FairMT: Fairness for Heterogeneous Multi-Task Learning

Guanyu Hu
Xi'an Jiaotong University
Queen Mary University of London

Tangzheng Lian
Kings College London

Na Yan
Kings College London

Dimitrios Kollias
Queen Mary University of London

Xinyu Yang
Xi'an Jiaotong University

Oya Celiktutan
Kings College London

Siyang Song
University of Exeter

Zeyu Fu
University of Exeter

Abstract

Fairness in machine learning has been extensively studied in single-task settings, while fair multi-task learning (MTL), especially with heterogeneous tasks (classification, detection, regression) and partially missing labels, remains largely unexplored. Existing fairness methods are predominantly classification-oriented and fail to extend to continuous outputs, making a unified fairness objective difficult to formulate. Further, existing MTL optimization is structurally misaligned with fairness: constraining only the shared representation, allowing task heads to absorb bias and leading to uncontrolled task-specific disparities. Finally, most work treats fairness as a zero-sum trade-off with utility, enforcing symmetric constraints that achieve parity by degrading well-served groups. We introduce FairMT*, a unified fairness-aware MTL framework that accommodates all three task types under incomplete supervision. At its core is an Asymmetric Heterogeneous Fairness Constraint Aggregation mechanism, which consolidates task-dependent asymmetric violations into a unified fairness constraint. Utility and fairness are jointly optimized via a primal–dual formulation, while a head-aware multi-objective optimization proxy provides a tractable descent geometry that explicitly accounts for head-induced anisotropy. Across three homogeneous and heterogeneous MTL benchmarks encompassing diverse modalities and supervision regimes,* FairMT *consistently achieves substantial fairness gains while maintaining superior task utility. Code will be released upon paper acceptance.*

1. Introduction

Fairness has become a fundamental requirement for machine learning systems deployed in socially sensitive domains such as healthcare, hiring, and credit lending [17, 20, 24]. However, research on algorithmic fairness has focused almost exclusively on the single-task setting [18, 29, 30], primarily in classification contexts. In contrast, fairness in *multi-task learning* (MTL) remains largely unexplored. The coexistence of multiple and often heterogeneous tasks makes it difficult to define and enforce fairness consistently across objectives, resulting in conflicting criteria and incompatible optimization goals. Thus, directly applying conventional fairness methods leads to degraded performance.

Achieving fairness in MTL is impeded by several challenges: **(1) Fairness by Suppression:** Prevailing fairness paradigms (**Fig. 1(c)**) can be broadly categorized into (i) adversarial representation learning[15, 31], (ii) correlation-independent regularization (e.g., HSIC fairness) [7, 21], and (iii) constraint-based optimization [1, 30], they address bias largely through suppression rather than enhancement. first two Representation-based approaches enforce fairness by decorrelating learned features from sensitive attributes, but this often induces representational collapse and weakens task-relevant semantics, degrading performance across all groups. Constraint-based formulations directly penalize disparity gaps yet tend to equalize outcomes by lowering the performance of well-served groups instead of elevating disadvantaged ones. The central challenge, therefore, lies in achieving fairness through *directional improvement*—enhancing underperforming groups while preserving representational integrity and overall utility.

(2) Incompatible fairness definitions. Fairness metrics are predominantly defined for discrete classification or detection [3, 18, 29], which can not naturally extend to regression outputs. Different tasks operate under *incompatible* fairness notions, and no unified mechanism exists to aggregate them into a single optimization objective; **(3) Head-induced residual bias of MTL.** Conventional MTL paradigms (**Fig. 1(d)**) constrain only the shared encoder while leaving task heads free to internalize residual bias.

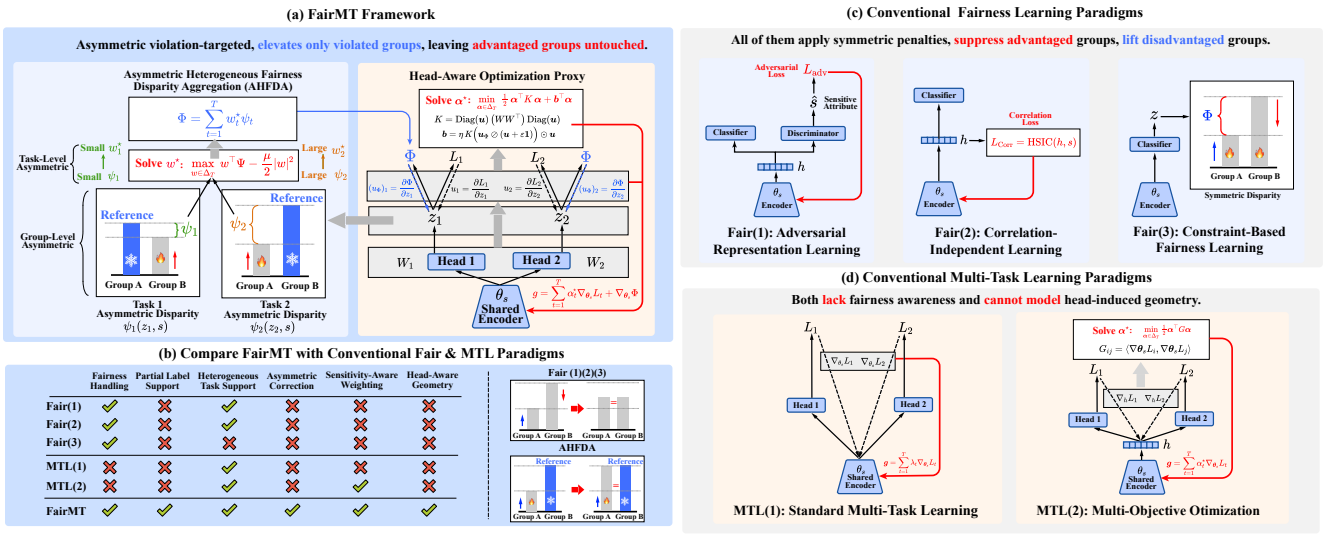

Figure 1. **Comparison of FairMT with Conventional Fair or MTL Paradigms.** (a) Our proposed FairMT framework. (b) Comparison between FairMT and conventional paradigms: the left panel highlights our advantages, and the right panel illustrates how AHFDA achieves fairness by asymmetrically lifting only violated groups without harming advantaged groups. (c) Conventional fairness learning paradigms, all applying symmetric penalties that simultaneously suppress advantaged groups and lift disadvantaged groups. (d) Conventional multi-task learning paradigms, both lacking fairness awareness and unable to model head-induced geometry.

As each head independently reshapes gradients, group disparities re-emerge at the prediction level and remain uncontrolled across tasks, revealing a lack of fairness awareness and an inability to model head-induced geometry; and **(4) Fairness under partial supervision.** Both fairness-aware and MTL methods implicitly assume *full annotations* are provided across all tasks [13, 16, 23, 28], whereas real-world datasets contain non-overlapping labels. Under partial annotation, models lack consistent supervision and joint utility–fairness optimization becomes ill-posed.

To address the above challenges in developing fair MTL, we propose FairMT, the *first* unified fairness-aware MTL framework that addresses all four challenges simultaneously. For **Challenges (1)**, we introduces task-dependent *asymmetric fairness disparity* metric that promote directional improvement for under-served groups rather than degrading dominant ones. Through an *Asymmetric Heterogeneous Fairness Disparity Aggregation (AHFDA)* mechanism, these task-specific disparity are consolidated into a single, optimizable objective, which addresses both **Challenges (2)** and **Challenges (4)** by providing a principled aggregation rule that accommodates heterogeneous outputs and partial supervision. **Challenge (3)** is solved by a *head-aware optimization proxy*, which not only corrects head-induced geometric bias but also reduces the full-parameter constrained optimization to an efficient head-space surrogate, preserving fairness at a fraction of the computational cost. These components are integrated into a tractable primal–dual formulation, enabling joint optimization of utility and fairness under a unified theoretical framework.

The main contributions of this paper are fourfold:

- **FairMT:** The first unified fairness-aware MTL framework that seamlessly supports classification, detection, and regression, even under partial task annotation.
- **AHFDA:** An asymmetric heterogeneous fairness aggregation mechanism that yields directional fairness gains for under-served groups without inducing utility collapse.
- **Head-aware multi-objective optimization:** Eliminates residual head-level bias and provides a tractable descent geometry for fairness optimization.
- **SOTA performance.** Extensive experiments across heterogeneous benchmarks, spanning datasets from different modalities and supervision regimes, shows substantial fairness gains with superiority in utility, consistently outperformed existing leading fairness and MTL baselines.

2. Preliminaries

We consider a fairness-aware multi-task learning (MTL) setting with T heterogeneous tasks, indexed by $[T] = \{1, \dots, T\}$. The training corpus is compiled from heterogeneous datasets (e.g., valence-arousal regression and action-unit detection), resulting in inherently *partial task supervision*—each sample provides labels for only a subset of tasks. Each sample is written as $(X, G, \{Y_t\}_{t\in\Omega})$, where $X \in \mathcal{X}$ is the input, $G \in \mathcal{G}$ is a sensitive attribute, and $\Omega \subseteq [T]$ denotes the subset of tasks with observed labels. The MTL model is parameterized by $\boldsymbol{\theta} = (\boldsymbol{\theta}_s, \{\boldsymbol{\theta}_t\}_{t=1}^T)$, where $\boldsymbol{\theta}_s$ denotes shared parameters and $\boldsymbol{\theta}_t$ denotes task-specific parameters. Each task t predicts via $f_{\boldsymbol{\theta}_s,\boldsymbol{\theta}_t} : \mathcal{X} \to \widehat{\mathcal{Y}}_t$ and incurs loss $\ell_{t,\boldsymbol{\theta}}(x, y_t) = M_t\, \ell_t(f_{\boldsymbol{\theta}_s,\boldsymbol{\theta}_t}(x), y_t)$, where

$M_t = \mathbb{1}[t \in \Omega]$ masks unobserved labels.

Conventional MTL. For each task t, assign a nonnegative weight c_t, with $\sum_{t=1}^{T} c_t > 0$, the MTL objective is

$$\min_{\boldsymbol{\theta}_s, \{\boldsymbol{\theta}_t\}} \sum_{t=1}^{T} c_t \, L_t(\boldsymbol{\theta}_s, \boldsymbol{\theta}_t),$$

where $L_t(\boldsymbol{\theta}_s, \boldsymbol{\theta}_t) := \mathbb{E}[\ell_{t,\boldsymbol{\theta}}(X, Y_t)]$ is task-specific risk.

Fairness in MTL. Classical fair learning imposes per-task constraints of the form $\psi_t(\boldsymbol{\theta}_s, \boldsymbol{\theta}_t) \le \epsilon_t$, where ψ_t measures group disparities (e.g., equalized odds, demographic-parity relaxations). A direct MTL extension yields

$$\min_{\boldsymbol{\theta}_s, \{\boldsymbol{\theta}_t\}} \sum_{t=1}^{T} c_t L_t(\boldsymbol{\theta}_s, \boldsymbol{\theta}_t) \quad \text{s.t.} \quad \sum_{t=1}^{T} \psi_t(\boldsymbol{\theta}_s, \boldsymbol{\theta}_t) \le \epsilon.$$

Differentiable Fairness Rates FAIRMT requires a unified formulation of differentiable fairness rates for both binary detection tasks and multi-class classification tasks, ensuring consistent gradient signals across heterogeneous task types. We therefore define soft population-level rates and their mini-batch estimators for each task type as follows.

Binary Detection Task. For each binary detection task $t \in [T]$ with partial-annotation mask $M_t = \mathbb{1}[t \in \Omega]$, let $p_t(x) \in [0,1]$ denote the *differentiable predicted-positive score*, typically parameterized as $p_t(x) = \sigma\big(f_{\boldsymbol{\theta}_s, \boldsymbol{\theta}_t}(x)\big)$ with $\sigma(\cdot)$ the logistic sigmoid. For a sensitive group $g \in \mathcal{G}$, define the *soft* true positive rate (TPR) and false positive rate (FPR) as the conditional expectations $\mathrm{TPR}_{g,t} = \mathbb{E}[p_t(X) \mid Y_t = 1, G = g]$, $\mathrm{FPR}_{g,t} = \mathbb{E}[p_t(X) \mid Y_t = 0, G = g]$. Their mini-batch estimates on $\mathcal{B}$ are naturally obtained via masked averages:

$$\widehat{\mathrm{TPR}}_{g,t} = \frac{1}{|\mathcal{B}^+_{g,t}|} \sum_{i \in \mathcal{B}^+_{g,t}} p_t(X^{(i)}),$$

$$\widehat{\mathrm{FPR}}_{g,t} = \frac{1}{|\mathcal{B}^-_{g,t}|} \sum_{i \in \mathcal{B}^-_{g,t}} p_t(X^{(i)}),$$

where $\mathcal{B}^+_{g,t} = \{i \in \mathcal{B} : G^{(i)} = g, Y_t^{(i)} = 1\}$ and $\mathcal{B}^-_{g,t} = \{i \in \mathcal{B} : G^{(i)} = g, Y_t^{(i)} = 0\}$ denote the sets of positive and negative samples for group g and task t, respectively.

Multi-Class Classification Task. For task $t \in [T]$ with partial-annotation mask $M_t = \mathbb{1}[t \in \Omega]$, let $f_{\boldsymbol{\theta}_s, \boldsymbol{\theta}_t}(x) \in \mathbb{R}^{C_t}$ denote the class logits, and define temperature-scaled probabilities $p_{t,k}(x) = \mathrm{softmax}(f_{\boldsymbol{\theta}_s, \boldsymbol{\theta}_t}(x))_k$, where $k \in \{1, \dots, C_t\}$. For each class k, we treat k as "positive" and all other classes as "negative", yielding class-wise soft rates

$$\widehat{\mathrm{TPR}}_{g,t,k} = \frac{1}{|\mathcal{B}^+_{g,t,k}|} \sum_{i \in \mathcal{B}^+_{g,t,k}} p_{t,k}(X^{(i)}),$$

$$\widehat{\mathrm{FPR}}_{g,t,k} = \frac{1}{|\mathcal{B}^-_{g,t,k}|} \sum_{i \in \mathcal{B}^-_{g,t,k}} p_{t,k}(X^{(i)}),$$

where $\mathcal{B}^+_{g,t,k}$ and $\mathcal{B}^-_{g,t,k}$ denote positive and negative group.

3. FAIRMT Framework

In this section, we present FAIRMT (**Fig.1(a)**), a fairness-aware Pareto-guided primal–dual framework.

3.1. Asymmetric Heterogeneous Fairness Disparity Aggregation (AHFDA)

For classification, detection, and regression tasks, we define task-specific asymmetric fairness disparities emphasizing the worst-performing group and integrate them through a tailored simplex-based optimization. Instead of averaging heterogeneous disparities, AHFCA concentrates constraint pressure on the most violated tasks, producing a unified fairness objective via projection-free simplex updates with sparse, vertex-structured iterates.

3.1.1. Fair Binary Detection.

For binary detection, group fairness is commonly measured by following metric: *(i) Demographic Parity* (DP) enforces equality of positive prediction rates across groups, but ignores correctness and is unsuitable in settings where fairness must be conditioned on true labels. *(ii) Equal Opportunity* (EO) instead requires equal true positive rates (TPR), while *(iii) Equalized Odds* (EOD) additionally equalizes false positive rates (FPR), In the sequel we adopt EOD as primary detection-fairness constraint (with EO recovered as a special case) and formalize its asymmetric variant below.

Asymmetric Equalized Odds (AEOD). Classical EOD [10, 25, 32] evaluate symmetric discrepancies such as $|\mathrm{TPR}_{g,t} - \mathrm{TPR}_{g',t}|$ and $|\mathrm{FPR}_{g,t} - \mathrm{FPR}_{g',t}|$, allowing fairness to be achieved by *lowering* the performance of well-served groups rather than improving under-served ones. To enforce *directional* improvement, each task t is anchored to a group-wise **reference** (advantated group), so only lagging groups are penalized while high-performing groups remain unaffected. For true positives and false positives we define

$$\tau_t = \max_{g \in \mathcal{G}_t^+} \widehat{\mathrm{TPR}}_{g,t}, \qquad \phi_t = \min_{g \in \mathcal{G}_t^-} \widehat{\mathrm{FPR}}_{g,t},$$

where τ_t represents the best attainable TPR among all valid groups, and ϕ_t is the lowest achievable FPR, $\mathcal{G}_t^+ = \{g : |\mathcal{B}^+_{g,t}| > 0\}$ and $\mathcal{G}_t^- = \{g : |\mathcal{B}^-_{g,t}| > 0\}$ collect valid groups for positives and negatives. These references prevent performance collapse, ensuring fairness improvements are driven toward the current strongest group rather than suppressing it. We then penalize *only* lagging TPRs and *only* exceeding FPRs via one-sided hinge gaps:

$$\Delta^{\mathrm{TPR}}_{g,t} = \big[\, \tau_t - \widehat{\mathrm{TPR}}_{g,t} \,\big]_+, \qquad \Delta^{\mathrm{FPR}}_{g,t} = \big[\, \widehat{\mathrm{FPR}}_{g,t} - \phi_t \,\big]_+,$$

where $[z]_+ = \max\{z, 0\}$. To aggregate group-wise gaps into a single per-task disparity, a worst-case reducer is used. The AEOD ψ_t^{AEOD} for task t is defined as

$$M_t \, \frac{1}{1+\beta} \left(\max_{g \in \mathcal{G}_t^+} \big(\Delta^{\mathrm{TPR}}_{g,t}\big)^2 + \eta \max_{g \in \mathcal{G}_t^-} \big(\Delta^{\mathrm{FPR}}_{g,t}\big)^2 \right), \tag{1}$$

where $\beta \in \{0, 1\}$ recovers EO ($\beta = 0$) or EOD ($\beta = 1$), and $M_t = \mathbb{1}[t \in \Omega]$ enforces partial annotation.

3.1.2. Fair Multi-class Classification.

Unlike the binary case, fairness in multi-class classification is evaluated in a one-vs-rest manner, treating each class independently against all others and reducing multi-class predictions to a sequence of aligned positive–negative decisions. The resulting disparities are typically measured using **Multi-EOD/EO**, extending equalized TPR/FPR criteria in the same spirit as fairness evaluation for detection tasks. we adapt Multi-EOD to construct our asymmetric variant.

Class-wise Asymmetric Equalized Odds (CAEOD). As in the binary case, to enforce directional fairness, each class k is anchored to its best attainable reference level:

$$\tau_{t,k} = \max_{g \in \mathcal{G}^{+}_{t,k}} \widehat{\mathrm{TPR}}_{g,t,k}, \qquad \phi_{t,k} = \min_{g \in \mathcal{G}^{-}_{t,k}} \widehat{\mathrm{FPR}}_{g,t,k},$$

and we penalize only lagging TPRs or exceeding FPRs through one-sided hinge gaps

$$\Delta^{\mathrm{TPR}}_{g,t,k} = [\,\tau_{t,k} - \widehat{\mathrm{TPR}}_{g,t,k}\,]_+,$$
$$\Delta^{\mathrm{FPR}}_{g,t,k} = [\,\widehat{\mathrm{FPR}}_{g,t,k} - \phi_{t,k}\,]_+.$$

For each class k, group discrepancies are reduced via a worst-case operator and squared to stabilize gradients:

$$\psi^{\mathrm{CAEOD}}_{t,k} = \frac{M_t}{1+\eta}\left(\max_g(\Delta^{\mathrm{TPR}}_{g,t,k})^2 + \eta \max_g(\Delta^{\mathrm{FPR}}_{g,t,k})^2\right), \quad (2)$$

where $k = 1, \ldots, C_t$, $\eta \in \{0, 1\}$ recovers EO ($\eta = 0$) or EOD ($\eta = 1$). Thus, a multi-class task produces a vector $[\psi^{\mathrm{CAEOD}}_{t,1}, \ldots, \psi^{\mathrm{CAEOD}}_{t,C_t}]$, preserving per-class granularity and avoiding any implicit averaging across classes.

3.1.3. Fair Regression.

Group fairness in regression lacks a universally adopted metric. In practice, several criteria are commonly used: **(i) KS** (KS Distributional Demographic Parity)[5, 6], requiring group-wise predictive distributions to match, i.e., $\sup_{C\subset\mathbb{R}} \left|P(\hat{y} \in C \mid g) - P(\hat{y} \in C \mid g')\right| = 0$, **(ii) CSP** (Conditional Statistical Parity)[2, 26], Split by a threshold τ and evaluate, for each stratum $S \in \{Y \geq \tau, Y < \tau\}$, $\Delta_S = \max_g \mathbb{E}[\hat{Y} \mid G{=}g, S] - \min_g \mathbb{E}[\hat{Y} \mid G{=}g, S]$, where smaller Δ_S implies greater fairness. **(iii) EP** (Error parity)[2], enforcing similarity of group-wise expected errors $e_g = \mathbb{E}[|y - \hat{y}| \mid g]$; and **(iv) Binned EO/EOD**[17], continuous predictions $\hat{y}$ are discretized into K bins, inducing a K-way one-vs-rest task with EO/EOD.

Asymmetric Error Parity (AEP). To unify all above heterogeneous metric within a single differentiable objective, we introduce **AEP**, which explicitly reduces group-wise error parity and thereby jointly tightens multiple fairness criteria. In particular, shrinking residual gaps between groups contracts marginal prediction gaps and improves distributional parity (KS), aligns group-conditioned prediction trends and reduces $\mathbb{E}[\hat{Y} \mid G, S]$ discrepancies (CSP), and limits miscalibrated score bins that induce inter-group TPR/FPR differences (EO/EOD).

Formally, for task $t \in [T]$ with partial-annotation mask $M_t = \mathbb{1}[t \in \Omega]$, AEP impose a *one-sided*, fully differentiable penalty that only encourages under-performing groups to catch up with the best-performing group. let $g_t(x) = f_{\boldsymbol{\theta}_s, \boldsymbol{\theta}_t}(x) \in \mathbb{R}$ denote the predictor, and $\ell(\cdot,\cdot) \in \{\mathrm{MAE}, \mathrm{MSE}\}$ be the regression loss. On a mini-batch $\mathcal{B}$, define sample sets $\mathcal{B}_{g,t} = \{\, i \in \mathcal{B} : G^{(i)} = g \,\}$, and empirical group errors $\widehat{\mathrm{Err}}_{g,t} = \frac{1}{|\mathcal{B}_{g,t}|}\sum_{i\in\mathcal{B}_{g,t}} \ell\Big(g_t(X^{(i)}), Y^{(i)}_t\Big)$, and collect valid groups as $\mathcal{G}_t = \{g : |\mathcal{B}_{g,t}| > 0\}$. The best achievable performance on task t is used as a reference, and only groups exceeding this are penalized:

$$\widehat{\Delta}_{g,t} = \left[\widehat{\mathrm{Err}}_{g,t} - \hat{e}^{\mathrm{ref}}_t\right]_+, \qquad \hat{e}^{\mathrm{ref}}_t = \min_{g\in\mathcal{G}_t} \widehat{\mathrm{Err}}_{g,t}.$$

Larger deviations indicate greater disparity, and a worst-case aggregation drives uniform improvement across lagging groups. The resulting constraint is

$$\psi^{\mathrm{AEP}}_t = M_t \max_{g\in\mathcal{G}_t} \left(\widehat{\Delta}_{g,t}\right)^2, \quad (3)$$

3.1.4. Fairness Disparity Aggregation

In FAIRMT, each task $t \in [T]$ produces a nonnegative violation score ψ_t from the asymmetric metric defined above. Simply summing or averaging $\{\psi_t\}$ is misaligned with multi-task fairness: small disparities can mask severe ones, effort may be wasted on already-fair tasks, and heterogeneity in task difficulty is ignored. An effective aggregator should therefore (i) concentrate pressure on most violated tasks, (ii) be directional, avoiding penalties on fair tasks, (iii) adaptively allocates optimal pressure across tasks.

AHFDA Objective Formulation. To aggregate per-task violations into a single fairness objective, we assign each ψ_t a weight w_t with $\sum_t w_t = 1$, ensuring a normalized and interpretable allocation of corrective pressure across tasks. Let $\boldsymbol{\Psi} = (\psi_1, \ldots, \psi_T) \in \mathbb{R}^T_{\geq 0}$ denote task-level violations and $\boldsymbol{w} \in \Delta_T = \{\boldsymbol{w} \geq 0,\ \mathbf{1}^\top \boldsymbol{w} = 1\}$ be simplex weights. We define the *directional, heterogeneity-aware* aggregator

$$\Phi(\boldsymbol{\Psi}) = \max_{\boldsymbol{w}\in\Delta_T} \boldsymbol{w}^\top \boldsymbol{\Psi} - \frac{\mu}{2}\|\boldsymbol{w}\|_2^2, \qquad \mu > 0. \quad (4)$$

The linear term concentrates mass on high-violation tasks, enforcing a worst-case bias; the quadratic term prevents collapse to a single task, yielding stable and smooth weights. μ controls the *aggressiveness* of weight reallocation: small μ concentrates on high-ψ tasks, sharpening worst-case correction, whereas larger μ distributes pressure more smoothly.

Projection View and Closed-Form Solution. To obtain an analytically tractable characterization of the optimal simplex weight, we rewrite the maximization in Eq. (4) as the equivalent strongly convex minimization

$$\boldsymbol{w}^\star = \frac{\mu}{2}\|\boldsymbol{w}\|_2^2 - \boldsymbol{\Psi}^\top \boldsymbol{w}, \qquad \mu > 0, \quad (5)$$

which yields interpolation form $\frac{\mu}{2}\left\|\boldsymbol{w}-\frac{1}{\mu}\boldsymbol{\Psi}\right\|_2^2+\text{const}$, Hence, the minimizer admits a **closed-form** *geometric characterization* as the Euclidean projection of $\frac{1}{\mu}\boldsymbol{\Psi}$ onto the probability simplex: $\boldsymbol{w}^\star=\mathrm{Proj}_{\Delta_T}\left(\frac{1}{\mu}\boldsymbol{\Psi}\right)$, there exists a scalar Lagrange multiplier $\tau\in\mathbb{R}$ such that

$$w_t^\star=\max\{0,\ \tfrac{\psi_t}{\mu}-\tau\},\qquad \sum_{t=1}^{T}w_t^\star=1. \tag{6}$$

This closed-form expression is the well-known *water-filling* structure: coordinates falling below the learned threshold τ are truncated to 0, while the remaining coordinates are linearly shifted to satisfy the simplex constraint. Consequently, AHFDA is intrinsically *directional*: tasks with negligible or vanishing violations (i.e., $\psi_t\approx 0$) are assigned zero weight, concentrating the fairness signal on the most severe violators (see ***Appendix A*** for a complete derivation).

Projection-Free Aggregation. While Eq.(6) gives an exact solution, repeated projections can be computationally heavy and may destroy the sparsity of weight iterates within a training loop. We instead adopt a *projection-free* conditional gradient (Frank–Wolfe, FW) solver tailored to Eq.(5). Writing Eq.(5) as a quadratic program

$$\min_{\boldsymbol{w}\in\Delta_T}\frac{1}{2}\boldsymbol{w}^\top(\mu I)\boldsymbol{w}-\boldsymbol{\Psi}^\top\boldsymbol{w},\quad \nabla f(\boldsymbol{w})=\mu\boldsymbol{w}-\boldsymbol{\Psi}, \tag{7}$$

it requires only a *linear minimization oracle (LMO)* over the simplex, keeping iterates as sparse convex combinations of vertices and yielding interpretable active sets of unfair tasks.

The full procedure of projection-free AHFDA is given in **Algorithm 1**. At iteration ℓ, AHFDA constructs a feasible search direction by linearizing the objective around the current iterate $\boldsymbol{w}^{(\ell)}\in\Delta_T$: $\min_{\boldsymbol{s}\in\Delta_T}\langle\nabla f(\boldsymbol{w}^{(\ell)}),\boldsymbol{s}\rangle$, where $\langle\cdot,\cdot\rangle$ denotes the Euclidean inner product and $\boldsymbol{s}$ is a feasible point on the simplex (i.e., a candidate descent direction). Since Δ_T is the convex hull of $\{e_1,\dots,e_T\}$, any linear objective over the simplex attains its minimum at one of these vertices. Consequently, the LMO returns

$$\boldsymbol{s}^{(\ell)}=e_{j_\ell},\qquad j_\ell\in\arg\max_{t\in\{1,\dots,T\}}\left(\psi_t-\mu w_t^{(\ell)}\right), \tag{8}$$

where e_{j_ℓ} is the simplex vertex corresponding to task j_ℓ. The iterate is then updated by a convex combination, with a step size $\gamma_\ell\in[0,1]$ to maintain feasibility:

$$\boldsymbol{w}^{(\ell+1)}=(1-\gamma_\ell)\,\boldsymbol{w}^{(\ell)}+\gamma_\ell\,\boldsymbol{s}^{(\ell)}\in\Delta_T. \tag{9}$$

For quadratic objective Eq. (5), step size admits a closed form. Let $\boldsymbol{d}^{(\ell)}=\boldsymbol{s}^{(\ell)}-\boldsymbol{w}^{(\ell)}$. Exact line search gives

$$\gamma_\ell=\Pi_{[0,1]}\left(-\frac{\langle\,\mu\boldsymbol{w}^{(\ell)}-\boldsymbol{\psi},\ \boldsymbol{d}^{(\ell)}\,\rangle}{\mu\,\|\boldsymbol{d}^{(\ell)}\|_2^2}\right), \tag{10}$$

Algorithm 1 Asymmetric Heterogeneous Fairness Disparity Aggregation (AHFDA)

Require: Violations $\boldsymbol{\Psi}\in\mathbb{R}^d_{\geq 0}$, $\mu>0$, tolerance ε
Ensure: $\boldsymbol{w}^\star\in\Delta_T$
Initialize $\boldsymbol{w}^{(0)}\in\Delta_T$ (e.g., uniform), set $t=0$
repeat
  $\nabla f(\boldsymbol{w}^{(t)})=\mu\boldsymbol{w}^{(t)}-\boldsymbol{\Psi}$
  $j_t\in\arg\max_{k\in[d]}(\psi_k-\mu w_k^{(t)})$, $\boldsymbol{s}^{(t)}=e_{j_t}$
  $\boldsymbol{d}^{(t)}=\boldsymbol{s}^{(t)}-\boldsymbol{w}^{(t)}$
  Exact line-search:

$$\gamma_t=\Pi_{[0,1]}\left(-\frac{\langle\mu\boldsymbol{w}^{(t)}-\boldsymbol{\Psi},\ \boldsymbol{d}^{(t)}\rangle}{\mu\,\|\boldsymbol{d}^{(t)}\|_2^2}\right)$$

  $\boldsymbol{w}^{(t+1)}=\boldsymbol{w}^{(t)}+\gamma_t\,\boldsymbol{d}^{(t)}\in\Delta_T$
  $g_t=\langle\nabla f(\boldsymbol{w}^{(t)}),\ \boldsymbol{w}^{(t)}-\boldsymbol{s}^{(t)}\rangle$
  $t\leftarrow t+1$
until $g_{t-1}\leq\varepsilon$
return $\boldsymbol{w}^\star=\boldsymbol{w}^{(t)}$

where $\Pi_{[0,1]}$ denotes projection onto $[0,1]$. Thus, each AHFDA iteration consists of a gradient evaluation, one coordinate maximization, and a scalar step-size update—an $O(T)$ procedure that preserves sparsity and interpretability of the fairness weights (Detailed derivations and convergence guarantees are provided in ***Appendix B***).

3.2. FairMT Framework

In conventional MTL without fairness, the training objective is cast as a multi-objective optimization (MOO) balancing task risks [8, 23]. Once the aggregated fairness disparity $\Phi(\boldsymbol{\theta})$ from AHFDA in Eq. (4) is imposed as a constraint $\Phi(\boldsymbol{\theta})\leq\epsilon$, the problem becomes a *constrained* MOO:

$$\min_{\boldsymbol{\theta}_s,\{\boldsymbol{\theta}_t\}}\ \big(L_1(\boldsymbol{\theta}_s,\boldsymbol{\theta}_1),\dots,L_T(\boldsymbol{\theta}_s,\boldsymbol{\theta}_T)\big),\quad \text{s.t. }\Phi(\boldsymbol{\theta})\leq\epsilon.$$

To address this, we introduce **FairMT**, a *Fairness-Aware Pareto-Guided Primal–Dual* framework, which introduces ***primal*** scalarization weights to balance the tasks and a ***dual*** multiplier to enforce the fairness constraint, relaxing the constrained formulation into an unconstrained one. This yields a saddle-point reformulation whose stationary solutions coincide with Pareto-stationary points.

Specifically, we seek to update the multi-task utility along a common descent direction that simultaneously improves all tasks. In line with AHFDA, we introduce the standard simplex $\Delta_T:=\{\boldsymbol{\alpha}\in\mathbb{R}^T_{\geq 0}\mid\sum_{t=1}^T\alpha_t=1\}$, and let $\boldsymbol{\alpha}\in\Delta_T$ denote the primal task weights that govern navigation along the Pareto frontier. The simplex constraint enforces nonnegativity and unit-sum normalization, ensuring that $\boldsymbol{\alpha}$ encodes valid convex combinations of task gradients.

Let the model parameters be $\boldsymbol{\theta}:=\big(\boldsymbol{\theta}_s,\{\boldsymbol{\theta}_t\}_{t=1}^T\big)$, aggregated fairness disparity $\Phi(\boldsymbol{\theta})$, and tolerance $\epsilon\geq 0$ for

admissible violations. The fairness constraint is therefore written as $h(\boldsymbol{\theta}) \triangleq \Phi(\boldsymbol{\theta}) - \epsilon \leq 0$. We form the Lagrangian

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\alpha}, \eta) = \underbrace{\sum_{t=1}^{T} \alpha_t \, L_t(\boldsymbol{\theta}_s, \boldsymbol{\theta}_t)}_{\text{primal (task risks)}} + \underbrace{\eta \, h(\boldsymbol{\theta})}_{\text{dual (fairness)}} . \tag{11}$$

Here, η is the dual multiplier associated with $h(\boldsymbol{\theta})$, adaptively regulating the influence of the fairness constraint and yielding a saddle-point formulation that harmonizes multi-task learning with constraint satisfaction. Consequently, the FAIRMT objective can be written in min–max form:

$$\min_{\boldsymbol{\theta}_s, \{\boldsymbol{\theta}_t\}_{t=1}^T, \boldsymbol{\alpha} \in \Delta_T} \; \max_{\eta \geq 0} \; \mathcal{L}\big(\boldsymbol{\theta}_s, \{\boldsymbol{\theta}_t\}, \boldsymbol{\alpha}, \eta\big). \tag{12}$$

KKT conditions and Pareto stationarity. An optimal saddle point $(\boldsymbol{\theta}_s^\star, \{\boldsymbol{\theta}_t^\star\}_{t=1}^T, \boldsymbol{\alpha}^\star, \eta^\star)$ satisfies:

(i) *Primal feasibility (utility):* $h(\boldsymbol{\theta}^\star) \leq 0$ and $\boldsymbol{\alpha}^\star \in \Delta_T$, i.e., $\Phi(\boldsymbol{\theta}^\star) \leq \epsilon$, $\alpha_t^\star \geq 0$, and $\sum_{t=1}^T \alpha_t^\star = 1$.

(ii) *Dual feasibility (fairness):* $\eta^\star \geq 0$. The dual multiplier assigns a nonnegative penalty to fairness violations, as required for inequality constraints.

(iii) *Complementary slackness (fairness):* $\eta^\star \, h(\boldsymbol{\theta}^\star) = 0$. If the constraint is active ($\Phi(\boldsymbol{\theta}^\star) = \epsilon$), then $\eta^\star \geq 0$; otherwise ($\Phi(\boldsymbol{\theta}^\star) < \epsilon$), multiplier vanishes ($\eta^\star = 0$).

(iv) *Stationarity (w.r.t. model parameters):* At optimality, first-order balance holds between aggregated task risks and fairness term. In *shared* parameter block:

$$\nabla_{\boldsymbol{\theta}_s} \left(\sum_{t=1}^{T} \alpha_t^\star L_t(\boldsymbol{\theta}_s^\star, \boldsymbol{\theta}_t^\star) \right) + \eta^\star \, \nabla_{\boldsymbol{\theta}_s} \Phi(\boldsymbol{\theta}^\star) = \mathbf{0}, \tag{13}$$

and in each *task-specific* block, for $\forall \, t \in \{1, \ldots, T\}$,

$$\alpha_t^\star \, \nabla_{\boldsymbol{\theta}_t} L_t(\boldsymbol{\theta}_s^\star, \boldsymbol{\theta}_t^\star) + \eta^\star \, \nabla_{\boldsymbol{\theta}_t} \Phi(\boldsymbol{\theta}^\star) = \mathbf{0}. \tag{14}$$

Under these conditions, $(\boldsymbol{\theta}^\star, \boldsymbol{\alpha}^\star, \eta^\star)$ constitutes a *fairness-constrained Pareto-stationary point*: the task risks are Pareto-balanced by $\boldsymbol{\alpha}^\star$ while feasibility of the fairness constraint is enforced via $\eta^\star$.

3.3. FAIRMT Optimization

Since FAIRMT jointly optimizes task risks under a global fairness constraint, exact stationary is generally intractable. We adopt an iterative primal–dual optimization scheme.

STEP 1: Primal Update of Task Weights. The task weights $\boldsymbol{\alpha}$ are explicit optimization variables in the KKT Eq. (13). They control how task gradients mix in the shared encoder and thus shape the descent geometry. However, the stationary point may not correspond to a valid simplex point $\boldsymbol{\alpha} \in \Delta_T$. We therefore approximate it by the *least-squares projection* of fairness-adjusted direction onto the *convex hull* of task gradients. Geometrically, with $\boldsymbol{\theta}$ fixed, the set $\{\nabla_{\boldsymbol{\theta}_s} L_t\}_{t=1}^T$ spans all feasible shared-parameter descent directions, and selecting best $\boldsymbol{\alpha}$ amounts to finding the convex combination closest to $-\eta \, \nabla_{\boldsymbol{\theta}_s} \Phi$, which is classical minimum-norm point problem [22, 27]. Thus, at iteration r we compute $\boldsymbol{\alpha}^{(r)}$ as the minimizer of

$$\min_{\boldsymbol{\alpha} \in \Delta_T} \tfrac{1}{2} \Big\| \sum_{t=1}^{T} \alpha_t \, \nabla_{\boldsymbol{\theta}_s} L_t(\boldsymbol{\theta}^{(r)}) + \eta^{(r)} \, \nabla_{\boldsymbol{\theta}_s} \Phi(\boldsymbol{\theta}^{(r)}) \Big\|_2^2. \tag{15}$$

When $\eta^{(r)}{=}0$, it reduces exactly to the MGDA update [23]. When T is large, solving the high-dimensional minimum-norm point exactly becomes costly and numerically fragile. We therefore introduce the **Head-Aware Proxy** in **Section. 3.4**, which recasts residual into convex quadratic program (QP) of Eq. (22) and yields a stable, efficient update.

STEP 2: Primal Update of Model Parameters. Given the current task weights $\boldsymbol{\alpha}^{(r)}$, we update the shared parameters $\boldsymbol{\theta}_s$ and the task-specific parameters $\{\boldsymbol{\theta}_t\}_{t=1}^T$ by a gradient step on the Lagrangian. With learning rate $\rho > 0$,

$$\boldsymbol{\theta}_s \leftarrow \boldsymbol{\theta}_s - \rho \Big(\sum_{t=1}^{T} \alpha_t^{(r)} \, \nabla_{\boldsymbol{\theta}_s} L_t(\boldsymbol{\theta}) + \eta^{(r)} \, \nabla_{\boldsymbol{\theta}_s} \Phi(\boldsymbol{\theta}) \Big), \tag{16}$$

$$\boldsymbol{\theta}_t \leftarrow \boldsymbol{\theta}_t - \rho \Big(\nabla_{\boldsymbol{\theta}_t} L_t(\boldsymbol{\theta}) + \eta^{(r)} \, \nabla_{\boldsymbol{\theta}_t} \Phi(\boldsymbol{\theta}) \Big), \quad \forall t \in [T]. \tag{17}$$

This corresponds to a primal descent step under the fairness-adjusted gradient field.

STEP 3: Dual Update of the Fairness Multiplier. The dual variable η is updated by a projected gradient-ascent step on the fairness constraint. With step size $\rho_\eta > 0$,

$$\eta \;\leftarrow\; \big[\eta + \rho_\eta \big(\Phi_\varepsilon(\boldsymbol{\theta}) - \epsilon \big) \big]_+, \tag{18}$$

where $[\cdot]_+$ projects onto $\mathbb{R}_{\geq 0}$. This update preserves dual feasibility and implicitly enforces complementary slackness: when $\Phi_\varepsilon(\boldsymbol{\theta}) < \epsilon$, η is driven toward 0; when the constraint is violated, η increases, strengthening the fairness term in the primal updates. Thus, the dual dynamics automatically regulate the degree of fairness enforcement.

3.4. Head-Aware Optimization Proxy.

In **STEP 1**, when T is large, obtaining a fairness-aware multi-task descent direction requires full-model gradients for all tasks and the fairness functional, which is prohibitive in high-dimensional spaces. Prior methods introduce surrogates such as MGDA-UB [23] construct a quadratic proxy by linearizing task losses in the shared representation, but implicitly assume that task heads are **i) geometrically neutral** and **ii) ignore the fairness constraint**. In practice, classifier heads apply heterogeneous linear transformations to shared features, distorting the effective descent geometry and potentially biasing updates. To address both issues, we introduce the ***Head-Aware Optimization Proxy***, which explicitly incorporates head-induced geometry of task heads and their data-dependent utility-fairness sensitivities, and recasts residual into convex quadratic program of Eq. (22).

Let $M_t = \mathbb{1}[t \in \Omega]$ indicate which samples in the mini-batch contribute to task t, and define $\mathcal{B}_t = \{i \in \mathcal{B} : M_t^{(i)} =$

1}. For each task t, let $z_t(X^{(i)})$ denote the logit from $f_{\boldsymbol{\theta}_s,\boldsymbol{\theta}_t}$. We compute the average logit-level sensitivities of the task losses and the fairness functional as

$$u_t = \frac{1}{|\mathcal{B}_t|}\sum_{i\in\mathcal{B}_t}\frac{\partial L_t(\boldsymbol{\theta})}{\partial z_t(X^{(i)})}, \quad (u_\Phi)_t = \frac{1}{|\mathcal{B}_t|}\sum_{i\in\mathcal{B}_t}\frac{\partial \Phi(\boldsymbol{\theta})}{\partial z_t(X^{(i)})}. \tag{19}$$

Collecting $\{u_t\}_{t=1}^T$, $\{(u_\Phi)_t\}_{t=1}^T$ yields vectors $\boldsymbol{u}$, $\boldsymbol{u}_\Phi$.

Let $W \in \mathbb{R}^{T\times D}$ stack the final linear head weights of the T task heads. The matrix $WW^\top \in \mathbb{R}^{T\times T}$ is a Gram matrix: its (i,j)-entry $(WW^\top)_{ij} = W_iW_j^\top$ equals the inner product between the head directions of tasks i and j in the shared representation space. Large off-diagonal values thus indicate that two heads produce similar linear responses to shared features, while small values indicate weak alignment. To incorporate data-dependent sensitivity, define $\mathrm{Diag}(\boldsymbol{u})$ as the diagonal matrix with (t,t)-entry u_t. We introduce the *head-aware Gram matrix*

$$K \;:=\; \mathrm{Diag}(\boldsymbol{u})\left(WW^\top\right)\mathrm{Diag}(\boldsymbol{u}). \tag{20}$$

The diagonal reweighting preserves the geometry of $WW^\top$ while amplifying tasks with higher logit sensitivity.

To inject *fairness* in the same geometry, we define

$$\boldsymbol{b} \;:=\; \eta\, K\Big(\boldsymbol{u}_\Phi \oslash (\boldsymbol{u}+\varepsilon\mathbf{1})\Big)\odot \boldsymbol{u}, \tag{21}$$

where $\eta \geq 0$ is the dual multiplier in Eq.(18), $\oslash$ denotes elementwise division, $\odot$ the Hadamard product, and $\varepsilon > 0$ avoids division by zero. Here $\boldsymbol{u}_\Phi \in \mathbb{R}^T$ is the fairness analogue of $\boldsymbol{u}$, whose t-th entry measures the average sensitivity of Φ to the logits of task t. Consequently, utility and fairness live in a common inner-product space, preventing misaligned gradient updates. To solve Eq. (15), we project the shared gradients onto the head space. Using the approximations $\nabla_{\boldsymbol{\theta}_s} L_t \approx u_t\, W_t$ and $\nabla_{\boldsymbol{\theta}_s}\Phi \approx (u_\Phi)_t\, W_t$, the residual becomes a quadratic in $\boldsymbol{\alpha}$ involving $(K,\boldsymbol{b})$. This gives head-aware convex QP

$$\min_{\boldsymbol{\alpha}\in\Delta_T}\ \tfrac{1}{2}\,\boldsymbol{\alpha}^\top K\,\boldsymbol{\alpha} + \boldsymbol{b}^\top\boldsymbol{\alpha}, \tag{22}$$

whose solution yields a head-aware approximation of the fairness-adjusted descent direction. The problem involves only T variables and admits projection-free Frank–Wolfe updates at $O(T^2)$ per iteration, negligible relative to backpropagation. When $WW^\top \propto I_T$ and $\boldsymbol{u} \propto \mathbf{1}$, the proxy reduces to MGDA-UB, recovering shared-only surrogate (Detailed derivations are provided in ***Appendix C***).

4. Experiments

MTL Settings We evaluate FAIRMT on three large-scale multi-task benchmarks spanning both vision and language modalities, mixing homogeneous and heterogeneous task structures, and covering both full and partial supervision.

Table 1. Comparison of FAIRMT with multi-task (†), fairness-only (*), and hybrid fair-MTL (†*) baselines. Utility metrics (Acc, F1, CCC) are denoted with ↑ (higher is better), while fairness metrics (EOD, EO, KS, CSP) are denoted with ↓ (lower is fairer).

(a) Attribute Detection (in %)

CelebA (Binary Detection)

Method	Acc ↑	Age		Acc ↑	Gender	
		EOD ↓	EO ↓		EOD ↓	EO ↓
MGDA†	87.7	11.4	10.4	87.7	22.5	22.3
GradNorm†	84.8	9.4	8.4	84.8	19.4	18.3
LNL*	80.8	7.8	7.5	79.7	13.6	10.6
UFaTE*	76.4	6.8	5.5	75.4	12.2	10.2
MGDA-Mean†*	76.9	7.7	6.3	74.6	17.8	17.2
MGDA-UFaTE†*	78.6	6.8	4.4	77.8	12.6	9.6
FairMT	**88.4**	**3.4**	**2.0**	**88.1**	**6.4**	**3.9**

(b) Affective Analysis (in %)

AffectNet (Classification)

Method	Acc ↑	Age		Acc ↑	Gender		Acc ↑	Race	
		EOD ↓	EO ↓		EOD ↓	EO ↓		EOD ↓	EO ↓
MGDA†	59.5	23.4	23.4	**59.5**	8.6	8.6	**59.5**	13.9	13.9
GradNorm†	57.5	20.3	19.6	57.5	9.1	9.1	57.5	14.9	14.8
LNL*	53.6	15.1	15.9	55.4	7.1	7.0	53.6	10.9	9.9
UFaTE*	51.1	12.8	12.4	53.4	6.0	6.0	54.0	8.7	8.5
MGDA-Mean†*	51.4	19.9	18.4	56.7	7.2	6.0	55.8	8.4	7.4
MGDA-UFaTE†*	55.6	15.0	13.2	56.4	6.7	6.8	56.2	7.4	7.4
FairMT	**60.6**	**9.9**	**8.3**	**59.5**	**4.0**	**3.9**	59.4	**5.6**	**4.8**

EmotioNet (Binary Detection)

Method	Acc ↑	Age		Acc ↑	Gender		Acc ↑	Race	
		EOD ↓	EO ↓		EOD ↓	EO ↓		EOD ↓	EO ↓
MGDA†	81.3	19.1	19.4	81.3	6.5	5.9	81.3	18.6	18.4
GradNorm†	78.4	21.4	20.4	78.4	5.8	5.6	78.4	18.5	18.4
LNL*	69.5	12.9	11.1	73.5	4.5	4.3	74.8	15.4	15.1
UFaTE*	67.3	11.1	10.9	77.4	3.3	4.8	77.8	14.3	13.3
MGDA-Mean†*	62.0	16.5	15.1	74.5	3.7	3.6	72.8	13.4	13.1
MGDA-UFaTE†*	69.6	12.5	11.2	75.4	3.9	3.5	74.7	14.9	13.7
FairMT	**81.6**	**8.8**	**7.5**	**81.8**	**2.3**	**1.8**	**81.8**	**8.1**	**6.8**

AffectNet-VA (Regression)

Method	CCC ↑	Age			CCC ↑	Gender			CCC ↑	Race		
		KS ↓	CSP ↓	EOD ↓		KS ↓	CSP ↓	EOD ↓		KS ↓	CSP ↓	EOD ↓
MGDA†	**69.3**	19.1	14.0	13.0	69.3	16.5	7.4	5.5	69.2	17.1	13.3	15.7
GradNorm†	67.2	17.8	14.4	13.1	67.5	17.7	7.1	6.7	67.3	18.2	13.0	14.1
LNL*	60.9	16.3	12.3	11.3	63.6	14.9	6.9	5.1	64.3	12.6	8.1	9.6
UFaTE*	58.5	14.3	12.9	8.6	63.0	13.1	6.1	5.3	63.8	13.1	7.9	10.9
MGDA-Mean†*	59.0	18.5	13.8	11.5	64.1	14.9	5.5	4.7	62.1	16.6	11.9	13.6
MGDA-UFaTE†*	61.9	17.8	13.4	12.1	65.9	13.8	6.8	5.8	64.3	10.0	7.1	8.7
FairMT	68.9	**13.9**	**7.1**	**9.6**	**69.6**	**10.7**	**3.7**	**2.2**	**69.7**	**6.2**	**3.3**	**4.9**

Toxicity & Social Bias Analysis (in %)

CiviComment (Detection Task)

Method	Acc ↑	LGBTQ+		Acc ↑	Race		Acc ↑	Religion	
		EOD ↓	EO ↓		EOD ↓	EO ↓		EOD ↓	EO ↓
MGDA†	87.1	14.3	14.1	81.6	22.9	22.2	**84.1**	28.1	27.9
GradNorm†	86.2	13.1	12.8	78.7	23.2	23.3	83.1	29.0	29.3
LNL*	82.6	7.6	7.6	74.1	14.2	14.7	77.6	17.7	19.7
UFaTE*	80.9	6.2	6.1	65.9	14.0	13.4	73.4	14.8	14.8
MGDA-Mean†*	81.9	11.8	11.9	74.4	17.8	16.5	75.1	15.7	14.8
MGDA-UFaTE†*	82.8	9.9	9.5	75.3	16.5	16.0	76.7	13.1	13.2
FairMT	**87.7**	**5.5**	**4.5**	**81.9**	**10.1**	**9.9**	83.8	**6.7**	**6.0**

CiviComment (Toxicity Regression)

Method	CCC ↑	LGBTQ+			CCC ↑	Race			CCC ↑	Religion		
		KS ↓	CSP ↓	EOD ↓		KS ↓	CSP ↓	EOD ↓		KS ↓	CSP ↓	EOD ↓
MGDA†	73.4	34.9	12.5	27.4	69.2	42.7	18.2	38.9	**72.0**	39.8	15.8	35.3
GradNorm†	71.5	39.5	14.9	26.8	68.4	41.1	18.8	35.6	69.4	39.9	22.3	38.3
LNL*	67.7	28.3	9.7	23.9	65.3	32.9	15.1	27.9	66.1	29.9	9.5	21.6
UFaTE*	65.9	25.5	7.3	21.7	62.6	27.8	15.2	26.1	63.7	25.4	12.8	24.3
MGDA-Mean†*	67.6	31.1	11.2	24.3	63.7	36.7	16.2	31.9	64.4	33.9	12.3	31.3
MGDA-UFaTE†*	68.9	24.9	6.0	22.8	64.1	29.2	14.5	27.3	66.5	28.4	11.8	23.3
FairMT	**73.7**	**21.7**	**4.9**	**19.5**	**69.3**	**23.9**	**10.2**	**21.1**	71.1	**20.7**	**7.2**	**19.9**

These settings jointly comprise detection, classification, and regression objectives, and require fairness assessment across diverse demographic axes including age, gender, race, religion, and sexual orientation, as summarized in Table 2. ***Please Refer to Appendix D for comprehensive descriptions of the datasets, implementation settings, evaluation metrics (Fairness & Utility), and SOTA baselines.***

Comparison with state-of-the-art (SOTA). As shown in Figure 2 and Table 1, FAIRMT consistently achieves su-

Table 2. Benchmark setups for fair multi-task learning across modalities, task regimes, supervision types, and sensitive groups.

Task Name	#Tasks	Modality	Task Regime	Supervision	Datasets	Task Types	Sensitive Groups
Attribute Detection	29	Visual	Homogeneous	Full labels	CelebA[14]	**Detection:** 29 Facial Attributes	**2 Age** {Young, Old}, **2 Gender** {Male, Female}
Affective Analysis	17	Visual	Heterogeneous	Partial labels	AffectNet[19] AffectNet-VA[19] EmotioNet[9]	**Classification:** 7 Expressions **Regression:** Valence & Arousal **Detection:** 8 Action-Unit Detection	**7 Age**: {00–09, 10–19, 20–29, 30–39,40–49, 50–59, 60+} **2 Gender**: {Male, Female} **4 Race**: {Asian, Black, Indian, White}
Toxicity & Social Bias Analysis	9	Text	Heterogeneous	Full labels	Civil Comments[12]	**Detection:** Rating, Identity Attack, Insult, Obscene, Likes, Disagree, Funny, Wow, Sad **Regression:** Toxicity Value	**2 Sexual Orientation**: {LGBTQ+, non-LGBTQ+} **3 Religions**: {Christian, Muslim, No-Religions} **5 Race**: {Asian, Black,White, Latino, Other}

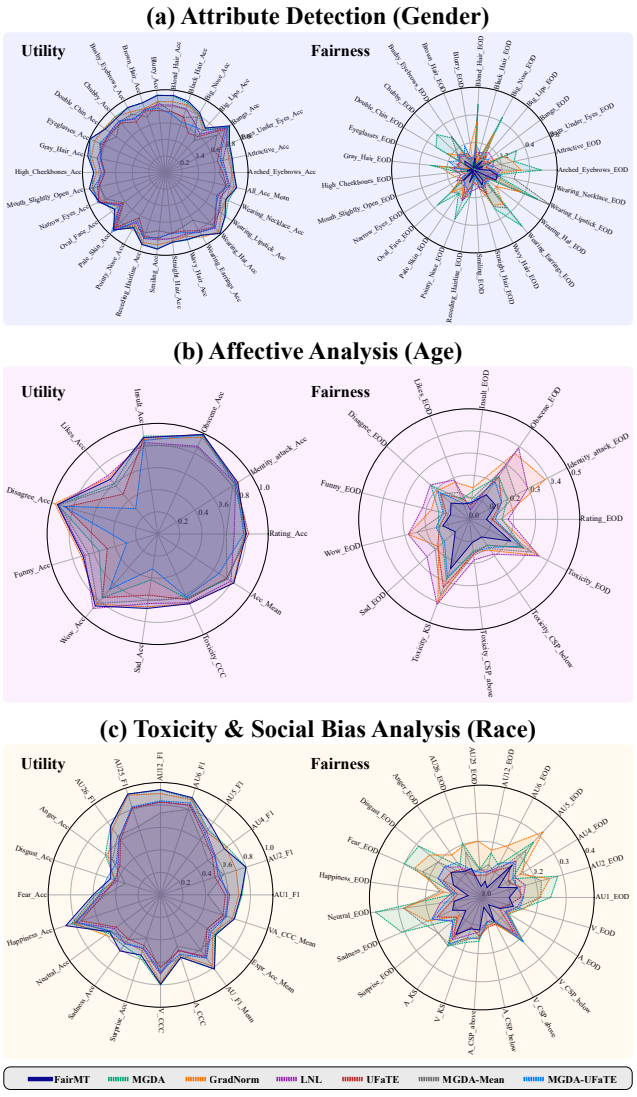

Figure 2. Comparison between **FairMT** and state-of-the-art baselines across three multi-task settings: (a) Attribute Detection, (b) Affective Analysis, and (c) Toxicity & Social Bias Analysis.

perior fairness *without incurring utility trade-off*, and even attains the best task performance across most benchmarks, modalities, and task compositions. We compare against representative SOTA baselines across three categories. For multi-task optimization, we include **MGDA** [23] and **GradNorm** [4]. For fairness-oriented learning, we adopt **LNL** [11] and **UFaTE** [7]. Since there is no comparable approach, we further construct two hybrid variants: **MGDA-Mean**, which applies MGDA under a naïve averaged disparity constraint by enforcing Eq. (2) on mean task disparities, and **MGDA-UFaTE**, which integrates MGDA with the UFaTE regularizer to jointly optimize utility and fairness. Multi-task optimizers (MGDA, GradNorm) improve overall utility but yield noticeable fairness gaps, whereas fairness-driven methods (LNL, UFaTE) reduce demographic disparities at the cost of predictive accuracy. Hybrid designs such as MGDA-Mean and MGDA-UFaTE partially alleviate this tension but remain inferior to FAIRMT on fairness/utility.

4.1. Ablation Study and Further Analysis

We comparing conventional MTL, utility-weighted baselines, MGDA, and their AHFDA-augmented variants. As shown in Fig. 3, FAIRMT attains a strictly better utility–fairness Pareto point. Adding AHFDA improves MT by lifting disadvantaged groups without degrading advantaged ones, and similarly strengthens MGDA, though its performance remains limited by MGDA's geometry-agnostic surrogate. Once MGDA's proxy is replaced with our head-aware formulation, both utility and fairness improve consistently, underscoring the importance of modeling head geometry for aligned fairness-aware MTL optimization **(Please see Appendix E for more detailed performance and efficiency analyses)**.

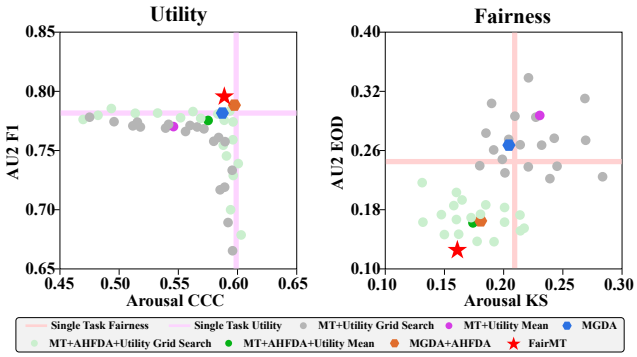

Figure 3. **Ablation on heterogeneous AU2–Arousal tasks.** We compare conventional MTL, utility-weighted baselines, MGDA, and their AHFDA-enhanced variants on both utility (left) and fairness (right).

5. Conclusion

We presented FAIRMT, a unified fairness-aware multi-task learning framework capable of handling heterogeneous tasks under partial supervision. By introducing the Asymmetric Heterogeneous Fairness Disparity Aggregation and a Pareto-guided optimization strategy, FAIRMT improves under-served groups without sacrificing overall task performance. Experiments across vision and language benchmarks show consistent gains in both utility and fairness across diverse datasets. This work illustrates that fairness and multi-task efficacy can be jointly optimized, opening avenues for extending FAIRMT to broader fairness notions and more complex real-world settings.

References

[1] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *International conference on machine learning*, pages 60–69. PMLR, 2018. 1

[2] Alekh Agarwal, Miroslav Dudík, and Zhiwei Steven Wu. Fair regression: Quantitative definitions and reduction-based algorithms. In *International conference on machine learning*, pages 120–129. PMLR, 2019. 4

[3] Richard Berk, Hoda Heidari, Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. A convex framework for fair regression. *arXiv preprint arXiv:1706.02409*, 2017. 1

[4] Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *International conference on machine learning*, pages 794–803. PMLR, 2018. 8

[5] Evgenii Chzhen, Christophe Denis, Mohamed Hebiri, Luca Oneto, and Massimiliano Pontil. Fair regression via plug-in estimator and recalibration with statistical guarantees. *Advances in Neural Information Processing Systems*, 33: 19137–19148, 2020. 4

[6] Evgenii Chzhen, Christophe Denis, Mohamed Hebiri, Luca Oneto, and Massimiliano Pontil. Fair regression via plug-in estimator and recalibration with statistical guarantees. *Advances in Neural Information Processing Systems*, 33: 19137–19148, 2020. 4

[7] Sepehr Dehdashtian, Bashir Sadeghi, and Vishnu Naresh Boddeti. Utility-fairness trade-offs and how to find them. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12037–12046, 2024. 1, 8

[8] Jean-Antoine Désidéri. Multiple-gradient descent algorithm (mgda) for multiobjective optimization. *Comptes Rendus Mathematique*, 350(5-6):313–318, 2012. 5

[9] C Fabian Benitez-Quiroz, Ramprakash Srinivasan, and Aleix M Martinez. Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5562–5570, 2016. 8

[10] Pratyush Garg, John Villasenor, and Virginia Foggo. Fairness metrics: A comparative analysis. In *2020 IEEE international conference on big data (Big Data)*, pages 3662–3666. IEEE, 2020. 3

[11] Byungju Kim, Hyunwoo Kim, Kyungsu Kim, Sungjin Kim, and Junmo Kim. Learning not to learn: Training deep neural networks with biased data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9012–9020, 2019. 8

[12] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International conference on machine learning*, pages 5637–5664. PMLR, 2021. 8

[13] Xi Lin, Hui-Ling Zhen, Zhenhua Li, Qing-Fu Zhang, and Sam Kwong. Pareto multi-task learning. *Advances in neural information processing systems*, 32, 2019. 2

[14] 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)*, 2015. 8

[15] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. In *International Conference on Machine Learning*, pages 3384–3393. PMLR, 2018. 1

[16] Debabrata Mahapatra and Vaibhav Rajan. Multi-task learning with user preferences: Gradient descent with controlled ascent in pareto optimization. In *International Conference on Machine Learning*, pages 6597–6607. PMLR, 2020. 2

[17] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54 (6):1–35, 2021. 1, 4

[18] Aditya Krishna Menon and Robert C Williamson. The cost of fairness in binary classification. In *Conference on Fairness, accountability and transparency*, pages 107–118. PMLR, 2018. 1

[19] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, 2017. 8

[20] Dana Pessach and Erez Shmueli. A review on fairness in machine learning. *ACM Computing Surveys (CSUR)*, 55(3): 1–44, 2022. 1

[21] Novi Quadrianto, Viktoriia Sharmanska, and Oliver Thomas. Discovering fair representations in the data domain. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8227–8236, 2019. 1

[22] Kazuyuki Sekitani and Yoshitsugu Yamamoto. A recursive algorithm for finding the minimum norm point in a polytope and a pair of closest points in two polytopes. *Mathematical Programming*, 61(1):233–249, 1993. 6

[23] Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. *Advances in neural information processing systems*, 31, 2018. 2, 5, 6, 8

[24] Ezekiel Soremekun, Mike Papadakis, Maxime Cordy, and Yves Le Traon. Software fairness: An analysis and survey. *ACM Computing Surveys*, 2022. 1

[25] Zeyu Tang and Kun Zhang. Attainability and optimality: The equalized odds fairness revisited. In *Conference on Causal Learning and Reasoning*, pages 754–786. PMLR, 2022. 3

[26] Shaokui Wei, Jiayin Liu, Bing Li, and Hongyuan Zha. Mean parity fair regression in rkhs. In *International conference on artificial intelligence and statistics*, pages 4602–4628. PMLR, 2023. 4

[27] Philip Wolfe. Finding the nearest point in a polytope. *Mathematical Programming*, 11(1):128–149, 1976. 6

[28] Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. *Advances in neural information processing systems*, 33:5824–5836, 2020. 2

[29] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P Gummadi. Fairness constraints: Mechanisms for fair classification. In *Artificial intelligence and statistics*, pages 962–970. PMLR, 2017. 1

[30] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P Gummadi. Fairness constraints: A flexible approach for fair classification. *Journal of Machine Learning Research*, 20(75):1–42, 2019. 1

[31] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340, 2018. 1

[32] Meiyu Zhong and Ravi Tandon. Intrinsic fairness-accuracy tradeoffs under equalized odds. In *2024 IEEE International Symposium on Information Theory (ISIT)*, pages 220–225. IEEE, 2024. 3