Selective Test-Time Debiasing for CLIP via Reward Gating

Anonymous ACL submission

Abstract

Vision language models (VLMs) demonstrate strong zero-shot performance, but often perpetuate social stereotypes in person-centric queries, yielding skewed demographic distributions. Current debiasing methods apply uniform bias corrections across all input queries regardless of their bias sensitivity, creating a fundamental fairness–utility trade-off. Strong debiasing distorts semantically meaningful information in bias-insensitive queries, while weak debiasing fails to mitigate stereotypes in bias-sensitive ones. This one-size-fits-all approach hampers simultaneously achieving high utility on bias-insensitive queries and fairness on bias-sensitive queries. We introduce **Reward-Gated Test-Time Adaptation (RG-TTA)**, a reinforcement learning-based test-time adaptation framework that selectively applies debiasing based on input sensitivity. RG-TTA adaptively triggers fairness regularization based on the bias sensitivity of each input during test-time policy adaptation, while focusing exclusively on optimizing cross-modal alignment for bias-insensitive inputs. Experiments on fairness benchmarks (e.g., FairFace, UTKFace) demonstrate substantial bias reduction while simultaneously improving zero-shot utility, resolving the trade-off of uniform debiasing. Code will be made publicly available.

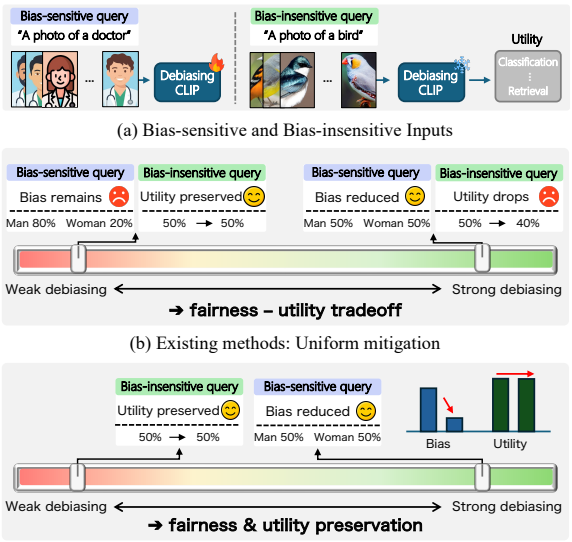

Figure 1: (a) We categorize inputs into **bias-sensitive** and **bias-insensitive**, where only the former requires debiasing intervention. (b) Existing methods apply **uniform mitigation**, creating a structural trade-off: weak debiasing retains bias in sensitive queries (left), while strong debiasing distorts insensitive queries, degrading utility (right). (c) Our approach employs **selective mitigation via reward-gating**, which applies strong debiasing only to bias-sensitive inputs while preserving insensitive ones, ensuring both fairness and utility.

1 Introduction

Vision Language Models (VLMs) have demonstrated exceptional zero-shot capabilities across a wide range of multimodal tasks (Deng et al., 2009; Plummer et al., 2015), reaching the stage of real-world applications. By learning joint representations from web-scale image-text pairs, these models achieve strong cross-modal alignment without task-specific fine-tuning. However, this same training paradigm causes VLMs to internalize social stereotypes (Birhane et al., 2021) present in their training data, leading to biased outputs that reflect and potentially amplify social prejudices (Hall et al., 2023; Hamidieh et al., 2024; Janghorbani and De Melo, 2023; Zhao et al., 2021). These biases manifest most critically in person-centric queries, where models produce skewed demographic distributions. For instance, querying "a photo of a doctor" yields disproportionately male images, or certain occupations become strongly associated with specific racial groups. Such behavior poses serious risk of reinforcing discriminatory decision-making.

Existing debiasing approaches for VLMs (Wang et al., 2021b; Chuang et al., 2023; Zhang et al., 2025) share common design philosophy: they ap-

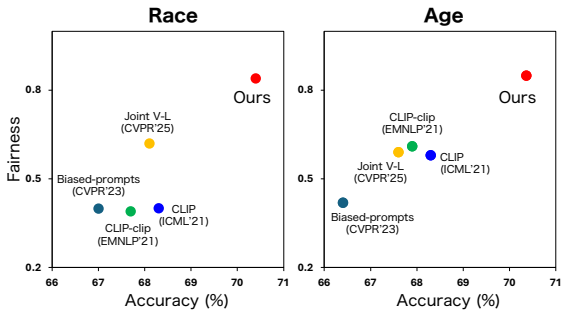

Figure 2: Fairness versus utility for Race and Age. **Accuracy** is measured as ImageNet zero-shot top-1 accuracy (%), and **Fairness** is measured as $1 - \text{MaxSkew@1000}$[1] (higher is better). Existing query-independent debiasing baselines (Chuang et al., 2023; Wang et al., 2021b; Zhang et al., 2025) exhibit a fairness–utility trade-off, whereas our method improves fairness while achieving higher accuracy.

ply fixed bias correction uniformly across all language queries, regardless of whether individual queries are sensitive to demographic biases. Although conceptually simple, this uniform mitigation strategy causes fundamental fairness–utility trade-off that we illustrated in Figure 1(b). Here, we use utility to refer to general-purpose zero-shot performance on downstream tasks (e.g., image classification and cross-modal retrieval).

For bias-sensitive queries, model's predictions are often entangled with demographic attributes, and strong debiasing is necessary for fair outcomes. In contrast for bias-insensitive queries, ground-truth semantics are largely orthogonal to demographic attributes, and the model's original predictions already reflect accurate cross-modal alignment. When a uniform debiasing framework is applied to both types of queries, one of two failure modes inevitably occurs. As we demonstrate in Figure 2, this inflexibility creates a trade-off where existing methods must compromise between fairness in sensitive queries and utility in general tasks, unable to excel at both simultaneously.

We believe that the aforementioned limitation is a consequence of the query-independent design paradigm. Since uniform debiasing methods cannot distinguish between inputs that require mitigation and those that do not, they are constrained to operate in a compromise regime. This observation motivates a paradigm shift toward adaptive debiasing that selectively activates mitigation based on the bias sensitivity of each input. To this end, we propose **Reward-Gated Test-Time Adaptation for CLIP (RG-TTA)**, a reinforcement learning (RL)-based framework designed for selective debiasing. Our key insight is that debiasing should be treated as a per-query decision rather than global transformation, enabling the model to adapt its behavior dynamically based on input characteristics. As illustrated in Figure 3, RG-TTA operates through an episodic test-time adaptation protocol. For each incoming query, we first assess its bias sensitivity by quantifying the alignment discrepancy between the query semantics and a set of demographic attributes. Based on this assessment, an adaptive reward-gating strategy dynamically triggers the fairness-regularized objective only for bias-sensitive queries, ensuring the preservation of the original cross-modal alignment for neutral inputs.

Empirical evaluations on multiple fairness benchmarks–including FairFace (Kärkkäinen and Joo, 2021), UTKFace (Zhang et al., 2017), and the challenging FACET (Gustafson et al., 2023) dataset–demonstrate that RG-TTA significantly reduces social bias across various demographic attributes. Notably, our framework effectively resolves the fairness–utility trade-off by achieving substantial bias reduction alongside higher accuracy on tasks such as ImageNet-1K (Deng et al., 2009) compared to existing query-independent baselines. By performing episodic optimization with this gated reward, RG-TTA provides a practical design principle for mitigating bias without broadly disrupting the alignment of vision-language models.

2 Related Work

Social debiasing in vision-language models. Social debiasing for CLIP-style VLMs is broadly categorized into (i) *training-based* approaches (Alabdulmohsin et al., 2024; Hirota et al., 2025; Zhang et al., 2025), which suppress sensitive-attribute signals by introducing additional objectives or modules during learning, and (ii) *training-free* approaches (Chuang et al., 2023; Gerych et al., 2024), which keep the foundation model fixed and apply post-hoc adjustments to embeddings or outputs. The former includes joint debiasing methods that align and remove biases across both modalities, while the latter estimates bias directions from attribute prompts and removes them via lightweight projections. However, many existing

[1]MaxSkew@1000 is computed from the protected-attribute distribution within the top-1000 retrieved samples for neutral queries; see Sec. 3 for details.

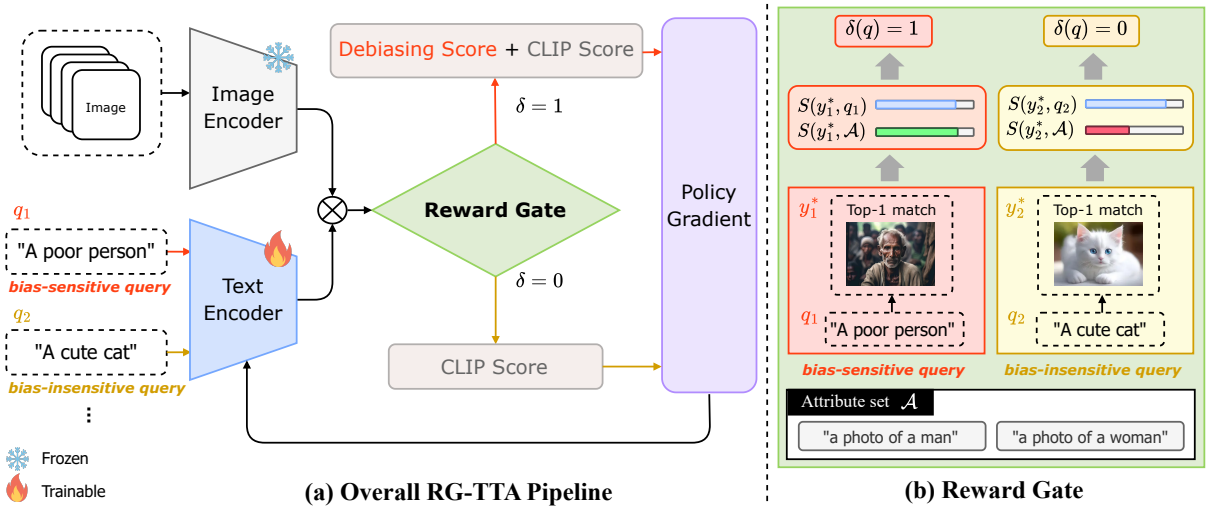

Figure 3: Overview of RG-TTA: RL-based episodic test-time adaptation with **reward gating** via the indicator $\delta(q)$. For each query, we update only the query-modality encoder (text encoder for text queries; image encoder for image queries) with a few policy-gradient steps on a truncated top-K candidate set. The indicator $\delta(q)$ controls the episode reward: when $\delta(q) = 0$, we optimize an alignment-only reward; when $\delta(q) = 1$, we add an attribute-balancing reward. The gate activates when the top-1 match y^* is sufficiently close to the attribute-alignment distribution, indicating elevated attribute entanglement.

methods remain *query-independent*, applying a single globally-defined transformation to all queries. Consequently, such fixed rules struggle to optimize for both objectives simultaneously, forcing a compromise between effective bias mitigation and the preservation of general model utility.

Test-Time Adaptation of CLIP via Reinforcement Learning. Test-time Adaptation (TTA) aims to improve model performance on unlabeled test inputs by performing parameter updates at inference time, enabling models to adapt to distribution shifts without retraining (Sun et al., 2020; Wang et al., 2021a). Early TTA methods relied on auxiliary self-supervised objectives such as entropy minimization (Sun et al., 2020; Liu et al., 2021; Wang et al., 2021a). However, these approaches often suffer from instability from objective mismatch or prediction collapse (Park et al., 2025), motivating recent interest in utilizing VLM's internal alignment signal as direct feedback. In particular, treating CLIP similarity as explicit reward and performing RL-based optimization (Zancato et al., 2023) at inference time (Zhao et al., 2024) has been shown to reduce the collapse behavior observed in entropy-based updates and to improve zero-shot generalization. We build upon the feedback-driven TTA paradigm and extend it to address social bias in VLMs. Our key contributions are (1) a bias-sensitivity gate that activates debiasing per input, and (2) a reward that combines cross-modal alignment with a bias-subspace debiasing signal. With episodic test-time updates and selective reward gating, RG-TTA mitigates the fairness–utility trade-off, improving fairness and general-purpose utility.

3 Method

We propose **Reward-Gated Test-Time Adaptation (RG-TTA)**, an RL-based framework for selective debiasing that adaptively updates model parameters during inference. RG-TTA is built on two key components. First, a **selective gating strategy** evaluates query sensitivity to trigger **fairness regularization** only when necessary. Second, an **adaptive reward function** balances CLIP alignment with an **attribute-balancing reward**, which encourages a uniform representation by favoring under-represented attributes. We update the model using a tractable approximation of the REINFORCE (Williams, 1992) algorithm. By focusing updates on the most relevant candidates through top-K truncation, this approach ensures that the optimization stays centered on query-specific semantics and limits drift via episodic resets.

3.1 Preliminaries

CLIP. A pretrained vision-language model (VLM) consists of an image encoder $f(\cdot)$ and a text

encoder $g(\cdot)$, which project both modalities into a shared embedding space (Radford et al., 2021). Given a query q and a candidate $y \in \mathcal{Y}$ selected from the opposite modality, we define an alignment score $S(q, y)$ as the cosine similarity between their embeddings. Specifically, in the text-to-image setting we use $S(q, y) = \cos\big(g(q), f(y)\big)$, whereas in the image-to-text setting we use $S(q, y) = \cos\big(f(q), g(y)\big)$. We use the policy score $S_\theta(q, y)$ for candidate ranking and parameterizing the policy. We use the reference score $S_{\text{ref}}(q, y)$ only to compute the CLIP reward; S_{ref} is obtained from a fixed CLIP ViT-L/14 model throughout all experiments.

Test-time adaptation in vision–language tasks. Test-time adaptation (TTA (Sun et al., 2020; Wang et al., 2021a)) updates a trained model at inference time using a few unlabeled steps. We follow an *episodic* protocol: each query q is adapted independently and the parameters are reset before the next query, which mitigates negative transfer. For vision–language retrieval, where q can be text or an image, we adapt only the query-modality encoder and keep the opposite-modality encoder fixed.

3.2 Selective Gating Strategy

Applying debiasing uniformly to all queries is unnecessary and can induce parameter drift on bias-insensitive queries. To mitigate the structural fairness–utility trade-off, we introduce a gating mechanism that measures the alignment discrepancy between query semantics and demographic attributes. For each query q, we select the top-1 candidate $y^* = \arg\max_y S_\theta(q, y)$ as an **anchor** and assess whether the query–candidate alignment is disproportionately driven by demographic associations by comparing y^* to the mean attribute similarity over a predefined attribute set $\mathcal{A}$ of protected-group exemplars. We instantiate $\mathcal{A}$ in the same modality as the query (e.g., attribute prompts for text and exemplar images for vision), so that $S_\theta(a, y^*)$ uses the same scoring function; concrete instantiations are given in Sec. 4.4.

The gating logic is based on the following intuition: a large discrepancy suggests that the match is driven by general, attribute-independent semantic alignment, in which case we deactivate fairness regularization ($\delta(q) = 0$) and focus on refining standard cross-modal alignment to enhance utility. Conversely, if y^* is close to the attribute alignment distribution (within a threshold ϵ), the signal is likely entangled with a specific attribute category, triggering an attribute-balancing reward ($\delta(q) = 1$). Accordingly, the gate is defined as:

$$\delta(q) = \mathbb{I}\Big[\, S(q, y^*) - \frac{1}{|A|} \sum_{a \in A} S(a, y^*) < \epsilon \Big]. \quad (1)$$

Here, $\mathbb{I}[\cdot]$ denotes the indicator function and ϵ is a fixed threshold (0.02) kept constant throughout the episode to prevent parameter drift.

3.3 Adaptive Reward Function

In this section, we define a candidate-wise reward $r(q, y_k)$ for each candidate $y_k \in \mathcal{Y}_K(q)$, where $\mathcal{Y}_K(q)$ denotes the candidate set used for adaptation. The construction of $\mathcal{Y}_K(q)$ is deferred to 3.4. Our reward always includes a CLIP-based alignment signal to preserve general cross-modal matching ability, and adds an attribute-balancing reward only when the gate is activated. This design maintains a balance between alignment and debiasing within each episode.

CLIP alignment reward. We follow prior work in defining a CLIP-based alignment reward (Zhao et al., 2024). Concretely, for each candidate y_k we compute a nonnegative alignment score as $s_k = \max\big(S_{\text{ref}}(q, y_k), 0\big)$, and use the episode mean $\bar{s} = \frac{1}{K} \sum_{j=1}^{K} s_j$ as a baseline. The resulting baseline-normalized alignment reward is

$$r_{\text{clip}}(q, y_k) = s_k - \bar{s}. \quad (2)$$

Bias subspace and attribute-balancing reward. When $\delta(q) = 1$, we incorporate an attribute-balancing reward computed in a predefined *bias subspace* into the reward. Let $\mu_c \in \mathbb{R}^D$ denote the class-mean embedding for class c in the shared embedding space, precomputed from a labeled source dataset and kept fixed during test-time adaptation. We further define a reference vector $\bar{\mu}$, which can be instantiated as the average of $\{\mu_c\}_{c=1}^{C}$, and construct the set of difference vectors $\{\mu_c - \bar{\mu}\}_{c=1}^{C}$. We then perform PCA on these differences and extract all nonzero principal components, yielding d components in total. The resulting orthonormal basis $U \in \mathbb{R}^{D \times d}$ defines the bias subspace.

For each selected candidate y_k, let z_k denote the embedding produced by the frozen encoder of the opposite modality. We then map both candidates and class prototypes into the bias subspace using the projection operator $P = UU^\top$, yielding $\tilde{z}_k = P(z_k - \bar{\mu})$ and $\tilde{\mu}_c = P(\mu_c - \bar{\mu})$. Next, we

compute a soft assignment over classes by normalizing Gaussian-kernel similarities with a temperature parameter γ. Defining the squared distance between the projected candidate and class prototype as $d_{kc} = \|\tilde{z}_k - \tilde{\mu}_c\|_2^2$, the soft assignment is given by:

$$\alpha_c^{(k)} = \frac{\exp(-d_{kc}/\gamma)}{\sum_{c'=1}^{C} \exp(-d_{kc'}/\gamma)}. \tag{3}$$

To estimate the attribute distribution at the episode level, we compute the popularity for each class as $p_c = \frac{1}{K}\sum_{j=1}^{K} \alpha_c^{(j)}$. We then define the attribute-balancing reward for each candidate as

$$d(q, y_k) = \sum_{c=1}^{C} \alpha_c^{(k)} \left(p_c - \frac{1}{C} \right). \tag{4}$$

This score reflects how strongly candidate y_k is associated with attribute classes that are over- or under-represented in the current episode, encouraging a more balanced set of selected candidates.

Final combined reward. Finally, we define the overall reward for each selected candidate by combining the CLIP alignment reward with the query-conditioned attribute-balancing reward:

$$r(q, y_k) = r_{\text{clip}}(q, y_k) - \delta(q)\, \lambda\, d(q, y_k), \tag{5}$$

where λ is a hyperparameter that controls the strength of the attribute-balancing reward. When $\delta(q) = 0$, the reward reduces to the CLIP alignment term alone. When $\delta(q) = 1$, candidates associated with over-represented classes (i.e., $p_c > 1/C$) tend to have larger $d(q, y_k)$ and thus incur a larger balancing penalty, while candidates linked to under-represented classes (i.e., $p_c < 1/C$) tend to have smaller or negative $d(q, y_k)$ and are relatively favored. As a result, the episode-level attribute distribution is encouraged toward the uniform prior.

3.4 Optimization and Episodic Update

We optimize the proposed objective at test time via episodic policy-gradient updates (Wang et al., 2021a; Shu et al., 2022), treating each input query q as an episode. We first evaluate the gating indicator $\delta(q)$ to determine whether to include the attribute-balancing term in the episode reward. Conditioned on this decision, we choose a candidate budget K (e.g., $K = 10$ when $\delta(q) = 0$ and $K = 1024$ when $\delta(q) = 1$), and construct a truncated candidate set $\mathcal{Y}_K(q)$ by selecting the top-K candidates from the fixed pool $\mathcal{Y}$ according to the alignment score $S_\theta(q, y)$. We then compute the candidate-wise reward $r(q, y_k)$ for each $y_k \in \mathcal{Y}_K(q)$ and update the query-modality encoder parameters θ with a small number of gradient steps. After the episode ends, we reset the parameters to their initial state before processing the next query, mitigating negative transfer across queries. We define the policy $\pi_\theta(y \mid q)$ over the full candidate pool $\mathcal{Y}$ using a softmax of the alignment score:

$$\pi_\theta(y \mid q) = \frac{\exp\big(S_\theta(q, y)\big)}{\sum_{y' \in \mathcal{Y}} \exp\big(S_\theta(q, y')\big)}. \tag{6}$$

Although π_θ is normalized over the full candidate pool $\mathcal{Y}$ (i.e., the denominator is computed over $\mathcal{Y}$), we approximate the policy-gradient objective by summing only over the truncated set $\mathcal{Y}_K(q)$ for the current query. Concretely, for each episode we minimize the following REINFORCE-style objective:

$$\mathcal{L}(q) = -\frac{1}{K} \sum_{y_k \in \mathcal{Y}_K(q)} r(q, y_k)\, \log \pi_\theta(y_k \mid q). \tag{7}$$

This top-K truncation based on $S(q, y)$ yields a tractable approximation to the full policy-gradient objective while focusing updates on the most relevant candidates for the query.

4 Experiments

4.1 Datasets

To comprehensively evaluate both debiasing performance and generalization, we used a diverse set of benchmarks. For fairness evaluation, we consider in-domain settings on FairFace (val) (Kärkkäinen and Joo, 2021) and UTKFace (Zhang et al., 2017), and an out-of-domain setting on FACET (Gustafson et al., 2023). To assess whether adaptation preserves general-purpose zero-shot utility, we additionally evaluate on ImageNet-1K (Deng et al., 2009) for image classification and Flickr1k (Plummer et al., 2015) for retrieval.

4.2 Metrics

Fairness metrics. Following prior work (Berg et al., 2022a; Seth et al., 2023a), we use retrieval-based metrics: MaxSkew@k and NDKL@k. These metrics quantify the disparity in the distribution of protected attributes (e.g., gender, age, and race) within the top-k retrieved images for neutral queries. MaxSkew measures the maximum representation of a dominant group, while NDKL mea-

Table 1: Gender and age debiasing performance across different source datasets. We report results on in/out-of-domain fairness benchmarks and zero-shot utility tasks. ABLE is calculated based on in-domain metrics. Best results are shown in **bold**, and second-best results are underlined.

Backbone	Biases	Methods	In-Domain		Out-of-Domain				IN1K		Flickr		ABLE (%)↑
			Source Dataset		Cross Dataset		FACET		Acc. (%)↑		R@5 (%)↑		
			MS↓	NDKL↓	MS↓	NDKL↓	MS↓	NDKL↓	Top-1	Top-5	TR	IR	
Source: UTKFace | Cross: FairFace													
ViT-B/16	Gender	Original CLIP	0.114	0.080	0.218	0.088	0.478	0.215	68.31	91.83	96.4	85.5	77.39
		CLIP-clip	0.070	0.055	0.133	0.038	0.459	0.190	67.81	91.42	95.4	83.0	78.52
		Biased-prompts	0.179	0.062	0.161	0.048	0.460	0.215	65.07	89.38	94.3	86.1	73.18
		Joint V-L	**0.048**	0.043	0.101	**0.032**	0.456	0.181	67.99	91.64	95.8	84.6	79.36
		Ours	0.051	**0.029**	**0.080**	0.035	**0.053**	**0.064**	**70.38**	**93.00**	**97.2**	**88.5**	**80.87**
	Age	Original CLIP	0.421	0.229	0.657	0.433	0.744	0.367	68.31	91.83	96.4	85.5	66.96
		CLIP-clip	0.393	**0.215**	0.643	0.430	0.745	0.364	67.93	91.58	96.1	84.5	67.70
		Biased-prompts	0.578	0.451	0.777	0.550	**0.635**	0.355	66.43	90.28	94.1	85.2	60.83
		Joint V-L	0.414	0.231	**0.606**	**0.410**	0.746	0.365	67.63	91.46	95.6	84.6	66.86
		Ours	**0.151**	0.226	0.641	0.420	0.742	**0.330**	**70.36**	**92.99**	**97.2**	**88.3**	**77.39**
ViT-B/32	Gender	Original CLIP	0.066	**0.032**	0.138	0.054	0.485	0.225	63.39	88.83	94.7	83.5	75.60
		CLIP-clip	0.098	0.045	0.253	0.105	0.500	0.240	62.21	88.23	93.0	81.0	73.79
		Biased-prompts	0.089	0.036	0.094	**0.027**	0.417	0.164	60.37	86.75	93.6	82.4	72.74
		Joint V-L	**0.043**	0.033	0.108	0.039	0.469	0.212	62.46	88.23	94.7	82.9	75.60
		Ours	0.054	0.038	**0.088**	0.035	**0.050**	**0.021**	**69.76**	**92.31**	**97.0**	**86.6**	**79.60**
	Age	Original CLIP	0.412	0.253	0.617	0.416	0.752	0.388	63.39	88.83	94.7	83.5	64.77
		CLIP-clip	0.415	0.264	0.659	0.435	0.754	0.397	62.70	88.31	94.2	83.2	64.34
		Biased-prompts	0.522	0.409	0.701	0.497	**0.663**	**0.366**	61.07	86.92	92.0	82.2	60.19
		Joint V-L	0.407	**0.252**	0.627	0.416	0.751	0.370	62.93	88.66	94.1	82.5	64.69
		Ours	**0.127**	0.283	**0.385**	**0.364**	0.741	0.369	**69.75**	**92.30**	**96.9**	**86.4**	**77.85**
Source: FairFace | Cross: UTKFace													
ViT-B/16	Gender	Original CLIP	0.218	0.088	0.114	0.080	0.478	0.215	68.31	91.83	96.4	85.5	73.87
		CLIP-clip	0.103	0.026	0.083	0.062	0.478	0.199	68.00	91.50	95.4	83.0	77.55
		Biased-prompts	0.161	0.048	0.179	0.062	0.460	0.215	65.07	89.38	94.3	86.1	73.78
		Joint V-L	**0.080**	**0.025**	0.040	0.023	0.446	0.170	68.05	91.63	96.6	84.3	78.35
		Ours	0.082	0.031	**0.030**	**0.022**	**0.114**	**0.040**	**70.32**	**92.98**	**97.2**	**88.5**	**79.75**
	Age	Original CLIP	0.657	0.433	0.421	0.229	0.744	0.367	68.31	91.83	96.4	85.5	58.94
		CLIP-clip	0.647	0.432	0.402	0.215	0.742	0.373	67.97	91.61	96.3	84.4	59.16
		Biased-prompts	0.777	0.550	0.578	0.451	**0.635**	0.355	66.43	90.28	94.1	85.2	54.33
		Joint V-L	0.608	**0.294**	0.377	**0.115**	0.738	0.341	68.34	91.74	96.0	84.0	60.61
		Ours	**0.526**	0.318	**0.245**	0.222	0.742	**0.265**	**70.36**	**92.99**	**97.2**	**88.3**	**64.24**
ViT-B/32	Gender	Original CLIP	0.138	0.054	0.066	0.032	0.485	0.225	63.39	88.83	94.7	83.5	73.37
		CLIP-clip	0.107	0.030	0.061	0.023	0.492	0.215	59.62	86.29	90.9	76.2	71.68
		Biased-prompts	0.094	**0.027**	0.089	0.036	0.417	0.164	60.37	86.75	93.6	82.4	72.59
		Joint V-L	0.090	0.030	**0.050**	**0.021**	0.466	0.204	62.52	88.56	94.9	82.9	74.24
		Ours	**0.078**	0.032	**0.050**	0.029	**0.137**	**0.036**	**69.76**	**92.30**	**97.0**	**86.6**	**79.54**
	Age	Original CLIP	0.617	0.416	0.412	0.253	0.752	0.388	63.39	88.83	94.7	83.5	58.29
		CLIP-clip	0.635	0.425	0.400	0.252	0.749	0.387	62.40	88.30	94.5	82.5	57.32
		Biased-prompts	0.701	0.497	0.522	0.409	**0.663**	0.366	61.07	86.92	92.0	82.2	54.76
		Joint V-L	0.572	0.364	0.385	**0.195**	0.750	0.381	63.13	88.71	94.1	82.8	59.60
		Ours	**0.523**	**0.229**	**0.245**	0.222	0.743	**0.265**	**69.75**	**92.30**	**96.9**	**86.4**	**64.09**

sures the divergence from uniform distribution. For both metrics, lower values indicate a fairer model.

Utility (V–L alignment) metrics. To ensure that our selective debiasing does not compromise the intrinsic V-L alignment of the pre-trained model, we evaluated zero-shot performance on standard benchmarks. We report Top-1 and Top-5 accuracy on ImageNet-1K(Deng et al., 2009) for classification, and Recall@5 for both Image-to-Text (TR) and Text-to-Image (IR) retrieval on Flickr1k(Plummer et al., 2015).

Alignment and Bias Level Evaluation (ABLE). Single metrics capture either fairness or utility, but not their trade-off. We use ABLE proposed by Zhang *et al.*(Zhang et al., 2025) for a holistic assessment. ABLE is defined as the harmonic mean of the zero-shot accuracy and the fairness score:

$$\text{ABLE} = \frac{2}{\frac{1}{acc} + \frac{1}{\exp(-\text{MaxSkew@}k)}} \tag{8}$$

where acc denotes the ImageNet Top-1 accuracy. Higher ABLE indicates a better balance between mitigating social bias and retaining zero-shot accuracy; we report ABLE$\times$100 (%) in tables.

4.3 Baselines

We compare RG-TTA against the **Original CLIP** and three representative debiasing baselines using ViT-B/16 and ViT-B/32 backbones.

Table 2: Race debiasing performance using UTKFace as the source.

Backbone	Methods	UTKFace		IN1K		Flickr		ABLE (%)↑
				Acc. (%)↑		R@5 (%)↑		
		MS↓	NDKL↓	Top-1	Top-5	TR	IR	
ViT-B/16	Original CLIP	0.575	0.137	68.31	91.83	96.4	85.5	63.31
	CLIP-clip	0.613	0.157	67.74	91.48	95.8	85.1	60.20
	Biased-prompts	0.604	0.208	67.00	90.72	94.1	85.8	60.21
	Joint V-L	0.378	0.069	68.07	91.64	96.5	83.8	68.30
	Ours	**0.150**	**0.022**	**70.35**	**93.00**	**97.2**	**88.3**	**77.42**
ViT-B/32	Original CLIP	0.698	0.213	63.39	88.83	94.7	83.5	55.75
	CLIP-clip	0.840	0.426	62.90	88.32	93.6	81.74	51.20
	Biased-prompts	**0.317**	**0.140**	61.80	87.46	91.9	83.34	60.21
	Joint V-L	0.638	0.230	63.01	88.56	94.6	82.4	57.48
	Ours	0.351	0.281	**69.74**	**92.34**	**97.0**	**86.5**	**70.07**

CLIP-clip (Wang et al., 2021b) removes bias-correlated embedding dimensions identified via mutual information with attribute labels. **Biased-prompts** (Chuang et al., 2023) neutralizes bias directions by projecting embeddings using prompt-derived attribute subspaces without retraining. **Joint V-L** (Zhang et al., 2025) jointly debiases image and text representations to mitigate over-debiasing effects. All methods are evaluated under identical settings for Gender, Age, and Race.

4.4 Implementation Details

We use the official pretrained CLIP checkpoints (Radford et al., 2021) with ViT-B/16, ViT-B/32, and ViT-L/14 backbones. Consistent with the episodic TTA protocol, we update only the encoder corresponding to the input query modality (e.g., the text encoder for text-to-image retrieval) while keeping the target modality encoder frozen. Optimization is performed using AdamW (Loshchilov and Hutter, 2019). Crucially, to balance efficiency and performance, we dynamically adjust the computational budget based on the gating decision $\delta(q)$: we perform a lightweight update with T=3 steps and K=10 candidates for bias-insensitive queries ($\delta(q)$=0), while expanding to T=10 steps and K=1024 candidates for bias-sensitive ones ($\delta(q)$=1). Detailed hyperparameters, including learning rates, gating thresholds, and prompt templates, are provided in Appendix A.

4.5 Results

Table 1 summarizes bias-mitigation results across different source datasets (UTKFace and FairFace) for Gender and Age. Overall, our method improves fairness in both in-domain and out-of-domain settings while enhancing ImageNet zero-shot performance. For Gender, we observe a clear in-domain improvement on UTKFace (MaxSkew@k: 0.114→0.051), with particularly large mitigation under distribution shift on FACET (MaxSkew@k: 0.478→0.053). Meanwhile, ImageNet Top-1 accuracy increases from 68.31 to 70.38, and ABLE also rises from 77.39 to 80.87, indicating that fairness gains do not come at the expense of utility (see Table 1 for full results).

This trend is consistent across attributes and source configurations: for Age and Race, we observe substantial in-domain fairness gains while maintaining (or slightly improving) zero-shot utility (Tables 1, 2). Switching the source dataset to FairFace shows the same behavior across in-domain and out-of-domain settings, aligning with our design—gating with episodic resets—that focuses updates on bias-sensitive queries while preventing drift on others.

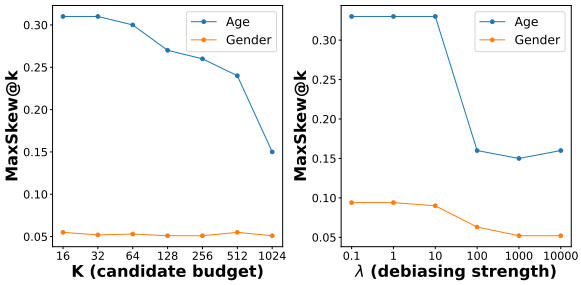

Figure 4: Ablation of K and λ showing their effect on MaxSkew@k.

Table 3: Ablation of reward variants for selective debiasing.

Reward	MS↓	IN1K Top-1↑	ABLE (%)↑
CLIP	0.604	70.41	61.54
CLIP+Debias	0.149	53.98	58.23
RG-TTA(Ours)	0.151	70.36	77.39

5 Discussion

5.1 Ablation Study

We ablate key design choices to examine how selective reward-gating mitigates the fairness–utility trade-off. Figure 4 ablates the hyperparameters used when debiasing is active, varying the candidate budget K and the debiasing strength λ in the $\delta(q) = 1$ regime. We observe that increasing K generally improves MaxSkew@1000 (notably for Age), suggesting that a larger truncated set yields a more stable estimate of the episode-level attribute distribution and thus a more reliable balancing signal. Varying λ shows a similar trend: weak values yield limited mitigation, while larger values

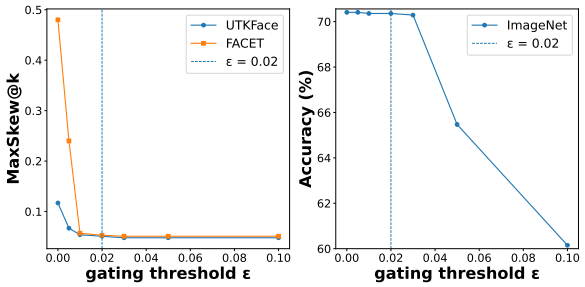

Figure 5: Sensitivity to the gating threshold ϵ. Left: MaxSkew@k (gender) on UTKFace/FACET. Right: ImageNet-1K Top-1 accuracy.

deliver consistent gains with diminishing returns, suggesting performance is not brittle once λ passes a minimal effective threshold.

We next examine when to apply the debiasing signal versus how strongly to apply it. Table 3 compares (i) an alignment-only reward, (ii) adding the attribute-balancing term uniformly to all queries, and (iii) activating the balancing term only when the gating indicator $\delta(q)$ triggers. While uniform debiasing can substantially reduce MaxSkew, it risks unnecessary adaptation and notably degrades zero-shot utility; in contrast, selective activation preserves utility while retaining the fairness gains, yielding a markedly better overall balance.

Finally, we characterize the gate through threshold sensitivity and activation patterns under a dataset-level proxy setting. Figure 5 shows that across a broad range of the threshold ϵ, fairness improves rapidly and remains largely stable, and utility is stable around the default ϵ=0.02, with noticeable degradation only when ϵ is overly large and activates debiasing too aggressively. We also evaluate the gate's activation behavior (Table 4) under a proxy labeling scheme: treating UTKFace/FACET queries as debiasing-needed and ImageNet-1K queries as debiasing-not-needed, the gate activates for almost all UTKFace/FACET queries while remaining inactive for nearly all ImageNet-1K queries, yielding low false negative/positive rates in this proxy setting.

5.2 Out-of-Domain Analysis on FACET

On FACET (out-of-domain), our method transfers well for Gender, maintaining lower MaxSkew@k than prior methods. We hypothesize that Gender is associated with multiple robust visual cues (e.g., face, hairstyle, clothing), which preserve group separation even after bias-subspace projection and lead to more stable debiasing signals. In contrast, Age relies on fine-grained facial cues that are sensitive to distance, resolution, and occlusion; in FACET's unconstrained images, these cues are weakened, reducing the reliability of the debiasing signal and ultimately limiting gains.

Table 4: FP/FN are computed under a proxy labeling scheme where UTKFace and FACET are treated as positives (debiasing-needed) and ImageNet-1K as negatives (debiasing-not-needed).

Dataset	$\delta = 1$ (%)	$\delta = 0$ (%)	FP/FN
UTKFace	99.9	0.1	FN = 0.1
FACET	99.8	0.2	FN = 0.2
ImageNet-1K	0.1	99.9	FP = 0.1

Figure 6: **Gating failure cases on FACET.** Bias-sensitive queries are incorrectly gated off ($\delta(q) = 0$), preventing fairness regularization from being activated.

5.3 Gating Failure Case Analysis

Figure 6 shows *false-negative* cases on FACET, where bias-sensitive queries are gated off ($\delta(q) = 0$) and thus do not trigger fairness regularization. These failures often arise when salient object and scene semantics in the top-1 anchor y^* overwhelm demographic-correlated cues, causing the semantic–attribute discrepancy to exceed the threshold.

6 Conclusion

We introduced Reward-Gated Test-Time Adaptation (RG-TTA), a selective test-time debiasing framework for CLIP-style vision–language models. RG-TTA uses an input-dependent reward gate to activate an attribute-balancing term only when necessary, together with episodic updates and parameter resets to limit unintended drift. Experiments across in-domain and out-of-domain fairness benchmarks as well as standard zero-shot utility tasks show that RG-TTA consistently reduces demographic skew while maintaining competitive utility. Overall, our results highlight selective, input-conditioned adaptation as a practical design principle for mitigating bias without broadly disrupting model behavior.

Limitations

Our approach relies on test-time adaptation (TTA) with per-query parameter updates, which can increase computation and latency, especially when using many update steps or large candidate sets. Because offline-trained debiasing methods amortize cost during training whereas TTA incurs cost online, direct runtime comparisons across these paradigms are not always apples-to-apples and can vary with the deployment scenario. Our experiments focus on a single protected attribute; extending the framework to multiple attributes requires more complex reward/constraint design and may introduce conflicting objectives. The method also assumes access to an external reward signal from a stronger model, which raises availability and cost considerations and may transfer the reward model's own biases into the adaptation signal. Moreover, our fairness objective implicitly targets proximity to a chosen reference distribution (e.g., uniform), and both evaluation and reward depend on the accuracy and domain robustness of attribute estimators; TTA behavior can further be sensitive to hyperparameters and may be unstable for some queries. Future work will develop more efficient update schemes to reduce online overhead. We will also explore scalable multi-attribute objectives/constraints and robustness techniques to mitigate sensitivity to reward sources and attribute estimators.

Ethics Statement

This work proposes a selective test-time adaptation (TTA) approach to mitigate distributional biases over protected attributes (e.g., gender, age, and race) that can arise for person-centric queries. Our experiments use publicly available fairness evaluation datasets (e.g., FairFace, UTKFace, and FACET) together with standard utility benchmarks, and quantify bias using distribution-based fairness metrics. Because face images and protected-attribute annotations can be sensitive, our study does not aim to identify individuals and assumes use strictly in accordance with the datasets' licenses and usage conditions. Our method further relies on an external reward signal from a stronger model (e.g., a fixed CLIP ViT-L/14 reference), which introduces practical considerations about the availability and cost of such signals and raises the possibility that biases present in the reward model could be propagated through the adaptation process. In addition, our fairness objective implicitly assumes a chosen target distribution (e.g., a uniform prior), which may not be appropriate for all tasks or domains. We therefore recommend that any real-world deployment be accompanied by careful auditing of the reward source and attribute estimators for bias and error, and that use in high-stakes decision-making contexts be avoided or subjected to additional, domain-specific validation and oversight.

References

Salma Abdel Magid, Jui-Hsien Wang, Kushal Kafle, and Hanspeter Pfister. 2024. They're all doctors: Synthesizing diverse counterfactuals to mitigate associative bias. *arXiv preprint arXiv:2406.11331*.

Ibrahim Alabdulmohsin, Xiao Wang, Andreas Steiner, Priya Goyal, Alexander D'Amour, and Xiaohua Zhai. 2024. Clip the bias: How useful is balancing data in multimodal learning? In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Hugo Berg, Siobhan Hall, Yash Bhalgat, Hannah Kirk, Aleksandar Shtedritski, and Max Bain. 2022a. A prompt array keeps the bias away: Debiasing vision-language models with adversarial learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 806–822. Association for Computational Linguistics.

Hugo Berg, Siobhan Mackenzie Hall, Yash Bhalgat, Hannah Rose Kirk, Aleksandar Shtedritski, and Max Bain. 2022b. 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*, pages 806–822. Association for Computational Linguistics.

Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. 2021. Multimodal datasets: Misogyny, pornography, and malignant stereotypes. *arXiv preprint arXiv:2110.01963*.

Ching-Yao Chuang, Varun Jampani, Yuanzhen Li, Antonio Torralba, and Stefanie Jegelka. 2023. Debiasing vision-language models via biased prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4007–4016.

Sepehr Dehdashtian, Lan Wang, and Vishnu Naresh Boddeti. 2024. Fairerclip: Debiasing clip's zero-shot predictions using functions in rkhss. In *International Conference on Learning Representations*. ArXiv:2403.15593.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale

hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255.

Walter Gerych, Haoran Zhang, Kimia Hamidieh, Eileen Pan, Maanas Sharma, Thomas Hartvigsen, and Marzyeh Ghassemi. 2024. Bendvlm: Test-time debiasing of vision-language embeddings. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Laura Gustafson, Chloe Rolland, Nikhila Ravi, Quentin Duval, Aaron Adcock, Cheng-Yang Fu, Melissa Hall, and Candace Ross. 2023. Facet: Fairness in computer vision evaluation benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20313–20325.

Melissa Hall, Laura Gustafson, Aaron Adcock, Ishan Misra, and Candace Ross. 2023. Vision-language models performing zero-shot tasks exhibit gender-based disparities. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2778–2785.

Kimia Hamidieh, Haoran Zhang, Walter Gerych, Thomas Hartvigsen, and Marzyeh Ghassemi. 2024. Identifying implicit social biases in vision-language models. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, pages 547–561.

Yusuke Hirota, Ryo Hachiuma, Boyi Li, Ximing Lu, Michael Ross Boone, Boris Ivanovic, Yejin Choi, Marco Pavone, Yu-Chiang Frank Wang, Noa Garcia, Yuta Nakashima, and Chao-Han Huck Yang. 2025. Bias in gender bias benchmarks: How spurious features distort evaluation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Yusuke Hirota, Yuta Nakashima, and Noa Garcia. 2023. Model-agnostic gender debiased image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15191–15200.

Phillip Howard, Avinash Madasu, Tiep Le, Gustavo Lujan Moreno, Anahita Bhiwandiwalla, and Vasudev Lal. 2024. Socialcounterfactuals: Probing and mitigating intersectional social biases in vision-language models with counterfactual examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*.

Sepehr Janghorbani and Gerard De Melo. 2023. Multimodal bias: Introducing a framework for stereotypical bias assessment beyond gender and race in vision–language models. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 1725–1735.

Kimmo Kärkkäinen and Jungseock Joo. 2021. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1547–1557.

Yi Li and Nuno Vasconcelos. 2024. Debias your vlm with counterfactuals: A unified approach. In *International Conference on Learning Representations*. ArXiv:2410.07593.

Yuejiang Liu, Parth Kothari, Bastien Van Delft, Baptiste Bellot-Gurlet, Taylor Mordan, and Alexandre Alahi. 2021. Ttt++: When does self-supervised test-time training fail or thrive? In *Advances in Neural Information Processing Systems (NeurIPS)*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Bo Pang, Tingrui Qiao, Caroline Walker, Chris Cunningham, and Yun Sing Koh. 2025. Cabin: Debiasing vision-language models using backdoor adjustments. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pages 484–492. International Joint Conferences on Artificial Intelligence Organization.

Mincheol Park, Heeji Won, Won Woo Ro, and Suhyun Kim. 2025. Rethinking entropy in test-time adaptation: The missing piece from energy duality. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2641–2649.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Ashish Seth, Mayur Hemani, and Chirag Agarwal. 2023a. Dear: Debiasing vision-language models with additive residuals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6820–6829.

Ashish Seth, Mayur Hemani, and Chirag Agarwal. 2023b. Dear: Debiasing vision-language models with additive residuals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6820–6829.

Xudong Shen, Chao Du, Tianyu Pang, Min Lin, Yongkang Wong, and Mohan Kankanhalli. 2023. Finetuning text-to-image diffusion models for fairness. *arXiv preprint arXiv:2311.07604*.

Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. 2022. Test-time prompt tuning for zero-shot generalization in vision-language models. In *Advances in Neural Information Processing Systems (NeurIPS)*. Curran Associates, Inc.

Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei A. Efros, and Moritz Hardt. 2020. Test-time training with self-supervision for generalization under distribution shifts. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno A. Olshausen, and Trevor Darrell. 2021a. Tent: Fully test-time adaptation by entropy minimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Jialu Wang, Yang Liu, and Xin Wang. 2021b. Are gender-neutral queries really gender-neutral? mitigating gender bias in image search. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1995–2008, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ronald J. Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3–4):229–256.

Luca Zancato, Alessandro Achille, Tian Yu Liu, Matthew Trager, Pramuditha Perera, and Stefano Soatto. 2023. Train/test-time adaptation with retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Haoyu Zhang, Yangyang Guo, and Mohan Kankanhalli. 2025. Joint vision-language social bias removal for clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4246–4255.

Zhifei Zhang, Yang Song, and Hairong Qi. 2017. Age progression/regression by conditional adversarial autoencoder. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 4352–4360. IEEE Computer Society.

Dachuan Zhao, Weiyue Li, Zhenda Shen, Yushu Qiu, Bowen Xu, Haoyu Chen, and Yongchao Chen. 2025. Bias is a subspace, not a coordinate: A geometric rethinking of post-hoc debiasing in vision-language models. *arXiv preprint arXiv:2511.18123*.

Dora Zhao, Angelina Wang, and Olga Russakovsky. 2021. Understanding and evaluating racial biases in image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14830–14840.

Shuai Zhao, Xiaohan Wang, Linchao Zhu, and Yi Yang. 2024. Test-time adaptation with clip reward for zero-shot generalization in vision-language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

A Implementation Details

In this section, we provide comprehensive details regarding the experimental setup, optimization hyperparameters, and the construction of the bias subspace, complementing the summary provided in the main text.

A.1 Computational Environment and Models

All experiments were conducted on a single **NVIDIA A6000 GPU** with automatic mixed precision (AMP) enabled to enhance efficiency. We utilized the official pre-trained CLIP checkpoints (Radford et al., 2021) for ViT-B/16, ViT-B/32, and ViT-L/14 backbones. Consistent with the episodic TTA protocol, the CLIP reward signal was computed using a separate, fixed **CLIP ViT-L/14** model to provide stable guidance.

A.2 Modality-Specific Update Strategy

During the episodic adaptation, we update only the encoder corresponding to the *input query modality* to align it with the frozen target modality:

- **Text-to-Image Retrieval:** We adapt the **text encoder** while keeping the image encoder frozen.
- **Image-to-Text Retrieval & Classification:** We adapt the **image encoder** while keeping the text encoder frozen.

A.3 Optimization Hyperparameters

We used the **AdamW** optimizer (Loshchilov and Hutter, 2019) with a learning rate of 1×10^{-4} and a weight decay of 0. The balancing coefficients for the reward function were set to $\lambda = 1000$ (debiasing penalty weight). Additionally, we set the Gaussian kernel temperature parameter $\gamma = 0.25$ to control the sharpness of soft assignments in the bias subspace.

A.4 Adaptive Computational Budgeting

The Rating Gate (RG) determines the bias sensitivity using a fixed threshold of $\epsilon = 0.02$. Based on the gate's output $\delta(q)$, we dynamically adjust the number of update steps (T) and the candidate pool size (K):

- **Bias-Insensitive ($\delta(q) = 0$):** We perform a lightweight update with $T = 3$ steps and $K = 10$ candidates.

Table 5: Race debiasing performance using FairFace as the source.

Backbone	Methods	Fairface		IN1K		Flickr		ABLE (%)↑
				Acc. (%)↑		R@5 (%)↑		
		MS↓	NDKL↓	Top-1	Top-5	TR	IR	
ViT-B/16	Original CLIP	0.528	0.182	68.31	91.83	96.4	85.5	63.31
	CLIP-clip	0.544	0.161	67.97	91.62	95.4	85.3	62.62
	Biased-prompts	0.518	0.219	67.00	90.72	94.1	85.8	63.07
	Joint V-L	**0.353**	**0.125**	68.07	91.64	96.5	83.8	69.14
	Ours	0.372	0.167	**70.32**	**92.96**	**97.2**	**88.4**	**69.62**
ViT-B/32	Original CLIP	0.568	0.165	63.39	88.83	94.7	83.5	59.84
	CLIP-clip	0.713	0.227	62.51	88.33	92.6	81.2	54.95
	Biased-prompts	0.595	0.282	61.80	87.46	91.9	83.3	58.29
	Joint V-L	0.503	0.149	63.07	88.61	94.2	83.0	61.74
	Ours	**0.389**	**0.148**	**69.73**	**92.26**	**97.0**	**86.5**	**68.74**

- **Bias-Sensitive ($\delta(q) = 1$):** We increase the budget to $T = 10$ steps and $K = 1024$ candidates. In this case, candidates are ranked by their CLIP alignment scores to filter the most relevant samples for the update.

A.5 Bias Subspace Construction Details

The bias subspace is constructed offline using the training split of the dataset:

- **Text Queries:** We use attribute-specific prompts formatted as "*a photo of a {attribute class} person*". One prompt is generated per attribute class.

- **Image Queries:** We construct the reference set A by sampling $M = 5$ images per attribute class uniformly from the **UTKFace** dataset. To ensure reproducibility and consistency, this reference set is fixed once and reused across all test episodes. The total size of the reference set is $|A| = C \times 5$, where C is the number of attribute classes.

B Additional Results

B.1 Additional Race Debiasing Results using FairFace

In the main text (Table 2), we utilized UTKFace as the source dataset for constructing the bias subspace to mitigate Race bias. To verify the robustness of our framework across different source domains, we conducted an additional experiment using **FairFace** as the source dataset.

The results are presented in Table 5. Consistent with the findings in the main text, our RG-TTA framework demonstrates a superior capability to balance fairness and utility. Regarding fairness, our method effectively mitigates racial bias compared to the Original CLIP; while **Joint V-L** shows competitive scores on ViT-B/16, our method achieves the best performance on ViT-B/32 (MaxSkew: 0.389). Crucially, in terms of utility preservation, our method consistently outperforms all baselines in zero-shot tasks (ImageNet and Flickr) across both backbones, confirming that our selective routing mechanism successfully prevents the over-debiasing observed in static approaches. Consequently, our method achieves the highest ABLE scores for both backbones (69.62% and 68.74%), proving that it maintains the optimal trade-off between fairness and utility regardless of the source dataset used.

Figure 7: **Gating failure cases on ImageNet.** In ImageNet, queries are expected to be bias-insensitive and thus gated off ($\delta(q) = 0$). Shown are false-positive cases where the gate is incorrectly activated ($\delta(q) = 1$), causing debiasing and alignment rewards to be jointly applied.

B.2 False-positive gating failures on ImageNet

Figure 7 illustrates *false-positive* cases on ImageNet-1K, where bias-insensitive object queries are incorrectly gated on ($\delta(q) = 1$), triggering unnecessary fairness regularization. These failures typically arise when the queried object strongly co-occurs with humans or human-like features in the top-1 retrieved anchor y^*. For instance, as shown in Figure 7, human figurines on a cake (left) or bystanders in the background (right) provide strong demographic signals that reduce the semantic-attribute discrepancy below the threshold ϵ. This misleads the gate into treating the object query as bias-sensitive. Although our hybrid reward design minimizes semantic drift even when the gate is mistakenly active, these cases represent a computational inefficiency.