# Test-Time Debiasing with Probabilistic Prompts via Wasserstein Distance in Vision-Language Models

Chengye Wang, Yuyuan Li, Xiaohua Feng, Xiaolin Zheng*, Chaochao Chen
ICML 2026

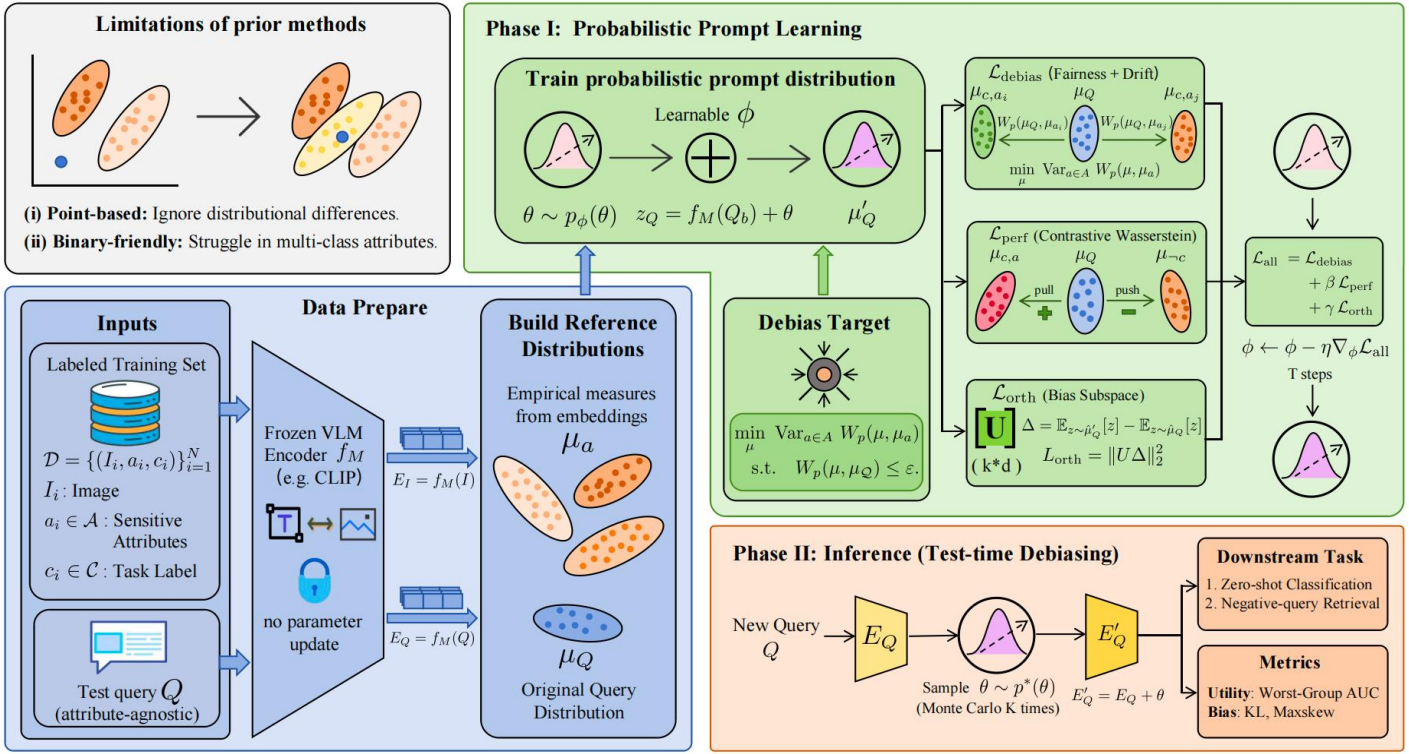
# Background

Vision-language models inherit social bias from large-scale pretraining and can skew retrieval even for attribute-agnostic queries.
Existing test-time debiasing is lightweight, but most methods correct a single point and weaken in multi-class settings.

## What makes W4D different?

1. Treat fairness as distribution alignment rather than point correction.
2. Use Wasserstein distance to capture both location and geometry of group clusters.
3. Naturally extends test-time debiasing to multi-class attributes such as race and skin tone.
4. Preserve semantics with contrastive and orthogonality regularization while keeping CLIP frozen.

# W4D Overview



# Key Formulation

## Distributional fairness

Debias target:
min Var_a W_p(mu, mu_a)  s.t. W_p(mu, mu_Q) <= eps
Equalize group distances while limiting semantic drift.

## Semantic preservation

L_perf pulls queries toward concept-consistent groups
and pushes them away from concept-inconsistent negatives.
L_orth keeps the update off the bias subspace U.

## Prompt learning

Optimize L_all = L_debias + beta L_perf + gamma L_orth.
Learn a prompt distribution p(theta) once.
At test time, sample K prompt shifts with no CLIP update.

# Experimental Setup

## Data, model, task

Datasets: FairFace, CelebA, UTKFace, Facet

Sensitive attributes: gender, race, skin tone

Backbones: CLIP-ViT-B/16 and CLIP-ViT-L/14

Tasks: zero-shot classification and stereotype-query retrieval

# Evaluation Protocol

## Metrics, baselines, goal

Utility: Worst-Group AUC-ROC

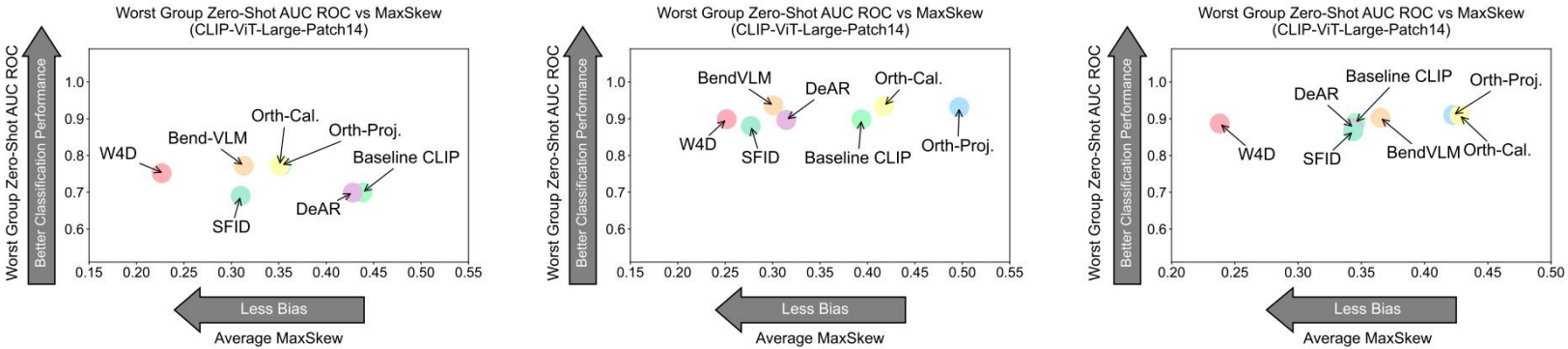
Bias: KL divergence and MaxSkew on top-m retrieved results

Baselines: CLIP, DeAR, Orth-Proj., Orth-Cal., BendVLM, SFID

Goal: better fairness-utility trade-off without finetuning the VLM

# Fairness-Utility Frontier

W4D achieves a better fairness–utility trade-off than prior test-time debiasing methods, reducing MaxSkew while maintaining competitive worst-group zero-shot AUC across all evaluated settings.




*(a)* CelebA – Gender *(b)* Facet – Gender *(c)* Facet – Skin

*Figure 2.* Comparison of W4D and prior test-time debiasing methods on CLIP-ViT-Large-Patch14.

# Negative-query Retrieval Results

W4D consistently achieves the lowest KL divergence and MaxSkew on negative-query retrieval, showing stronger debiasing performance across gender and race settings.

*Table 1.* Debiasing UTKFACE and FAIRFACE dataset with respect to `gender` and `race` for STEREOTYPE queries.

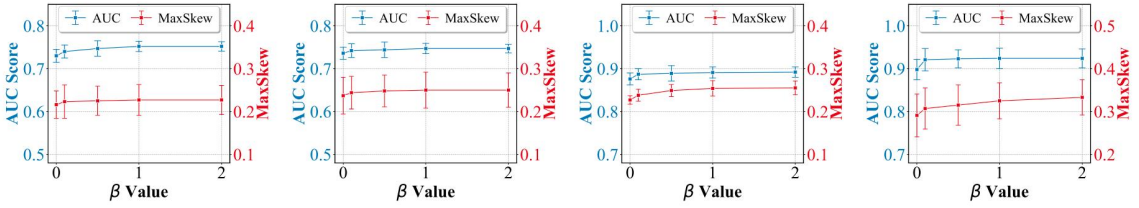
| Attribute | Method | UTKFACE | | | | FAIRFACE | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | CLIP-ViT-B-P16 | | CLIP-ViT-L-P14 | | CLIP-ViT-B-P16 | | CLIP-ViT-L-P14 | |
| | | KL Div.↓ | MaxSkew↓ | KL Div.↓ | MaxSkew↓ | KL Div.↓ | MaxSkew↓ | KL Div.↓ | MaxSkew↓ |
| Gender | Baseline CLIP | 0.1575 ± 0.0131 | 0.8204 ± 0.0571 | 0.0738 ± 0.0065 | 0.4757 ± 0.0290 | 0.1179 ± 0.0101 | 0.6662 ± 0.0444 | 0.0813 ± 0.0079 | 0.5011 ± 0.0377 |
| | Orth-Proj. | 0.1470 ± 0.0036 | 0.7092 ± 0.0148 | 0.0351 ± 0.0074 | 0.2981 ± 0.0413 | 0.3155 ± 0.0115 | 1.3983 ± 0.0612 | 0.0257 ± 0.0014 | 0.2005 ± 0.0150 |
| | Orth-Cal. | 0.2048 ± 0.0040 | 0.9348 ± 0.0184 | 0.0181 ± 0.0033 | 0.1776 ± 0.0267 | 0.4043 ± 0.0106 | 1.8997 ± 0.0523 | 0.0373 ± 0.0046 | 0.2422 ± 0.0159 |
| | BendVLM | 0.0096 ± 0.0022 | 0.1248 ± 0.0147 | 0.0182 ± 0.0054 | 0.1756 ± 0.0325 | 0.0267 ± 0.0030 | 0.2063 ± 0.0068 | **0.0057 ± 0.0013** | **0.0952 ± 0.0166** |
| | DeAR | 0.1315 ± 0.0629 | 0.7020 ± 0.2855 | 0.0735 ± 0.0057 | 0.4745 ± 0.0252 | 0.0861 ± 0.0109 | 0.4853 ± 0.0618 | 0.0667 ± 0.0221 | 0.4234 ± 0.1168 |
| | SFID | 0.0391 ± 0.0102 | 0.3074 ± 0.0555 | 0.0039 ± 0.0014 | 0.0790 ± 0.0186 | 0.0548 ± 0.0123 | 0.3729 ± 0.0587 | 0.0368 ± 0.0087 | 0.2547 ± 0.0403 |
| | W4D | **0.0048 ± 0.0038** | **0.0833 ± 0.0401** | **0.0036 ± 0.0042** | **0.0674 ± 0.0479** | **0.0150 ± 0.0091** | **0.1614 ± 0.0702** | 0.0097 ± 0.0072 | 0.1324 ± 0.0712 |
| Race | Baseline CLIP | 0.1227 ± 0.0059 | 1.0074 ± 0.0615 | 0.0984 ± 0.0096 | 0.8773 ± 0.0655 | 0.1882 ± 0.0067 | 1.3405 ± 0.0625 | 0.1865 ± 0.0120 | 1.2872 ± 0.0733 |
| | Orth-Proj. | 0.1794 ± 0.0258 | 1.4795 ± 0.2020 | 0.1246 ± 0.0132 | 1.3198 ± 0.1190 | 0.2624 ± 0.0240 | 1.4336 ± 0.0807 | 0.1617 ± 0.0092 | 1.1865 ± 0.0438 |
| | Orth-Cal. | 0.1739 ± 0.0258 | 1.3721 ± 0.1885 | 0.1150 ± 0.0122 | 1.1798 ± 0.1081 | 0.2558 ± 0.0227 | 1.4099 ± 0.0541 | 0.1582 ± 0.0094 | 1.2596 ± 0.0592 |
| | BendVLM | 0.0724 ± 0.0094 | 0.7140 ± 0.0794 | 0.0629 ± 0.0154 | 0.6207 ± 0.0546 | 0.0879 ± 0.0091 | 0.8586 ± 0.1161 | 0.0946 ± 0.0050 | 0.9714 ± 0.0501 |
| | DeAR | 0.1160 ± 0.0093 | 0.9568 ± 0.0599 | 0.0931 ± 0.0187 | 0.8563 ± 0.0916 | 0.1738 ± 0.0257 | 1.2641 ± 0.1593 | 0.1780 ± 0.0267 | 1.2581 ± 0.1199 |
| | SFID | 0.0838 ± 0.0072 | 0.8302 ± 0.0435 | 0.0387 ± 0.0110 | 0.5419 ± 0.0991 | 0.1360 ± 0.0048 | 1.0626 ± 0.0589 | 0.1420 ± 0.0062 | 1.1098 ± 0.0334 |
| | W4D | **0.0201 ± 0.0079** | **0.3312 ± 0.1075** | **0.0218 ± 0.0008** | **0.3133 ± 0.0378** | **0.0383 ± 0.0080** | **0.5083 ± 0.0881** | **0.0466 ± 0.0113** | **0.6639 ± 0.1178** |

# Sensitivity Analysis

Increasing β improves utility but may slightly increase bias.

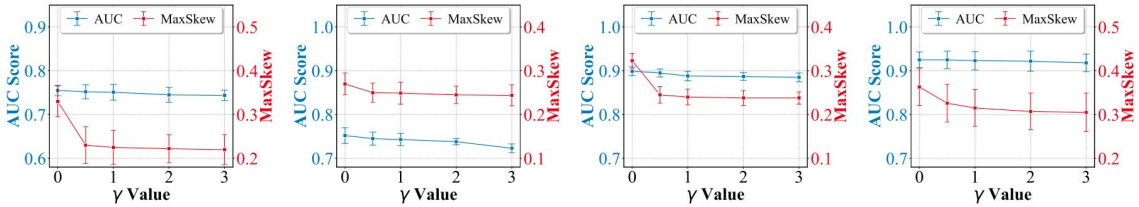
Increasing γ reduces MaxSkew with a modest AUC trade-off.

Larger K provides more stable performance.



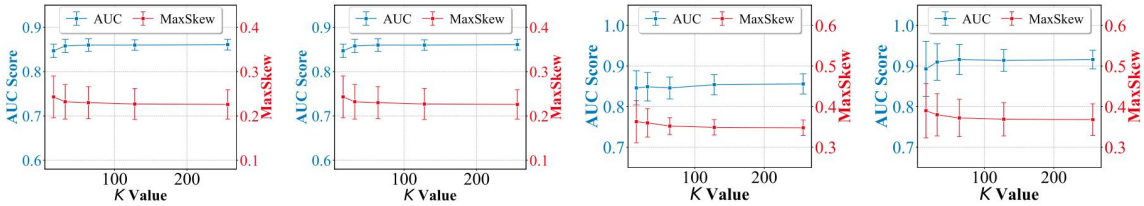

(a) CelebA-Gender-CLIP14 (b) CelebA-Gender-CLIP16 (c) Facet-Skin-CLIP14 (d) Facet-Skin-CLIP16

*Figure 4.* Effect of $\beta$ on debiasing and performance.




(a) CelebA-Gender-CLIP14 (b) CelebA-Gender-CLIP16 (c) Facet-Skin-CLIP14 (d) Facet-Skin-CLIP16

*Figure 5.* Effect of $\gamma$ on debiasing and performance.




(a) CelebA-Gender-CLIP14 (b) CelebA-Gender-CLIP16 (c) Facet-Skin-CLIP14 (d) Facet-Skin-CLIP16

*Figure 6.* Effect of Monte Carlo sample size $K$ on debiasing and performance.

## Takeaways

1. W4D reframes test-time debiasing from point correction to distribution alignment.
2. It is especially strong on multi-class attributes, where binary-friendly methods degrade.
3. Probabilistic prompts expose a practical fairness / utility / stability trade-off without updating CLIP.

Codebase: https://github.com/QDRhhhh/W4D
Contact: wangchengye@zju.edu.cn