FairPO: Fair Preference Optimization for Multi-Label Learning

Soumen Kumar Mondal   SOUMENKM@IITB.AC.IN

Prateek Chanda[†]   PRATEEK.CHANDA@IITB.AC.IN

Akshit Varmora[†]   AKSHIT.VARMORA@IITB.AC.IN

Ganesh Ramakrishnan   GANRAMK@IITB.AC.IN

IIT Bombay, India

Abstract

Multi-label classification (MLC) often suffers from performance disparities across labels. We propose **FairPO**, a framework combining preference-based loss and group-robust optimization to improve fairness by targeting underperforming labels. FairPO partitions labels into a *privileged* set for targeted improvement and a *non-privileged* set to maintain baseline performance. For privileged labels, a DPO-inspired preference loss addresses hard examples by correcting ranking errors between true labels and their confusing counterparts. A constrained objective maintains performance for non-privileged labels, while a Group Robust Preference Optimization (GRPO) formulation adaptively balances both objectives to mitigate bias. We also demonstrate FairPO's versatility with reference-free variants using Contrastive (CPO) and Simple (SimPO) Preference Optimization[1].

1. Introduction

The Challenge of Performance Disparity in MLC. Standard multi-label classification (MLC) models assume all labels are equal, minimizing an aggregate loss over the entire label set [27, 31, 34]. This assumption fails in real-world scenarios where data asymmetries cause significant performance disparities. These disparities stem from several factors, including **class imbalance**, where models favor frequent labels over rare ones [1, 35], and varying **real-world importance**, such as distinguishing a critical disease from a benign one [21, 28]. Disparities also arise from **semantic complexity**, where models learn simple concepts but fail on nuanced ones. Consequently, training becomes dominated by the easy majority, yielding systems that are unreliable for the very cases that matter most [19, 29].

Limitations of Conventional Methods. Conventional methods are limited by the aggregate nature of the Binary Cross-Entropy (BCE) loss, which allows the learning signal for underrepresented labels to be drowned out by easy ones [1, 24]. While solutions like Focal Loss help, they are fundamentally restricted because they treat each label's prediction in isolation [11]. They do not teach the model to perform *discrimination* by resolving confusion between a true label and a specific negative alternative. This reveals a gap for a framework that can learn *relative preferences*, improving performance on difficult labels while maintaining it for others. A preference-based objective forces the model to learn that a true label's score is explicitly *preferred* over that of a confusing competitor.

---

0. [†] Equal contribution

1. Our code is available at GitHub: https://github.com/soumenkm/FairPO

Our Approach: The FairPO Framework. We introduce **FairPO** (Fair Preference Optimization), a framework to manage performance disparities by partitioning labels into a *privileged* group for targeted improvement and a *non-privileged* group for performance maintenance. We prioritize learning on the *privileged* group using a preference-based objective, a novel adaptation of techniques from generative model alignment for a discriminative task. For each privileged true label, FairPO dynamically identifies its *confusing counterparts*, incorrect labels with misleadingly high scores and trains the model to prefer the true label, creating a large discriminative margin. Concurrently, a constrained objective safeguards the *non-privileged* group against performance degradation relative to a reference model. These competing objectives are dynamically balanced using a Group Robust Preference Optimization (GRPO) formulation [22].

Application and Contributions. Our general FairPO framework can address any MLC task requiring label prioritization due to factors like class imbalance or real-world importance. To demonstrate its effectiveness, we apply it to **class imbalance**, defining infrequent tail labels as the *privileged* group and frequent head labels as the *non-privileged* one. Our main contributions are:

1. We propose **FairPO**, a novel framework for targeted performance improvement in MLC applicable to various sources of disparity.

2. We are the first to adapt preference optimization techniques [17, 32] from generative modeling to create discriminative objectives for MLC.

3. We successfully apply GRPO to solve the trade-off between objectives for the *privileged* and *non-privileged* groups, a novel application in this context.

Experiments show FairPO significantly boosts performance on infrequent labels by up to 3.59% mAP while robustly maintaining performance on frequent ones, outperforming several baselines.

Figure 1: FairPO Framework

2. Methodology: The FairPO Framework

Our methodology builds upon a foundation of preference optimization techniques. To provide the necessary background on these methods, namely DPO, CPO, SimPO, and GRPO, we have moved the detailed review of their core formulations to Appendix A.

2.1. Problem Setup and Fairness Goals

Let us consider a standard multi-label classification setting with a dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$, where $\mathbf{x}_i$ is an input instance and $y_i \in \{0, 1\}^{|\mathcal{T}|}$ is the corresponding multi-label vector over a universe of $|\mathcal{T}|$ labels. The goal is to train a model, parameterized by weights $\mathbf{w}$, that produces a set of per-label scores $m(\mathbf{x}_i; \mathbf{w}_t)$ for each label $t \in \mathcal{T}$. Our framework also utilizes a pre-trained reference model with parameters $\hat{\mathbf{w}}_t$.
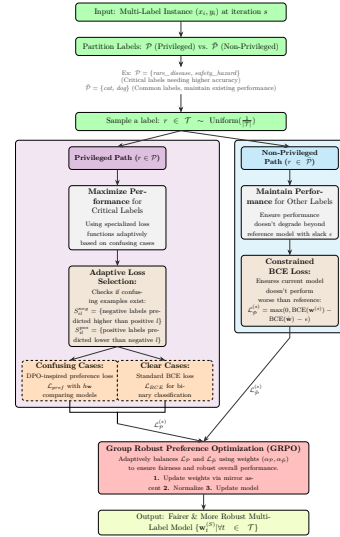
Conventional methods often train such models by minimizing a single, aggregate loss where each label contributes equally. However, as motivated in our introduction, real-world applications frequently demand a more nuanced approach where certain labels are prioritized. Our framework, FairPO, is explicitly designed for these scenarios. The core idea is to partition the total label set $\mathcal{T}$ into two disjoint subsets:

- A *privileged* **group** $\mathcal{P} \subset \mathcal{T}$, which contains labels for which we seek to significantly enhance the model's performance.

- A *non-privileged* **group** $\bar{\mathcal{P}} = \mathcal{T} \setminus \mathcal{P}$, which contains the remaining labels, for which our objective is to robustly maintain at least a baseline level of performance.

This partitioning is a key strength of our framework's general design. The criteria for assigning labels to the *privileged* group are flexible and can be adapted to the specific problem at hand, such as using label frequency for class imbalance, domain knowledge for real-world importance, or annotation consistency for data quality issues. In this work, to demonstrate FairPO's efficacy, we focus on the class imbalance problem, where the *privileged* group consists of the least frequent labels. We defer the precise details of this setup to Section 3.

2.2. FairPO Objectives

FairPO's methodology is built on three core components: a conditional objective for privileged labels, a constrained objective for non-privileged labels, and a robust optimization framework to balance them. The detailed mathematical formulations for each component are provided in Appendix B.

Objective for Privileged Labels ($l \in \mathcal{P}$)**:** To improve performance on critical labels, we employ a conditional objective designed to target hard-to-discriminate cases. When a *confusing set* of labels exists for an instance (i.e., high-scoring negatives or low-scoring positives), a preference loss inspired by DPO [20] is applied to enforce correct relative rankings. In the absence of such confusing examples, the objective reverts to a standard BCE loss, which acts as a crucial fallback for stable training. Our framework is versatile and also supports reference-free preference loss variants inspired by CPO [32] and SimPO [17].

Objective for Non-Privileged Labels ($j \in \bar{\mathcal{P}}$)**:** To maintain baseline performance on the remaining labels, we use a constrained objective. This objective employs a hinge mechanism that only incurs a penalty if the model's performance on a label drops significantly below that of a reference model by a predefined slack margin ϵ. This acts as a protective measure, preventing substantial performance degradation for the non-privileged group.

Group Robust Optimization: These two distinct objectives for the privileged and non-privileged groups are adaptively balanced using the Group Robust Preference Optimization (GRPO) framework [22]. The overall learning problem is formulated as the following minimax objective:

$$\min_{\{\mathbf{w}_t | t \in \mathcal{T}\}} \max_{\alpha_{\mathcal{P}} + \alpha_{\bar{\mathcal{P}}} = 1} \left[\alpha_{\mathcal{P}} \mathcal{L}_{\mathcal{P}}(\cdot) + \alpha_{\bar{\mathcal{P}}} \mathcal{L}_{\bar{\mathcal{P}}}(\cdot) \right]. \tag{1}$$

This formulation dynamically adjusts the training focus between the privileged and non-privileged groups, seeking a solution that is robust to the worst-case group loss and thereby managing the fairness-performance trade-off.

Algorithm 1 FairPO Training Overview (DPO-inspired)

1: **Initialize:** Model parameters $\{\mathbf{w}_t | t \in \mathcal{T}\}^{(0)}$ (*e.g.*, from supervised fine-tuning), set group weights $\alpha_{\mathcal{P}}^{(0)}, \alpha_{\bar{\mathcal{P}}}^{(0)}$

2: **Input:** Dataset $\mathcal{D}$, reference parameters $\{\hat{\mathbf{w}}_t | t \in \mathcal{T}\}$, hyperparameters $\beta, \epsilon, \eta_{\mathbf{w}}, \eta_{\alpha}$.

3: **For** each training iteration $s = 0, \dots, S - 1$:

4:     Sample instance $(x_i, y_i) \sim \mathcal{D}$ and a label $r \in \mathcal{T}$.

5:     **If** $r \in \mathcal{P}$:

6:         Compute privileged loss $\mathcal{L}_{\mathcal{P}}^{(s)}$ for (x_i, r) (DPO Eq. 13, or BCE Eq. 11).

7:     **Else if** $r \in \bar{\mathcal{P}}$:

8:         Compute non-privileged loss $\mathcal{L}_{\bar{\mathcal{P}}}^{(s)}$ for (x_i, r) (Eq. 16).

9:     Update group weights $\alpha^{(s+1)}$ via mirror ascent using $\mathcal{L}_{\mathcal{P}}^{(s)}, \mathcal{L}_{\bar{\mathcal{P}}}^{(s)}$ (GRPO step).

10:     Update model parameters $\{\mathbf{w}_t | t \in \mathcal{T}\}^{(s+1)}$ via mirror descent using weighted gradients.

11: **End For**

12: **Return** Optimized parameters $\{\mathbf{w}_t | t \in \mathcal{T}\}^{(S)}$.

    Full details are in Algorithm 2 (see Appendix E).

2.3. Optimization Algorithm

We solve the minimax objective in Eq. 17 iteratively using alternating mirror descent (Algorithm 1). A key component for stability is **loss scaling** during the update of the group weights α. Since the group losses, $\mathcal{L}_{\mathcal{P}}$ and $\mathcal{L}_{\bar{\mathcal{P}}}$, can have different scales and variances, using raw loss values for the mirror ascent step is unstable. To mitigate this, we update α based on the relative change of each group's loss from its running average, $\bar{\mathcal{L}}_g^{\text{avg}}$, ensuring a more stable optimization. Further details are in Appendix E.

3. Experimental Setup

We evaluate FairPO on two multi-label benchmarks: **MS-COCO 2014** [10] and **NUS-WIDE** [3]. To address class imbalance, we partition labels by frequency: the 20) and the remaining 80). While this partitioning is effective, the FairPO framework is general and could use other criteria like domain importance. We assess performance using mAP, Sample F1, and Accuracy, focusing on the mAP gain on the privileged set, , relative to the **BCE-SFT** baseline. We compare against baselines including **Group DRO + BCE** [25] and **Focal Loss** [11]. Our base model is a frozen Vision Transformer (ViT) [5] with separate non-linear MLP heads per label, trained with AdamW [15]. The trained BCE-SFT baseline provides the reference model parameters $\hat{\mathbf{w}}$. We test three framework variants: **FairPO-DPO**, **FairPO-SimPO**, and **FairPO-CPO**. Full details are in Appendix F.

4. Results and Analysis

As shown in Tables 1 and 2, FairPO-CPO consistently achieves the highest performance on the privileged group ($\mathcal{P}$), confirming its effectiveness. On MS-COCO, it delivers a statistically significant gain of **+3.44% ΔmAP($\mathcal{P}$)** over strong baselines like BCE-SFT. Crucially, this targeted improvement comes with only a minor and statistically insignificant performance drop on the non-privileged group ($\bar{\mathcal{P}}$), demonstrating FairPO's ability to reallocate model capacity without causing measurable harm.

Table 1: 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%). Results are mean $\pm$ std. dev. over 3 runs. The best result for each metric is in **bold**. ΔmAP is calculated relative to the strongest baseline for each respective group (Focal Loss for $\mathcal{P}$, BCE-SFT for $\bar{\mathcal{P}}$). [†]Indicates a statistically significant improvement (p < 0.05), while [‡]indicates the difference is not statistically significant (p > 0.1).

Method	mAP		Sample F1		Accuracy		Δ mAP ($\mathcal{P}$)	Δ mAP ($\bar{\mathcal{P}}$)
	$\mathcal{P}$	$\bar{\mathcal{P}}$	$\mathcal{P}$	$\bar{\mathcal{P}}$	$\mathcal{P}$	$\bar{\mathcal{P}}$	(vs. Focal)	(vs. BCE)
BCE-SFT [24]	$86.32_{\pm0.11}$	$\mathbf{90.65}_{\pm0.08}$	$61.43_{\pm0.15}$	$\mathbf{64.89}_{\pm0.12}$	$94.89_{\pm0.09}$	$\mathbf{98.12}_{\pm0.05}$	-2.03	Ref
GDRO + BCE [25]	$87.92_{\pm0.12}$	$90.41_{\pm0.09}$	$62.31_{\pm0.20}$	$64.75_{\pm0.13}$	$95.72_{\pm0.10}$	$98.05_{\pm0.05}$	-0.43	$-0.24^{\ddagger}$
Focal Loss [11]	$88.35_{\pm0.14}$	$89.81_{\pm0.12}$	$63.15_{\pm0.16}$	$64.18_{\pm0.15}$	$96.11_{\pm0.12}$	$97.90_{\pm0.07}$	Ref	-0.84
FairPO-DPO	$88.02_{\pm0.15}$	$89.97_{\pm0.11}$	$63.45_{\pm0.17}$	$63.65_{\pm0.16}$	$97.89_{\pm0.13}$	$98.04_{\pm0.06}$	-0.33	-0.68
FairPO-SimPO	$87.67_{\pm0.18}$	$88.76_{\pm0.21}$	$62.34_{\pm0.22}$	$63.12_{\pm0.19}$	$95.69_{\pm0.15}$	$97.78_{\pm0.09}$	-0.68	-1.89
FairPO-CPO	$\mathbf{89.76}_{\pm0.09}$	$90.34_{\pm0.07}$	$\mathbf{64.01}_{\pm0.13}$	$64.32_{\pm10}$	$\mathbf{98.03}_{\pm0.07}$	$98.06_{\pm0.05}$	$\mathbf{+1.41}^{\dagger}$	$\mathbf{-0.31}^{\ddagger}$

Table 2: Performance comparison on NUS-WIDE. Notations are similar to Table 1.

Method	mAP		Sample F1		Accuracy		Δ mAP ($\mathcal{P}$)	Δ mAP ($\bar{\mathcal{P}}$)
	$\mathcal{P}$	$\bar{\mathcal{P}}$	$\mathcal{P}$	$\bar{\mathcal{P}}$	$\mathcal{P}$	$\bar{\mathcal{P}}$	(vs. Focal)	(vs. BCE)
BCE SFT [24]	$63.53_{\pm0.15}$	$\mathbf{70.24}_{\pm0.11}$	$48.12_{\pm0.21}$	$\mathbf{55.83}_{\pm0.14}$	$91.51_{\pm0.12}$	$\mathbf{95.22}_{\pm0.08}$	-2.28	Ref
GDRO + BCE [25]	$64.84_{\pm0.16}$	$69.91_{\pm0.12}$	$49.13_{\pm0.22}$	$55.62_{\pm0.15}$	$92.11_{\pm0.13}$	$95.13_{\pm0.09}$	-0.97	-0.33
Focal Loss [11]	$65.81_{\pm0.17}$	$68.95_{\pm0.15}$	$50.33_{\pm0.20}$	$54.31_{\pm0.18}$	$93.05_{\pm0.11}$	$94.75_{\pm0.11}$	Ref	-1.29
FairPO-DPO	$66.34_{\pm0.20}$	$69.05_{\pm0.14}$	$51.71_{\pm0.25}$	$54.52_{\pm0.19}$	$93.92_{\pm0.16}$	$95.04_{\pm0.09}$	+0.53	-1.19
FairPO-SimPO	$64.11_{\pm0.22}$	$68.03_{\pm0.24}$	$48.82_{\pm0.28}$	$53.81_{\pm0.22}$	$91.94_{\pm0.18}$	$94.52_{\pm0.13}$	-1.70	-2.21
FairPO-CPO	$\mathbf{67.12}_{\pm0.14}$	$69.83_{\pm0.10}$	$\mathbf{52.21}_{\pm0.19}$	$55.24_{\pm0.13}$	$\mathbf{94.31}_{\pm0.10}$	$95.12_{\pm0.08}$	$\mathbf{+1.31}^{\dagger}$	$\mathbf{-0.41}^{\ddagger}$

The superiority of the CPO variant stems from its robust design: unlike DPO, it is reference-free, and its integrated BCE regularizer provides a crucial signal for absolute correctness that SimPO lacks. This dual objective of learning both relative preferences and absolute scores yields a more stable and effective training process, making FairPO-CPO the most practical choice.

5. Ablation Studies

Ablation studies on MS-COCO with FairPO-CPO confirm that each component is critical (Tables 3 & 4). Removing the preference loss, the non-privileged constraint, or the adaptive GRPO balancing all lead to significant performance degradation. A non-targeted *Global CPO* variant, while better than standard BCE, is substantially less effective than the full FairPO model, highlighting the necessity of our targeted approach. The design of the preference objective itself is also vital: restricting the preference signal to only *Confusing Negatives* or removing the *BCE Fallback* for non-confusing cases both degrade performance and stability. These results demonstrate that while preference optimization is powerful, its effectiveness hinges on the complete, balanced FairPO framework.

6. Related Work

Our work is situated at the intersection of three research areas, which we detail further in Appendix C. First, while recent efforts in **fair MLC** have addressed challenges such as tail labels [8] and subjective

Table 3: 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		ΔmAP$(\mathcal{P})$
	$\mathcal{P}$	$\bar{\mathcal{P}}$	$\mathcal{P}$	$\bar{\mathcal{P}}$	$\mathcal{P}$	$\bar{\mathcal{P}}$	
FairPO-CPO (Full)	**89.76**	**90.34**	**64.01**	**64.32**	**98.03**	**98.06**	**+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)	+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)	+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)	+2.16
Global CPO (on all labels) (No $\mathcal{P}/\bar{\mathcal{P}}$ split or GRPO)	88.55 (-1.21)	90.68 (+0.34)	62.75 (-1.26)	64.85 (+0.47)	96.95 (-1.08)	98.11 (+0.05)	+2.23

Table 4: Ablation on preference formulation (FairPO-CPO, MS-COCO). ΔmAP$(\mathcal{P})$ vs BCE SFT. Parentheses show change vs Full FairPO-CPO.

Preference Detail Variant	mAP		Sample F1		Accuracy		ΔmAP$(\mathcal{P})$
	$\mathcal{P}$	$\bar{\mathcal{P}}$	$\mathcal{P}$	$\bar{\mathcal{P}}$	$\mathcal{P}$	$\bar{\mathcal{P}}$	
FairPO-CPO (Full) (Conf. Neg & Pos, BCE Fallback)	**89.76**	**90.34**	**64.01**	**64.32**	**98.03**	**98.06**	**+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)	-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)	+2.73

fairness [12], FairPO contributes a novel approach by partitioning labels into privileged ($\mathcal{P}$) and non-privileged ($\bar{\mathcal{P}}$) sets and applying targeted, distinct objectives to each. Second, we adapt modern **preference optimization** techniques, originally developed for aligning LLMs like DPO [20], CPO [32], and SimPO [17]. Instead of ranking entire outputs, we repurpose these methods to resolve ambiguities between true labels and their dynamically identified *confusing* counterparts. Finally, to manage the trade-off between our objectives, we employ **Group Robust Optimization**. Inspired by Group DRO [26] and GRPO [22], we uniquely define the groups by our label partitions ($\mathcal{P}$ and $\bar{\mathcal{P}}$) and use GRPO's adaptive weighting to balance their custom-formulated losses.

7. Discussion

In conclusion, we introduced FairPO, a novel framework that effectively integrates preference optimization with group robustness to enhance fairness in MLC, with our CPO variant proving particularly effective. While promising, FairPO has limitations, including the potential instability of its dynamic 'confusing set', the DPO variant's reliance on a reference model, and the need for careful tuning of GRPO's balancing act. These challenges directly motivate our future work, which will focus on extending FairPO to attribute generation (Appendix D), performing broader empirical validation, exploring alternative label partitioning strategies, and developing theoretical insights into the framework's convergence.

References

[1] Francisco Charte, Antonio Jesús Rivera, María José del Jesús, and Francisco Herrera. Addressing imbalance in multilabel classification: Measures and random resampling algorithms. *Neurocomputing*, 163:3–16, 2015. URL https://api.semanticscholar.org/CorpusID:207107609.

[2] Paul Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences, 2023. URL https://arxiv.org/abs/1706.03741.

[3] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. Nuswide: a real-world web image database from national university of singapore, 2009. URL https://api.semanticscholar.org/CorpusID:6483070.

[4] Songlin Dong, Yuhang He, Zhengdong Zhou, Haoyu Luo, Xing Wei, Alex C. Kot, and Yihong Gong. Class-independent increment: An efficient approach for multi-label class-incremental learning, 2025. URL https://arxiv.org/abs/2503.00515.

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. URL https://arxiv.org/abs/2010.11929.

[6] Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. Kto: Model alignment as prospect theoretic optimization, 2024. URL https://arxiv.org/abs/2402.01306.

[7] Bram L. Gorissen, İhsan Yanıkoğlu, and Dick den Hertog. A practical guide to robust optimization, June 2015. ISSN 0305-0483. URL http://dx.doi.org/10.1016/j.omega.2014.12.006.

[8] Hao Guo and Song Wang. Long-tailed multi-label visual recognition by collaborative training on uniform and re-balanced samplings, 06 2021.

[9] Ke Ji, Jiahao Xu, Tian Liang, Qiuzhi Liu, Zhiwei He, Xingyu Chen, Xiaoyuan Liu, Zhijie Wang, Junying Chen, Benyou Wang, Zhaopeng Tu, Haitao Mi, and Dong Yu. The first few tokens are all you need: An efficient and effective unsupervised prefix fine-tuning method for reasoning models, 2025. URL https://rgdoi.net/10.13140/RG.2.2.33772.07043.

[10] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. URL https://arxiv.org/abs/1405.0312.

[11] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection, 2018. URL https://arxiv.org/abs/1708.02002.

[12] Tianci Liu, Haoyu Wang, Yaqing Wang, Xiaoqian Wang, Lu Su, and Jing Gao. Simfair: A unified framework for fairness-aware multi-label classification, 2023. URL https://arxiv.org/abs/2302.09683.

[13] Yongshuai Liu, Jiaxin Ding, and Xin Liu. Ipo: Interior-point policy optimization under constraints, 2019. URL https://arxiv.org/abs/1910.09615.

[14] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. URL https://arxiv.org/abs/1711.05101.

[15] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. URL https://arxiv.org/abs/1711.05101.

[16] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning, 2022. URL https://arxiv.org/abs/1908.09635.

[17] Yu Meng, Mengzhou Xia, and Danqi Chen. Simpo: Simple preference optimization with a reference-free reward, 2024. URL https://arxiv.org/abs/2405.14734.

[18] Long Ouyang and Others. Training language models to follow instructions with human feedback, 2022. URL https://arxiv.org/abs/2203.02155.

[19] Geoff Pleiss, Tianyi Zhang, Ethan R. Elenberg, and Kilian Q. Weinberger. Identifying mislabeled data using the area under the margin ranking, 2020. URL https://arxiv.org/abs/2001.10528.

[20] 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, 2024. URL https://arxiv.org/abs/2305.18290.

[21] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, Matthew P. Lungren, and Andrew Y. Ng. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning, 2017. URL https://arxiv.org/abs/1711.05225.

[22] 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, 2024. URL https://arxiv.org/abs/2405.20304.

[23] Leslie Rice, Anna Bair, Huan Zhang, and J. Zico Kolter. Robustness between the worst and average case, 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/ea4c796cccfc3899b5f9ae2874237c20-Paper.pdf.

[24] Usha Ruby and Vamsidhar Yendapalli. Binary cross entropy with deep learning technique for image classification, 10 2020.

[25] Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization, 2020. URL https://arxiv.org/abs/1911.08731.

[26] Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization, 2020. URL https://arxiv.org/abs/1911.08731.

[27] Leander Schietgat, Celine Vens, Jan Struyf, Hendrik Blockeel, Dragi Kocev, and Sašo Džeroski. Predicting gene function using hierarchical multi-label decision tree ensembles. *BMC Bioinformatics*, 11(1):2, 2010. ISSN 1471-2105. doi: 10.1186/1471-2105-11-2. URL https://doi.org/10.1186/1471-2105-11-2.

[28] Panakanti Shashanka and Tatireddy Subba Reddy. Dermatologist-level classification of skin cancer using cascaded ensembling of convolutional neural network. In *2023 International Conference on Research Methodologies in Knowledge Management, Artificial Intelligence and Telecommunication Engineering (RMKMATE)*, pages 1–5, 2023. doi: 10.1109/RMKMATE59243.2023.10368872.

[29] Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. Dataset cartography: Mapping and diagnosing datasets with training dynamics, 2020. URL https://arxiv.org/abs/2009.10795.

[30] Chakkrit Tantithamthavorn, Ahmed E. Hassan, and Kenichi Matsumoto. The impact of class rebalancing techniques on the performance and interpretation of defect prediction models, 2018. URL https://arxiv.org/abs/1801.10269.

[31] Jiang Wang, Yi Yang, Junhua Mao, Zhiheng Huang, Chang Huang, and Wei Xu. Cnn-rnn: A unified framework for multi-label image classification, 2016. URL https://arxiv.org/abs/1604.04573.

[32] 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, 2024. URL https://arxiv.org/abs/2401.08417.

[33] Shusheng Xu, Wei Fu, Jiaxuan Gao, Wenjie Ye, Weilin Liu, Zhiyu Mei, Guangju Wang, Chao Yu, and Yi Wu. Is dpo superior to ppo for llm alignment? a comprehensive study, 2024. URL https://arxiv.org/abs/2404.10719.

[34] Alessandro Zangari, Matteo Marcuzzo, Matteo Rizzo, Lorenzo Giudice, Andrea Albarelli, and Andrea Gasparetto. Hierarchical text classification and its foundations: A review of current research, 2024. ISSN 2079-9292. URL https://www.mdpi.com/2079-9292/13/7/1199.

[35] Min-Ling Zhang and Zhi-Hua Zhou. A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 26(8):1819–1837, 2014. doi: 10.1109/TKDE.2013.39.

Appendix A. Preliminaries: Preference Optimization Methods

Our framework builds upon recent preference optimization techniques. We briefly review the key formulations.

Direct Preference Optimization (DPO): DPO [20] 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. Assuming a Bradley-Terry preference model tied to an implicit reward function related to π_θ and a reference policy π_{ref}, DPO maximizes the likelihood of observed preferences, resulting in the loss:

$$h_{\pi_\theta}(x, y_w, y_l) = \beta \log \frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta \log \frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}, \tag{2}$$

$$L_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x,y_w,y_l)\sim D}\left[\log \sigma\left(h_{\pi_\theta}(x, y_w, y_l)\right)\right]. \tag{3}$$

where σ is the sigmoid function and β controls the deviation from π_{ref}.

Group Robust Preference Optimization (GRPO): GRPO [22] 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:

$$\min_{\pi_\theta} \max_{\alpha\in\Delta_{K-1}} \sum_{g=1}^K \alpha_g L_{\text{Pref}}(\pi_\theta; \pi_{\text{ref}}, D_g), \tag{4}$$

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}. The optimization dynamically increases weights α_g for groups with higher current loss, focusing learning on the worst-performing groups.

Simple Preference Optimization (SimPO): SimPO [17] 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:

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

Contrastive Preference Optimization (CPO): CPO [32] 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:

$$L_{\text{prefer}}(\pi_\theta) = -\mathbb{E}_{(x,y_w,y_l)\sim D}\left[\log \sigma\left(\beta \log \pi_\theta(y_w|x) - \beta \log \pi_\theta(y_l|x)\right)\right] \tag{6}$$

$$L_{\text{NLL}}(\pi_\theta) = -\mathbb{E}_{(x,y_w)\sim D}[\log \pi_\theta(y_w|x)] \tag{7}$$

$$L_{\text{CPO}}(\pi_\theta) = L_{\text{prefer}} + L_{\text{NLL}}. \tag{8}$$

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

Appendix B. FairPO: Detailed Methodology

B.1. Objective for the Privileged Group: Preference-Based Discrimination

For the *privileged* group $\mathcal{P}$, simply adjusting the weight of a standard BCE loss is insufficient. The primary challenge for these labels is often not basic classification, but fine-grained discrimination against closely related, incorrect alternatives. To address this directly, we reformulate the learning objective for this group as a preference learning task as motivated in our introduction.

Defining Confusing Counterparts. For a given instance $\mathbf{x}_i$ and a privileged label $l \in \mathcal{P}$, we dynamically identify a set of *confusing counterparts* based on the model's current predictions. We define two such sets:

- When the true label is positive ($y_{il} = 1$), the set of **confusing negatives** is composed of incorrect labels that the model scores higher than or equal to the true label:

$$S_{il}^{\text{neg}} = \{k \in \mathcal{T} \mid y_{ik} = 0 \text{ and } m(\mathbf{x}_i; \mathbf{w}_k) \geq m(\mathbf{x}_i; \mathbf{w}_l)\}. \tag{9}$$

- When the true label is negative ($y_{il} = 0$), the set of **confusing positives** is composed of correct labels that the model scores lower than or equal to the true negative:

$$S_{il}^{\text{pos}} = \{k \in \mathcal{T} \mid y_{ik} = 1 \text{ and } m(\mathbf{x}_i; \mathbf{w}_k) \leq m(\mathbf{x}_i; \mathbf{w}_l)\}. \tag{10}$$

The overall confusing set for (i, l) is denoted by $S_{il} = S_{il}^{\text{neg}} \cup S_{il}^{\text{pos}}$.

Conditional Objective with a Stablility Fallback. Our objective for the *privileged* group is designed to be adaptive. When a confusing set S_{il} is non-empty, we apply a preference loss $\mathcal{L}_{\text{pref}}$ to directly target these hard discriminative cases. However, as the model learns and becomes more accurate, the confusing set for many examples will naturally become sparse or empty. Relying solely on the preference loss in such a scenario would lead to an unstable, sparse gradient signal, effectively stalling the training process. To ensure continuous and stable learning, we incorporate a crucial **stability fallback**. If $S_{il} = \emptyset$ for a given instance, we revert to a standard Binary Cross-Entropy (BCE) loss:

$$\ell_{\text{BCE}}(\mathbf{x}_i, y_{il}; \mathbf{w}_l) = -[y_{il} \log m(\mathbf{x}_i; \mathbf{w}_l) + (1 - y_{il}) \log(1 - m(\mathbf{x}_i; \mathbf{w}_l))]. \tag{11}$$

This fallback mechanism is a critical design choice for robustness. Initially, the model makes many mistakes, leading to large confusing sets and frequent application of the preference loss. As training progresses, the preference loss successfully resolves these hard cases, and the BCE fallback takes over to fine-tune the decision boundaries on the now *easier* examples. Our analysis of the training dynamics confirms that a sufficient preference signal exists throughout training, with the fallback rate gradually increasing as the model improves. The total loss for the privileged group, $\mathcal{L}_{\mathcal{P}}$, is the expectation over this adaptive, conditional choice. In our framework, we explore several modern preference optimization techniques to instantiate $\mathcal{L}_{\text{pref}}$.

FairPO-DPO Variant. To define our preference losses consistently, let us denote the score of the *preferred* label in a pair as $m(\mathbf{x}_i; \mathbf{w}_p)$ and the score of the *dispreferred* label as $m(\mathbf{x}_i; \mathbf{w}_d)$. The assignment of these roles depends on the ground truth. For example, if $y_{il} = 1$ and $k \in S_{il}^{\text{neg}}$, the true label l is preferred ($p = l$) and the confusing negative k is dispreferred ($d = k$). Conversely, if $y_{il} = 0$ and $k \in S_{il}^{\text{pos}}$, the true negative l is preferred ($p = l$) and the confusing positive k is dispreferred ($d = k$).

Inspired by DPO [20], this variant computes a preference loss relative to a reference model $\hat{\mathbf{w}}$. The preference loss is the negative log-likelihood of the preference, modeled by the difference between the log-probability ratios:

$$\mathcal{L}_{\text{pref}}^{\text{DPO}}(\mathbf{x}_i, p, d) = -\log \sigma \left(\beta \left(\log \frac{m(\mathbf{x}_i; \mathbf{w}_p)}{m(\mathbf{x}_i; \hat{\mathbf{w}}_p)} - \log \frac{m(\mathbf{x}_i; \mathbf{w}_d)}{m(\mathbf{x}_i; \hat{\mathbf{w}}_d)} \right) \right). \tag{12}$$

The full DPO-based loss for the privileged group, $\mathcal{L}_{\mathcal{P}}^{\text{DPO}}$, combines this preference loss with the BCE fallback:

$$\mathcal{L}_{\mathcal{P}}^{\text{DPO}} = \mathbb{E}_{(\mathbf{x}_i, l) \text{ s.t. } l \in \mathcal{P}} \left[\mathbf{1}_{S_{il} \neq \emptyset} \cdot \mathcal{L}_{\text{pref}}^{\text{DPO}}(\mathbf{x}_i, p, d) + \mathbf{1}_{S_{il} = \emptyset} \cdot \ell_{\text{BCE}}(\mathbf{x}_i, y_{il}; \mathbf{w}_l) \right], \qquad (13)$$

where for each sampled privileged label l with a non-empty confusing set, a confusing counterpart k is sampled from S_{il}, and the pair (p, d) is determined based on (l, k) and the ground truth to compute the loss.

FairPO-SimPO Variant (Reference-Free with Margin). This variant adapts SimPO [17] to create a reference-free preference loss that incorporates an explicit target margin $\gamma > 0$. We adapt SimPO's core concept, which was originally designed for sequences, to our per-label score setting. The resulting preference loss is:

$$\mathcal{L}_{\text{pref}}^{\text{SimPO}}(\mathbf{x}_i, p, d) = -\log \sigma \left(\beta \log \frac{m(\mathbf{x}_i; \mathbf{w}_p)}{m(\mathbf{x}_i; \mathbf{w}_d)} - \gamma \right). \qquad (14)$$

The term $\beta \log(\cdot)$ captures the relative preference between the scores, while the margin term $-\gamma$ enforces a stronger separation, pushing the model to not just rank the labels correctly but to do so by a significant amount. This preference loss is used in place of the DPO term within the overall conditional objective for the *privileged* group (analogous to Eq. 13).

FairPO-CPO Variant (Reference-Free with BCE Regularization). Our final and most effective variant adapts CPO [32]. This approach is unique in that it integrates a BCE regularizer directly into the preference objective to ensure model stability and a sense of absolute correctness. When a confusing set is found, the preference loss is defined as:

$$\mathcal{L}_{\text{pref}}^{\text{CPO}}(\mathbf{x}_i, p, d, l) = -\log \sigma \left(\beta \log \frac{m(\mathbf{x}_i; \mathbf{w}_p)}{m(\mathbf{x}_i; \mathbf{w}_d)} \right) + \lambda_{\text{CPO}} \cdot \ell_{\text{BCE}}(\mathbf{x}_i, y_{il}; \mathbf{w}_l). \qquad (15)$$

For each sampled label l, the pair (p, d) is determined from a confusing counterpart $k \in S_{il}$ if the set is non-empty. Here, the first term is a margin-free preference objective, while the second term, weighted by a hyperparameter λ_{CPO}, is the standard BCE loss for the privileged label l. This BCE component is crucial, acting as a regularizer that grounds the model in learning absolute scores. The full privileged loss objective is then analogous to the DPO formulation (Eq. 13): the comprehensive $\mathcal{L}_{\text{pref}}^{\text{CPO}}$ is used when a confusing set is found, and only the standard BCE term (Eq. 11), is used as a fallback otherwise.

Advantages of the Preference-Based Objective. Our choice to adapt modern preference optimization techniques is motivated by the fundamental limitations of standard point-wise losses like **BCE** and **Focal Loss** for fine-grained discrimination. These losses evaluate each label's score independently, answering the question, *"Is the score for label l correct in isolation?"*. While effective for general classification, this approach provides an indirect and often insufficient signal for resolving confusion between two closely-scored labels [11]. For example, a model might correctly learn to assign a high score (e.g., 0.8) to a true positive label and a slightly lower score (e.g., 0.7) to a confusing negative. A point-wise loss would provide only a small error signal for this negative case. The preference-based objective in FairPO is fundamentally different. It operates on *pairs* of labels and directly answers a more relational question: *"Is the score for the correct label l decisively*

higher than the score for its specific confusing competitor k?". By optimizing the log-ratio of these scores, our framework generates a strong, targeted gradient to explicitly drive their values apart. This provides a far more direct and effective mechanism for resolving the most challenging discriminative cases within the *privileged* group, a capability that point-wise losses lack by design.

B.2. Objective for the Non-Privileged Group: Constrained Performance

For the *non-privileged* group $\bar{\mathcal{P}}$, the primary objective is different. While the model dedicates its capacity to improving the challenging *privileged* labels, we must ensure that this focus does not come at an unacceptable cost to the performance on the remaining labels. Therefore, our goal for this group is not to aggressively maximize performance, but to *preserve* it by preventing significant degradation relative to a reliable baseline.

To achieve this, we introduce a constrained objective that leverages a pre-trained reference model with parameters $\hat{\mathbf{w}}$. We use the standard BCE loss (Eq. 11), as our objective for a given label $j \in \bar{\mathcal{P}}$. The model is penalized only if the BCE loss of the current model, $\ell_{\text{BCE}}(\mathbf{x}_i, y_{ij}; \mathbf{w}_j)$, exceeds the loss of the reference model, $\ell_{\text{BCE}}(\mathbf{x}_i, y_{ij}; \hat{\mathbf{w}}_j)$, by more than a small, predefined slack margin $\epsilon \geq 0$. This is implemented using a hinge loss mechanism:

$$\mathcal{L}_{\bar{\mathcal{P}}} = \mathbb{E}_{(\mathbf{x}_i, j) \text{ s.t. } j \in \bar{\mathcal{P}}} \left[\max(0, \ell_{\text{BCE}}(\mathbf{x}_i, y_{ij}; \mathbf{w}_j) - \ell_{\text{BCE}}(\mathbf{x}_i, y_{ij}; \hat{\mathbf{w}}_j) - \epsilon) \right]. \tag{16}$$

The gradient for any non-privileged label is zero as long as its performance is *good enough* (i.e., close to or better than the reference model). A learning signal is generated only when performance on a label j drops below this safety threshold.

B.3. Group Robust Optimization Formulation

Having defined two distinct objectives, a preference-based loss $\mathcal{L}_{\mathcal{P}}$ for the *privileged* group and a constrained loss $\mathcal{L}_{\bar{\mathcal{P}}}$ for the *non-privileged* group, the final challenge is to balance them. As these objectives are often in competition, a simple weighted sum is insufficient. Instead, we formulate the overall training objective as a minimax game, adapting the principles of Group Robust Preference Optimization (GRPO) [22]. We treat the *privileged* and *non-privileged* label sets as two distinct groups. The goal is to train a model, parameterized by $\mathbf{w}$, that is robust to the worst-case distribution of losses across these groups. This is expressed as the following minimax problem:

$$\min_{\{\mathbf{w}_t\}} \max_{\alpha \in \Delta} \left[\alpha_{\mathcal{P}} \mathcal{L}_{\mathcal{P}} + \alpha_{\bar{\mathcal{P}}} \mathcal{L}_{\bar{\mathcal{P}}} \right], \tag{17}$$

where $\mathcal{L}_{\mathcal{P}}$ and $\mathcal{L}_{\bar{\mathcal{P}}}$ represent the expected losses for their respective groups, and $\alpha = (\alpha_{\mathcal{P}}, \alpha_{\bar{\mathcal{P}}})$ is a vector of adaptive weights in the probability simplex Δ. This formulation provides a principled way to manage the performance trade-off between the two groups.

Appendix C. Extended Related Works

Fairness in Multi-Label Classification: Ensuring fairness in MLC [16, 30] is complex due to multi-faceted label relationships. Recent efforts address MLC-specific fairness challenges, such as tackling label imbalance impacting tail labels [8], learning instance and class-level subjective fairness [12], or incorporating fairness in dynamic learning settings like class-incremental MLC [4]. FairPO contributes by explicitly partitioning labels into privileged ($\mathcal{P}$) and non-privileged ($\bar{\mathcal{P}}$) sets,

applying distinct fairness-motivated objectives to each—notably using preference signals for $\mathcal{P}$—and managing them via a robustness framework. This targeted approach to enhancing performance for pre-defined critical labels, while safeguarding others, differentiates our work.

Preference Optimization: Preference optimization, especially for aligning LLMs [2, 18], has rapidly advanced. Direct Preference Optimization (DPO) [20] and its reference-free variants like CPO [32] and SimPO [17] optimize policies directly from preference pairs. The field continues to evolve with methods like Identity Preference Optimization (IPO) for stability [13], Kahneman-Tversky-based optimization (KTO) [6], simple yet effective Rejection Sampling Fine-Tuning (RFT) [9], and ongoing theoretical analyses of these methods [33]. We use preferences not to rank entire outputs, but to specifically differentiate true label scores from those of their *confusing* positive/negative counterparts within the privileged label set, thereby sharpening the model's decision boundaries for critical distinctions.

Group Robust Optimization: Distributionally Robust Optimization (DRO) [7], particularly Group DRO [25], aims to improve worst-case performance across predefined data groups, enhancing fairness and robustness. This concept sees continued development, for instance, in improving its practicality [23]. Group Robust Preference Optimization (GRPO) [22] extended this to LLM preference learning, balancing performance across preference groups. FairPO directly employs GRPO's adaptive optimization strategy. However, our groups are defined by the label partition ($\mathcal{P}$ and $\bar{\mathcal{P}}$), and GRPO balances their distinct, custom-formulated loss objectives. This provides a principled mechanism for robustly managing the specific fairness-performance trade-offs in our MLC context.

Appendix D. Adapting FairPO for Multi-Attribute Generation

This section outlines our planned extension of FairPO to multi-attribute generation, a conceptual direction for future work. The goal is to generate a sequence y from a prompt x using a policy $\pi_{\mathbf{w}}(y|x)$ that aligns with fairness goals over a set of attributes $\mathcal{A}$. The core idea involves partitioning $\mathcal{A}$ into privileged $\mathcal{P}$ and non-privileged $\bar{\mathcal{P}}$ sets and retaining the GRPO minimax structure (Eq. 17). The group losses would be defined over a preference dataset $\mathcal{D}_{pref} = \{(x_i, y_{wi}, y_{li}, j_i)\}_{i=1}^{M}$, with preference losses like DPO applied to the log-probabilities of entire generated sequences rather than individual label scores.

Proposed Privileged Loss ($\mathcal{L}_{\mathcal{P}}$): For privileged attributes $j \in \mathcal{P}$, the goal is to ensure the learned policy $\pi_{\mathbf{w}}$ strongly reflects preferences $y_w \succ y_l$ established by that attribute. This is achieved using a standard DPO loss, averaged over the privileged subset of the preference data:

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

Minimizing this loss directly encourages the model to favor preferred sequences for preferences driven by privileged attributes, relative to the reference policy π_{ref}.

Proposed Non-Privileged Loss ($\mathcal{L}_{\bar{\mathcal{P}}}$): For non-privileged attributes $k \in \bar{\mathcal{P}}$, the objective remains analogous to the classification setting: preventing significant performance degradation. This is accomplished with a hinge formulation based on the DPO loss:

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

This penalizes the model only if its preference modeling for non-privileged attributes degrades substantially beyond baseline performance (represented by $\log 2$ for random preference) plus a slack ϵ'. The overall FairPO objective would then use GRPO to balance these two losses.

Appendix E. FairPO Algorithm

The FairPO framework is trained iteratively to solve the minimax objective presented in Eq. 17. The detailed procedure, which is inspired by the DPO-based variant of FairPO, is provided in Algorithm 2.

Initialization: The training process begins by initializing the model parameters $\{\mathbf{w}_t | t \in \mathcal{T}\}$, for instance by copying them from a pre-trained reference model $\{\hat{\mathbf{w}}_t | t \in \mathcal{T}\}$. The adaptive group weights, $\alpha_{\mathcal{P}}$ and $\alpha_{\bar{\mathcal{P}}}$, are typically set to uniform values such as 0.5 each.

Iterative Training Loop: The core of the framework is an iterative training loop. In each step, an instance (x_i, y_i) is sampled from the dataset $\mathcal{D}$, and a single label $r \in \mathcal{T}$ is randomly selected from that instance for processing. The subsequent steps depend on whether this sampled label belongs to the privileged or non-privileged set.

If the sampled label r is in the **privileged set** $\mathcal{P}$, the algorithm first identifies if a *confusing set* S_{il} exists for that label (where $l = r$), as detailed in Algorithm 2. The loss computation is then conditional on this set:

- If confusing examples exist ($S_{il} \neq \emptyset$), a DPO-inspired preference loss is computed between label l and a randomly sampled confusing example $k \in S_{il}$. This preference loss directly encourages the model to improve its ranking of l relative to its specific confounder k.

- If no confusing examples are found ($S_{il} = \emptyset$), the algorithm reverts to a standard base classification loss (e.g., BCE, Eq. 11) for label l. This fallback is crucial as it ensures the model continues to receive a learning signal on *easier* instances, promoting stable training.

The loss calculated from either of these cases contributes to the current step's privileged group loss, $\mathcal{L}_{\mathcal{P}}^{(s)}$. Conversely, if the sampled label r belongs to the *non-privileged set* $\bar{\mathcal{P}}$, the constrained loss $\mathcal{L}_{\bar{\mathcal{P}}}^{(s)}$ is computed according to Eq. 16. This loss penalizes the model only if its performance on the label $j = r$ deviates from the reference model's performance by more than a predefined slack margin ϵ.

After the appropriate group loss is computed, the GRPO mechanism performs two key updates. First, the *Adaptive Weight Update* adjusts the group weights $\alpha_{\mathcal{P}}$ and $\alpha_{\bar{\mathcal{P}}}$ using a mirror ascent step. This step uses an exponential weighting based on the current (and scaled) group losses, dynamically increasing the focus on the group that is currently performing worse (Lines 39-41). Second, the *Model Parameter Update* updates all model parameters $\mathbf{w}_t$ via a mirror descent step, using a combined gradient that is weighted by the newly updated adaptive weights $\alpha_{\mathcal{P}}$ and $\alpha_{\bar{\mathcal{P}}}$.

This entire process repeats for a predefined number of iterations or until convergence, allowing FairPO to dynamically balance its objectives to achieve robust fairness. For variants like FairPO-SimPO or FairPO-CPO, the core logic remains identical; only the DPO-inspired preference loss component is replaced with their respective preference formulations (e.g., Eq. 14 or 15). The overall GRPO structure and non-privileged handling are consistent across all variants.

Appendix F. Dataset and Preprocessing Details

MS-COCO 2014 [10]: We used the official 2014 train/val splits. The training set contains 82,783 images and the validation set (used as our test set) contains 40,504 images. There are 80 object categories. The privileged set $\mathcal{P}$ consisted of the 16 labels (20% of 80) with the lowest frequency in the training set. The remaining 64 labels formed $\bar{\mathcal{P}}$.

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

1: **Initialize:** $\{\mathbf{w}_t^{(0)} \in \mathbb{R}^d | \forall t \in \mathcal{T}\}$ (e.g., copy $\{\hat{\mathbf{w}}_t | \forall t \in \mathcal{T}\}$), $\alpha_{\mathcal{P}}^{(0)} \leftarrow 0.5, \alpha_{\bar{\mathcal{P}}}^{(0)} \leftarrow 0.5$.

2: **Choose:** $\eta_{\mathbf{w}}, \eta_{\alpha}, \beta, \{\hat{\mathbf{w}}_t | \forall t \in \mathcal{T}\}, \epsilon$.

3: **for** $s = 0$ to S (MaxIterations) **do**

4:     Sample an example: $(x_i, [y_{i1}, \dots, y_{iT}]) \in \mathcal{D} \sim p_{\mathcal{D}}(.)$.

5:     Initialize group losses for this step: $\mathcal{L}_{\mathcal{P}}^{(s)} \leftarrow 0, \mathcal{L}_{\bar{\mathcal{P}}}^{(s)} \leftarrow 0$.

6:     Initialize gradients: $g_{\mathcal{P}}^t \leftarrow \mathbf{0}, g_{\bar{\mathcal{P}}}^t \leftarrow \mathbf{0} \quad \forall t \in \mathcal{T}$.

7:     Forward pass: $m(x_i; \mathbf{w}_t^{(s)}) \leftarrow \sigma(\mathbf{w}_t^{(s)^T} \mathbf{z}_i)$ where $\mathbf{z}_i \leftarrow \pi_\theta(x_i) \quad \forall t \in \mathcal{T}$.

8:     Sample a label: $r \in \mathcal{T} \sim Uniform(\frac{1}{|\mathcal{T}|})$.

9:     **if** $r \in \mathcal{P}$ **then**   ▷ Handle privileged label

10:         $l \leftarrow r, S_{il}^{\text{neg}} \leftarrow \emptyset, S_{il}^{\text{pos}} \leftarrow \emptyset$

11:         **if** $y_{il} = +1$ **then**   ▷ True Positive case for privileged label l

12:             $S_{il}^{\text{neg}} \leftarrow \{k \in \mathcal{T} \mid y_{ik} = 0 \text{ and } m(x_i; \mathbf{w}_k^{(s)}) \geq m(x_i; \mathbf{w}_l^{(s)})\}, S_{il} \leftarrow S_{il}^{\text{neg}}$

13:         **else if** $y_{il} = 0$ **then**   ▷ True Negative case for privileged label l

14:             $S_{il}^{\text{pos}} \leftarrow \{k \in \mathcal{T} \mid y_{ik} = +1 \text{ and } m(x_i; \mathbf{w}_k^{(s)}) \leq m(x_i; \mathbf{w}_l^{(s)})\}, S_{il} \leftarrow S_{il}^{\text{pos}}$

15:         **end if**

16:         **if** $S_{il} \neq \emptyset$ **then**   ▷ Confusing examples exist, use DPO-inspired loss

17:             Sample $k \in S_{il} \sim Uniform(\frac{1}{|S_{il}|})$

18:             **if** $y_{il} = +1$ **then**   ▷ DPO for True Positive l vs Confusing Negative k

19:                 $h_{\mathbf{w}^{(s)}}(x_i, l, k) \leftarrow \left(\log \frac{m(x_i; \mathbf{w}_l^{(s)})}{m(x_i; \hat{\mathbf{w}}_l)}\right) - \left(\log \frac{m(x_i; \mathbf{w}_k^{(s)})}{m(x_i; \hat{\mathbf{w}}_k)}\right).$

20:                 $\mathcal{L}_{\text{pref}} \leftarrow -\log \sigma \left(\beta \cdot h_{\mathbf{w}^{(s)}}(x_i, l, k)\right)$

21:             **else if** $y_{il} = 0$ **then**   ▷ DPO for True Negative l vs Confusing Positive k

22:                 $h_{\mathbf{w}^{(s)}}(x_i, k, l) \leftarrow \left(\log \frac{m(x_i; \mathbf{w}_k^{(s)})}{m(x_i; \hat{\mathbf{w}}_k)}\right) - \left(\log \frac{m(x_i; \mathbf{w}_l^{(s)})}{m(x_i; \hat{\mathbf{w}}_l)}\right).$

23:                 $\mathcal{L}_{\text{pref}} \leftarrow -\log \sigma \left(\beta \cdot h_{\mathbf{w}^{(s)}}(x_i, k, l)\right)$

24:             **end if**

25:             $\mathcal{L}_{\mathcal{P}}^{(s)} \leftarrow \mathcal{L}_{\text{pref}}, g_{\mathcal{P}}^t \leftarrow g_{\mathcal{P}}^t + \nabla_{\mathbf{w}_t} \mathcal{L}_{\text{pref}}|_{\mathbf{w}_t^{(s)}} \quad \forall t \in \mathcal{T}$.

26:         **else**   ▷ No confusing examples, use BCE loss for privileged label l

27:             $\mathcal{L}_{\text{BCE}} \leftarrow -[y_{il} \log m(x_i; \mathbf{w}_l^{(s)}) + (1 - y_{il}) \log(1 - m(x_i; \mathbf{w}_l^{(s)}))]$

28:             $\mathcal{L}_{\mathcal{P}}^{(s)} \leftarrow \mathcal{L}_{\text{BCE}}, g_{\mathcal{P}}^t \leftarrow g_{\mathcal{P}}^t + \nabla_{\mathbf{w}_t} \mathcal{L}_{\text{BCE}}|_{\mathbf{w}_t^{(s)}} \quad \forall t \in \mathcal{T}$.

29:         **end if**

30:     **else if** $r \in \bar{\mathcal{P}}$ **then**   ▷ Handle non-privileged label

31:         $j \leftarrow r$

32:         $\ell(\mathbf{w}_j^{(s)}) \leftarrow -[y_{ij} \log(m(x_i; \mathbf{w}_j^{(s)})) + (1 - y_{ij}) \log(1 - m(x_i; \mathbf{w}_j^{(s)}))]$

33:         $\ell(\hat{\mathbf{w}}_j) \leftarrow -[y_{ij} \log(m(x_i; \hat{\mathbf{w}}_j)) + (1 - y_{ij}) \log(1 - m(x_i; \hat{\mathbf{w}}_j))]$

34:         $\mathcal{L}_{\bar{\mathcal{P}}}^{(s)} \leftarrow \max\left(0, \ell(\mathbf{w}_j^{(s)}) - \ell(\hat{\mathbf{w}}_j) - \epsilon\right), g_{\bar{\mathcal{P}}}^t \leftarrow g_{\bar{\mathcal{P}}}^t + \nabla_{\mathbf{w}_t} \mathcal{L}_{\bar{\mathcal{P}}}^{(s)}|_{\mathbf{w}_t^{(s)}} \quad \forall t \in \mathcal{T}$.

35:     **end if**

36:     $\alpha_{\mathcal{P}}^{(s+1)} \leftarrow \alpha_{\mathcal{P}}^{(s)} \exp(\eta_\alpha \mathcal{L}_{\mathcal{P}, scaled}^{(s)})$ 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

37:     $Z \leftarrow \alpha_{\mathcal{P}}^{(s+1)} + \alpha_{\bar{\mathcal{P}}}^{(s+1)}, \alpha_{\mathcal{P}}^{(s+1)} \leftarrow \frac{\alpha_{\mathcal{P}}^{(s+1)}}{Z}$ and $\alpha_{\bar{\mathcal{P}}}^{(s+1)} \leftarrow \frac{\alpha_{\bar{\mathcal{P}}}^{(s+1)}}{Z}$ ▷ Weight normalization

38:     $\mathbf{w}_t^{(s+1)} \leftarrow \mathbf{w}_t^{(s)} - \eta_{\mathbf{w}}(\alpha_{\mathcal{P}}^{(s+1)} g_{\mathcal{P}}^t + \alpha_{\bar{\mathcal{P}}}^{(s+1)} g_{\bar{\mathcal{P}}}^t) \quad \forall t \in \mathcal{T}$ ▷ Mirror descent

39: **end for**

40: **return** $\{\mathbf{w}_t^{(S)} | \forall t \in \mathcal{T}\}$

NUS-WIDE [3]: This dataset contains 269,648 images with 81 concept labels. We used the common split of 161,789 images for training and 107,859 for testing. The privileged set $\mathcal{P}$ consisted of the 16 labels (approx. 20% of 81) with the lowest frequency in the training set. The remaining 65 labels formed $\bar{\mathcal{P}}$.

Image Preprocessing: For both datasets, images were resized to 224×224 pixels and normalized using the standard ImageNet mean and standard deviation, consistent with the ViT pretraining. Standard data augmentations like random horizontal flips and random resized crops were applied during training.

Appendix G. FairPO Experimental Details

G.1. Common Setup for All FairPO Variants

Unless specified otherwise, a common setup was used for all FairPO variants to ensure fair comparison. The base model for feature extraction was a Vision Transformer (ViT), specifically `vit-base-patch16-224` [5], which was pre-trained on ImageNet-21k and fine-tuned on ImageNet-1k. During our fine-tuning, the ViT backbone was kept frozen, with the exception of its final encoder layer, which was made trainable to allow for adaptation of higher-level features. All experiments were conducted on the MS-COCO 2014 [10] and NUS-WIDE [3] datasets. Images were resized to 224×224 pixels, normalized using ImageNet statistics, and augmented with standard techniques like random horizontal flips and resized crops. The AdamW optimizer [14] was used to update all trainable parameters. Models were trained for a maximum of 25 epochs with a batch size of 32, and we employed an early stopping strategy with a patience of 5 epochs based on the overall mAP on the validation set.

G.2. Per-Label Non-Linear MLP Classifier Head

For each of the T labels in a dataset, we employed a dedicated and independent non-linear Multi-Layer Perceptron (MLP) head to predict the probability of that label being positive. Using separate MLP heads allows for more complex, non-linear decision boundaries tailored to each label's specific characteristics, which is particularly beneficial for labels with varying difficulty. Each MLP head takes the d-dimensional feature vector (where $d = 768$ for ViT-Base) from the ViT's [CLS] token as input and outputs a single logit. The final probability score $m(x_i; \mathbf{w}_t)$ is obtained by applying a sigmoid function to this logit. The specific architecture for each MLP head is as follows:

1. Linear Layer: $d \rightarrow 256$ neurons, followed by ReLU Activation
2. Linear Layer: $256 \rightarrow 64$ neurons, followed by ReLU Activation
3. Linear Layer: $64 \rightarrow 16$ neurons, followed by ReLU Activation
4. Linear Layer: $16 \rightarrow 4$ neurons, followed by ReLU Activation
5. Linear Layer (Output): $4 \rightarrow 1$ neuron (producing the logit)

The parameters $\mathbf{w}_t$ for each label t's MLP head are unique to that label. All parameters within these MLP heads were fully trainable during both the SFT pre-training (for the reference model) and the final FairPO fine-tuning.