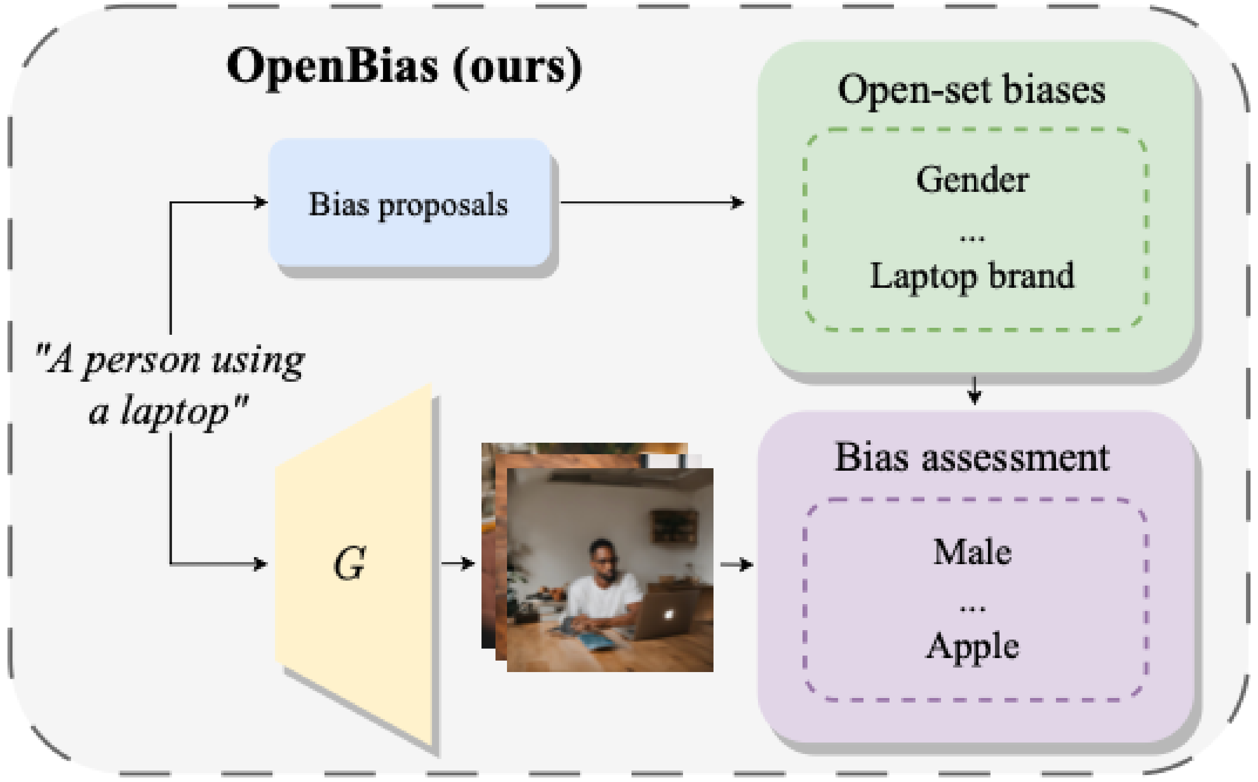
# CLIP Debiasing

Computer Vision
A.Y. 2024/2025

**Andrea Baldi**
id 1966232

# Bias Definition

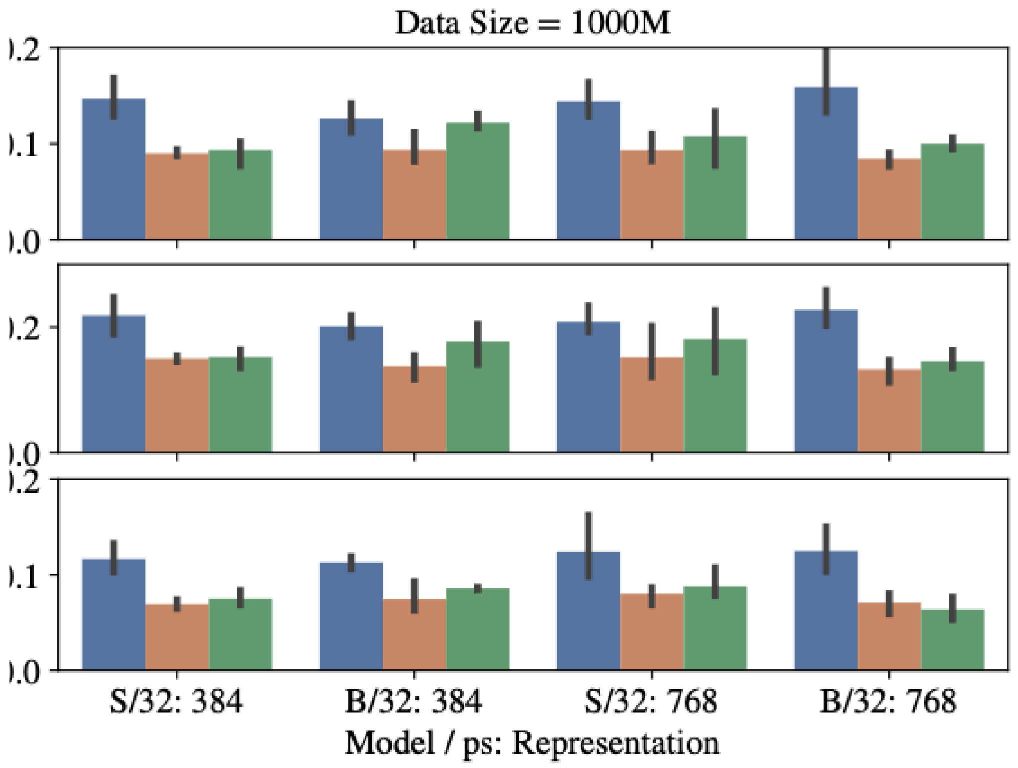
Before Deepening in Clip Debiasing, a definition of Bias is needed, in this project we specifically refer to these two types of Bias:

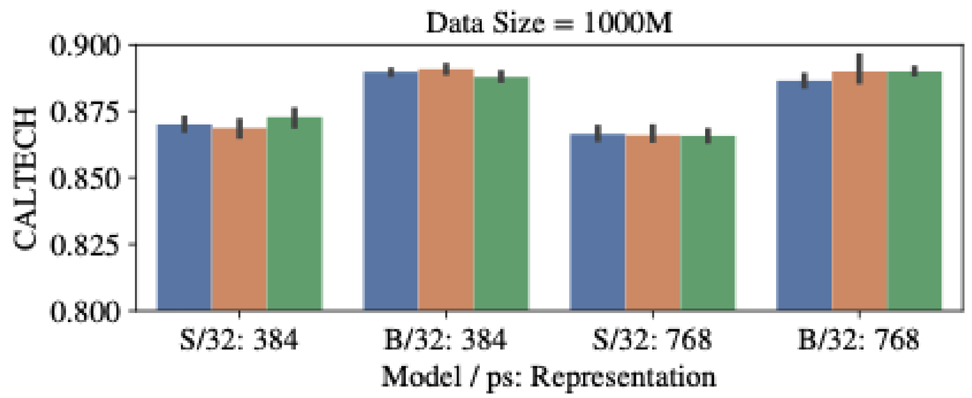



- **Representation Bias (RB)**: it measures whether certain groups are over- or under-represented compared to an ideal reference (e.g., uniform or demographic parity).
- **Association Bias (AB)**: The degree to which a model's predictions are spuriously correlated with sensitive attributes. It quantifies how much the probability of a non-sensitive label (e.g., an occupation) changes across sensitive groups, indicating learned associations that reflect societal stereotypes.

**In this project we focus on AB on gender/profession relationships**

# Clip Bias




Several papers has shown what types of **Bias affect CLIP** :

**Representation Bias : E[p(man) − p(woman)] = 0.20**
**Association Bias: |p(y|man) − p(y|woman)|= 0.20**
*(with y profession)*

Additional consideration should be made: Aggressively Debiasing CLIP may lead to a **decrease of zero-shot and inference capabilities**.

Finding the **trade-off** between these two values is crucial.

# Debising CLIP SOTA

The current state of the art in debiasing CLIP is **Multi-Modal Moment Matching (M4)** by DeepMind (ICLR 2024).
It reweights training samples to align both first-order (representation) and second-order (association) statistics across modalities.

**M4** reduces representation bias of:

**RB**: mean parity from ~0.20 to ~0.05

**AB**: FairFace 38.8 → 29.9, MIAP 28.4 → 20.5 while slightly improving performance (ImageNet 0-shot 77.0 → 77.5, COCO retrieval@5 86% → 87%).

$$\underset{\mathbb{E}[\mathbf{q}]=\eta \,\wedge\, 0\leq\mathbf{q}\leq Q}{\text{minimize}} \quad \left\{ \frac{1}{2}\mathbb{E}_{\mathcal{D}}\left[\mathbf{u}\cdot(\mathbf{q}-\eta)^2\right] + V\cdot\left(\sum_{k\in[m]} l_k^R + \sum_{k\in[m],\, r\in[c]} l_{k,r}^D\right)\right\},$$

# METHODS DEEPENING

01

## PROMPT BANKING

An hand-crafted and scraped Dataset has no corpus but just gender and profession values. Assuring a good corpus through Prompt Banking reduces the overfitting issue

02

## Loss Redefinition

As stated before, the goal is to reduce bias without dropping zero-shot performances so Loss is rewritten as follows:

LOSS = BIAS_LOSS + λ * ANCHOR_LOSS

03

## LoRa

LoraConfig =

- r=8,
- lora_alpha=16,
- target_modules=["q_proj", "k_proj", "v_proj", "o_proj"],
- lora_dropout=0.1,
- bias="none",

# Common Dataset

**01**

## FairFace

This dataset has been used just for the evaluation.  It is our dataset to evaluate Representation Bias.

**02**

## OxfordIIITPet

This dataset has been used to evaluate zero-shot capacity.

**03**

## COCO

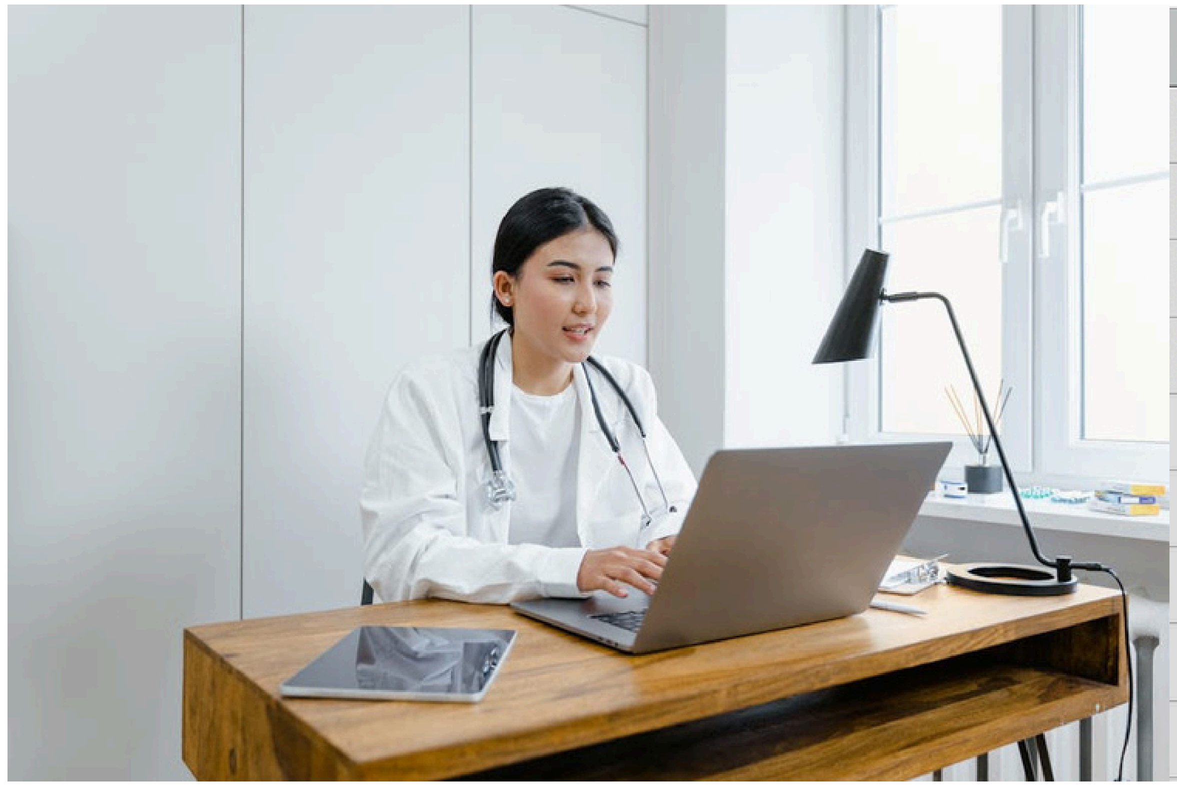
This dataset was considered since it provides good notations and image with context, in the end, it has not been included due to a high RB and few gender releted annotations

# Dataset Hand-Crafted

04

## Pexels

This dataset contains 180 elements and has been scraped. It contains gender and profession information to be embedded with prompt banking. It shows clearly profession and gender



| file_name | profession | query_gender |
|---|---|---|
| engineer_male_34 | engineer | male |
| engineer_male_34 | engineer | male |
| engineer_male_75 | engineer | male |
| engineer_male_92 | engineer | male |
| engineer_male_38 | engineer | male |
| engineer_male_39 | engineer | male |
| engineer_male_89 | engineer | male |
| engineer_male_38 | engineer | male |
| engineer_male_67 | engineer | male |

| file_name | profession | query_gender |
|---|---|---|
| fe_01.jpeg | engineer | female |
| fe_02.jpeg | engineer | female |
| fe_03.jpeg | engineer | female |
| fe_04.jpeg | engineer | female |
| fe_05.jpeg | engineer | female |
| fe_06.jpeg | engineer | female |
| fe_07.jpeg | engineer | female |
| fe_08.jpeg | engineer | female |
| fe_09.jpeg | engineer | female |
| fe_10.jpeg | engineer | female |
| fe_11.jpeg | engineer | female |

05

## challenge

This dataset is purely hand-crafted, it contains 180 images as Pexels but images has a higher context difficulty.

# EXPERIMENTAL SETUP 1/2

01

## Model

**Environment:**

- Google Colab + GPU (CUDA)
- Libraries: transformers, peft, torchvision, scikit-learn, datasets

**Model:**

- Base: openai/clip-vit-base-patch32
- Fine-tuned via LoRA (Low-Rank Adaptation) applied to the text encoder
- Vision encoder frozen to preserve visual embeddings

**LoRa:**

- Applied to the text encoder of CLIP to adapt language embeddings for fairness objectives.
- The vision encoder remains frozen, preserving pretrained visual representations.
- Enables bias correction without degrading CLIP's zero-shot generalization.

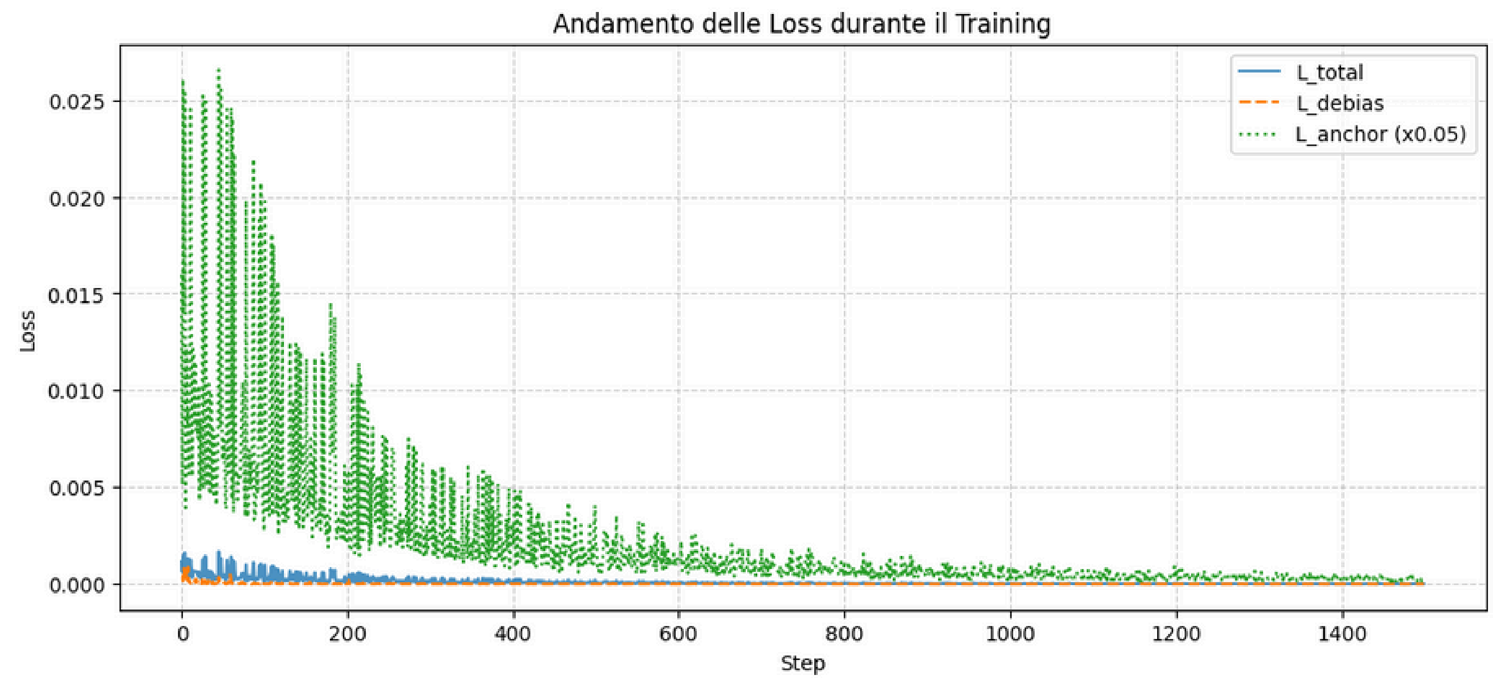
# EXPERIMENTAL SETUP 2/2

02

## TRAINING

**Training procedure**:

- Compute anchor embeddings (mean text features per profession)
- Dual-loss objective:
  - L_debias: enforces equal distance between male/female prompt embeddings
  - L_anchor: prevents semantic drift from class anchors
- Total Loss: L_total = L_debias + λ * L_anchor
- Optimizer: AdamW
- Batch size: 64, random prompt sampling per step



# Evaluation Pre / Post FineTune

Evaluation in computed over : **RB (FairFace), AB (Pexels and Challenge) and zero-shot (OxfordIIITPet)**.
It is computed firstly o**n original CLIP, then on Fine-Tuned One**.

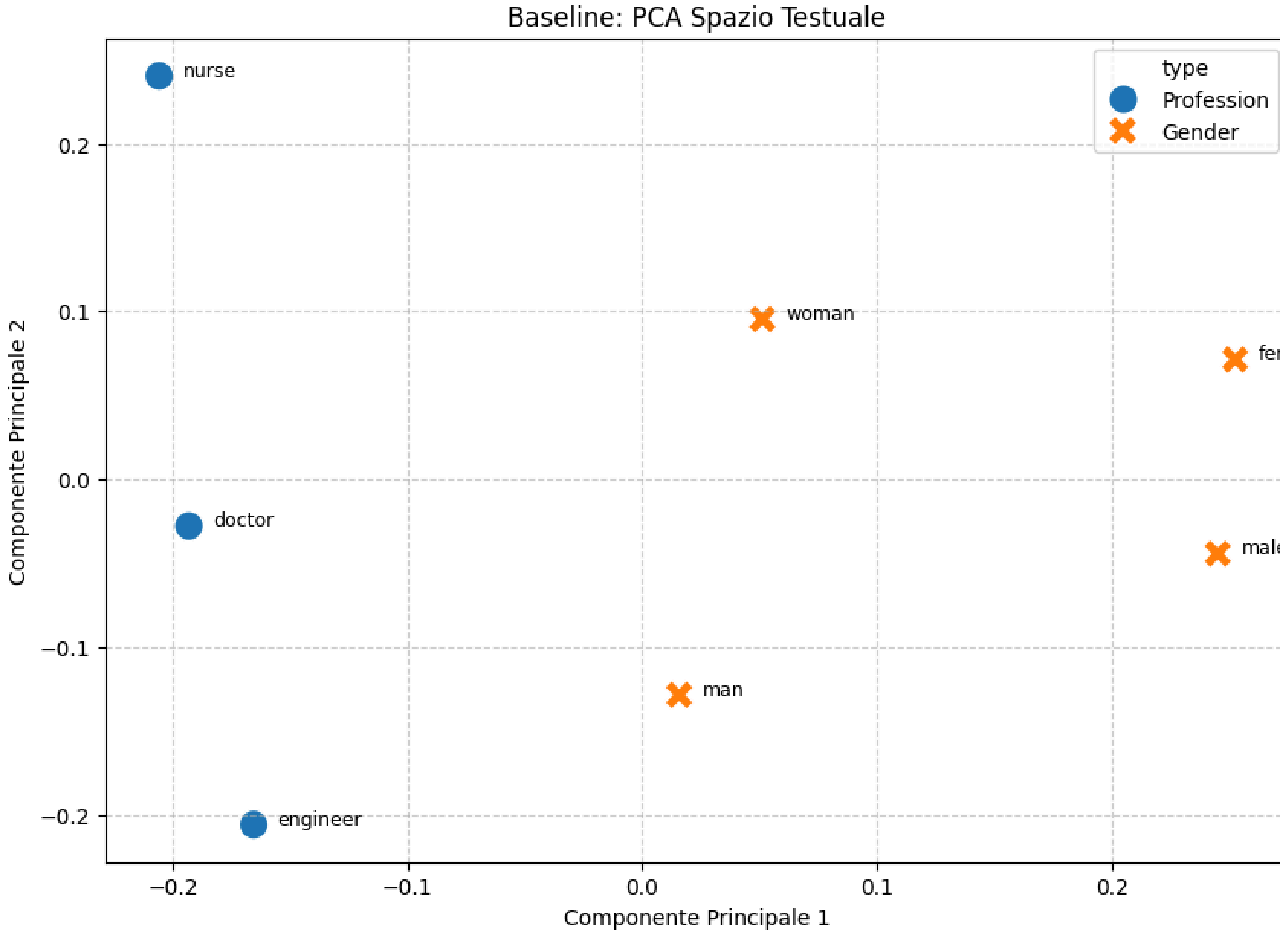
## 01 Original

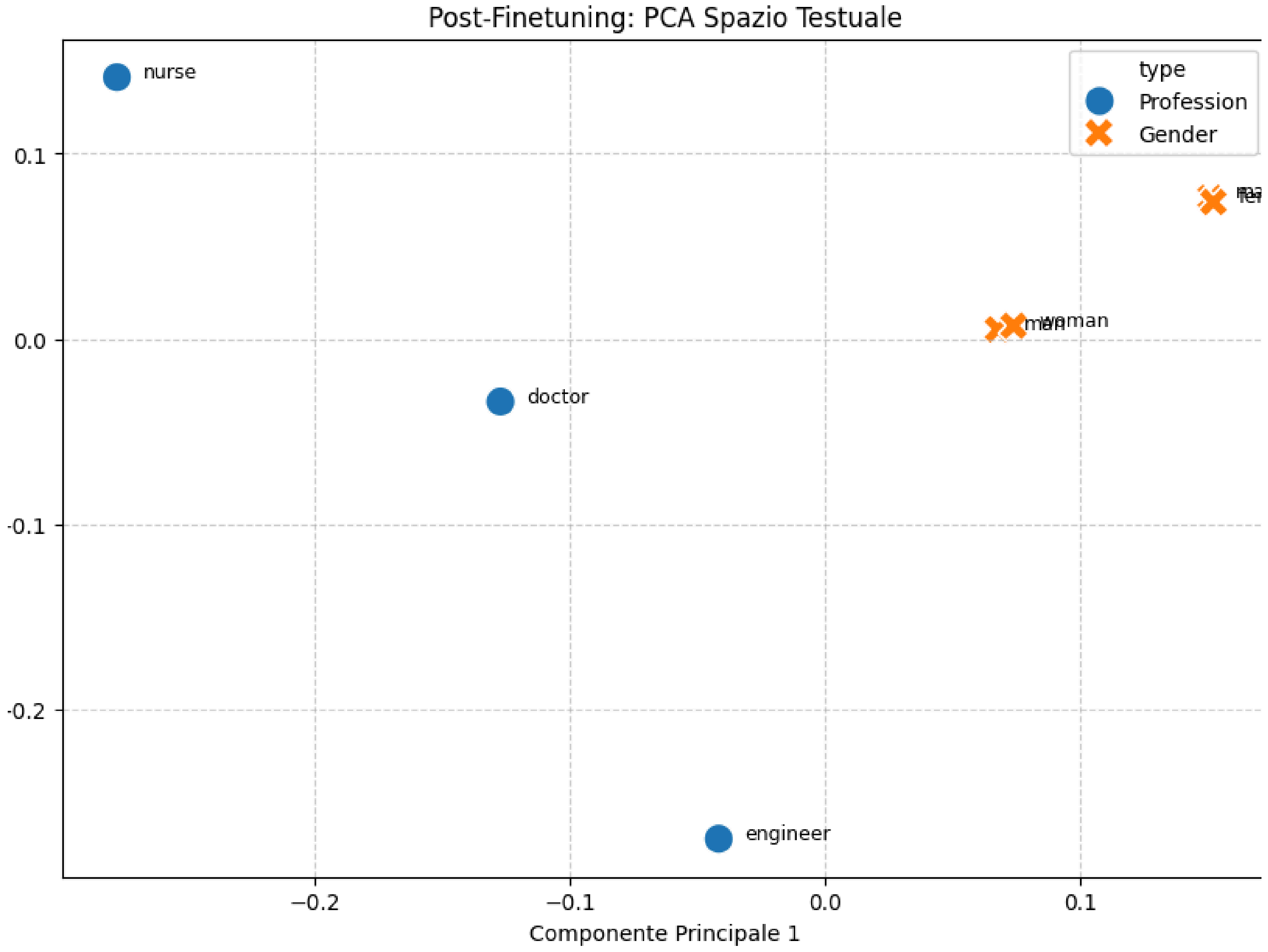
Baseline Results:
"**pets**": {
"top1_accuracy": 0.8765, "top5_accuracy": 0.9931}
"**pexels_bias**": {
"overall_abs_bias":0.0196 },
"**fairface_bias**": {
"Black": 0.8881,
"Southeast Asian": 0.9222,
"East Asian": 0.9367,
"Indian": 0.93865,
"Latino_Hispanic": 0.9556,
"White": 0.9568,
"Middle Eastern": 0.9677 },

## 02 Fine-Tuned

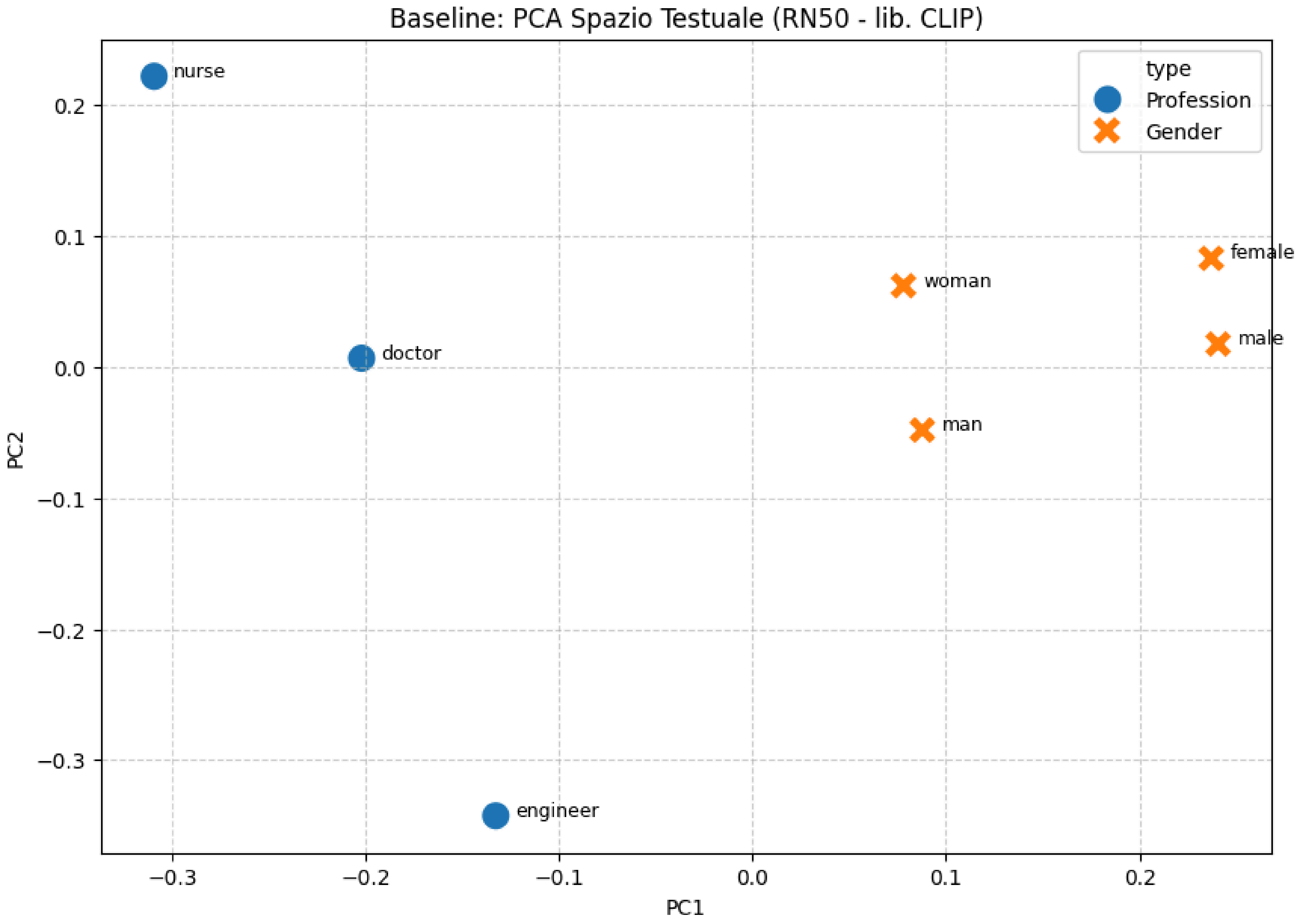
Fine_Tuned Results:
**pets**: Top-1=0.8689, Top-5=0.9929
**pexels_bias** = Overall Absolute Bias Score: 0.0173
**fairface_bias**:
Black 0.890746
Southeast Asian 0.923675
Indian 0.939314
East Asian 0.940645
Latino_Hispanic 0.949476
White 0.957794
Middle Eastern 0.964433

# PCA ANALYSIS - PRE/POST FINE-TUNE



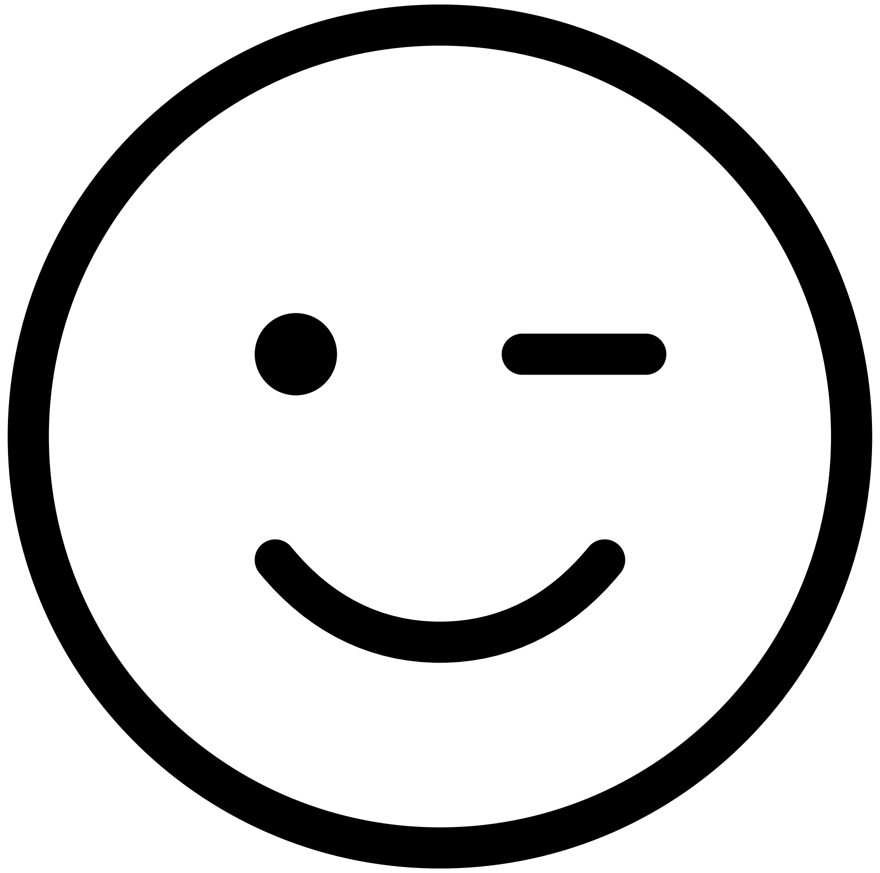



# PCA ANALYSIS - RN50



# BIBLIOGRAPHY


- [1] I. Alabdulmohsin, X. Wang, A. Steiner, P. Goyal, A. D’Amour, X. Zhai, “CLIP the Bias: How Useful is Balancing Data in Multimodal Learning?”, Proceedings of ICLR 2024, Google DeepMind.
- [2] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, “Learning Transferable Visual Models From Natural Language Supervision,” ICML 2021 (OpenAI CLIP Paper).
- [3] M. D’Incà, E. Peruzzo, M. Mancini, D. Xu, V. Goel, X. Xu, Z. Wang, H. Shi, N. Sebe, “OpenBias: Open-set Bias Detection in Text-to-Image Generative Models,” arXiv preprint arXiv:2404.07990v2, August 2024.
- [4] (Debiasing general) – No specific metadata found in the provided PDF “debiasing.pdf.” Presumable citation: Anonymous, “Bias Mitigation in Multimodal and Vision-Language Models,” Technical Report, 2023.
- [5] (Debiasing2 general) – No metadata provided; suggested placeholder: Anonymous, “Methods for Debiasing Multimodal Models,” Internal Study, 2023.



# CLIP Debiasing

Computer Vision
A.Y. 2024/2025

**Andrea Baldi**
id 1966232