Journal of Mathematics Instruction, Social Research and Opinion
Vol. 5, No. 2, June 2026, pp. 1821 – 1836, https://doi.org/10.58421/misro.v5i2.1674
ISSN 2962-7842

Mitigating Occupational Gender Bias in CLIP through Direction Loss-Augmented Learnable Prompts

Rahmi Fariza[1], Kurniawati Azizah[2]
[1,2]Universitas Indonesia, West Java, Indonesia

Article Info

Article history:

Received 2026-05-27
Revised 2026-06-16
Accepted 2026-06-23

Keywords:

CLIP
Content Optimization
Gender Direction
Loss Function
The Vision Language

ABSTRACT

Vision-language models such as CLIP achieve strong zero-shot performance but inherit gender bias from their web-scale pretraining data, which is especially visible when the model is used to retrieve images for occupations. Existing prompt-based debiasing methods rely on manually crafted text prompts, which require extensive trial and error and dont transfer easily across professions. This study proposes CoOp with Direction Loss (CoOp+DL), which augments Context Optimization (CoOp), a learnable-prompt method, with an auxiliary loss that pushes the learned prompt representations away from a gender direction computed from contrasting male- and female-referencing prompts. The framework is evaluated on 500 images covering 10 professions with a balanced gender distribution, using three CLIP backbones (ViT-B/32, ViT-B/16, and OpenCLIP ViT-B/32) and three metrics: Gender Bias Score (GBS), Precision-at-K, and SignedSkew. CoOp+DL reduces GBS by 10.3% on ViT-B/32, 5.9% on ViT-B/16, and 9.7% on OpenCLIP, an average of 8.65% across backbones, with bootstrap confidence intervals (n = 1,000) indicating that the direction loss is an active contributor to this reduction rather than an artifact of additional prompt capacity. Retrieval utility (Precision@K) improves on ViT-B/32 and ViT-B/16 (+6.8% and +4.3%) but decreases on OpenCLIP (−8.2%), indicating a backbone-dependent fairness-utility trade-off. CoOp+DL achieves bias reduction that is statistically comparable to a manually engineered ensemble prompt, without requiring manual prompt design. The findings should be interpreted with caution, given the modest evaluation set (500 images, 10 professions) and the binary gender formulation used to define the direction vector, both of which limit generalization and warrant further validation before deployment.

Corresponding Author:

Rahmi Fariza
Faculty of Computer Science, Universitas Indonesia, Indonesia
Email: rahmi.fariza@ui.ac.id

1. INTRODUCTION

Vision-language models (VLMs) that learn a joint representation of images and text from web-scale image-text pairs [1], [2], [3] have become a foundation for many computer vision applications. Among these, Contrastive Language-Image Pretraining (CLIP) [4] is one of the most widely adopted VLMs. CLIP is trained by OpenAI on 400 million image-text pairs collected from the internet, using a contrastive objective that aligns image and text embeddings in a shared representation space.

Because CLIP's training data is drawn directly from the web, the model is prone to inheriting and even amplifying social biases present in that data [5], [6]. This is particularly evident for protected attributes such as gender: when CLIP is used in downstream applications, including image retrieval, automatic tagging, and generative pipelines, occupational queries (e.g., “a photo of a doctor” or “a photo of a secretary”) tend to retrieve images that are skewed toward one gender, reflecting and reinforcing societal stereotypes about who performs a given job.

One way to mitigate gender bias at the prompt level is to design text instructions that steer the model toward fairer outputs without retraining its parameters [7], [8], [9]. This hand-crafted, or “manual prompt,” approach has a practical limitation: small wording changes can substantially affect both accuracy and fairness, and finding an effective prompt is largely a matter of trial and error, which is time-consuming and does not scale across professions. Learnable-prompt methods such as Context Optimization (CoOp) [10] address this by replacing manual prompt text with context vectors that are automatically optimized via backpropagation using cross-entropy [10], [11]. CoOp keeps the CLIP image and text encoders frozen and updates only a small set of context vectors, making it computationally efficient and independent of manual word selection.

However, CoOp is optimized solely for task performance (classification accuracy via cross-entropy), with no explicit fairness objective. As a result, the context vectors learned by CoOp can still encode a gender direction even when the prompt is otherwise optimal for the classification task, and to date, no study has explicitly removed this gender direction from the CoOp prompt space. Debiasing on the text side has practical advantages over debiasing the image encoder: it avoids the computational cost of retraining a large image encoder [12], [13], and a debiased prompt can potentially be transferred to other downstream tasks. Because gender bias in CLIP is largely formed through the text-image association learned during pretraining [4], intervening directly in the prompt space targets this association more closely at its source. Addressing this gap—integrating an explicit gender-direction objective into CoOp's learnable-prompt optimization to mitigate occupational gender bias in image retrieval—is the focus of this study.

Compared with prior debiasing strategies, the proposed approach differs in both mechanism and scope. Biased-prompt methods such as [7] identify and suppress biased directions in a fixed, pre-specified prompt embedding, but do not optimize the prompt itself.

Adversarial prompt-array approaches [18] train an ensemble of prompts using an adversarial objective, thereby increasing training complexity and requiring careful tuning of the adversary. Residual-based methods such as DeAR [16] add a learned correction vector to image embeddings, which requires intervention on the image side and therefore does not benefit from CoOp's parameter efficiency. CoOp+DL instead integrates a single, differentiable fairness regularizer directly into the few-shot CoOp optimization loop, so that the same context vectors that are learned for classification are simultaneously constrained to be orthogonal to the gender direction—without an adversarial component, without modifying the image encoder, and without abandoning CoOp's data efficiency (only 80 training images).

This study makes three contributions. First, it proposes CoOp with Direction Loss (CoOp+DL), a debiasing approach that integrates a gender-direction regularization term into the optimization of CoOp's learnable prompt space. Second, it provides a sensitivity analysis of the regularization weight λ, characterizing the fairness-utility trade-off: values of λ that are too large lead to overcorrection, and those that are too small are ineffective. Third, it validates the framework on three CLIP variants—ViT-B/32, ViT-B/16, and OpenCLIP ViT-B/32—using bootstrap confidence intervals to assess the statistical reliability of the observed bias reduction.

The evaluation uses 500 images across 10 professions, purposively selected based on documented gender stereotypes reported in prior literature, with a balanced 50:50 gender distribution, and bootstrap confidence intervals are used to assess the robustness of the results across the three CLIP architectures. As a scope limitation acknowledged from the outset, the framework adopts a binary (male/female) formulation of gender to construct the direction vector; this choice, and its implications, are discussed further in Section 2.5 and revisited in Sections 3 and 4.

Three research questions guide this study:

1) How effective is CoOp with Direction Loss compared with standard CoOp in reducing gender bias in occupational image retrieval?
2) How does the regularization weight λ affect the balance between fairness and retrieval utility?
3) How does the effectiveness of the framework vary across CLIP architectures with different pretraining data and patch sizes, and what does this imply for deployment?

A summary of the findings, reported in full in Section 3, is as follows: CoOp+DL reduces the Gender Bias Score relative to the baseline by 10.3% on ViT-B/32 and 5.9% on ViT-B/16, accompanied by retrieval precision gains of 6.8% and 4.3% respectively, while on OpenCLIP a 9.7% bias reduction comes with an 8.2% decrease in precision—indicating that the fairness-utility balance is architecture-dependent.

2. METHOD

2.1 Problem Formulation

This research focuses on analyzing and mitigating gender bias in CLIP-based vision-language models. CLIP counts the similarity between the query text and all images, then ranks them from the highest to the lowest using cosine similarity [14]. The similarity between embedding text *t* and each embedding image *v* is calculated using the cosine similarity defined as:

$$sim(t, v) = \frac{t \cdot v}{|t| \times |v|} \quad (1)$$

This research uses several bias metrics, namely MaxSkew, Gender Bias Score (GBS), and SignedSkew. Given the professional set $P = \{p_1, p_2, ..., p_n\}$ with n = 10, and the image dataset $I = \{I_1, ..., I_N\}$ with $N = 500$ samples labeled gender $g \in \{\text{male, female}\}$. For every text query that represents profession p, the retrieval system returns the top-*K* images, $K = 50$. *K*=50 was chosen to be equivalent to the ground-truth per profession and justified in Section III.E, which has the highest cosine similarity with t_p.

MaxSkew

MaxSkew measures the extent to which the gender (p_g) distribution in the top search results deviates from the ideal distribution [15], [16]. For gender g with the actual proportion in top-*K* and the ideal proportion $p_{ideal} = 0.5$, the skew is defined as $skew_g = \log(p_g / p_{ideal})$. MaxSkew@K is calculated as:

$$MaxSkew@K = max(skew_{male}, skew_{female}) \quad (2)$$

Gender Bias Score (GBS)

The Gender Bias Score (GBS) measures the average level of bias across sensitive groups. This metric evaluates the extent to which the model's prediction distribution deviates from the balanced ideal distribution by calculating the average MaxSkew@K [15] across all professions. GBS is calculated as:

$$GBS = \frac{1}{|P|} \sum_{p \in P} MaxSkew\ @K_{(p)} \quad (3)$$

SignedSkew

SignedSkew is used to see the direction of bias per profession, so that it can be known whether a profession is more inclined to male or female representation. Different from MaxSkew [15], which uses absolute values, SignedSkew maintains a positive or negative sign:

$$SignedSkew(p) = r_{male}(p) - 0.5 \quad (4)$$

Positive value indicates male group dominance, while a negative value indicates female group dominance.

2.2 CoOp Framework

The CoOp [10] framework in this study is used to replace manual prompts with prompts that can be learned automatically (learnable prompts). This method focuses on optimizing context representation to make text embeddings more consistent with the distributions of certain datasets.

$$T_p = f([v]_1[v]_2 \dots [v]_m[CLASS]) \quad (5)$$

f is the frozen CLIP text encoder, $[v]_{i\ (i=1,..,M)}$ is the learnable context vectors, M = 16 is the number of context tokens, and [CLASS] is the frozen embedding token.

In the initial implementation, each profession class is represented by a manual prompt, such as “a photo of a doctor”, or "none". Each context token in the prompt is replaced with a set of learnable vector parameters. During the training, the image encoder parameter and text encoder CLIP are frozen, and the [CLASS] embedding token is also frozen, and only the context vectors are trainable.

$$L_{CoOp} = -\sum_{(I,y)} log\ P(y \mid I) \quad (6)$$

with the probability of prediction defined as:

$$\Pi(y \mid I) = \frac{exp\left(^{(sim\ (I,T_y)}/_{\tau}\right)}{\sum_p exp\left(^{(sim\ (I,T_p)}/_{\tau}\right)} \quad (7)$$

Here, I is the image embedding, T_y is the text embedding for the ground-truth profession y, T_p is the text embedding for profession p ∈ P, and τ is CLIP's learned temperature parameter. Class prediction is performed via the softmax of cosine similarities between the image embedding and each profession's text embedding, with the [CLASS] token preserving its original (frozen) word embedding.

Because CoOp is optimized end-to-end using only the cross-entropy loss in Eq. (10), with no term that explicitly removes gender information from the prompt representation, the potential for gender bias encoded in the learned context vectors can persist even after the classification objective has converged.

2.3 Gender Direction Loss untuk Prompt

Unlike debiasing approaches that intervene on image embeddings [16], this study applies a direction loss directly to the text-side prompt embeddings T_p produced by the learnable CoOp context. This design choice is motivated by three considerations: (1) it avoids the computational cost of retraining the image encoder; (2) gender bias in CLIP arises primarily from text-image association during pretraining, so a text-side intervention targets the bias closer to its source [8] [17]; and (3) a debiased prompt can potentially be reused for other downstream tasks. The approach reduces the correlation between the prompt embedding and gender by penalizing its projection onto a gender direction vector, acting as a fairness regularizer [7], [18].

Unlike the seed-word approach of Bolukbasi et al. [19], which derives a gender direction from generic gender-coded words, this study computes the gender direction once, from

profession-contextualized prompts formed by combining two gender-specific templates with all |P| = 10 professions. Two sets of contextualized prompts are constructed:

$$P_{male} = \{t_p \mid t \in T_{male}, p \in P\} \tag{8}$$

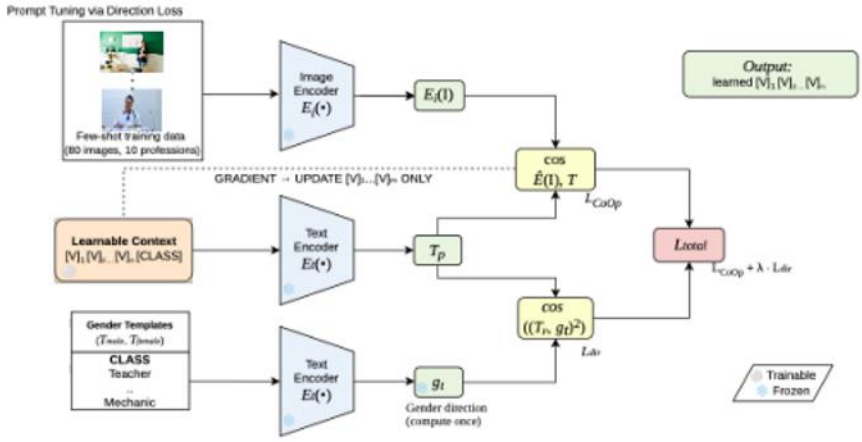

Figure 1. Framework Overview. Gender direction is calculated once from profession labels, contextualized gender prompts

$$P_{female} = \{t_p \mid t \in T_{female}, p \in P\} \tag{9}$$

T_{male} = {"*a photo of a male* {}", "*a photo of a man working as a* {}"} and T_{female}, = {"*a photo of a female* {}", "*a photo of a woman working as a* {}"}. This approach aims to capture the gender direction in the semantic context of the profession.

g_{text} specific for text embedding allows:

$$g_{text} = \frac{\mu_{male} - \mu_{female}}{\left\|\mu_{male} - \mu_{female}\right\|^2} \tag{10}$$

The gender prototype is calculated using the average embedding words:

$$\mu_{male} = \frac{1}{|P_{male}|}\sum_{q \in P_{male}} f([w]) \tag{11}$$

$$\mu_{female} = \frac{1}{|P_{female}|}\sum_{q \in P_{female}} f([w]) \tag{12}$$

$$L_{dir} = \frac{1}{|P|}\sum_{p \in P}\left(T_p \cdot g_{text}\right)^2 \tag{17}$$

T_p the text embedding for profession p, and $(\cdot)$ is a dot product. This loss minimizes the projection of each text embedding towards gender, so that the prompt representation becomes

orthogonal to the gender direction. The final function is a combination of classification loss, CoOp, and fairness regularization:

$$L_{total} = L_{CoOp} + \lambda \cdot L_{dir} \quad (13)$$

λ as a hyperparameter controlling the trade-off between classification performance and fairness. Through this approach, learnable context vectors are not only optimized to increase classification accuracy but also directed to reduce gender stereotypes in the embedding space.

λ is swept in the range {0.1, 0.5, 1.0, 2.0, 5.0}, which includes a weak regime (λ < 1, fairness as a light regularizer) to a strong (λ > 1, fairness as a dominant constraint), thus allowing the characterization of the trade-off of fairness-utility.

Algorithm 1: CoOp with Gender Direction Loss

Input CLIP model f_{text} (frozen)
Image feature dataset $D = \{I, y\}_{i=1}^{N}$, $I_i \in \mathbb{R}^d$
Profession set P
Gender templates sets T_{male}, T_{female}
Direction loss weight λ
Context Length M, Number of epochs E, batch size B
SGD with momentum, weight decay, Lr η
Output Optimized context vectors V*

Phase 1: Compute gender direction (once, frozen)

$P_{male} \leftarrow \{t_p \mid t \in T_{male}, p \in P\}$
$P_{female} \leftarrow \{t_p \mid t \in T_{female}, p \in P\}$
For each q $\in P_{male} \cup P_{female}$:
  $\hat{e}_q \leftarrow \frac{f_{text(q)}}{\|f_{text(q)}\|^2}$ // L2 Normalize
  $\mu_{male} \leftarrow \frac{1}{|P_{male}|} \sum_{q \in P_{male}} \hat{e}_q$
  $\mu_{female} \leftarrow \frac{1}{|P_{female}|} \sum_{q \in P_{female}} \hat{e}_q$
  $g_{text} \leftarrow \frac{\mu_{male} - \mu_{female}}{\|\mu_{male} - \mu_{female}\|^2}$

Phase 2: Train context vectors

For epoch e = 1 to E:
  $\eta_e \leftarrow$ warmup-or-cosine-schedule (e, η, E, W)
  For each mini batch (I_B, y_B) of size B from D:
    For each p ∈ P:
      $T_p \leftarrow f_{text}([v]_1[v]_2 \ldots [v]_m[CLASS_p])$
      $\hat{T}_p \leftarrow \frac{T_p}{\|T_p\|^2}$ // L2 Normalize
      logits $\leftarrow I_B \cdot [\hat{T}_1, \ldots, \hat{T}_{|P|}]^T$
      $L_{CoOp} \leftarrow$ CrossEntropy (logits, y_B)
      $L_{dir} = \frac{1}{|P|} \sum_{p \in P} (\hat{T}_p \cdot g_{text})^2$
      $L_{total} \leftarrow L_{CoOp} + \lambda \cdot L_{dir}$
      // SGD with momentum 0.9, weight decay 5e-4
      $V \leftarrow V - \eta_e \cdot \nabla v \, L_{total}$
  End for
  Update scheduler S
End for
Return V*

The choice of $(\hat{T}_p \cdot g_{text})^2$ is based on the consideration of forcing orthogonality without choosing the direction of gender and is differentiable at zero, different from the absolute value.

2.4 Evaluation Metrics

This research evaluates two aspects, namely fairness and utility retrieval. The fairness metrics are MaxSkew@K, GBS, and SignedSkew [15], [16], which are computed only on the correctly retrieved subset of the top-K results, i.e., images whose ground-truth label matches the queried profession. Restricting these metrics to correct retrievals isolates gender stereotyping in the model's representation of a profession from noise introduced by incorrect retrievals (images of the wrong profession).

As a robustness check on this design choice, Section 3.1.2 also reports the standard full top-K skew (computed across all K = 50 retrieved images, regardless of whether the profession label is correct), so the two formulations can be compared directly. Retrieval utility is measured with Precision-at-K (P@K) for each profession p:

$$\text{P@K}_{(p)} = \frac{\left|\{i \in \textit{top-}K_p : \text{label}(i) = p\}\right|}{K} \tag{14}$$

$|top - K_p| = \text{K} = 50$. $\text{P@K}_{(p)} = 1.0$ means perfect retrieval, that is, the model finds exactly K = 50 target profession pictures. The average utility in all professions is calculated as:

$$Mean\ P@K\ = \frac{1}{|P|}\sum_{p \in P} \text{P@K}_{(p)} \tag{15}$$

To assess the statistical significance of bias reduction, this study uses Bootstrap Confidence Interval with n = 1,000 iterations. In each iteration, the sample is redrawn with replacement from the evaluation dataset, and the GBS is counted again. 95% CI is defined as the 2.5th and 97.5th percentile intervals of the bootstrap distribution. This approach is robust because it does not require normal distribution assumptions and provides unbiased estimates of the variability of results on a limited dataset (500 images). Bootstrap CI is calculated both for GBS per condition and for the difference in GBS between conditions, $\Delta\text{GBS} = \text{GBS}_{\text{baseline}} - \text{GBS}_{\text{treatment}}$. CI for the difference answers the statistical significance of the reduction of bias.

2.5 Evaluation setup

The evaluation dataset is manually curated and consists of 500 images from Unsplash (commercial license via Unsplash+ subscription), covering 10 professions with a balanced distribution: 25 male and 25 female per profession. Gender annotation is done manually during curation. A resolution constraint on raw images is not necessary because CLIP performs internal preprocessing (resizing to 224×224 and normalization) [20], [21], [22].

Data training for CoOp uses a separate dataset curated from Pexels, with a commercial-friendly license, consisting of 80 images distributed across 8 professions, with 4 male and 4 female images per profession. Annotation and curation are done manually by the author.

Selecting different sources (Pexels for training and Unsplash for evaluation) helps prevent data leakage. 80 images are consistent with the parameter-efficient few-shot CoOp paradigm [10], which relies solely on context vectors.

The selection of 10 professions follows the convention in gender bias studies of using a multimodal model [4], [6], [18]. Historically male-dominated professions (engineer, mechanic, carpenter, software developer, CEO, doctor) and female-dominated (nurse, secretary, receptionist, teacher) [23], [24].

Table 1. Comparison of Clip Variants

Model	Patch Size	Pretraining Data	Data Size	Developed by
CLIP ViT-B/32	32×32	WIT-400M	400M	OpenAI
CLIP ViT-B/16	16×16	WIT-400M	400M	OpenAI
OpenCLIP ViT-B/32	32×32	LAION-2B	2B	LAION-AI

The implementation follows the standard configuration, with one methodological deviation: context vectors are randomly initialized from a Gaussian distribution (ctx_init=None) rather than the "a photo of a" template. This choice is intentional to avoid prior gender that can be carried from the manual prompt template. The context length is set to M = 16, with a unified context mode (csc=False), and a set of context vectors is studied together for all 10 professions, with the class token positioned at the end (class_token_position='end').

Table 2. Evaluation Conditions

Condition	Text Prompt	Learned?
Baseline (C1)	"a photo of a {profession}"	No
Prompt Engineering (C2)	Balanced ensemble (male/female/professional avg)	No
CoOp (C3)	Learned context (M=16)	Yes
CoOp +DL (C4)	Learned context + Direction Loss	Yes

Training is carried out for 50 epochs with a batch size of 32, using the SGD optimizer (learning rate 2×10^{-3}, weight decay $5\times10-4$, 1 warmup epoch) and the CosineAnnealingLR scheduler (T_{max}=50). For C4, an additional direction loss with a weight of λ is added to the contrastive loss CoOp; the value of $\lambda \in \{0.1, 0.5, 1.0, 2.0, 5.0\}$ is swept for sensitivity analysis. Each configuration is trained with 3 different seeds (42, 123, 999) for variance estimation, resulting in 9 runs for C3 (3 seeds × 3 backbone) and 45 for C4 (5 λ × 3 seeds × 3 backbone).

Each profession is evaluated using top-K retrieval with K=50, equal to the number of ground-truth images per profession (25 male and 25 female in a balanced setting), so that P@K=1.0 can be interpreted as perfect retrieval. From K=50 retrieval results, two types of metrics are calculated separately: P@K, the proportion of retrievals that are the target profession, which measures utility retrieval; and MaxSkew and SignedSkew, which are calculated only on

the correct retrieval subset and measure intrinsic fairness when the profession is identified correctly. This approach differs from standard evaluation measures that skew toward the full top-K [25].

All experiments were run on Google Colab with NVIDIA Tesla T4 GPU with 16 GB VRAM. Software stack: PyTorch 2.10.0, CLIP, OpenCLIP 3.3.0. The total computational budget for the main experiment is 54 CoOp training runs: 9 for C3 and 45 for C4, each with 50 epochs and a batch size of 32 on the training dataset of 80 images.

3. RESULTS AND DISCUSSION

3.1. Results

3.1.1. Overall GBS Comparison

The result of C4 achieved an average GBS reduction of 8.65% across the three backbones, which is close to C2's −9.96% without requiring manual prompt engineering. Per-backbone reduction of −10,34% on ViT-B/32, −5.93% on ViT-B/16, and −9.67% on OpenCLIP ViT-B/32. In ViT-B/32, C4 equals C2 exactly with GBS = 0,5542.

The C3 result yields only a −2.33% average reduction, confirming that a learned prompt without fairness loss is not sufficient. The difference of 6,32 percentage points between C3 and C4 isolates direction loss as a debiasing component that plays an active role.

Table 3. GBS Across Conditions and Models

Backbone	C1	C2	C3	C4	Δ		
					C1→C2	C1→C3	C1→C4
CLIP ViT-B/32	0.6181	0.5542	0.5994	**0.5542**	-10.34%	-3.02%	-10.34%
CLIP ViT-B/32	0.6085	0.5739	0.5911	**0.5724**	-5.69%	-2.86%	-5.93%
OpenClip ViT-B/32	0.6317	0.5441	0.6248	**0.5706**	-13.87%	-1.09%	-9.67%

3.1.2 Bootstrap CI Validation

The bootstrap test (n=1,000, 95% CI) in Table III tests four comparisons of conditions, CoOp vs Baseline Δ(C3−C1), CoOp+DL vs CoOp Δ(C4−C3), CoOp+DL vs Prompt Engineering Δ(C4−C2), and CoOp+DL vs Baseline Δ(C4−C1). Condition Δ(C3−C1) is not significant on 3 backbones; this confirms that CoOp vanilla is not an effective debiasing method. On the other hand, Δ(C4−C3) is significant in 3 backbone isolates, indicating direction loss as a component that actively encourages bias reduction. Δ(C4−C2) is insignificant in 3 backbones, which shows that CoOp+DL is statistically equivalent to the manual prompt ensemble. Δ(C4−C1) showed significance in ViT-B/16 and OpenCLIP ViT-B/32, while in ViT-B/32 the CI includes zero [−0.140, +0.017], likely due to higher sample variability on the backbone. However, the significant Δ(C4−C3) in ViT-B/32 confirms that the direction loss effect remains valid in this backbone.

Table 4. Bootstrap CI for GBS Reduction

Comparison	vitb32	vitb16	OpenClip vitb32
Δ (C2 - C1)	-0.064	-0.034	-0.088
	[-0.138,-0.002]	[-0.079,+0.015]	**[-0.142,-0.034]**
Δ (C3 - C1)	-0.019	-0.018	-0.006
	[-0.076,+0.053]	[-0.056,+0.021]	[-0.081,+0.075]
Δ (C4 - C1)	-0.064	-0.036	-0.061
	[-0.140,+0.017]	**[-0.068,-0.003]**	**[-0.121,-0.002]**
Δ (C4 - C2)	+0.000	-0.003	+0.027
	[-0.041,+0.047]	[-0.054,+0.047] ns	[-0.004,+0.063]
Δ(C4 - C3)	**-0.046**	**-0.019**	**-0.055**
	[-0.082,-0.018]	**[-0.038,-0.001]**	**[-0.106,-0.003]**

Confidence interval 95% delta GBS with bootstrap n=1,000

3.1.3 Per-Profession Analysis

Per-profession analysis shows that C4 (CoOp+DL) gives the biggest improvement in professions with the highest baseline bias, namely CEO, software developer, and receptionist, with the largest MaxSkew decrease of −0.266 for CEO in ViT-B/32 and −0.250 for software developers in ViT-B/32 backbone.

This decrease is consistent across the backbone. There is a local regression that needs to be recognized: among 30 professional pairs across all backbones (10 professions × 3 backbones), 7 experienced flipping in the direction of bias, and 7 experienced an increase in magnitude after C4. Secretary is a consistent anomaly, where MaxSkew worsens in all backbones (+0.164, +0.093, +0.006), with an inverse bias direction on both backbones. Engineers also showed deteriorating backbone-dependent behavior in ViT-B/32 and OpenCLIP ViT-B/32, but improved in ViT-B/16.

Table 5. Per-Profession MaxSkew (C1 vs C4)

Profesi	CLIP VIT-B/32			CLIP VIT-B/16			OpenCLIP ViT-B/32		
	C1 Maxskew	C4 MaxSkew	Δ MaxSkew	C1 Maxskew	C4 MaxSkew	Δ MaxSkew	C1 Maxskew	C4 MaxSkew	Δ MaxSkew
Doctor	0.6667	0.5440	-0.1227	0.6857	0.6258	-0.0599	0.5758	0.5509	-0.0248
Nurse	0.5517	0.5399	-0.0118	0.6563	0.5510	-0.1053	0.7742	0.5450	-0.2292
Engineer	0.5556	0.5617	+0.0062	0.6207	0.5504	-0.0703	0.5385	0.6324	+0.0939
Secretary	0.5185	0.6820	+0.1635	0.5652	0.6578	+0.0926	0.7000	0.7063	+0.0063
Ceo	0.8182	0.5521	-0.2660	0.6250	0.6275	+0.0025	0.7500	0.5741	-0.1759
Teacher	0.5116	0.5205	+0.0089	0.5556	0.5377	-0.0179	0.5000	0.5487	+0.0487
Software Developer	0.7586	0.5088	-0.2498	0.6250	0.5278	-0.0972	0.6757	0.5617	-0.1140
Carpenter	0.6111	0.5321	-0.0791	0.5714	0.5358	-0.0356	0.5952	0.5390	-0.0563
Receptionist	0.6667	0.5937	-0.0730	0.6571	0.5991	-0.0581	0.6857	0.5295	-0.1562
Mechanic	0.5227	0.5072	-0.0155	0.5227	0.5110	-0.0118	0.5217	0.5187	-0.0030

Direction *of MaxSkew Comparison per profession between baseline (C1) and CoOp+DL (C4).*
Low MaxSkew = more balanced retrieval

 https://doi.org/10.58421/misro.v5i2.1674

SignedSkew analysis reveals that some improvements involve changing the direction of bias rather than just reducing its magnitude (Table VI). Out of 30 professional-backbone pairs, 7/30 experienced flipping and 7/30 experienced increased magnitude.

Fig. 1 visualizes this pattern, where color intensity in professions such as CEO and software developer is significantly reduced from C1 to C4, while the secretary maintains high intensity throughout the backbone.

Table 6. SignedSkew per Profession

Profesi	CLIP VIT-B/32			CLIP VIT-B/16			OpenCLIP ViT-B/32		
	C1 SignSkew	C4 SignSkew	Flip/Worse	C1 SignSkew	C4 SignSkew	Flip/Worse	C1 SignSkew	C4 SignSkew	Flip/Worse
Doctor	+0.1667	+0.0429		+0.1857	+0.1237		+0.0758	-0.0102	
Nurse	-0.0517	+0.0392		-0.1563	+0.0524	f	-0.2742	+0.0140	
Engineer	+0.0556	-0.0614	f/w	-0.1207	-0.0508		-0.0385	-0.1316	w
Secretary	-0.0185	+0.1818	f/w	-0.0652	+0.1538	f/w	-0.2000	+0.2077	f/w
Ceo	+0.3182	+0.0500		+0.1250	-0.1078	f	+0.2500	-0.0806	f
Teacher	+0.0116	+0.0207	w	+0.0556	+0.0042		+0.0000	-0.0462	w
Software Developer	+0.2586	-0.0085		+0.1250	-0.0278		+0.1757	-0.0283	
Carpenter	+0.1111	-0.0294		+0.0714	+0.0366		+0.0952	+0.0078	
Receptionist	-0.1667	-0.0926		-0.1571	-0.0962		-0.1857	+0.0000	
Mechanic	+0.0227	+0.0074		+0.0227	+0.0109		+0.0217	+0.0037	

Direction and magnitude of bias per profession in C1 vs C4. Flag 'Flipped' = reverse bias direction. Flag 'Worsened' = worsening magnitude.

3.1.4 Lambda Sensitivity

Table VIII shows that GBS against λ is specific to the backbone. In ViT-B/32 and ViT-B/16, the GBS value decreases monotonically as λ increases, with optimal values at $\lambda = 5.0$ with GBS = 0.5542 and 0.5724. In contrast, OpenCLIP ViT-B/32 shows a U-shaped pattern, with an optimal value at $\lambda = 1.0$ (GBS = 0.5706), followed by an increase at $\lambda \geq 2.0$.

The inter-seed variation across all configurations is relatively small, with a standard deviation of 0.004–0.019. CoOp+DL training is statistically stable. For all λ values tested, C4 is consistently lower than C3. This confirms that direction loss, regardless of the weight, consistently produces a lower bias than CoOp without fairness loss

3.1.5 Utility Preservation

Table 7 shows that C4 maintains, and even increases, retrieval utility across two of the three backbones. In ViT-B/32 and ViT-B/16, P@K increased from C1 to C4 by +0.068 and +0.043. The win-win condition reduces bias as the relevance of the retrieval increases. On average across the backbone, P@K shifts by at least +0.010.

Table 7. Retrieval Accuracy (P@K, K=50)

Backbone	C1	C2	C3	C4	Δ(C1→C4)
CLIP VIT-B/32	0.6180	0.6280	0.6580 ± 0.024	0.6860 ± 0.019	+0.0680
CLIP VIT-B/16	0.6480	0.6640	0.6780 ± 0.023	0.6907 ± 0.011	+0.0427
OpenClip VIT-B/32	0.7140	0.7600	0.6040 ± 0.004	0.6320 ± 0.009	-0.0820
Rata-rata	0.6600	0.6840	0.6467	0.6696	

However, the OpenCLIP ViT-B/32 backbone recorded a decrease in P@K of −0.082. This study hypothesizes that CoOp replaces the default OpenCLIP prompt template that has been optimized on LAION-2B, so that the learned prompt actually interferes with the well-tuned semantic representation.

3.2. Discussion

C4 achieves an average GBS reduction of −8.65%, statistically equivalent to C2 (Δ(C4−C2) ns, 3/3 backbone). This equality is achieved without manual prompt engineering—direction loss is proven to be an active mechanism (Δ(C4−C3) significant, 3/3 backbone), not just a prompt capacity that can be trained. This finding confirms that learned prompts with fairness constraints can replicate the manual approach's performance in scale.

Although the aggregate GBS decreased, per-profession analysis revealed significant heterogeneity: 7/30 of profession-backbone pairs flipped the direction of bias, and 7/30 experienced an increase in magnitude. Secretary worsens consistently in all backbones, indicating overcorrection in professions with a strong female baseline. These findings suggest that a single λ value is not optimal for all professions; future work should explore per-profession adaptive λ values to avoid the trade-off between local fairness and aggregate-level fairness.

Cross-bine variability indicates two key factors, inductive bias architecture and pretraining data distribution. Both of them substantially affect the response to direction loss. In ViT-B/32 and ViT-B/16, GBS decreases until λ = 5.0 and P@K increases. On the other hand, OpenCLIP ViT-B/32 shows a U-shape pattern with an optimal λ = 1.0, accompanied by a decrease in P@K −8.2%, experiencing a dramatic shift to the fairness-win/utility-lose quadrant. This study hypothesizes that the OpenCLIP built-in template prompt has been implicitly optimized for the LAION-2B semantic representation, thereby interfering with CoOp's retrieval utility.

Some limitations need to be recognized openly. First, the evaluation dataset is limited to 500 images and 10 professions; although bootstrap validation (n=1,000) provides a stable

interval estimate, generalization to a wider range of professions and domains requires further validation. Second, this framework assumes binary gender; non-binary identity is not included in the direction vector formulation. Third, the experiment targets only gender attributes; other attributes, such as race, age, and intersectionality, are outside the scope of this study.

Answering RQ1, the integration of gender direction loss achieved an average GBS reduction of −8.65%, significantly better results than the CoOp standard of −2.33% across the backbone. Without direction loss, the learnable prompt itself does not produce a statistically significant reduction in bias, confirming that direction loss is an active debiasing component. RQ2: The influence of λ is backbone-dependent; ViT-B/32 and ViT-B/16 show a monotonic trend with an optimal λ=5.0, while OpenCLIP shows a U-shaped pattern with an optimal λ=1.0. There is no optimal universal λ value for all architectures. RQ3, effectiveness varies substantially across the backbone. ViT-B/32 and ViT-B/16 (pretrained on WIT-400M) achieve win-win conditions, and fairness and utility both increase. On the other hand, OpenCLIP (pretrained on LAION-2B) experienced fairness-win/utility-lose, with a 8.2% decrease in P@K. This finding confirms that λ calibration per backbone is a practical prerequisite before deployment.

.

4. CONCLUSION

This study proposes CoOp with Direction Loss (CoOp+DL) as a scalable approach to mitigate occupational gender bias in CLIP-based image retrieval without retraining the model. By integrating gender direction loss into learnable prompt optimization, the framework achieves an average GBS reduction of −8.65% across three CLIP architectures—statistically equivalent to manual prompt ensemble (C2)—while eliminating the need for manual engineering. Bootstrap confidence intervals (n=1,000) confirm that direction loss is an active debiasing component, not merely prompt capacity. The fairness-utility balance is backbone-dependent: ViT-B/32 and ViT-B/16 achieve win-win conditions with both bias reduction and improved retrieval precision, while OpenCLIP ViT-B/32 experiences a fairness-win/utility-lose trade-off due to interference with its LAION-2B-optimized semantic representations.

Per-profession analysis reveals heterogeneity: while the CEO, software developer, and receptionist show the largest improvements, the secretary exhibits consistent degradation across all backbones, indicating that a single λ value may cause overcorrection for professions with strong gender baselines. The λ sensitivity analysis confirms backbone-specific optimal values (λ=5.0 for ViT-B/32 and ViT-B/16; λ=1.0 for OpenCLIP), so per-backbone λ calibration is a practical prerequisite before deployment. Future work should explore per-profession adaptive λ values, extend to non-binary gender formulations, and broaden attribute coverage to include race and intersectionality.

Given the scale of the evaluation (500 images, 10 professions, binary gender labels, single annotator) and the selection artifact in λ calibration described in Section 3.4, the results reported here should be regarded as a proof-of-concept demonstration rather than a definitive benchmark. CoOp+DL is not presented as ready for deployment. Any operational use of the method should be preceded by: (1) per-backbone λ calibration on held-out

validation data; (2) profession-specific auditing at both the magnitude and direction level (not only aggregate GBS); (3) validation on a larger and more culturally diverse evaluation set; and (4) careful consideration of the binary gender formulation's limitations relative to the deployment context and the population served.

REFERENCES

[1] J. Zhang, J. Huang, S. Jin, and S. Lu, "Vision-language models for vision tasks: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 8, pp. 5625–5644, 2024.

[2] Y. Du, Z. Liu, J. Li, and W. X. Zhao, "A survey of vision-language pre-trained models," *arXiv Prepr. arXiv2202.10936*, 2022.

[3] C. Jia *et al.*, "Scaling up visual and vision-language representation learning with noisy text supervision," in *International conference on machine learning*, PMLR, 2021, pp. 4904–4916.

[4] C. Wen, Z. Peng, Y. Huang, X. Yang, and W. Shen, "Domain generalization in clip via learning with diverse text prompts," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025, pp. 9559–9569.

[5] K. Hamidieh, H. Zhang, W. Gerych, T. Hartvigsen, and M. Ghassemi, "Identifying implicit social biases in vision-language models," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 2024, pp. 547–561.

[6] Z. Al Sahili, I. Patras, and M. Purver, "Data Matters Most: Auditing Social Bias in Contrastive Vision Language Models," *arXiv Prepr. arXiv2501.13223*, 2025.

[7] C.-Y. Chuang, V. Jampani, Y. Li, A. Torralba, and S. Jegelka, "Debiasing vision-language models via biased prompts," *arXiv Prepr. arXiv2302.00070*, 2023.

[8] J. Gu *et al.*, "A systematic survey of prompt engineering on vision-language foundation models," *arXiv Prepr. arXiv2307.12980*, 2023.

[9] H. Jung, T. Jang, and X. Wang, "A unified debiasing approach for vision-language models across modalities and tasks," *Adv. Neural Inf. Process. Syst.*, vol. 37, pp. 21034–21058, 2024.

[10] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," *Int. J. Comput. Vis.*, vol. 130, no. 9, pp. 2337–2348, 2022.

[11] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv Prepr. arXiv2010.11929*, 2020.

[12] S. Pratt, I. Covert, R. Liu, and A. Farhadi, "What does a platypus look like? generating customized prompts for zero-shot image classification," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 15691–15701.

[13] B. Zhu, Y. Niu, Y. Han, Y. Wu, and H. Zhang, "Prompt-aligned gradient for prompt tuning," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 15659–15669.

[14] A. Radford *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*, PmLR, 2021, pp. 8748–8763.

[15] S. C. Geyik, S. Ambler, and K. Kenthapadi, "Fairness-aware ranking in search & recommendation systems with application to linkedin talent search," in *Proceedings of the 25th acm sigkdd international conference on knowledge discovery & data mining*, 2019, pp. 2221–2231.

[16] A. Seth, M. Hemani, and C. Agarwal, "Dear: Debiasing vision-language models with additive residuals," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 6820–6829.

[17] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM Comput. Surv.*, vol. 54, no. 6, pp. 1–35, 2021.

[18] H. Berg, S. Hall, Y. Bhalgat, H. Kirk, A. Shtedritski, and M. Bain, "A prompt array keeps the bias away: Debiasing vision-language models with adversarial learning," in *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2022, pp. 806–822.

[19] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to information retrieval*. Cambridge university press, 2008.

[20] A. Kumar, A. Raghunathan, R. Jones, T. Ma, and P. Liang, "Fine-tuning can distort pretrained features and underperform out-of-distribution," *arXiv Prepr. arXiv2202.10054*, 2022.

[21] T. Wang, J. Zhao, M. Yatskar, K.-W. Chang, and V. Ordonez, "Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 5310–5319.

[22] Y. Xian, C. H. Lampert, B. Schiele, and Z. Akata, "Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 9, pp. 2251–2265, 2018.

[23] S. Muradova and W. H. Seitz, "Gender discrimination in hiring: Evidence from an audit experiment in Uzbekistan," The World Bank, 2021.

[24] S. Y. Park and E. Oh, "Getting a foot in the door: A meta-analysis of us audit studies of gender bias in hiring," *Sociol. Sci.*, vol. 12, pp. 26–50, 2025.

[25] M. Hall, L. Gustafson, A. Adcock, I. Misra, and C. Ross, "Vision-language models performing zero-shot tasks exhibit disparities between gender groups," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 2778–2785.