 ± 0.002
	DIS	**0.951 ± 0.001**	0.986 ± 0.001	0.919 ± 0.000	**0.987 ± 0.000**	0.986 ± 0.002
	ENT	0.947 ± 0.007	**0.988 ± 0.001**	**0.922 ± 0.001**	**0.987 ± 0.000**	**0.994 ± 0.001**
ΔROC (↓)	ERM	0.048 ± 0.006	0.031 ± 0.002	0.019 ± 0.001	0.036 ± 0.004	**0.004 ± 0.001**
	UNL	0.049 ± 0.007	**0.026 ± 0.002**	0.023 ± 0.002	0.037 ± 0.003	0.007 ± 0.002
	ADV	**0.038 ± 0.002**	**0.026 ± 0.002**	0.020 ± 0.000	0.031 ± 0.003	0.007 ± 0.003
	DIS	0.042 ± 0.006	0.028 ± 0.001	**0.018 ± 0.001**	0.029 ± 0.003	0.005 ± 0.001
	ENT	0.046 ± 0.008	0.027 ± 0.001	**0.018 ± 0.001**	**0.028 ± 0.002**	**0.004 ± 0.001**
ROC ratio (↑)	ERM	0.950 ± 0.006	0.968 ± 0.002	0.979 ± 0.001	0.964 ± 0.004	**0.996 ± 0.001**
	UNL	0.948 ± 0.007	**0.974 ± 0.002**	0.975 ± 0.002	0.962 ± 0.003	0.993 ± 0.002
	ADV	**0.960 ± 0.002**	0.973 ± 0.002	0.978 ± 0.001	0.969 ± 0.003	0.993 ± 0.003
	DIS	0.956 ± 0.006	0.972 ± 0.002	**0.981 ± 0.001**	0.971 ± 0.003	0.995 ± 0.001
	ENT	0.953 ± 0.008	0.973 ± 0.001	**0.981 ± 0.001**	**0.972 ± 0.002**	**0.996 ± 0.001**
PR (↑)	ERM	0.676 ± 0.010	0.978 ± 0.002	0.558 ± 0.003	0.982 ± 0.000	0.951 ± 0.004
	UNL	0.707 ± 0.013	0.978 ± 0.003	0.555 ± 0.004	0.979 ± 0.000	0.961 ± 0.003
	ADV	0.814 ± 0.002	0.979 ± 0.001	0.575 ± 0.000	0.984 ± 0.000	0.947 ± 0.004
	DIS	0.815 ± 0.014	0.981 ± 0.002	0.583 ± 0.004	**0.986 ± 0.000**	0.962 ± 0.004
	ENT	**0.822 ± 0.006**	**0.983 ± 0.002**	**0.594 ± 0.002**	0.984 ± 0.000	**0.980 ± 0.002**
ΔPR (↓)	ERM	0.237 ± 0.017	0.039 ± 0.010	0.099 ± 0.006	0.031 ± 0.004	0.052 ± 0.017
	UNL	0.190 ± 0.010	0.037 ± 0.011	0.107 ± 0.006	0.033 ± 0.002	0.035 ± 0.012
	ADV	**0.105 ± 0.009**	0.030 ± 0.004	0.104 ± 0.006	0.027 ± 0.004	0.067 ± 0.019
	DIS	0.164 ± 0.046	0.031 ± 0.008	0.107 ± 0.005	0.025 ± 0.003	0.031 ± 0.012
	ENT	0.144 ± 0.038	**0.028 ± 0.002**	**0.079 ± 0.019**	**0.025 ± 0.002**	**0.021 ± 0.014**
PR ratio (↑)	ERM	0.736 ± 0.015	0.960 ± 0.010	0.835 ± 0.009	0.969 ± 0.004	0.947 ± 0.018
	UNL	0.768 ± 0.007	0.962 ± 0.012	0.825 ± 0.007	0.967 ± 0.002	0.964 ± 0.012
	ADV	**0.881 ± 0.008**	0.970 ± 0.004	0.834 ± 0.009	0.973 ± 0.004	0.931 ± 0.019
	DIS	0.817 ± 0.050	0.969 ± 0.008	0.831 ± 0.007	0.975 ± 0.003	0.968 ± 0.012
	ENT	0.840 ± 0.040	**0.972 ± 0.002**	**0.875 ± 0.029**	**0.975 ± 0.002**	**0.979 ± 0.014**

Table 1. Performance evaluation of fine-tuning strategies on UTKFace, FairFace, and WaterBird datasets, averaged over three random seeds. The best values are shown in **bold**. The scores that improve upon ERM by more than 0.005 are highlighted in green, while degradations greater than 0.005 are shown in red. Changes within ±0.005 of ERM score are considered insignificant.

5.1. Experimental Setup

Datasets and Base Model. We evaluate on widely used image datasets in the algorithmic fairness literature: CelebA (Liu et al., 2015), UTK-Face (Zhang et al., 2017), FairFace (Karkkainen & Joo, 2021), and WaterBirds (Sagawa et al., 2020). CelebA comprises 202,599 face images with 40 binary attribute labels. UTK-Face contains 20,000 face images annotated with gender, age, and race. FairFace contains 108,501 face images with gender, age, and race annotations and features a race-balanced distribution. WaterBird contains 11,708 bird images annotated with bird type and background. The downstream and sensitive attributes of each dataset are summarized in Table 4 in the appendix of the supplementary material. All images are resized to 224×224 to match our base model's input and are normalized prior to training. We randomly split each dataset into training (70%), validation (15%), and test (15%) sets. We use an ImageNet-pretrained Vision Transformer (ViT-Base) (Dosovitskiy et al., 2021), specifically the `vit-base-patch16-224-in21k` version with 86.6M parameters[1].

[1] https://huggingface.co/google/vit-base-patch16-224-in21k

Training Details. We employ a class-balanced sampling strategy and optimize the models using the AdamW optimizer (Loshchilov & Hutter, 2024) with a batch size of 256, a learning rate of 1e-4, and a weight decay of 5e-4. Training runs for 10 epochs with a CosineAnnealingLR scheduler, where $T_{\max} = 5$. We enable early stopping with a patience of 5 consecutive epochs if validation accuracy remains within a 2% band. For the LoRA adapters, we set the rank to 4, the alpha parameter to 8, and apply a dropout rate of 0.1. Our implementation utilizes the PEFT library from Hugging Face (Mangrulkar et al., 2022). Method-specific hyperparameters are fixed across datasets as follows: $\lambda_{\text{NORM}} = 1\text{e}{-5}$, $\lambda_{\text{UNL}} = 1.0$, $\lambda_{\text{ADV}} = 0.1$, $\lambda_{\text{DIS}} = 0.1$, and $\lambda_{\text{ENT}} = 1.0$.

5.2. Results

We evaluate the fine-tuning strategies from Section 4 using both utility and fairness metrics, implemented with `scikit-learn` (Pedregosa et al., 2011) and `Fairlearn` (Weerts et al., 2023). Precision, recall, and false-positive rate (FPR) are macro-averaged across classes to accommodate both binary and multiclass settings. All results are averaged over three random seeds.

Metric	Method	CelebA			
		Bald × Gender	Black hair × Gender	Eyeglasses × Gender	Smiling × Gender
ROC (↑)	ERM	0.995 ± 0.000	0.965 ± 0.000	0.998 ± 0.000	0.982 ± 0.000
	UNL	0.994 ± 0.001	0.963 ± 0.000	0.998 ± 0.000	0.981 ± 0.000
	ADV	**0.997 ± 0.000**	0.967 ± 0.000	**0.999 ± 0.000**	0.983 ± 0.000
	DIS	**0.997 ± 0.000**	**0.968 ± 0.000**	**0.999 ± 0.000**	**0.984 ± 0.000**
	ENT	0.990 ± 0.001	**0.968 ± 0.000**	0.998 ± 0.001	0.984 ± 0.001
ΔROC (↓)	ERM	0.008 ± 0.003	0.016 ± 0.002	0.003 ± 0.001	0.008 ± 0.001
	UNL	0.008 ± 0.004	0.017 ± 0.001	0.002 ± 0.001	0.010 ± 0.001
	ADV	0.019 ± 0.012	0.015 ± 0.001	**0.001 ± 0.000**	0.008 ± 0.001
	DIS	**0.005 ± 0.002**	0.015 ± 0.001	**0.001 ± 0.000**	**0.007 ± 0.001**
	ENT	0.010 ± 0.002	**0.014 ± 0.001**	0.003 ± 0.001	**0.007 ± 0.001**
ROC ratio (↑)	ERM	0.992 ± 0.003	0.983 ± 0.002	0.997 ± 0.001	0.991 ± 0.001
	UNL	0.992 ± 0.004	0.982 ± 0.001	0.998 ± 0.001	0.990 ± 0.001
	ADV	0.981 ± 0.012	0.984 ± 0.001	**0.999 ± 0.000**	0.992 ± 0.001
	DIS	**0.995 ± 0.002**	0.985 ± 0.001	**0.999 ± 0.000**	**0.993 ± 0.001**
	ENT	0.990 ± 0.002	**0.986 ± 0.001**	0.997 ± 0.001	0.992 ± 0.001
PR (↑)	ERM	0.828 ± 0.010	0.894 ± 0.001	0.987 ± 0.001	0.982 ± 0.000
	UNL	0.808 ± 0.005	0.888 ± 0.001	0.985 ± 0.001	0.982 ± 0.000
	ADV	0.861 ± 0.008	0.900 ± 0.001	**0.993 ± 0.001**	0.984 ± 0.000
	DIS	**0.869 ± 0.012**	**0.903 ± 0.001**	0.993 ± 0.002	**0.985 ± 0.000**
	ENT	0.851 ± 0.004	0.900 ± 0.001	0.988 ± 0.001	0.984 ± 0.001
ΔPR (↓)	ERM	0.227 ± 0.109	0.013 ± 0.003	0.004 ± 0.003	0.019 ± 0.000
	UNL	0.230 ± 0.077	**0.009 ± 0.004**	0.004 ± 0.002	0.021 ± 0.001
	ADV	**0.092 ± 0.047**	0.010 ± 0.002	**0.004 ± 0.001**	0.018 ± 0.000
	DIS	0.127 ± 0.013	0.011 ± 0.003	0.005 ± 0.002	**0.017 ± 0.000**
	ENT	0.136 ± 0.015	0.015 ± 0.002	0.009 ± 0.003	0.017 ± 0.001
PR ratio (↑)	ERM	0.743 ± 0.130	0.986 ± 0.003	0.996 ± 0.003	0.981 ± 0.000
	UNL	0.718 ± 0.098	**0.990 ± 0.004**	0.996 ± 0.002	0.979 ± 0.001
	ADV	**0.902 ± 0.049**	0.989 ± 0.002	**0.996 ± 0.001**	0.982 ± 0.000
	DIS	0.868 ± 0.012	0.987 ± 0.003	0.995 ± 0.002	**0.983 ± 0.000**
	ENT	0.858 ± 0.010	0.983 ± 0.002	0.991 ± 0.003	0.983 ± 0.001

Table 2. Performance evaluation of fine-tuning strategies on CelebA dataset, averaged over three random seeds. The best values are shown in **bold**. The scores that improve upon ERM by more than 0.005 are highlighted in green, while degradations greater than 0.005 are shown in red. Changes within ±0.005 of the ERM score are considered insignificant.

Tables 1, 2, and 3 report AUC-based results, covering both utility (overall AUC) and fairness (subgroup-wise AUC ratio and difference). For AUROC, the ERM baseline exhibits strong utility across datasets (overall AUROC > 0.90) with small subgroup disparities (difference < 0.05; ratio > 0.95). Averaging method-wise AUROC scores and their rankings across datasets (Figure 2), we observe the ordering DIS $\succ$ ENT $\succ$ ADV $\succ$ ERM $\succ$ UNL for both overall AUROC and subgroup-wise AUROC ratio. Thus, aside from UNL, all fairness-aware fine-tuning methods improve upon ERM. Notably, UNL—effective in generative detoxification–does not translate to gains for classification.

For AUPRC, ERM remains generally strong but shows utility dips for UTKFace (Age), FairFace (Age), and CelebA (Bald, Black Hair), alongside larger subgroup disparities for UTKFace (Age), FairFace (Age), WaterBird (Type), and CelebA (Bald, Attractive, Blond Hair). In contrast to AUROC, DIS, ENT, and ADV yield more pronounced gains over ERM in both utility and fairness. Aggregating AUPRC scores and rankings across datasets (Figure 3) gives DIS $\succ$ ENT $\succ$ ADV $\succ$ ERM $\succ$ UNL for overall AUPRC (utility) and ENT $\succ$ DIS $\succ$ ADV $\succ$ ERM $\succ$ UNL for the subgroup-wise AUPRC ratio (fairness), again showing all fairness-aware methods outperform ERM except UNL.

Additional metrics appear in the appendix due to space constraints (Tables 5, 6, and 7), reporting overall utility and subgroup-wise ratio/difference for accuracy, precision, recall, and FPR. For accuracy, ERM is typically strong in utility but drops below 0.90 on UTKFace (Age), FairFace (Age), and CelebA (Black Hair, Young, Attractive), with larger disparities (difference > 0.05, ratio < 0.95) for UTKFace (Age, Gender) and FairFace (Age, Gender). For precision, similar utility dips occur on UTKFace (Age), FairFace (Age), and CelebA (Bald, Black Hair, Wearing Hat, Young, Attractive, Blond Hair), with larger disparities for UTKFace (Age, Gender), FairFace (Age, Gender), and CelebA (Bald, Wearing Hat, Blond Hair). For recall, utility drops are observed on UTKFace (Age), FairFace (Age), and CelebA (Young, Attractive), with larger disparities for UTKFace (Age, Gender), FairFace (Age, Gender), and CelebA (Bald, Blond Hair). Across these metrics, DIS and ENT generally improve upon or match ERM in both utility and fairness, while ADV is competitive but less consistent.

FPR reveals an interesting pattern. In terms of overall FPR (utility), ERM is mostly solid but shows higher error (FPR

Metric	Method	CelebA			
		Wearing hat × Gender	Young × Gender	Attractive × Gender	Blond hair × Gender
ROC (↑)	ERM	0.997 ± 0.000	0.941 ± 0.001	0.917 ± 0.001	0.986 ± 0.000
	UNL	0.996 ± 0.000	0.940 ± 0.001	0.915 ± 0.000	0.986 ± 0.001
	ADV	0.998 ± 0.000	0.946 ± 0.001	0.919 ± 0.001	0.988 ± 0.000
	DIS	**0.999 ± 0.000**	**0.948 ± 0.001**	**0.922 ± 0.001**	**0.989 ± 0.000**
	ENT	0.996 ± 0.000	0.948 ± 0.002	0.921 ± 0.001	0.988 ± 0.000
ΔROC (↓)	ERM	0.002 ± 0.001	0.006 ± 0.003	0.004 ± 0.002	0.005 ± 0.002
	UNL	**0.000 ± 0.000**	**0.005 ± 0.001**	0.005 ± 0.002	0.005 ± 0.002
	ADV	0.001 ± 0.001	0.006 ± 0.003	0.004 ± 0.002	0.002 ± 0.002
	DIS	**0.000 ± 0.000**	0.007 ± 0.003	0.004 ± 0.002	**0.002 ± 0.001**
	ENT	0.003 ± 0.001	0.006 ± 0.003	**0.004 ± 0.001**	0.005 ± 0.002
ROC ratio (↑)	ERM	0.998 ± 0.001	0.994 ± 0.003	0.995 ± 0.002	0.994 ± 0.002
	UNL	**1.000 ± 0.000**	**0.995 ± 0.001**	0.994 ± 0.002	0.995 ± 0.003
	ADV	0.999 ± 0.001	0.994 ± 0.003	0.995 ± 0.002	0.998 ± 0.002
	DIS	**1.000 ± 0.000**	0.993 ± 0.003	**0.996 ± 0.002**	**0.998 ± 0.001**
	ENT	0.997 ± 0.001	0.994 ± 0.003	0.995 ± 0.001	0.995 ± 0.002
PR (↑)	ERM	0.961 ± 0.002	0.980 ± 0.001	0.922 ± 0.001	0.931 ± 0.002
	UNL	0.958 ± 0.001	0.980 ± 0.001	0.920 ± 0.001	0.928 ± 0.003
	ADV	0.971 ± 0.001	0.982 ± 0.001	0.924 ± 0.001	0.936 ± 0.002
	DIS	**0.975 ± 0.001**	0.983 ± 0.001	**0.927 ± 0.001**	**0.939 ± 0.001**
	ENT	0.942 ± 0.005	**0.983 ± 0.000**	0.926 ± 0.000	0.931 ± 0.002
ΔPR (↓)	ERM	0.037 ± 0.003	0.034 ± 0.001	0.170 ± 0.003	0.356 ± 0.016
	UNL	0.038 ± 0.007	0.035 ± 0.000	0.165 ± 0.004	0.366 ± 0.014
	ADV	0.026 ± 0.005	0.031 ± 0.000	0.163 ± 0.003	0.330 ± 0.011
	DIS	**0.023 ± 0.003**	0.031 ± 0.001	0.160 ± 0.003	0.314 ± 0.013
	ENT	0.043 ± 0.010	**0.029 ± 0.000**	**0.158 ± 0.003**	**0.294 ± 0.021**
PR ratio (↑)	ERM	0.962 ± 0.003	0.965 ± 0.001	0.821 ± 0.003	0.623 ± 0.017
	UNL	0.961 ± 0.007	0.964 ± 0.000	0.825 ± 0.004	0.611 ± 0.015
	ADV	0.973 ± 0.005	0.969 ± 0.000	0.828 ± 0.003	0.651 ± 0.011
	DIS	**0.976 ± 0.003**	0.969 ± 0.001	0.832 ± 0.003	0.670 ± 0.013
	ENT	0.955 ± 0.011	**0.970 ± 0.000**	**0.834 ± 0.003**	**0.687 ± 0.022**

Table 3. Performance evaluation of fine-tuning strategies on CelebA dataset, averaged over three random seeds. The best values are shown in **bold**. The scores that improve upon ERM by more than 0.005 are highlighted in green, while degradations greater than 0.005 are shown in red. Changes within ±0.005 of the ERM score are considered insignificant.

> 0.10) for CelebA (Young, Attractive), and larger subgroup differences (difference > 0.05) for UTKFace (Gender), FairFace (Gender), and CelebA (Bald, Blond Hair). However, for the subgroup-wise FPR ratio, substantial bias (ratio < 0.95) appears broadly across datasets and target–sensitive pairs. Moreover, fairness-aware methods often underperform ERM on this particular ratio metric, suggesting a tension among fairness objectives: improving one criterion can degrade another. Since we keep method-specific hyperparameters fixed across datasets, targeted tuning may help balance these trade-offs.

Overall, DIS and ENT maintain high downstream performance while substantially reducing bias where ERM shows disparities; ADV is competitive but less stable. In contrast, UNL offers limited benefits over ERM in classification, despite prior success in generative detoxification.

6. Conclusion

This work presents four fairness-aware fine-tuning methods designed for the practically relevant scenario where downstream task and sensitive attribute datasets are available separately. Through comprehensive experiments on standard fairness datasets, we demonstrate that while adversarial debiasing yields inconsistent improvements and sensitive unlearning proves ineffective for classification tasks, orthogonality-based disentanglement and entropy maximization approaches consistently reduce bias and often enhance overall utility compared to standard fine-tuning. However, our findings also reveal fundamental incompatibilities between fairness metrics, underscoring the inherent challenges of simultaneously optimizing multiple fairness objectives in parameter-efficient adaptation.

These promising findings open several avenues for future investigation. First, extending fairness-aware fine-tuning to intersectional fairness settings (Gohar & Cheng, 2023) by incorporating separate classification heads and sensitive adapters for each attribute in the intersection could address more complex bias patterns. Second, investigating the impact of distribution shift between task and sensitive attribute datasets, and developing robust fairness-aware methods that maintain performance under such shifts (Shao et al., 2024), represents a critical practical consideration. Finally, studying fairness-aware fine-tuning under privacy-constrained scenarios by incorporating differentially private methods for sharing LoRA adapters (Yu et al., 2022; Sun et al., 2024) would enable broader deployment in sensitive applications.

References

Adel, T., Valera, I., Ghahramani, Z., and Weller, A. One-Network Adversarial Fairness. In *AAAI*, 2019.

Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., and Wallach, H. A Reductions Approach to Fair Classification. In *ICML*, 2018.

Agarwal, A., Dudík, M., and Wu, Z. S. Fair Regression: Quantitative Definitions and Reduction-based Algorithms. In *ICML*, 2019.

Ali, J., Kleindessner, M., Wenzel, F., Budhathoki, K., Cevher, V., and Russell, C. Evaluating the Fairness of Discriminative Foundation Models in Computer Vision. In *AIES*, 2023.

Ashurst, C. and Weller, A. Fairness Without Demographic Data: A Survey of Approaches. In *EAAMO*, 2023.

Awais, M. et al. Foundation Models Defining a New Era in Vision: a Survey and Outlook. *IEEE TPAMI*, 2025.

Barocas, S., Hardt, M., and Narayanan, A. *Fairness and Machine Learning: Limitations and Opportunities*. MIT press, 2023.

Bellamy, R. K. et al. AI Fairness 360: An Extensible Toolkit for Detecting and Mitigating Algorithmic Bias. *IBM Journal of Research and Development*, 2019.

Berk, R., Heidari, H., Jabbari, S., Kearns, M., and Roth, A. Fairness in Criminal Justice Risk Assessments: The State of the Art. *Sociological Methods & Research*, 2021.

Bommasani, R. et al. On the Opportunities and Risks of Foundation Models. *arXiv preprint arXiv:2108.07258*, 2021.

Calmon, F., Wei, D., Vinzamuri, B., Natesan Ramamurthy, K., and Varshney, K. R. Optimized Pre-Processing for Discrimination Prevention. *NeurIPS*, 2017.

Caton, S. and Haas, C. Fairness in Machine Learning: A Survey. *ACM Comput. Surv.*, 2024.

Chouldechova, A. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big data*, 2017.

Creager, E., Madras, D., Jacobsen, J.-H., Weis, M., Swersky, K., Pitassi, T., and Zemel, R. Flexibly Fair Representation Learning by Disentanglement. In *ICML*, 2019.

Ding, Z., Liu, K. Z., Peetathawatchai, P., Isik, B., and Koyejo, S. On Fairness of Low-Rank Adaptation of Large Models. *arXiv preprint arXiv:2405.17512*, 2024.

Dosovitskiy, A. et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *ICLR*, 2021.

Dutt, R., Bohdal, O., Tsaftaris, S. A., and Hospedales, T. FairTune: Optimizing Parameter Efficient Fine Tuning for Fairness in Medical Image Analysis. In *ICLR*, 2024.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. Fairness Through Awareness. In *ITCS*, 2012.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., and Venkatasubramanian, S. Certifying and Removing Disparate Impact. In *KDD*, 2015.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., March, M., and Lempitsky, V. Domain-Adversarial Training of Neural Networks. *JMLR*, 2016.

Gohar, U. and Cheng, L. A Survey on Intersectional Fairness in Machine Learning: Notions, Mitigation, and Challenges. In *IJCAI*, 2023.

Han, Z., Gao, C., Liu, J., Zhang, J., and Zhang, S. Q. Parameter-Efficient Fine-Tuning for Large Models: A Comprehensive Survey. *arXiv preprint arXiv:2403.14608*, 2024.

Hardt, M., Price, E., and Srebro, N. Equality of Opportunity in Supervised Learning. *NeurIPS*, 2016.

Hort, M., Chen, Z., Zhang, J. M., Harman, M., and Sarro, F. Bias Mitigation for Machine Learning Classifiers: A Comprehensive Survey. *ACM Journal on Responsible Computing*, 2024.

Hu, E. J., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al. LoRA: Low-Rank Adaptation of Large Language Models. In *ICLR*, 2022.

Kamiran, F. and Calders, T. Data Pre-Processing Techniques for Classification without Discrimination. *Knowledge and Information Systems*, 2012.

Kamiran, F., Karim, A., and Zhang, X. Decision Theory for Discrimination-Aware Classification. In *ICDM*, 2012.

Kamishima, T., Akaho, S., Asoh, H., and Sakuma, J. Fairness-Aware Classifier with Prejudice Remover Regularizer. In *ECML PKDD*, 2012.

Karkkainen, K. and Joo, J. FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age for Bias Measurement and Mitigation. In *WACV*, 2021.

Kim, J. S., Chen, J., and Talwalkar, A. FACT: A Diagnostic for Group Fairness Trade-offs. In *ICML*, 2020.

Liu, Z., Luo, P., Wang, X., and Tang, X. Deep Learning Face Attributes in the Wild. In *ICCV*, 2015.

Locatello, F., Abbati, G., Rainforth, T., Bauer, S., Schölkopf, B., and Bachem, O. On the Fairness of Disentangled Representations. *NeurIPS*, 2019.

Loshchilov, I. and Hutter, F. Decoupled Weight Decay Regularization. In *ICLR*, 2024.

Mangrulkar, S., Gugger, S., Debut, L., Belkada, Y., Paul, S., and Bossan, B. PEFT: State-of-the-art Parameter-Efficient Fine-Tuning methods, 2022.

Mao, Y., Deng, Z., Yao, H., Ye, T., Kawaguchi, K., and Zou, J. Last-Layer Fairness Fine-tuning is Simple and Effective for Neural Networks. *arXiv preprint arXiv:2304.03935*, 2023.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. A Survey on Bias and Fairness in Machine Learning. *ACM computing surveys (CSUR)*, 2021.

Pedregosa, F. et al. Scikit-learn: Machine learning in Python. *JMLR*, 2011.

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., and Weinberger, K. Q. On Fairness and Calibration. *NeurIPS*, 2017.

Roy, P. C. and Boddeti, V. N. Mitigating Information Leakage in Image Representations: A Maximum Entropy Approach. In *CVPR*, 2019.

Sagawa, S., Koh, P. W., Hashimoto, T. B., and Liang, P. Distributionally Robust Neural Networks for Group Shifts: On the Importance of Regularization for Worst-Case Generalization. In *ICLR*, 2020.

Shao, M., Li, D., Zhao, C., Wu, X., Lin, Y., and Tian, Q. Supervised Algorithmic Fairness in Distribution Shifts: A Survey. In *IJCAI*, 2024.

Sukumaran, R., Feizi, A., Romero-Sorian, A., and Farnadi, G. FairLoRA: Unpacking Bias Mitigation in Vision Models with Fairness-Driven Low-Rank Adaptation. *arXiv preprint arXiv:2410.17358*, 2024.

Sun, Y., Li, Z., Li, Y., and Ding, B. Improving LoRA in Privacy-Preserving Federated Learning. In *ICLR*, 2024.

Verma, S. and Rubin, J. Fairness Definitions Explained. In *Proceedings of the International Workshop on Software Fairness*, 2018.

Wan, M., Zha, D., Liu, N., and Zou, N. In-Processing Modeling Techniques for Machine Learning Fairness: A Survey. *ACM Transactions on Knowledge Discovery from Data*, 2023.

Weerts, H., Dudík, M., Edgar, R., Jalali, A., Lutz, R., and Madaio, M. Fairlearn: Assessing and Improving Fairness of AI Systems. *JMLR*, 2023.

Yu, D. et al. Differentially Private Fine-tuning of Language Models. In *ICLR*, 2022.

Zhang, B. H., Lemoine, B., and Mitchell, M. Mitigating Unwanted Biases with Adversarial Learning. In *AIES*, 2018.

Zhang, J., Chen, S., Liu, J., and He, J. Composing Parameter-Efficient Modules with Arithmetic Operations. *NeurIPS*, 2023a.

Zhang, J., You, J., Panda, A., and Goldstein, T. LoRI: Reducing Cross-Task Interference in Multi-Task LowRank Adaptation. *arXiv preprint arXiv:2504.07448*, 2025.

Zhang, Q., Chen, M., Bukharin, A., He, P., Cheng, Y., Chen, W., and Zhao, T. Adaptive Budget Allocation for Parameter-Efficient Fine-Tuning. In *ICLR*, 2023b.

Zhang, Z., Song, Y., and Qi, H. Age Progression/Regression by Conditional Adversarial Autoencoder. In *CVPR*, 2017.

Zhao, W. X. et al. A Survey of Large Language Models. *arXiv preprint arXiv:2303.18223*, 2023.

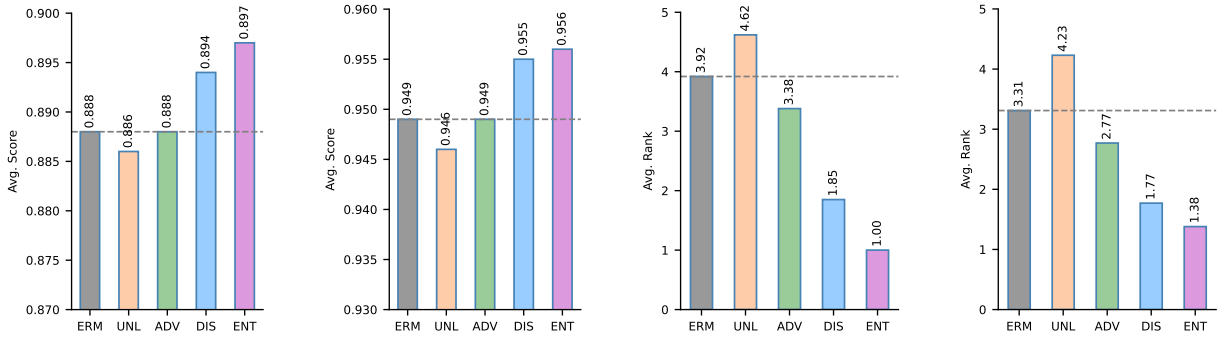
Zhou, C. et al. A Comprehensive Survey on Pretrained Foundation Models: A History from BERT to ChatGPT. *International Journal of Machine Learning and Cybernetics*, 2024.

A. Additional Details and Results

Table 4 summarizes the downstream and sensitive attributes of the datasets used in our experiments. Tables 5, 6, and 7 report overall utility and subgroup-wise ratio/difference for accuracy, precision, recall, and FPR. Figures 4, 5, 6, and 7 present method-wise accuracy, precision, recall, and FPR scores and their rankings averaged across datasets (also see Table 8). Tables 9 and 10 report sensitivity analysis of method-specific hyperparameters.

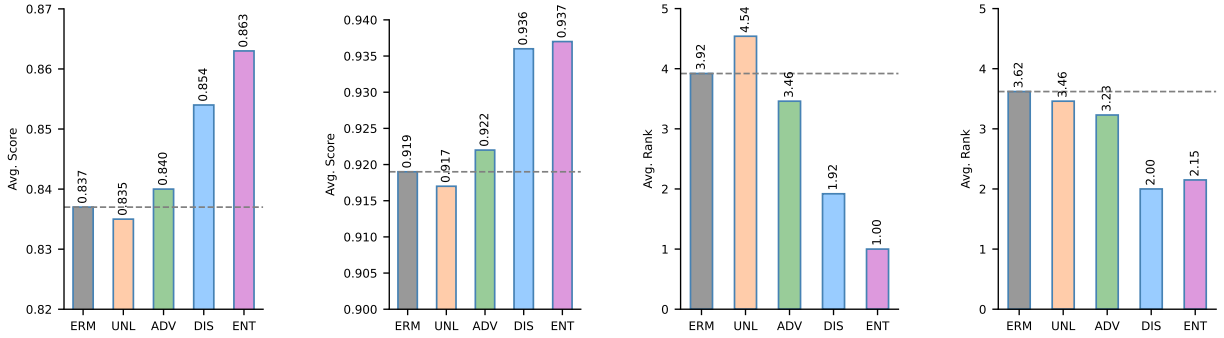
Dataset	Downstream Task Attribute(s)	Sensitive Attribute(s)
CelebA	Bald; Black hair; Eyeglasses; Smiling; Wearing hat; Young; Attractive; Blond hair	Gender (Male, Female)
UTKFace	Age (5 bins) Gender (Male, Female)	Race (White, Black, Asian, Indian, and Others)
FairFace	Age (9 bins) Gender (Male, Female)	Race (White, Southeast Asian, Middle Eastern, Black, Indian, Latino Hispanic, East Asian)
WaterBird	Bird type (Water bird, Land bird)	Background (Water, Land)

Table 4. Summary of downstream labels and sensitive attributes for each dataset, covering both binary and multi-valued cases.



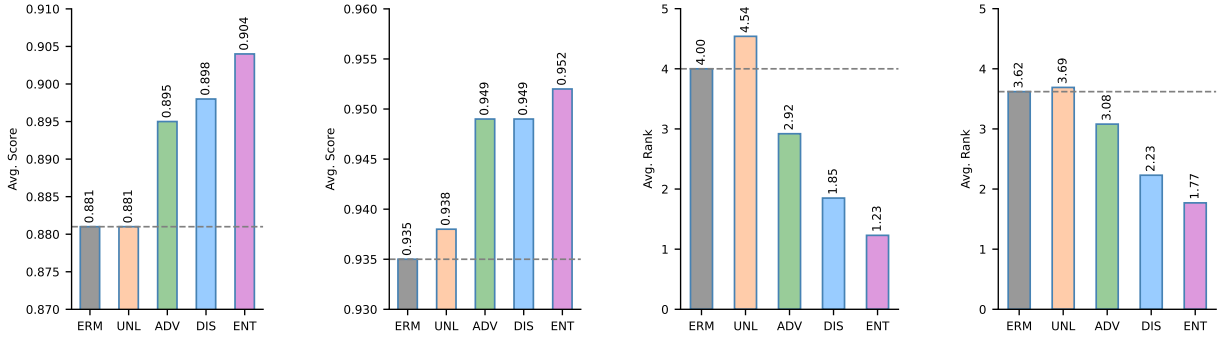
(a) Overall Acc Score (↑) *(b)* Acc Ratio Score (↑) *(c)* Overall Acc Rank (↓) *(d)* Acc Ratio Rank (↓)

Figure 4. Average overall accuracy and group-wise accuracy ratio (ratio of the best- to worst-performing subgroup) scores and ranks for each method, aggregated over all experiments.



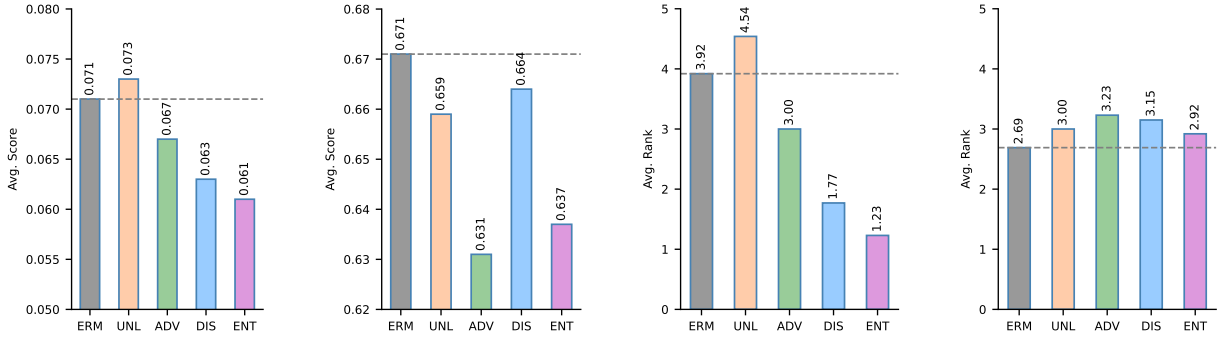
(a) Overall PPV Score (↑) *(b)* PPV Ratio Score (↑) *(c)* Overall PPV Rank (↓) *(d)* PPV Ratio Rank (↓)

Figure 5. Average overall precision and group-wise precision ratio (ratio of the best- to worst-performing subgroup) scores and ranks for each method, aggregated over all experiments.



(a) Overall TPR Score (↑) *(b)* TPR Ratio Score (↑) *(c)* Overall TPR Rank (↓) *(d)* TPR Ratio Rank (↓)

Figure 6. Average overall recall and group-wise recall ratio (ratio of the best- to worst-performing subgroup) scores and ranks for each method, aggregated over all experiments.



(a) Overall FPR Score (↑) *(b)* FPR Ratio Score (↑) *(c)* Overall FPR Rank (↓) *(d)* FPR Ratio Rank (↓)

Figure 7. Average overall FPR and group-wise FPR ratio (ratio of the best- to worst-performing subgroup) scores and ranks for each method, aggregated over all experiments.

Metric	Method	UTKFace		FairFace		WaterBird
		Age × Race	Gender × Race	Age × Race	Gender × Race	Type × Background
ACC (↑)	ERM	0.757 ± 0.004	0.949 ± 0.001	0.532 ± 0.002	0.934 ± 0.001	0.953 ± 0.004
	UNL	0.750 ± 0.004	0.946 ± 0.002	0.529 ± 0.004	0.928 ± 0.001	0.957 ± 0.001
	ADV	0.765 ± 0.002	0.946 ± 0.003	0.533 ± 0.007	0.938 ± 0.001	0.935 ± 0.012
	DIS	0.767 ± 0.002	0.954 ± 0.001	0.546 ± 0.004	0.942 ± 0.001	0.961 ± 0.004
	ENT	**0.778 ± 0.006**	**0.956 ± 0.001**	**0.553 ± 0.005**	**0.943 ± 0.000**	**0.971 ± 0.002**
ΔACC (↓)	ERM	0.097 ± 0.006	0.063 ± 0.002	0.055 ± 0.001	0.075 ± 0.005	0.036 ± 0.003
	UNL	0.108 ± 0.005	**0.046 ± 0.008**	0.071 ± 0.012	0.077 ± 0.006	0.035 ± 0.004
	ADV	**0.072 ± 0.014**	0.060 ± 0.004	0.068 ± 0.006	0.073 ± 0.005	0.044 ± 0.026
	DIS	0.073 ± 0.007	0.061 ± 0.002	**0.053 ± 0.004**	0.066 ± 0.004	0.027 ± 0.003
	ENT	0.086 ± 0.012	0.056 ± 0.004	0.054 ± 0.004	**0.065 ± 0.004**	**0.009 ± 0.005**
ACC ratio (↑)	ERM	0.883 ± 0.007	0.934 ± 0.002	0.903 ± 0.002	0.922 ± 0.005	0.963 ± 0.003
	UNL	0.868 ± 0.007	**0.953 ± 0.008**	0.873 ± 0.020	0.920 ± 0.006	0.964 ± 0.004
	ADV	0.911 ± 0.017	0.937 ± 0.004	0.881 ± 0.011	0.924 ± 0.005	0.954 ± 0.027
	DIS	**0.911 ± 0.008**	0.937 ± 0.002	**0.907 ± 0.007**	0.932 ± 0.004	0.972 ± 0.003
	ENT	0.897 ± 0.013	0.942 ± 0.004	**0.907 ± 0.007**	**0.933 ± 0.004**	**0.990 ± 0.005**
PPV (↑)	ERM	0.625 ± 0.025	0.949 ± 0.001	0.502 ± 0.004	0.933 ± 0.001	0.926 ± 0.006
	UNL	0.642 ± 0.025	0.946 ± 0.002	0.499 ± 0.005	0.928 ± 0.001	0.934 ± 0.002
	ADV	0.687 ± 0.019	0.946 ± 0.003	0.513 ± 0.006	0.938 ± 0.001	0.896 ± 0.015
	DIS	0.715 ± 0.009	0.954 ± 0.001	0.525 ± 0.005	0.942 ± 0.001	0.938 ± 0.006
	ENT	**0.751 ± 0.023**	**0.956 ± 0.001**	**0.547 ± 0.004**	**0.943 ± 0.000**	**0.952 ± 0.003**
ΔPPV (↓)	ERM	0.212 ± 0.061	0.062 ± 0.002	0.099 ± 0.007	0.065 ± 0.006	0.018 ± 0.008
	UNL	0.236 ± 0.039	**0.044 ± 0.008**	0.095 ± 0.004	0.069 ± 0.006	**0.011 ± 0.004**
	ADV	**0.147 ± 0.067**	0.059 ± 0.005	0.102 ± 0.012	0.064 ± 0.006	0.041 ± 0.020
	DIS	0.169 ± 0.051	0.060 ± 0.002	0.091 ± 0.013	**0.057 ± 0.005**	0.020 ± 0.006
	ENT	0.177 ± 0.055	0.054 ± 0.003	**0.089 ± 0.008**	**0.057 ± 0.005**	0.027 ± 0.013
PPV ratio (↑)	ERM	0.728 ± 0.071	0.935 ± 0.002	0.821 ± 0.010	0.932 ± 0.006	0.981 ± 0.009
	UNL	0.683 ± 0.037	**0.954 ± 0.008**	0.826 ± 0.006	0.927 ± 0.007	**0.988 ± 0.005**
	ADV	**0.810 ± 0.078**	0.938 ± 0.005	0.822 ± 0.019	0.933 ± 0.006	0.955 ± 0.020
	DIS	0.786 ± 0.064	0.938 ± 0.002	0.843 ± 0.019	**0.941 ± 0.005**	0.979 ± 0.006
	ENT	0.779 ± 0.064	0.944 ± 0.003	**0.850 ± 0.012**	**0.941 ± 0.005**	0.972 ± 0.013
TPR (↑)	ERM	0.662 ± 0.025	0.949 ± 0.001	0.582 ± 0.001	0.934 ± 0.001	0.944 ± 0.002
	UNL	0.681 ± 0.020	0.947 ± 0.002	0.581 ± 0.002	0.929 ± 0.001	0.945 ± 0.002
	ADV	0.768 ± 0.006	0.946 ± 0.003	0.606 ± 0.002	0.938 ± 0.001	0.935 ± 0.010
	DIS	0.755 ± 0.011	0.954 ± 0.001	0.609 ± 0.004	0.942 ± 0.001	0.956 ± 0.001
	ENT	**0.789 ± 0.003**	**0.956 ± 0.001**	**0.627 ± 0.003**	**0.943 ± 0.000**	**0.969 ± 0.002**
ΔTPR (↓)	ERM	0.270 ± 0.053	0.064 ± 0.003	0.083 ± 0.014	0.073 ± 0.005	**0.007 ± 0.002**
	UNL	0.215 ± 0.018	**0.047 ± 0.007**	0.079 ± 0.012	0.073 ± 0.006	0.013 ± 0.000
	ADV	**0.161 ± 0.024**	0.062 ± 0.004	0.066 ± 0.002	0.069 ± 0.005	0.021 ± 0.006
	DIS	0.205 ± 0.009	0.062 ± 0.002	**0.064 ± 0.015**	**0.063 ± 0.004**	0.008 ± 0.007
	ENT	0.168 ± 0.018	0.058 ± 0.005	0.076 ± 0.009	**0.063 ± 0.004**	0.012 ± 0.004
TPR ratio (↑)	ERM	0.682 ± 0.050	0.933 ± 0.003	0.865 ± 0.021	0.924 ± 0.006	**0.993 ± 0.003**
	UNL	0.728 ± 0.032	**0.951 ± 0.007**	0.871 ± 0.018	0.923 ± 0.007	0.986 ± 0.000
	ADV	**0.809 ± 0.027**	0.936 ± 0.004	0.896 ± 0.003	0.928 ± 0.005	0.977 ± 0.006
	DIS	0.761 ± 0.013	0.936 ± 0.002	**0.901 ± 0.023**	**0.935 ± 0.004**	0.991 ± 0.007
	ENT	0.808 ± 0.020	0.940 ± 0.005	0.885 ± 0.013	**0.935 ± 0.004**	0.987 ± 0.004
FPR (↓)	ERM	0.073 ± 0.001	0.051 ± 0.001	0.063 ± 0.000	0.066 ± 0.001	0.056 ± 0.002
	UNL	0.076 ± 0.001	0.053 ± 0.002	0.063 ± 0.001	0.071 ± 0.001	0.055 ± 0.002
	ADV	0.072 ± 0.000	0.054 ± 0.003	0.062 ± 0.001	0.062 ± 0.001	0.065 ± 0.010
	DIS	0.071 ± 0.001	0.046 ± 0.001	0.061 ± 0.000	0.058 ± 0.001	0.044 ± 0.001
	ENT	**0.067 ± 0.002**	**0.044 ± 0.001**	**0.060 ± 0.000**	**0.057 ± 0.000**	**0.031 ± 0.002**
ΔFPR (↓)	ERM	0.042 ± 0.010	0.064 ± 0.003	0.006 ± 0.000	0.073 ± 0.005	**0.007 ± 0.002**
	UNL	0.053 ± 0.005	**0.047 ± 0.007**	0.008 ± 0.002	0.073 ± 0.006	0.013 ± 0.000
	ADV	0.044 ± 0.010	0.062 ± 0.004	0.007 ± 0.001	0.069 ± 0.005	0.021 ± 0.006
	DIS	0.043 ± 0.011	0.062 ± 0.002	0.006 ± 0.000	**0.063 ± 0.004**	0.008 ± 0.007
	ENT	**0.041 ± 0.007**	0.058 ± 0.005	**0.005 ± 0.001**	**0.063 ± 0.004**	0.012 ± 0.004
FPR ratio (↑)	ERM	**0.631 ± 0.055**	0.364 ± 0.019	0.909 ± 0.006	0.343 ± 0.024	**0.907 ± 0.032**
	UNL	0.560 ± 0.030	**0.461 ± 0.039**	0.885 ± 0.023	**0.393 ± 0.034**	0.810 ± 0.005
	ADV	0.606 ± 0.054	0.399 ± 0.009	0.889 ± 0.009	0.360 ± 0.025	0.772 ± 0.061
	DIS	0.611 ± 0.059	0.340 ± 0.015	0.901 ± 0.001	0.367 ± 0.024	0.876 ± 0.094
	ENT	0.599 ± 0.046	0.337 ± 0.036	**0.915 ± 0.009**	0.344 ± 0.023	0.731 ± 0.059

Table 5. Performance evaluation of fine-tuning strategies on UTKFace, FairFace, and WaterBird datasets, averaged over three random seeds. The best values are shown in **bold**. The scores that improve upon ERM by more than 0.005 are highlighted in green, while degradations greater than 0.005 are shown in red. Changes within ±0.005 of ERM score are considered insignificant.

Metric	Method	CelebA			
		Bald × Gender	Black hair × Gender	Eyeglasses × Gender	Smiling × Gender
ACC (↑)	ERM	0.986 ± 0.000	0.891 ± 0.000	0.995 ± 0.000	0.928 ± 0.000
	UNL	0.984 ± 0.000	0.885 ± 0.001	0.994 ± 0.000	0.925 ± 0.002
	ADV	0.983 ± 0.002	0.892 ± 0.005	0.996 ± 0.000	0.930 ± 0.001
	DIS	0.990 ± 0.000	0.894 ± 0.002	**0.997 ± 0.000**	**0.933 ± 0.001**
	ENT	**0.991 ± 0.000**	**0.896 ± 0.004**	**0.997 ± 0.000**	0.933 ± 0.002
ΔACC (↓)	ERM	0.034 ± 0.001	**0.039 ± 0.001**	0.007 ± 0.001	0.013 ± 0.000
	UNL	0.038 ± 0.000	0.042 ± 0.001	0.008 ± 0.000	0.015 ± 0.001
	ADV	0.039 ± 0.004	0.039 ± 0.003	0.007 ± 0.000	**0.012 ± 0.001**
	DIS	0.025 ± 0.001	0.039 ± 0.002	0.005 ± 0.000	0.013 ± 0.001
	ENT	**0.021 ± 0.001**	0.039 ± 0.003	**0.004 ± 0.000**	**0.012 ± 0.001**
ACC ratio (↑)	ERM	0.966 ± 0.001	**0.957 ± 0.001**	0.993 ± 0.001	0.986 ± 0.000
	UNL	0.962 ± 0.000	0.954 ± 0.001	0.992 ± 0.000	0.984 ± 0.001
	ADV	0.961 ± 0.004	0.957 ± 0.004	0.993 ± 0.000	**0.987 ± 0.001**
	DIS	0.975 ± 0.001	0.957 ± 0.002	0.995 ± 0.000	0.986 ± 0.001
	ENT	**0.979 ± 0.001**	0.957 ± 0.003	**0.996 ± 0.000**	**0.987 ± 0.001**
PPV (↑)	ERM	0.808 ± 0.004	0.841 ± 0.001	0.972 ± 0.002	0.928 ± 0.000
	UNL	0.793 ± 0.001	0.834 ± 0.001	0.968 ± 0.001	0.926 ± 0.002
	ADV	0.788 ± 0.014	0.842 ± 0.006	0.971 ± 0.000	0.930 ± 0.001
	DIS	0.844 ± 0.004	0.844 ± 0.002	0.979 ± 0.002	**0.933 ± 0.001**
	ENT	**0.860 ± 0.005**	**0.847 ± 0.004**	**0.985 ± 0.000**	0.933 ± 0.002
ΔPPV (↓)	ERM	0.102 ± 0.021	0.002 ± 0.000	0.004 ± 0.001	0.015 ± 0.000
	UNL	0.101 ± 0.059	**0.001 ± 0.000**	0.005 ± 0.003	0.016 ± 0.001
	ADV	0.129 ± 0.036	0.003 ± 0.001	**0.001 ± 0.001**	**0.013 ± 0.002**
	DIS	**0.027 ± 0.006**	0.004 ± 0.002	**0.001 ± 0.001**	0.014 ± 0.001
	ENT	0.041 ± 0.030	0.007 ± 0.001	0.002 ± 0.001	0.015 ± 0.001
PPV ratio (↑)	ERM	0.874 ± 0.026	0.997 ± 0.001	0.996 ± 0.001	0.984 ± 0.000
	UNL	0.872 ± 0.075	**0.999 ± 0.000**	0.994 ± 0.003	0.982 ± 0.001
	ADV	0.836 ± 0.047	0.996 ± 0.002	**0.999 ± 0.001**	**0.986 ± 0.002**
	DIS	**0.968 ± 0.008**	0.996 ± 0.002	**0.999 ± 0.001**	0.985 ± 0.001
	ENT	0.952 ± 0.036	0.991 ± 0.001	0.998 ± 0.001	0.984 ± 0.001
TPR (↑)	ERM	0.957 ± 0.003	0.901 ± 0.001	0.988 ± 0.001	0.927 ± 0.001
	UNL	0.949 ± 0.006	0.898 ± 0.001	0.988 ± 0.001	0.925 ± 0.002
	ADV	**0.985 ± 0.001**	0.904 ± 0.001	0.994 ± 0.000	0.929 ± 0.001
	DIS	0.983 ± 0.001	**0.907 ± 0.001**	**0.995 ± 0.000**	**0.933 ± 0.001**
	ENT	0.980 ± 0.002	**0.907 ± 0.001**	0.994 ± 0.000	0.933 ± 0.002
ΔTPR (↓)	ERM	0.076 ± 0.019	0.029 ± 0.002	0.005 ± 0.000	0.018 ± 0.001
	UNL	0.097 ± 0.041	0.032 ± 0.002	0.005 ± 0.002	0.021 ± 0.001
	ADV	0.053 ± 0.025	0.029 ± 0.005	0.006 ± 0.002	0.018 ± 0.003
	DIS	**0.050 ± 0.025**	0.029 ± 0.001	**0.001 ± 0.000**	0.020 ± 0.001
	ENT	0.051 ± 0.025	**0.028 ± 0.000**	0.004 ± 0.001	**0.016 ± 0.002**
TPR ratio (↑)	ERM	0.920 ± 0.021	0.968 ± 0.002	0.995 ± 0.000	0.980 ± 0.002
	UNL	0.898 ± 0.046	0.965 ± 0.002	0.995 ± 0.002	0.978 ± 0.001
	ADV	0.946 ± 0.025	0.968 ± 0.006	0.994 ± 0.002	0.980 ± 0.004
	DIS	**0.949 ± 0.026**	0.969 ± 0.001	**0.999 ± 0.000**	0.979 ± 0.001
	ENT	0.948 ± 0.026	**0.970 ± 0.000**	0.996 ± 0.001	**0.983 ± 0.002**
FPR (↓)	ERM	0.043 ± 0.003	0.099 ± 0.001	0.012 ± 0.001	0.073 ± 0.001
	UNL	0.051 ± 0.006	0.102 ± 0.001	0.012 ± 0.001	0.075 ± 0.002
	ADV	**0.015 ± 0.001**	0.096 ± 0.001	0.006 ± 0.000	0.071 ± 0.001
	DIS	0.017 ± 0.001	**0.093 ± 0.001**	**0.005 ± 0.000**	**0.067 ± 0.001**
	ENT	0.020 ± 0.002	**0.093 ± 0.001**	0.006 ± 0.000	0.067 ± 0.002
ΔFPR (↓)	ERM	0.076 ± 0.019	0.029 ± 0.002	0.005 ± 0.000	0.018 ± 0.001
	UNL	0.097 ± 0.041	0.032 ± 0.002	0.005 ± 0.002	0.021 ± 0.001
	ADV	0.053 ± 0.025	0.029 ± 0.005	0.006 ± 0.002	0.018 ± 0.003
	DIS	**0.050 ± 0.025**	0.029 ± 0.001	**0.001 ± 0.000**	0.020 ± 0.001
	ENT	0.051 ± 0.025	**0.028 ± 0.000**	0.004 ± 0.001	**0.016 ± 0.002**
FPR ratio (↑)	ERM	**0.258 ± 0.132**	**0.752 ± 0.017**	0.645 ± 0.010	0.786 ± 0.014
	UNL	0.096 ± 0.093	0.740 ± 0.012	0.682 ± 0.135	0.766 ± 0.010
	ADV	0.067 ± 0.059	0.746 ± 0.041	0.358 ± 0.157	0.782 ± 0.032
	DIS	0.064 ± 0.063	0.742 ± 0.009	**0.781 ± 0.066**	0.754 ± 0.008
	ENT	0.065 ± 0.063	0.751 ± 0.005	0.475 ± 0.138	**0.793 ± 0.022**

Table 6. Performance evaluation of fine-tuning strategies on CelebA dataset, averaged over three random seeds. The best values are shown in **bold**. The scores that improve upon ERM by more than 0.005 are highlighted in green, while degradations greater than 0.005 are shown in red. Changes within ±0.005 of the ERM score are considered insignificant.

Metric	Method	CelebA			
		Wearing hat × Gender	Young × Gender	Attractive × Gender	Blond hair × Gender
ACC (↑)	ERM	0.986 ± 0.000	0.868 ± 0.001	0.827 ± 0.001	0.940 ± 0.001
	UNL	0.987 ± 0.000	0.867 ± 0.001	0.826 ± 0.000	0.937 ± 0.001
	ADV	0.988 ± 0.001	0.869 ± 0.003	0.828 ± 0.001	0.941 ± 0.003
	DIS	0.990 ± 0.000	0.876 ± 0.001	0.831 ± 0.001	0.942 ± 0.002
	ENT	**0.991 ± 0.001**	**0.878 ± 0.005**	**0.832 ± 0.001**	**0.947 ± 0.002**
ΔACC (↓)	ERM	0.008 ± 0.001	0.079 ± 0.002	**0.011 ± 0.003**	0.057 ± 0.000
	UNL	0.007 ± 0.001	0.080 ± 0.002	0.015 ± 0.002	0.060 ± 0.002
	ADV	0.007 ± 0.000	**0.075 ± 0.002**	0.013 ± 0.003	0.058 ± 0.003
	DIS	**0.006 ± 0.001**	0.076 ± 0.003	0.013 ± 0.002	0.057 ± 0.001
	ENT	0.007 ± 0.000	0.075 ± 0.003	0.012 ± 0.002	**0.055 ± 0.003**
ACC ratio (↑)	ERM	0.992 ± 0.001	0.913 ± 0.003	**0.987 ± 0.004**	0.941 ± 0.000
	UNL	0.993 ± 0.001	0.912 ± 0.003	0.982 ± 0.003	0.938 ± 0.002
	ADV	0.993 ± 0.000	0.916 ± 0.002	0.985 ± 0.004	0.941 ± 0.003
	DIS	**0.994 ± 0.001**	**0.917 ± 0.003**	0.985 ± 0.002	0.942 ± 0.001
	ENT	0.993 ± 0.000	**0.917 ± 0.003**	0.985 ± 0.002	**0.944 ± 0.003**
PPV (↑)	ERM	0.896 ± 0.002	0.808 ± 0.002	0.827 ± 0.001	0.860 ± 0.002
	UNL	0.898 ± 0.002	0.807 ± 0.002	0.826 ± 0.000	0.855 ± 0.003
	ADV	0.905 ± 0.005	0.810 ± 0.002	0.829 ± 0.000	0.861 ± 0.005
	DIS	0.918 ± 0.003	0.818 ± 0.002	0.831 ± 0.001	0.863 ± 0.005
	ENT	**0.925 ± 0.007**	**0.820 ± 0.007**	**0.833 ± 0.001**	**0.872 ± 0.004**
ΔPPV (↓)	ERM	0.056 ± 0.006	0.042 ± 0.002	0.005 ± 0.001	0.159 ± 0.004
	UNL	0.054 ± 0.008	0.038 ± 0.000	0.007 ± 0.004	0.165 ± 0.004
	ADV	0.048 ± 0.006	0.047 ± 0.004	0.009 ± 0.001	0.152 ± 0.007
	DIS	0.042 ± 0.003	0.039 ± 0.004	0.005 ± 0.001	0.150 ± 0.006
	ENT	**0.031 ± 0.005**	**0.037 ± 0.009**	**0.003 ± 0.001**	**0.135 ± 0.004**
PPV ratio (↑)	ERM	0.939 ± 0.007	0.948 ± 0.002	0.994 ± 0.002	0.818 ± 0.004
	UNL	0.941 ± 0.008	0.953 ± 0.001	0.992 ± 0.005	0.810 ± 0.005
	ADV	0.948 ± 0.006	0.943 ± 0.005	0.989 ± 0.002	0.825 ± 0.008
	DIS	0.955 ± 0.003	0.953 ± 0.004	0.994 ± 0.001	0.829 ± 0.008
	ENT	**0.967 ± 0.005**	**0.955 ± 0.011**	**0.996 ± 0.001**	**0.847 ± 0.005**
TPR (↑)	ERM	0.977 ± 0.001	0.865 ± 0.002	0.827 ± 0.001	0.944 ± 0.001
	UNL	0.977 ± 0.000	0.860 ± 0.001	0.826 ± 0.000	0.943 ± 0.001
	ADV	0.989 ± 0.001	0.871 ± 0.002	0.828 ± 0.001	0.947 ± 0.001
	DIS	0.989 ± 0.001	0.873 ± 0.003	0.831 ± 0.001	0.950 ± 0.001
	ENT	**0.990 ± 0.000**	**0.875 ± 0.003**	**0.832 ± 0.001**	**0.953 ± 0.003**
ΔTPR (↓)	ERM	0.009 ± 0.002	0.009 ± 0.007	0.008 ± 0.005	0.067 ± 0.005
	UNL	0.009 ± 0.004	**0.004 ± 0.002**	0.007 ± 0.003	0.066 ± 0.010
	ADV	**0.002 ± 0.001**	0.013 ± 0.006	0.014 ± 0.006	0.057 ± 0.010
	DIS	0.003 ± 0.001	0.011 ± 0.007	0.015 ± 0.003	**0.049 ± 0.009**
	ENT	0.003 ± 0.002	0.008 ± 0.002	**0.002 ± 0.002**	0.054 ± 0.013
TPR ratio (↑)	ERM	0.991 ± 0.002	0.989 ± 0.008	0.990 ± 0.006	0.928 ± 0.006
	UNL	0.990 ± 0.004	**0.995 ± 0.002**	0.991 ± 0.004	0.929 ± 0.010
	ADV	**0.997 ± 0.001**	0.985 ± 0.007	0.982 ± 0.007	0.939 ± 0.011
	DIS	**0.997 ± 0.001**	0.988 ± 0.008	0.981 ± 0.004	**0.947 ± 0.010**
	ENT	0.997 ± 0.002	0.990 ± 0.003	**0.997 ± 0.002**	0.943 ± 0.014
FPR (↓)	ERM	0.023 ± 0.001	0.135 ± 0.002	0.173 ± 0.001	0.056 ± 0.001
	UNL	0.023 ± 0.000	0.140 ± 0.001	0.174 ± 0.000	0.057 ± 0.001
	ADV	0.011 ± 0.001	0.129 ± 0.002	0.172 ± 0.001	0.053 ± 0.001
	DIS	0.011 ± 0.001	0.127 ± 0.003	0.169 ± 0.001	0.050 ± 0.001
	ENT	**0.010 ± 0.000**	**0.125 ± 0.003**	**0.168 ± 0.001**	**0.047 ± 0.003**
ΔFPR (↓)	ERM	0.009 ± 0.002	0.009 ± 0.007	0.008 ± 0.005	0.067 ± 0.005
	UNL	0.009 ± 0.004	**0.004 ± 0.002**	0.007 ± 0.003	0.066 ± 0.010
	ADV	**0.002 ± 0.001**	0.013 ± 0.006	0.014 ± 0.006	0.057 ± 0.010
	DIS	0.003 ± 0.001	0.011 ± 0.007	0.015 ± 0.003	**0.049 ± 0.009**
	ENT	0.003 ± 0.002	0.008 ± 0.002	**0.002 ± 0.002**	0.054 ± 0.013
FPR ratio (↑)	ERM	0.713 ± 0.043	0.944 ± 0.043	0.966 ± 0.020	0.504 ± 0.015
	UNL	0.714 ± 0.109	**0.976 ± 0.011**	0.968 ± 0.013	0.521 ± 0.033
	ADV	**0.822 ± 0.029**	0.917 ± 0.041	0.940 ± 0.025	0.541 ± 0.046
	DIS	0.770 ± 0.074	0.931 ± 0.044	0.931 ± 0.014	**0.568 ± 0.049**
	ENT	0.794 ± 0.102	0.946 ± 0.015	**0.990 ± 0.008**	0.540 ± 0.049

Table 7. Performance evaluation of fine-tuning strategies on CelebA dataset, averaged over three random seeds. The best values are shown in **bold**. The scores that improve upon ERM by more than 0.005 are highlighted in green, while degradations greater than 0.005 are shown in red. Changes within ±0.005 of the ERM score are considered insignificant.

Metric	Method	Avg. Score			Avg. Rank (↓)		
		Overall (↑)	Diff. (↓)	Ratio (↑)	Overall (↑)	Diff. (↓)	Ratio (↑)
AUROC	ERM	0.968	0.015	0.985	3.77	3.08	3.38
	UNL	0.968	0.015	0.984	4.23	3.54	3.46
	ADV	0.970	0.014	0.986	2.62	2.46	2.54
	DIS	**0.972**	**0.013**	**0.987**	**1.23**	**1.92**	**1.92**
	ENT	**0.972**	**0.013**	0.986	2.00	2.31	2.38
AUPRC	ERM	0.895	0.101	0.886	3.92	3.77	3.77
	UNL	0.895	0.098	0.887	4.46	3.85	3.92
	ADV	0.912	**0.077**	0.914	2.54	2.38	2.38
	DIS	**0.915**	0.080	0.911	**1.31**	2.31	2.31
	ENT	0.913	**0.077**	**0.916**	2.08	**2.15**	**2.15**
ACC	ERM	0.888	0.044	0.949	3.92	3.23	3.31
	UNL	0.886	0.046	0.946	4.62	4.23	4.23
	ADV	0.888	0.044	0.949	3.38	2.77	2.77
	DIS	0.894	0.040	0.955	1.85	2.15	1.77
	ENT	**0.897**	**0.038**	**0.956**	**1.00**	**1.46**	**1.38**
PPV	ERM	0.837	0.065	0.919	3.92	3.62	3.62
	UNL	0.835	0.065	0.917	4.54	3.38	3.46
	ADV	0.840	0.063	0.922	3.46	3.31	3.23
	DIS	0.854	**0.052**	0.936	1.92	2.23	**2.00**
	ENT	**0.863**	**0.052**	**0.937**	**1.00**	**2.15**	2.15
TPR	ERM	0.881	0.055	0.935	4.00	3.54	3.62
	UNL	0.881	0.051	0.938	4.54	3.54	3.69
	ADV	0.895	0.044	0.949	2.92	3.00	3.08
	DIS	0.898	0.045	0.949	1.85	2.31	2.23
	ENT	**0.904**	**0.042**	**0.952**	**1.23**	**1.85**	**1.77**
FPR	ERM	0.071	0.032	**0.671**	3.92	3.08	**2.69**
	UNL	0.073	0.033	0.659	4.54	3.69	3.00
	ADV	0.067	0.030	0.631	3.00	3.38	3.23
	DIS	0.063	0.028	0.664	1.77	2.38	3.15
	ENT	**0.061**	**0.027**	0.637	**1.23**	**1.62**	2.92

Table 8. Average utility and fairness scores and ranks for each method, aggregated over all experiments. The best values are shown in **bold**. Average scores that improve upon ERM by more than 0.005 are highlighted in green, while degradations greater than 0.005 are shown in red. Changes within ±0.005 of the ERM score are considered insignificant.

Method	Value	UTKFace (Age × Race)					
		ROC (↑)	ΔROC (↓)	ROC ratio (↑)	PR (↑)	ΔPR (↓)	PR ratio (↑)
UNL (λ_{UNL})	- 0.25	0.882 ± 0.008	0.088 ± 0.006	0.904 ± 0.007	0.531 ± 0.022	0.212 ± 0.020	0.657 ± 0.023
	- 0.50	0.938 ± 0.001	0.073 ± 0.007	0.924 ± 0.007	0.694 ± 0.032	0.225 ± 0.005	0.722 ± 0.012
	- 0.75	**0.949 ± 0.002**	**0.048 ± 0.007**	**0.950 ± 0.008**	**0.744 ± 0.015**	0.234 ± 0.044	0.729 ± 0.049
	- 1.00	0.945 ± 0.000	0.049 ± 0.007	0.948 ± 0.007	0.707 ± 0.013	**0.190 ± 0.010**	**0.768 ± 0.007**
	- 1.25	0.938 ± 0.007	0.056 ± 0.005	0.942 ± 0.006	0.695 ± 0.003	0.218 ± 0.029	0.736 ± 0.034
ADV (λ_{ADV})	0.01	0.949 ± 0.001	0.033 ± 0.005	0.965 ± 0.005	**0.818 ± 0.010**	0.191 ± 0.020	0.784 ± 0.025
	0.05	0.949 ± 0.001	**0.033 ± 0.004**	**0.966 ± 0.005**	0.816 ± 0.010	0.193 ± 0.022	0.782 ± 0.027
	0.10	**0.950 ± 0.000**	0.038 ± 0.002	0.960 ± 0.002	0.814 ± 0.002	**0.105 ± 0.009**	**0.881 ± 0.008**
	0.50	0.948 ± 0.001	**0.033 ± 0.004**	0.965 ± 0.005	0.816 ± 0.010	0.186 ± 0.022	0.789 ± 0.028
	1.00	0.949 ± 0.001	**0.033 ± 0.004**	0.965 ± 0.005	0.814 ± 0.012	0.187 ± 0.020	0.788 ± 0.026
DIS (λ_{DIS})	0.01	**0.952 ± 0.000**	**0.032 ± 0.004**	0.966 ± 0.005	**0.836 ± 0.011**	0.190 ± 0.018	0.786 ± 0.023
	0.05	**0.952 ± 0.000**	**0.032 ± 0.004**	0.966 ± 0.005	0.835 ± 0.011	0.190 ± 0.021	0.785 ± 0.026
	0.10	0.951 ± 0.001	0.042 ± 0.006	0.956 ± 0.006	0.815 ± 0.014	**0.164 ± 0.046**	**0.817 ± 0.050**
	0.50	**0.952 ± 0.000**	0.032 ± 0.005	**0.967 ± 0.005**	0.835 ± 0.011	0.192 ± 0.019	0.783 ± 0.025
	1.00	**0.952 ± 0.000**	0.032 ± 0.005	0.966 ± 0.005	**0.836 ± 0.011**	0.192 ± 0.017	0.784 ± 0.022
ENT (λ_{ENT})	0.10	**0.955 ± 0.001**	0.031 ± 0.005	0.968 ± 0.005	0.817 ± 0.012	0.223 ± 0.029	0.752 ± 0.029
	0.50	0.953 ± 0.002	**0.030 ± 0.005**	**0.969 ± 0.005**	0.813 ± 0.012	0.215 ± 0.023	0.758 ± 0.025
	1.00	0.947 ± 0.007	0.046 ± 0.008	0.953 ± 0.008	**0.822 ± 0.006**	**0.144 ± 0.038**	**0.840 ± 0.040**
	5.00	0.949 ± 0.003	0.033 ± 0.007	0.965 ± 0.007	0.800 ± 0.006	0.219 ± 0.016	0.753 ± 0.026
	10.00	0.952 ± 0.001	0.037 ± 0.005	0.962 ± 0.005	0.805 ± 0.003	0.201 ± 0.008	0.774 ± 0.009

Table 9. Performance evaluation of fine-tuning strategies on UTKFace dataset, averaged over three random seeds (for varying method-specific hyperparameter values). The best values for each method are shown in **bold**.

Method	Value	UTKFace (Gender × Race)					
		ROC (↑)	ΔROC (↓)	ROC ratio (↑)	PR (↑)	ΔPR (↓)	PR ratio (↑)
UNL (λ_{UNL})	- 0.25	0.915 ± 0.005	0.083 ± 0.005	0.911 ± 0.005	0.907 ± 0.003	0.066 ± 0.003	0.929 ± 0.003
	- 0.50	0.975 ± 0.003	0.035 ± 0.002	0.964 ± 0.003	0.968 ± 0.004	**0.027 ± 0.008**	**0.972 ± 0.008**
	- 0.75	0.984 ± 0.002	0.030 ± 0.005	0.970 ± 0.005	0.977 ± 0.004	0.039 ± 0.014	0.960 ± 0.015
	- 1.00	**0.985 ± 0.001**	0.026 ± 0.002	0.974 ± 0.002	**0.978 ± 0.003**	0.037 ± 0.011	0.962 ± 0.012
	- 1.25	0.983 ± 0.002	**0.025 ± 0.004**	**0.975 ± 0.004**	0.975 ± 0.002	0.030 ± 0.011	0.970 ± 0.012
ADV (λ_{ADV})	0.01	**0.986 ± 0.000**	**0.026 ± 0.002**	0.973 ± 0.002	**0.981 ± 0.001**	**0.025 ± 0.004**	**0.975 ± 0.004**
	0.05	**0.986 ± 0.000**	0.027 ± 0.002	0.973 ± 0.002	**0.981 ± 0.001**	0.026 ± 0.004	0.974 ± 0.004
	0.10	0.985 ± 0.001	**0.026 ± 0.002**	0.973 ± 0.002	0.979 ± 0.001	0.030 ± 0.004	0.970 ± 0.004
	0.50	**0.986 ± 0.000**	**0.026 ± 0.002**	**0.974 ± 0.002**	**0.981 ± 0.001**	**0.025 ± 0.004**	0.974 ± 0.004
	1.00	**0.986 ± 0.000**	0.027 ± 0.002	0.973 ± 0.002	**0.981 ± 0.001**	**0.025 ± 0.004**	0.974 ± 0.004
DIS (λ_{DIS})	0.01	**0.989 ± 0.000**	**0.023 ± 0.005**	**0.977 ± 0.005**	0.985 ± 0.001	**0.029 ± 0.006**	**0.971 ± 0.006**
	0.05	**0.989 ± 0.000**	**0.023 ± 0.005**	**0.977 ± 0.005**	**0.985 ± 0.000**	**0.029 ± 0.006**	**0.971 ± 0.006**
	0.10	0.986 ± 0.001	0.028 ± 0.001	0.972 ± 0.002	0.981 ± 0.002	0.031 ± 0.008	0.969 ± 0.008
	0.50	**0.989 ± 0.000**	**0.023 ± 0.005**	**0.977 ± 0.005**	**0.985 ± 0.000**	**0.029 ± 0.006**	**0.971 ± 0.006**
	1.00	**0.989 ± 0.000**	**0.023 ± 0.005**	**0.977 ± 0.005**	**0.985 ± 0.000**	**0.029 ± 0.006**	**0.971 ± 0.006**
ENT (λ_{ENT})	0.10	**0.990 ± 0.000**	**0.021 ± 0.007**	**0.979 ± 0.007**	**0.987 ± 0.001**	0.029 ± 0.007	0.971 ± 0.007
	0.50	**0.990 ± 0.000**	0.022 ± 0.006	0.978 ± 0.006	**0.987 ± 0.001**	0.032 ± 0.003	0.968 ± 0.003
	1.00	0.988 ± 0.001	0.027 ± 0.001	0.973 ± 0.001	0.983 ± 0.002	**0.028 ± 0.002**	**0.972 ± 0.002**
	5.00	0.988 ± 0.001	0.025 ± 0.007	0.975 ± 0.007	0.985 ± 0.002	0.035 ± 0.008	0.965 ± 0.008
	10.00	0.984 ± 0.001	0.027 ± 0.005	0.973 ± 0.005	0.981 ± 0.002	0.032 ± 0.009	0.967 ± 0.009

Table 10. Performance evaluation of fine-tuning strategies on UTKFace dataset, averaged over three random seeds (for varying method-specific hyperparameter values). The best values for each method are shown in **bold**.