FairPO: Robust Preference Optimization for Fair Multi-Label Learning

Soumen Kumar Mondal[‡], **Prateek Chanda**[‡], **Akshit Varmora**[‡], **Ganesh Ramakrishnan**[‡]

[‡] Indian Institute of Technology Bombay, India
{soumenkm, prateekc, akshitv, ganeshr}@iitb.ac.in

December 15, 2025

Motivation: Why Do We Need Fair Multi-Label Learning?

MLC is Ubiquitous, but Standard Training Can Be Unfair:

Real-World Impact: MLC models deployed in sensitive areas: Content Tagging (News, Social Media), Medical Diagnosis Assistance (Symptoms $\rightarrow$ Conditions), Skill Identification (Resumes $\rightarrow$ Job Skills).

The Risk of Bias: Models optimizing aggregate metrics (e.g., average F1, mAP) can:

- Perform poorly on rare but important labels often termed as privileged labels ($\mathcal{P}$) whereas performs well on frequent labels or non privileged labels ($\bar{\mathcal{P}}$).
- Exhibit large performance gaps: $perf(\mathcal{P}) \ll perf(\bar{\mathcal{P}})$.
- Example: A medical system consistently missing rare disease labels ($\in \mathcal{P}$).

Limitations of Current Methods:

- Standard losses (e.g., BCE) are fairness-agnostic.
- Simple re-weighting might not robustly balance groups or handle confusing labels effectively.

Why FairPO?

FairPO is needed to **explicitly** address these MLC fairness challenges by:

- Directly targeting performance disparities between label groups ($\mathcal{P}, \bar{\mathcal{P}}$).
- Employing robust optimization for principled balancing, ensuring fairness isn't sacrificed for overall gains.

FairPO: Robust Preference Optimisation for Fair Multilabel Learning

FairPO Objective & Fairness Goals

- We consider a multi-label classification problem with T possible labels, $\mathcal{T} = \{1, 2, \ldots, T\}$. We have a dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$, where x_i is an input and $y_i = [y_{i1}, \ldots, y_{iT}]$ is the ground truth label vector, with $y_{il} = +1$ if label l applies to x_i, and $y_{il} = 0$ otherwise.
- Given instance x_i, labels $\mathcal{T}$. Partition $\mathcal{T}$ into Privileged $\mathcal{P}$ and Non-Privileged $\bar{\mathcal{P}}$. Model score for label t: $m(x_i; \mathbf{w}_t)$. Reference score: $m(x_i; \hat{\mathbf{w}}_t)$.
- Confusing Sets for Privileged Label $l \in \mathcal{P}$:
  - If $y_{il} = +1$: Set of *confusing negatives*, $S_{il}^{\text{neg}} = \{k \in \mathcal{T} \mid y_{ik} = 0 \text{ and } m(x_i; \mathbf{w}_k) \geq m(x_i; \mathbf{w}_l)\}$
  - If $y_{il} = 0$: Set of *confusing positives*, $S_{il}^{\text{pos}} = \{k' \in \mathcal{T} \mid y_{ik'} = +1 \text{ and } m(x_i; \mathbf{w}_{k'}) \leq m(x_i; \mathbf{w}_l)\}$

  Overall confusing set: $S_{il} = S_{il}^{\text{neg}} \cup S_{il}^{\text{pos}}$.
- Fairness Goals:
  - Privileged ($\mathcal{P}$): If $S_{il} \neq \emptyset$, encourage $m(x_i; \mathbf{w}_l) \gg m(x_i; \mathbf{w}_k)$ for $k \in S_{il}^{\text{neg}}$, and $m(x_i; \mathbf{w}_l) \ll m(x_i; \mathbf{w}_{k'})$ for $k' \in S_{il}^{\text{pos}}$. Else, use standard BCE.
  - Non-Privileged ($\bar{\mathcal{P}}$): Maintain performance: $\ell(\mathbf{w}_j) \leq \ell(\hat{\mathbf{w}}_j) + \epsilon$.

Preliminaries: DPO and GRPO

DPO: Direct Preference Optimisation

DPO directly optimizes a policy π_θ using preference pairs (x, y_w, y_l), where y_w is preferred over y_l for prompt x. DPO maximizes the likelihood of preferring winning response over losing response, resulting in the loss:

$$h_{\pi_\theta}(x, y_w, y_l) = \beta \log \frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta \log \frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}, \tag{1}$$

$$L_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x,y_w,y_l)\sim D} \left[\log \sigma \left(h_{\pi_\theta}(x, y_w, y_l)\right)\right]. \tag{2}$$

GRPO: Group Robust Preference Optimisation

GRPO extends preference optimization to handle diverse preference groups $\{D_g\}_{g=1}^K$. Instead of minimizing the average loss, GRPO minimizes the worst-case loss across groups using a robust objective:

$$\min_{\pi_\theta} \max_{\alpha \in \Delta_{K-1}} \sum_{g=1}^{K} \alpha_g L_{\text{pref}}(\pi_\theta; \pi_{\text{ref}}, D_g), \tag{3}$$

where L_{pref} is a base preference loss (like L_{DPO}), and $\alpha = (\alpha_1, ..., \alpha_K)$ are adaptive weights in the probability simplex Δ_{K-1} which is optimized by Mirror Descent algorithm.

Preliminaries: SimPO and CPO

SimPO: Simple Preference Optimisation

SimPO aims to align the implicit reward with generation metrics and eliminates the need for π_{ref}. It uses the length-normalized average log-likelihood as the reward: $r_{\text{SimPO}}(x, y) = \frac{\beta}{|y|} \log \pi_\theta(y|x)$. It also introduces a target margin $\gamma > 0$ into the preference model. The resulting SimPO loss is:

$$L_{\text{SimPO}}(\pi_\theta) = -\mathbb{E}_{(x,y_w,y_l)\sim D}\left[\log \sigma\left(\frac{\beta}{|y_w|}\log \pi_\theta(y_w|x) - \frac{\beta}{|y_l|}\log \pi_\theta(y_l|x) - \gamma\right)\right]. \tag{4}$$

CPO: Contrastive Preference Optimisation

CPO also removes the dependency on π_{ref} for efficiency, approximating the DPO objective. It combines a reference-free preference loss with a negative log-likelihood (NLL) regularizer on preferred responses y_w to maintain generation quality:

$$L_{\text{prefer}}(\pi_\theta) = -\mathbb{E}_{(x,y_w,y_l)\sim D}\left[\log \sigma\left(\beta \log \pi_\theta(y_w|x) - \beta \log \pi_\theta(y_l|x)\right)\right] \tag{5}$$

$$L_{\text{NLL}}(\pi_\theta) = -\mathbb{E}_{(x,y_w)\sim D}[\log \pi_\theta(y_w|x)] \tag{6}$$

$$L_{\text{CPO}}(\pi_\theta) = L_{\text{prefer}} + L_{\text{NLL}}. \tag{7}$$

This formulation avoids the computational cost of the reference model pass.

FairPO: Preference-based Privileged Loss

Motivation behind the Privileged Loss: Addressing Hard Examples in Privileged Labels ($\mathcal{P}$)

- For privileged labels $l \in \mathcal{P}$, achieving high accuracy and robust discriminative power is paramount. Standard classification losses (like BCE) may not provide sufficient signal for "hard" examples.
- **Key Challenge 1: True Positives vs. Confusing Negatives.** When a label l is truly positive ($y_{il} = +1$), the model might still assign a low score $m(x_i; \mathbf{w}_l)$, or critically, assign a higher score to an incorrect label $k \in S_{il}^{\text{neg}}$ (a confusing negative). We need to ensure $m(x_i; \mathbf{w}_l) \gg m(x_i; \mathbf{w}_k)$.
- **Key Challenge 2: True Negatives vs. Confusing Positives.** Conversely, if label l is truly negative ($y_{il} = 0$), the model might erroneously assign it a high score, or more critically, assign a lower score than an actual positive label $k' \in S_{il}^{\text{pos}}$ (a confusing positive). We need to ensure $m(x_i; \mathbf{w}_l) \ll m(x_i; \mathbf{w}_{k'})$.

FairPO: Preference-based Privileged Loss [contd.]

Motivation behind the Privileged Loss: Addressing Hard Examples in Privileged Labels ($\mathcal{P}$) [contd.]

- **Preference Learning Framework:** Inspired by DPO, we frame these challenges as learning explicit preferences:
  - For $y_{il} = +1$ and $k \in S_{il}^{\text{neg}}$: Prefer true label l over confusing negative k ($l \succ k$).
  - For $y_{il} = 0$ and $k' \in S_{il}^{\text{pos}}$: Prefer true negative l over confusing positive k' (conceptually, the absence of l is preferred over the misattribution of k', or $l \prec k'$ in terms of scores).
- **Quantifying Preference (DPO-inspired):** We use a comparison term $h_{\mathbf{w}}(x_i, \text{preferred}, \text{dispreferred})$ which measures how much better the current model $\mathbf{w}$ separates the preferred from the dispreferred, relative to a reference $\hat{\mathbf{w}}$.
  - If $y_{il} = +1$, preferred is l, dispreferred is $k \in S_{il}^{\text{neg}}$.
  - If $y_{il} = 0$, preferred is l (target: low score), dispreferred is $k' \in S_{il}^{\text{pos}}$ (target: high score relative to l). The DPO formulation handles this by aiming for $m(x_i; w_l) \ll m(x_i; w_{k'})$.
- **Fallback:** If no such confusing examples exist for (x_i, l) (i.e., $S_{il} = \emptyset$), a standard BCE loss for label l is applied to ensure consistent learning.

FairPO: Preference-based Privileged Loss [contd.]

Privileged Loss

$$h_{\mathbf{w}}(x_i, l, k) = \left(\log \frac{m(x_i; \mathbf{w}_l)}{m(x_i; \hat{\mathbf{w}}_l)}\right) - \left(\log \frac{m(x_i; \mathbf{w}_k)}{m(x_i; \hat{\mathbf{w}}_k)}\right). \tag{8}$$

Our goal is to maximize this relative separation. We achieve this by minimizing the negative log-likelihood of the preference $l \succ k$, formulated as the **Privileged Loss** $\mathcal{L}_{\mathcal{P}}$:

$$\mathcal{L}_{\mathcal{P}}(\{\mathbf{w}_t, \hat{\mathbf{w}}_t | \exists t \in \mathcal{T}\}) = \mathbb{E}_{(x_i, l, k) \text{ s.t. } l \in \mathcal{P}, y_{il} = +1, k \in S_{il}} \left[-\log \sigma\left(\beta \cdot h_{\mathbf{w}}(x_i, l, k)\right)\right]. \tag{9}$$

Minimizing $\mathcal{L}_{\mathcal{P}}$ encourages $h_{\mathbf{w}}(x_i, l, k)$ to be large and positive, thereby pushing the model $\mathbf{w}$ to assign significantly higher scores to true positive privileged labels compared to their confusing negative, relative to the reference model.

Practical Implementation

- In a practical implementation using a pretrained model π_θ (like an LLM), $\pi_\theta(x_i)$ can produce an input representation $\mathbf{z}_i \in \mathbb{R}^d$ (e.g., an embedding, where d is the embedding dimension).
- A classifier head (e.g., a FFNN), can then operate on $\mathbf{z}_i$. In this context, $\mathbf{w}_t$ represents the parameters (e.g., a weight vector) within this classifier head responsible for producing the score $m(x_i; \mathbf{w}_t)$ for label t.
- The reference parameters $\hat{\mathbf{w}}$ can represent a baseline, obtained by SFT using MLC objective.

FairPO: Margin-based Non-Privileged Loss

Non-Privileged Loss

- For the non-privileged labels $j \in \bar{\mathcal{P}}$, the primary goal is fairness through maintaining baseline performance.
- We aim to prevent the model's focus on privileged labels from causing significant degradation in performance on these non-privileged labels compared to the reference model $\hat{\mathbf{w}}_j$.
- We enforce this using a constraint-based loss. We first define a standard base classification loss, specifically binary cross-entropy (BCE), for a label j and instance x_i:

$$loss(\mathbf{w}_j) = -y_{ij} \log(m(x_i; \mathbf{w}_j)) - (1 - y_{ij}) \log(1 - m(x_i; \mathbf{w}_j)), \tag{10}$$

$$loss(\hat{\mathbf{w}}_j) = -y_{ij} \log(m(x_i; \hat{\mathbf{w}}_j)) - (1 - y_{ij}) \log(1 - m(x_i; \hat{\mathbf{w}}_j)). \tag{11}$$

- The **Non-Privileged Loss** $\mathcal{L}_{\bar{\mathcal{P}}}$ uses a hinge mechanism to penalize the model only when the performance degradation exceeds a predefined slack margin $\epsilon \geq 0$:

$$\mathcal{L}_{\bar{\mathcal{P}}}(\{\mathbf{w}_t, \hat{\mathbf{w}}_t | \exists t \in \mathcal{T}\}) = \mathbb{E}_{(x_i, \{y_{ij} | j \in \bar{\mathcal{P}}\}) \text{ s.t. } y_{ij} \in \{+1, 0\}} \left[L_h(\mathbf{w}_j, \hat{\mathbf{w}}_j; \epsilon) \right], \tag{12}$$

$$L_h(\mathbf{w}_j, \hat{\mathbf{w}}_j; \epsilon) = \max\left(0,\, loss(\mathbf{w}_j) - loss(\hat{\mathbf{w}}_j) - \epsilon\right). \tag{13}$$

- This ensures that the model parameters $\mathbf{w}$ are only adjusted for non-privileged labels if their performance drops substantially below the reference performance plus the allowed slack ϵ.

FairPO: Learning Algorithm

Algorithm 2 FairPO Algorithm for Multi-Label Classification (DPO-inspired)

Initialize: model parameters $\{\mathbf{w}_t^{(0)} \in \mathbb{R}^d | \forall t \in \mathcal{T}\}$ (e.g., copy $\{\hat{\mathbf{w}}_t | \forall t \in \mathcal{T}\}$), group weights $\alpha_{\mathcal{P}}^{(0)} \leftarrow 0.5, \alpha_{\bar{\mathcal{P}}}^{(0)} \leftarrow 0.5$.
Choose: learning rates $\eta_{\mathbf{w}}, \eta_\alpha$, DPO hyperparameter β, reference parameters $\{\hat{\mathbf{w}}_t | \forall t \in \mathcal{T}\}$, constraint slack ϵ.
for $s = 0$ to S (MaxIterations) **do**
  Sample an example: $(x_i, [y_{i1}, \ldots, y_{iT}]) \in \mathcal{D} \sim p_{\mathcal{D}}(.)$.
  Initialize group losses for this step: $\mathcal{L}_{\mathcal{P}}^{(s)} \leftarrow 0$, $\mathcal{L}_{\bar{\mathcal{P}}}^{(s)} \leftarrow 0$.
  Initialize gradients: $g_{\mathcal{P}}^t \leftarrow \vec{0}$, $g_{\bar{\mathcal{P}}}^t \leftarrow \vec{0}$ $\forall t \in \mathcal{T}$.
  Forward pass: $m(x_i; \mathbf{w}_t^{(s)}) \leftarrow \sigma(\mathbf{w}_t^{(s)^T} \mathbf{z}_i)$ where $\mathbf{z}_i \leftarrow \pi_\theta(x_i)$ $\forall t \in \mathcal{T}$.
  Sample a label: $r \in \mathcal{T} \sim Uniform(\frac{1}{|\mathcal{T}|})$.
  if $r \in \mathcal{P}$ **then** ▷ Handle privileged label
    $l \leftarrow r$
    $S_{il}^{\text{neg}} \leftarrow \emptyset$, $S_{il}^{\text{pos}} \leftarrow \emptyset$
    if $y_{il} = +1$ **then** ▷ True Positive case for privileged label l
      $S_{il}^{\text{neg}} \leftarrow \{k \in \mathcal{T} \mid y_{ik} = 0 \text{ and } m(x_i; \mathbf{w}_k^{(s)}) \geq m(x_i; \mathbf{w}_l^{(s)})\}$
      $S_{il} \leftarrow S_{il}^{\text{neg}}$
    else if $y_{il} = 0$ **then** ▷ True Negative case for privileged label l
      $S_{il}^{\text{pos}} \leftarrow \{k \in \mathcal{T} \mid y_{ik} = +1 \text{ and } m(x_i; \mathbf{w}_k^{(s)}) \leq m(x_i; \mathbf{w}_l^{(s)})\}$
      $S_{il} \leftarrow S_{il}^{\text{pos}}$
    end if
    if $S_{il} \neq \emptyset$ **then** ▷ Confusing examples exist, use DPO-inspired loss
      Sample $k \in S_{il} \sim Uniform(\frac{1}{|S_{il}|})$
      if $y_{il} = +1$ **then** ▷ DPO for True Positive l vs Confusing Negative k
        $h_{\mathbf{w}^{(s)}}(x_i, l, k) \leftarrow \left(\log \frac{m(x_i; \mathbf{w}_l^{(s)})}{m(x_i; \hat{\mathbf{w}}_l)}\right) - \left(\log \frac{m(x_i; \mathbf{w}_k^{(s)})}{m(x_i; \hat{\mathbf{w}}_k)}\right)$.
        $\mathcal{L}_{\text{pref}} \leftarrow -\log \sigma\left(\beta \cdot h_{\mathbf{w}^{(s)}}(x_i, l, k)\right)$
      else if $y_{il} = 0$ **then** ▷ DPO for True Negative l vs Confusing Positive k
        $h_{\mathbf{w}^{(s)}}(x_i, k, l) \leftarrow \left(\log \frac{m(x_i; \mathbf{w}_k^{(s)})}{m(x_i; \hat{\mathbf{w}}_k)}\right) - \left(\log \frac{m(x_i; \mathbf{w}_l^{(s)})}{m(x_i; \hat{\mathbf{w}}_l)}\right)$.
        $\mathcal{L}_{\text{pref}} \leftarrow -\log \sigma\left(\beta \cdot h_{\mathbf{w}^{(s)}}(x_i, k, l)\right)$
      end if
      $\mathcal{L}_{\mathcal{P}}^{(s)} \leftarrow \mathcal{L}_{\text{pref}}$
      $g_{\mathcal{P}}^t \leftarrow g_{\mathcal{P}}^t + \nabla_{\mathbf{w}_t} \mathcal{L}_{\text{pref}}|_{\mathbf{w}_t^{(s)}}$ $\forall t \in \mathcal{T}$.
    else ▷ No confusing examples, use BCE loss for privileged label l
      $\mathcal{L}_{\text{BCE}} \leftarrow -[y_{il} \log m(x_i; \mathbf{w}_l^{(s)}) + (1 - y_{il}) \log(1 - m(x_i; \mathbf{w}_l^{(s)}))]$
      $\mathcal{L}_{\mathcal{P}}^{(s)} \leftarrow \mathcal{L}_{\text{BCE}}$
      $g_{\mathcal{P}}^t \leftarrow g_{\mathcal{P}}^t + \nabla_{\mathbf{w}_t} \mathcal{L}_{\text{BCE}}|_{\mathbf{w}_t^{(s)}}$ $\forall t \in \mathcal{T}$.
    end if
  else if $r \in \bar{\mathcal{P}}$ **then** ▷ Handle non-privileged label
    $j \leftarrow r$
    $\ell(\mathbf{w}_j^{(s)}) \leftarrow -[y_{ij} \log(m(x_i; \mathbf{w}_j^{(s)})) + (1 - y_{ij}) \log(1 - m(x_i; \mathbf{w}_j^{(s)}))]$
    $\ell(\hat{\mathbf{w}}_j) \leftarrow -[y_{ij} \log(m(x_i; \hat{\mathbf{w}}_j)) + (1 - y_{ij}) \log(1 - m(x_i; \hat{\mathbf{w}}_j))]$
    $\mathcal{L}_{\bar{\mathcal{P}}}^{(s)} \leftarrow \max\left(0, \ell(\mathbf{w}_j^{(s)}) - \ell(\hat{\mathbf{w}}_j) - \epsilon\right)$
    $g_{\bar{\mathcal{P}}}^t \leftarrow g_{\bar{\mathcal{P}}}^t + \nabla_{\mathbf{w}_t} \mathcal{L}_{\bar{\mathcal{P}}}^{(s)}|_{\mathbf{w}_t^{(s)}}$ $\forall t \in \mathcal{T}$.
  end if
  $\alpha_{\mathcal{P}}^{(s+1)} \leftarrow \alpha_{\mathcal{P}}^{(s)} \exp(\eta_\alpha \mathcal{L}_{\mathcal{P},scaled}^{(s)})$ and $\alpha_{\bar{\mathcal{P}}}^{(s+1)} \leftarrow \alpha_{\bar{\mathcal{P}}}^{(s)} \exp(\eta_\alpha \mathcal{L}_{\bar{\mathcal{P}},scaled}^{(s)})$ ▷ Mirror ascent
  $Z \leftarrow \alpha_{\mathcal{P}}^{(s+1)} + \alpha_{\bar{\mathcal{P}}}^{(s+1)}$
  $\alpha_{\mathcal{P}}^{(s+1)} \leftarrow \frac{\alpha_{\mathcal{P}}^{(s+1)}}{Z}$ and $\alpha_{\bar{\mathcal{P}}}^{(s+1)} \leftarrow \frac{\alpha_{\bar{\mathcal{P}}}^{(s+1)}}{Z}$ ▷ Weight normalization
  $\mathbf{w}_t^{(s+1)} \leftarrow \mathbf{w}_t^{(s)} - \eta_{\mathbf{w}}(\alpha_{\mathcal{P}}^{(s+1)} g_{\mathcal{P}}^t + \alpha_{\bar{\mathcal{P}}}^{(s+1)} g_{\bar{\mathcal{P}}}^t)$ $\forall t \in \mathcal{T}$ ▷ Mirror descent
end for
return $\{\mathbf{w}_t^{(S)} | \forall t \in \mathcal{T}\}$

FairPO: SimPO and CPO Integration

SimPO-Inspired Privileged Loss

SimPO's core elements are its reference-free nature and the inclusion of a target margin γ. Adapting this to our setting involves comparing the model's confidence (represented by $m(x_i; \mathbf{w}_t)$) for the true positive label l against the confusing negative label k. The SimPO-inspired privileged loss, $\mathcal{L}_{\mathcal{P}}^{\text{SimPO}}$, is defined as:

$$\mathcal{L}_{\mathcal{P}}^{\text{SimPO}}(\{\mathbf{w}_t | \exists t \in \mathcal{T}\}) = \mathbb{E}_{\substack{(x_i, l, k) \text{ s.t.} \\ l \in \mathcal{P},\, y_{il}=+1, \\ k \in S_{il}}} \left[-\log \sigma \left(\beta \left(\log \frac{m(x_i; \mathbf{w}_l)}{m(x_i; \mathbf{w}_k)} \right) - \gamma \right) \right]. \tag{14}$$

Replacing $\mathcal{L}_{\mathcal{P}}$ with $\mathcal{L}_{\mathcal{P}}^{\text{SimPO}}$ in FairPO equation yields a FairPO variant leveraging the SimPO formulation for privileged labels.

FairPO: SimPO and CPO Integration [contd.]

CPO-Inspired Privileged Loss

CPO also offers a reference-free preference objective but includes a regularization term to prevent the model from deviating too far from a reasonable distribution. Adapting CPO involves two components replacing the single $\mathcal{L}_{\mathcal{P}}$ term.

1. **Preference Component ($L_{\mathcal{P}}^{\text{CPO-prefer}}$):** Similar to the SimPO adaptation but without the margin γ, this term compares the logits of the preferred label l and the confusing label k:

$$L_{\mathcal{P}}^{\text{CPO-prefer}}(\{\mathbf{w}_t|\exists t \in \mathcal{T}\}) = \mathbb{E}_{\substack{(x_i,\, l,\, k) \text{ s.t.} \\ l\in\mathcal{P},\, y_{il}=+1, \\ k\in S_{il}}} \left[-\log \sigma \left(\beta \left(\log \frac{m(x_i; \mathbf{w}_l)}{m(x_i; \mathbf{w}_k)} \right) \right) \right]. \tag{15}$$

2. **NLL Regularizer Component ($L_{\mathcal{P}}^{\text{CPO-NLL}}$):** CPO uses a negative log-likelihood loss on the preferred outputs to regularize. In our context, the "preferred output" corresponds to correctly classifying the true positive label $l \in \mathcal{P}$.

$$L_{\mathcal{P}}^{\text{CPO-NLL}}(\{\mathbf{w}_t|\exists t \in \mathcal{T}\}) = \mathbb{E}_{(x_i,l) \text{ s.t. } l\in\mathcal{P}, y_{il}=+1}[-\log m(x_i; \mathbf{w}_l)]. \tag{16}$$

The combined CPO-inspired privileged loss, $\mathcal{L}_{\mathcal{P}}^{\text{CPO}}$, is a weighted sum of these two components:

$$\mathcal{L}_{\mathcal{P}}^{\text{CPO}}(\{\mathbf{w}_t|\exists t \in \mathcal{T}\}) = L_{\mathcal{P}}^{\text{CPO-prefer}}(\{\mathbf{w}_t|\exists t \in \mathcal{T}\}) + \lambda L_{\mathcal{P}}^{\text{CPO-NLL}}(\{\mathbf{w}_t|\exists t \in \mathcal{T}\}), \tag{17}$$

FairPO: Multi-Attribute Generation

Core Idea of FairPO on Generation Task

- We adapt the FairPO framework to multi-attribute generation tasks, where the goal is to generate an output sequence y from a prompt x using a single generative policy $\pi_{\mathbf{w}}(y|x)$, such that the generation aligns with fairness goals defined over a set of attributes $\mathcal{A}$.
- Similar to the classification setting, we partition $\mathcal{A}$ into privileged $\mathcal{P}$ and non-privileged $\bar{\mathcal{P}}$ sets.
- The core GRPO minimax structure remains the same, but the group losses are defined based on the preference dataset $\mathcal{D}_{pref} = \{(x_i, y_{wi}, y_{li}, j_i)\}_{i=1}^{M}$ and the attributes driving the preferences.
- The primary difference lies in how the preference losses are applied, particularly for the privileged group, as DPO now operates on the log-probabilities of entire generated sequences rather than individual label scores.

FairPO: Multi-Attribute Generation [contd.]

Privileged Loss

- For generation tasks, the goal for privileged attributes $j \in \mathcal{P}$ is to ensure that the learned policy $\pi_{\mathbf{w}}$ strongly reflects observed preferences $y_w \succ y_l$ when the preference was established based on such an attribute.
- This translates to assigning significantly higher relative log-probability to the preferred sequence y_w compared to the dispreferred sequence y_l, relative to the reference policy π_{ref}.
- We achieve this using the standard DPO loss formulation, averaged over the subset of the preference data $\mathcal{D}_{pref}$ where the driving attribute j belongs to the privileged set $\mathcal{P}$. The privileged loss is thus defined as:

$$\mathcal{L}_{\mathcal{P}}(\pi_{\mathbf{w}}, \pi_{\text{ref}}) = \mathbb{E}_{(x,y_w,y_l,j)\sim\mathcal{D}_{pref}|j\in\mathcal{P}}\left[-\log\sigma\left(\beta \cdot h_{\pi_{\mathbf{w}}}(x, y_w, y_l)\right)\right] \tag{18}$$

Non-Privileged Loss

The objective for non-privileged attributes $k \in \bar{\mathcal{P}}$ remains analogous to the classification setting: prevent significant degradation compared to a baseline. The non-privileged loss uses the hinge formulation based on the DPO loss for preferences driven by attributes $k \in \bar{\mathcal{P}}$:

$$\mathcal{L}_{\bar{\mathcal{P}}}(\pi_{\mathbf{w}}, \pi_{\text{ref}}) = \mathbb{E}_{(x,y_w,y_l,k)\sim\mathcal{D}_{pref}|k\in\bar{\mathcal{P}}}\left[\max\left(0, \mathcal{L}_{DPO}(\pi_{\mathbf{w}}, \pi_{\text{ref}}; x, y_w, y_l) - (\log 2) - \epsilon'\right)\right]. \tag{19}$$

FairPO: Experimental Setup

Datasets and Preprocessing

- **MS-COCO 2014:** 80 object categories. Used official train/val splits.
  - Privileged ($\mathcal{P}$): 16 labels (20% least frequent in training set).
  - Non-Privileged ($\bar{\mathcal{P}}$): Remaining 64 labels.
- **NUS-WIDE:** 81 concept labels. Used common split (161,789 train / 107,859 test).
  - Privileged ($\mathcal{P}$): 16 labels (approx. 20% least frequent in training set).
  - Non-Privileged ($\bar{\mathcal{P}}$): Remaining 65 labels.
- **Image Preprocessing:** Images resized to 224×224 pixels, normalized using ImageNet mean/std (consistent with ViT pretraining).

Model and Implementation Details

- **Base Model:** Vision Transformer (ViT) pretrained on ImageNet.
- **Fine-tuning:** FairPO framework, including classifier head and label-specific parameters $\mathbf{w}_t$, fine-tuned on top of frozen ViT features.
- **Reference Parameters ($\hat{\mathbf{w}}$):** Obtained by standard Supervised Fine-Tuning (SFT) of the same ViT architecture with BCE loss on the target MLC dataset.

FairPO: Experimental Setup [contd.]

Evaluation Metrics

- **mAP (Mean Average Precision):** Average precision across all labels, sensitive to ranking quality.
- **Sample F1 Score:** F1 computed per sample and averaged (Sample-F1). Reflects label prediction accuracy.
- **Accuracy:** Element-wise accuracy across all labels and instances (proportion of correct y_{it} predictions).
- **Exact Match Ratio:** Proportion of samples where the predicted label set exactly matches the ground truth set.

Baseline Experiments

1. **BCE-SFT:** The standard approach, fine-tuning the classifier head with Binary Cross-Entropy (BCE) loss, treating labels independently. This serves as our reference model ($\hat{\mathbf{w}}$).
2. **BCE-SFT + Re-Weighting (Privileged):** Same as BCE-SFT, but statically increasing the BCE loss weight for labels in the privileged group ($\mathcal{P}$) during training to encourage focus on them.
3. **Group DRO + BCE:** Applies the Group DRO dynamic weighting mechanism to balance the standard BCE loss calculated separately for group $\mathcal{P}$ and group $\bar{\mathcal{P}}$, aiming to minimize the maximum group loss.

FairPO: Experimental Setup [contd.]

Main Experiments

1. **FairPO-DPO:** Using the DPO-inspired preference loss (Eq. 9) for $\mathcal{P}$.
2. **FairPO-SimPO:** Using the SimPO-inspired reference-free loss (Eq. 14) for $\mathcal{P}$.
3. **FairPO-CPO:** Using the CPO-inspired reference-free loss with NLL regularization (Eq. 17) for $\mathcal{P}$.

FairPO: Main Results

Results on COCO Dataset

Table: Performance comparison on MS-COCO. $\mathcal{P}$ denotes the privileged label set (20% least frequent), and $\bar{\mathcal{P}}$ denotes the non-privileged set (remaining 80%). Best results for $\mathcal{P}$ metrics and $\Delta\text{mAP}(\mathcal{P})$ are in **bold**.

Method	mAP		Sample F1		Accuracy		EMR		$\Delta\text{mAP}(\mathcal{P})$
	$\mathcal{P}$	$\bar{\mathcal{P}}$	$\mathcal{P}$	$\bar{\mathcal{P}}$	$\mathcal{P}$	$\bar{\mathcal{P}}$	$\mathcal{P}$	$\bar{\mathcal{P}}$	
BCE SFT	86.32	**90.65**	61.43	**64.89**	94.89	**98.12**	35.78	**36.98**	Ref
BCE SFT + RW	87.25	89.85	62.68	64.11	95.95	97.93	47.43	33.81	+0.93
GDRO + BCE	87.92	90.41	62.31	64.75	95.72	98.05	46.12	35.16	+1.60
FairPO-DPO	88.02	89.97	63.45	63.65	97.89	98.04	62.19	35.12	+1.70
FairPO-SimPO	87.67	88.76	62.34	63.12	95.69	97.78	45.32	32.34	+1.35
FairPO-CPO	**89.76**	90.34	**64.01**	64.32	**98.03**	98.06	**65.43**	35.23	**+3.44**

FairPO: Main Results [contd.]

Results on NUS-WIDE Dataset

Table: Performance comparison on NUS-WIDE. $\mathcal{P}$ denotes the privileged label set (20% least frequent), and $\bar{\mathcal{P}}$ denotes the non-privileged set (remaining 80%). Best results for $\mathcal{P}$ metrics and $\Delta\text{mAP}(\mathcal{P})$ are in **bold**.

Method	mAP		Sample F1		Accuracy		EMR		$\Delta\text{mAP}(\mathcal{P})$
	$\mathcal{P}$	$\bar{\mathcal{P}}$	$\mathcal{P}$	$\bar{\mathcal{P}}$	$\mathcal{P}$	$\bar{\mathcal{P}}$	$\mathcal{P}$	$\bar{\mathcal{P}}$	
BCE SFT	63.53	**70.24**	48.12	**55.83**	91.51	**95.22**	19.32	**11.56**	Ref
BCE SFT + RW	65.12	69.14	49.51	54.73	92.33	94.81	21.23	10.32	+1.59
GDRO + BCE	64.84	69.91	49.13	55.62	92.11	95.13	21.02	11.34	+1.31
FairPO-DPO	66.34	69.05	51.71	54.52	93.92	95.04	27.91	11.21	+2.81
FairPO-SimPO	64.11	68.03	48.82	53.81	91.94	94.52	20.18	10.19	+0.58
FairPO-CPO	**67.12**	69.83	**52.21**	55.24	**94.31**	95.12	**31.87**	11.25	**+3.59**

FairPO: Ablation Studies

Ablation on Core Components

- **w/o Preference Loss for $\mathcal{P}$:** The preference loss for privileged labels ($\mathcal{L}_{\mathcal{P}}$) is replaced with standard BCE. *(Assesses criticality of preference signals for $\mathcal{P}$ labels.)*
- **w/o $\bar{\mathcal{P}}$ Constraint (using BCE for $\bar{\mathcal{P}}$):** The constrained loss for non-privileged labels ($\mathcal{L}_{\bar{\mathcal{P}}}$) is replaced with standard BCE. *(Evaluates the protective role of the constraint for $\bar{\mathcal{P}}$ labels.)*
- **w/o GRPO (Fixed 0.5/0.5 Weights):** The GRPO adaptive balancing is removed, and losses for $\mathcal{P}$ and $\bar{\mathcal{P}}$ groups are combined with fixed equal weights. *(Highlights the benefit of GRPO's adaptive balancing.)*

Ablation on Preference Formulation Details

- **Only Confusing Negatives for $\mathcal{P}$:** The preference learning for privileged labels ($y_{il} = +1$) only considers confusing negatives (S_{il}^{neg}), ignoring confusing positives when $y_{il} = 0$. *(Tests the impact of not addressing true negatives vs. confusing positives.)*
- **w/o BCE Fallback for $\mathcal{P}$ (No loss if $S_{il} = \emptyset$):** The standard BCE loss is omitted for privileged labels when no confusing examples (S_{il}) are identified. *(Assesses the importance of the fallback classification signal for model stability and learning on "easier" privileged instances.)*

FairPO: Ablation Results

Results

Table: Ablation on core components of FairPO-CPO (MS-COCO). ΔmAP($\mathcal{P}$) vs BCE SFT. Parentheses show change vs Full FairPO-CPO.

Method Variant	mAP		Sample F1		Accuracy		EMR		ΔmAP($\mathcal{P}$)
	$\mathcal{P}$	$\bar{\mathcal{P}}$	$\mathcal{P}$	$\bar{\mathcal{P}}$	$\mathcal{P}$	$\bar{\mathcal{P}}$	$\mathcal{P}$	$\bar{\mathcal{P}}$	
FairPO-CPO (Full)	**89.76**	**90.34**	**64.01**	**64.32**	**98.03**	**98.06**	**65.43**	**35.23**	**+3.44**
w/o Preference Loss ($\mathcal{L}_{\mathcal{P}}$ is BCE)	88.12 (-1.64)	90.45 (+0.11)	62.45 (-1.56)	64.80 (+0.48)	95.80 (-2.23)	98.09 (+0.03)	48.51 (-16.92)	35.30 (+0.07)	+1.80
w/o $\bar{\mathcal{P}}$ Constraint ($\mathcal{L}_{\bar{\mathcal{P}}}$ is BCE)	89.55 (-0.21)	88.98 (-1.36)	63.70 (-0.31)	62.95 (-1.37)	97.90 (-0.13)	97.55 (-0.51)	63.12 (-2.31)	31.95 (-3.28)	+3.23
w/o GRPO (Fixed 0.5/0.5 weights)	88.48 (-1.28)	89.75 (-0.59)	62.88 (-1.13)	63.50 (-0.82)	96.53 (-1.50)	97.88 (-0.18)	56.70 (-8.73)	34.15 (-1.08)	+2.16

FairPO: Ablation Results [contd.]

Results

Table: Ablation on preference formulation (FairPO-CPO, MS-COCO). ΔmAP($\mathcal{P}$) vs BCE SFT (86.32). Parentheses show change vs Full FairPO-CPO.

Preference Detail Variant	mAP		Sample F1		Accuracy		EMR		ΔmAP($\mathcal{P}$)
	$\mathcal{P}$	$\bar{\mathcal{P}}$	$\mathcal{P}$	$\bar{\mathcal{P}}$	$\mathcal{P}$	$\bar{\mathcal{P}}$	$\mathcal{P}$	$\bar{\mathcal{P}}$	
FairPO-CPO (Full) (Conf. Neg & Pos, BCE Fallback)	**89.76**	**90.34**	**64.01**	**64.32**	**98.03**	**98.06**	**65.43**	**35.23**	**+3.44**
Only Confusing Negatives	73.15 (-16.61)	90.25 (-0.09)	47.88 (-16.13)	64.20 (-0.12)	94.67 (-3.36)	98.01 (-0.05)	22.54 (-42.89)	35.10 (-0.13)	-13.17
w/o BCE Fallback (No loss if $S_{il} = \emptyset$)	89.05 (-0.71)	90.21 (-0.13)	63.20 (-0.81)	64.10 (-0.22)	97.55 (-0.48)	97.99 (-0.07)	60.75 (-4.68)	34.90 (-0.33)	+2.73

FairPO: Conclusions and Future Works

Conclusions

- Proposed **FairPO**, a novel framework to enhance fairness in Multi-Label Classification (MLC).
- Leverages *preference optimization* (initially DPO-inspired) to better distinguish true labels from confusing labels for a *privileged* label set ($\mathcal{P}$).
- Employs a *constrained objective* to maintain performance on a *non-privileged* set ($\bar{\mathcal{P}}$) relative to a reference model.
- Integrates *Group Robust Optimization (GRPO)* to dynamically balance the objectives for $\mathcal{P}$ and $\bar{\mathcal{P}}$, promoting robust fairness.
- Provides a principled approach to mitigate performance disparities across label groups in MLC.

Future Works

- **Multi-Label Generation:** Extend FairPO to generate fair and coherent sets of labels/attributes, particularly for ambiguous inputs, leveraging underlying sequence generation capabilities (Appendix A).
- **Partition Sensitivity:** Analyze the impact of different strategies for partitioning labels into $\mathcal{P}$ and $\bar{\mathcal{P}}$.
- **Hyperparameter Tuning:** Investigate the influence of FairPO-specific hyperparameters ($\beta, \epsilon, \eta_\alpha$).

References

- Paul F Christiano, Jan Leike, Tom B Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, volume 30, 2017.
- Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.
- Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. In *Advances in Neural Information Processing Systems*, volume 36, 2023.
- Shyam Sundhar Ramesh, Yifan Hu, Iason Chaimalas, Viraj Mehta, Pier Giuseppe Sessa, Haitham Bou Ammar, and Ilija Bogunovic. Group Robust Preference Optimization in Reward-free RLHF. In *Advances in Neural Information Processing Systems*, volume 37, 2024.
- Mengzhou Xia, Yu Meng, and Danqi Chen. SimPO: Simple Preference Optimization with a Reference-Free Reward. In *Advances in Neural Information Processing Systems*, volume 37, 2024.
- Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. Contrastive Preference Optimization: Pushing the Boundaries of LLM Performance in Machine Translation. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pages 55480–55514. PMLR, 2024.

Questions?

Thank You!