A Closed-Form Solution for Debiasing Vision-Language Models with Utility Guarantees Across Modalities and Tasks

Supplementary Material

Overview

1. **Appendix A**: Proofs of the propositions, lemmas, and theorems.
2. **Appendix B**: Details of the LLM-guided group prototype construction module.
3. **Appendix C**: The occupation list and its group-explicit variants used in text-to-image generation.
4. **Appendix D**: The evaluation process in text-to-image generation.
5. **Appendix E**: Detailed descriptions of the datasets used.
6. **Appendix F**: Experimental results for CLIP (ResNet50) and BLIP.
7. **Appendix G**: More illustrative examples for text-to-image generation.

A. Proofs

Proof of Proposition 1. Denote the self-utility losses for the image and text as $\ell_{\text{self}}^{(I)} := 1 - \langle \vec{u}_I, \vec{e}_I \rangle$ and $\ell_{\text{self}}^{(T)} := 1 - \langle \vec{u}_T, \vec{e}_T \rangle$. For the cross-utility loss, we have

$$\begin{aligned}
\ell_{\text{cross}} = \big|\langle \vec{u}_I, \vec{u}_T \rangle - \langle \vec{e}_I, \vec{e}_T \rangle\big| &= \big|\langle \vec{u}_I, \vec{u}_T \rangle - \langle \vec{e}_I, \vec{u}_T \rangle + \langle \vec{e}_I, \vec{u}_T \rangle - \langle \vec{e}_I, \vec{e}_T \rangle\big| \\
&\overset{(1)}{\leq} \big|\langle \vec{u}_I, \vec{u}_T \rangle - \langle \vec{e}_I, \vec{u}_T \rangle\big| + \big|\langle \vec{e}_I, \vec{u}_T \rangle - \langle \vec{e}_I, \vec{e}_T \rangle\big| \\
&= \big|\langle \vec{u}_I - \vec{e}_I, \vec{u}_T \rangle\big| + \big|\langle \vec{e}_I, \vec{u}_T - \vec{e}_T \rangle\big| \\
&\overset{(2)}{\leq} \|\vec{u}_I - \vec{e}_I\| \, ||\vec{u}_T|| + ||\vec{e}_I|| \, \|\vec{u}_T - \vec{e}_T\| \\
&= \|\vec{u}_I - \vec{e}_I\| + \|\vec{u}_T - \vec{e}_T\|.
\end{aligned}$$

Here, $\overset{(1)}{\leq}$ and $\overset{(2)}{\leq}$ are because of the triangle inequality and Cauchy–Schwarz inequalities. For any unit vectors $\vec{u}, \vec{e}$, we have

$$\|\vec{u} - \vec{e}\,\|^2 = \|\vec{u}\|^2 + \|\vec{e}\,\|^2 - 2\langle \vec{u}, \vec{e} \rangle = 2\big(1 - \langle \vec{u}, \vec{e} \rangle\big) = 2\,\ell_{\text{self}},$$

Applying this to image and text gives

$$\|\vec{u}_I - \vec{e}_I\| = \sqrt{2\,\ell_{\text{self}}^{(I)}}, \qquad \|\vec{u}_T - \vec{e}_T\| = \sqrt{2\,\ell_{\text{self}}^{(T)}}.$$

Then

$$\ell_{\text{cross}} \;\leq\; \|\vec{u}_I - \vec{e}_I\| + \|\vec{u}_T - \vec{e}_T\| \;=\; \sqrt{2\,\ell_{\text{self}}^{(I)}} + \sqrt{2\,\ell_{\text{self}}^{(T)}}$$

□

Proof of Lemma 1. The objective function in **Problem (1)** can be rewritten as

$$\begin{aligned}
G(\vec{u}) &= w_1 ||\vec{u}_{\mathcal{A}_\parallel}|| + w_2(1 - \langle \vec{u}, \vec{e} \rangle) \\
&= w_1 ||\vec{u}_{\mathcal{A}_\parallel}|| + w_2\Big[1 - \big(\langle \vec{u}_{\mathcal{A}_\parallel}, \vec{e}_{\mathcal{A}_\parallel} \rangle + \langle \vec{u}_{\mathcal{A}_\perp}, \vec{e}_{\mathcal{A}_\perp} \rangle\big)\Big] \\
&\overset{(1)}{\geq} w_1 ||\vec{u}_{\mathcal{A}_\parallel}|| + w_2\Big[1 - \big(||\vec{u}_{\mathcal{A}_\parallel}|| \, ||\vec{e}_{\mathcal{A}_\parallel}|| + ||\vec{u}_{\mathcal{A}_\perp}|| \, ||\vec{e}_{\mathcal{A}_\perp}||\big)\Big]
\end{aligned}$$

Here, $\overset{(1)}{\geq}$ is because the Cauchy-Schwarz inequality $\langle \vec{u}_{\mathcal{A}_\parallel}, \vec{e}_{\mathcal{A}_\parallel} \rangle + \langle \vec{u}_{\mathcal{A}_\perp}, \vec{e}_{\mathcal{A}_\perp} \rangle \leq ||\vec{u}_{\mathcal{A}_\parallel}|| \, ||\vec{e}_{\mathcal{A}_\parallel}|| + ||\vec{u}_{\mathcal{A}_\perp}|| \, ||\vec{e}_{\mathcal{A}_\perp}||$ with the equality holds *iff* both components of the orthogonal decompositions of $\vec{u}$ and $\vec{e}$ are collinear, i.e., $\vec{u}_{\mathcal{A}_\parallel} \parallel \vec{e}_{\mathcal{A}_\parallel}$ and

$\vec{u}_{\mathcal{A}_\perp} \parallel \vec{e}_{\mathcal{A}_\perp}$. Since we are minimizing $G(\vec{u})$, any strict inequality in $\overset{(1)}{\geq}$ would mean that the right-hand collinear case produces a strictly smaller value of $G(\vec{u})$. Hence, no non-aligned $\vec{u}$ can be optimal. At optimality, $\overset{(1)}{\geq}$ must be tight and therefore $\vec{u}^\star \in \text{span}\{\vec{e}_{\mathcal{A}_\parallel}, \vec{e}_{\mathcal{A}_\perp}\}$.

□

Proof of Proposition 2. There are two edge cases where the **Problem (2)** degenerates into 1D.

Case A ($\|\vec{e}_{\mathcal{A}_\parallel}\| = 0$ and $\|\vec{e}_{\mathcal{A}_\perp}\| = 1$). This means $\vec{e} = \vec{e}_{\mathcal{A}_\perp}$ is already fair. Then F becomes

$$F(\alpha, \beta) = w_1|\alpha| + w_2\big(1 - \beta\big).$$

Here, no debiasing is needed, and we set $\alpha = 0$ and $\beta = \|\vec{e}_{\mathcal{A}_\perp}\|$ to fully preserve utility, which gives $\vec{u} = \vec{e} = \vec{e}_{\mathcal{A}_\perp}$ and $\ell_{\text{self}} = 1 - ||\vec{e}_{\mathcal{A}_\perp}|| = 0$.

Case B ($\|\vec{e}_{\mathcal{A}_\perp}\| = 0$ and $\|\vec{e}_{\mathcal{A}_\parallel}\| = 1$). This means $\vec{e} = \vec{e}_{\mathcal{A}_\parallel}$ contains only attribute information (e.g., prompts like "a photo of a male" or "a photo of a female"). In this case, F reduces to

$$F(\alpha) = w_1|\alpha| + w_2(1 - \alpha).$$

Since $\vec{e}$ reflects purely attribute-related semantics, debiasing is unnecessary. Therefore, we fully preserve its utility by setting $\alpha = ||\vec{e}_{\mathcal{A}_\parallel}||, \beta = 0$, which yields $\vec{u} = \vec{e} = \vec{e}_{\mathcal{A}_\parallel}$ and $\ell_{\text{self}} = 1 - ||\vec{e}_{\mathcal{A}_\parallel}|| = 0$.

□

Proof of Lemma 2. Assume $\beta < 0$ and define a sign-flipped feasible point $(\alpha, \tilde{\beta}) = (\alpha, -\beta)$, where $\tilde{\beta} > 0$. The objective difference between the two points is

$$\begin{aligned}
F(\alpha, \tilde{\beta}) - F(\alpha, \beta) &= w_1|\alpha| + w_2(1 - \alpha||\vec{e}_{\mathcal{A}_\parallel}|| - \tilde{\beta}||\vec{e}_{\mathcal{A}_\perp}||) - \Big[w_1|\alpha| + w_2(1 - \alpha||\vec{e}_{\mathcal{A}_\parallel}|| - \beta||\vec{e}_{\mathcal{A}_\perp}||)\Big] \\
&= w_2\big(- \tilde{\beta}||\vec{e}_{\mathcal{A}_\perp}|| + \beta||\vec{e}_{\mathcal{A}_\perp}||\big) \\
&= w_2\big(- (-\beta)||\vec{e}_{\mathcal{A}_\perp}|| + \beta||\vec{e}_{\mathcal{A}_\perp}||\big) \\
&= 2w_2\beta||\vec{e}_{\mathcal{A}_\perp}|| \\
&\overset{(1)}{\leq} 0.
\end{aligned}$$

Here, $\overset{(1)}{\leq}$ because $w_2 \geq 0, ||\vec{e}_{\mathcal{A}_\perp}|| > 0$ and $\beta < 0$. Therefore, $F(\alpha, \tilde{\beta}) \leq F(\alpha, \beta)$, meaning that a negative β cannot yield a smaller objective value. Thus, the optimal solution must satisfy $\beta \geq 0$. A similar argument applies to α. Assume $\alpha < 0$ and define $(\tilde{\alpha}, \beta) = (-\alpha, \beta)$, where $\tilde{\alpha} > 0$. Then,

$$\begin{aligned}
F(\tilde{\alpha}, \beta) - F(\alpha, \beta) &= w_1|\tilde{\alpha}| + w_2(1 - \tilde{\alpha}||\vec{e}_{\mathcal{A}_\parallel}|| - \beta||\vec{e}_{\mathcal{A}_\perp}||) - \Big[w_1|\alpha| + w_2(1 - \alpha||\vec{e}_{\mathcal{A}_\parallel}|| - \beta||\vec{e}_{\mathcal{A}_\perp}||)\Big] \\
&\overset{(1)}{=} w_2\Big[- \tilde{\alpha}||\vec{e}_{\mathcal{A}_\parallel}|| + \alpha||\vec{e}_{\mathcal{A}_\parallel}||\Big] \\
&= w_2\Big[- (-\alpha)||\vec{e}_{\mathcal{A}_\parallel}|| + \alpha||\vec{e}_{\mathcal{A}_\parallel}||\Big] \\
&= 2w_2\alpha||\vec{e}_{\mathcal{A}_\parallel}|| \\
&\overset{(2)}{\leq} 0.
\end{aligned}$$

Here, $\overset{(1)}{=}$ because $|\tilde{\alpha}| = |-\alpha| = |\alpha|$ and $\overset{(2)}{\leq}$ because $w_2 \geq 0, ||\vec{e}_{\mathcal{A}_\parallel}|| > 0$ and $\alpha < 0$. Therefore, a negative α also cannot provide a smaller objective value, implying that the optimal solution must satisfy $\alpha \geq 0$. Hence, all optimal solutions of **Problem (2)** lie in the first quadrant of the unit circle. We can then write the **Problem (2)** as

$$\min_{0 \leq \alpha \leq 1} F(\alpha) := w_1 L(\alpha) + w_2 V(\alpha),$$

where $L(\alpha) = \alpha$ is the attribute leakage and $V(\alpha) = 1 - \alpha||\vec{e}_{\mathcal{A}_\parallel}|| - \sqrt{1-\alpha^2}||\vec{e}_{\mathcal{A}_\perp}||$ is the self-utility loss.

We further show that no optimal solution can lie in the interval $\|\vec{e}_{\mathcal{A}_\parallel}\| < \alpha \leq 1$. Differentiating $F(\alpha)$ gives

$$\frac{dF}{d\alpha} = w_1 \frac{dL}{d\alpha} + w_2 \frac{dV}{d\alpha} = w_1 + w_2 \frac{dV}{d\alpha}.$$

The derivative of $V(\alpha)$ is

$$\frac{dV}{d\alpha} = -\frac{d}{d\alpha}\Big(\alpha \, \|\vec{e}_{\mathcal{A}_\parallel}\| + \sqrt{1-\alpha^2} \, \|\vec{e}_{\mathcal{A}_\perp}\|\Big) = \frac{\alpha}{\sqrt{1-\alpha^2}} \, \|\vec{e}_{\mathcal{A}_\perp}\| - \|\vec{e}_{\mathcal{A}_\parallel}\|. \qquad (\alpha \neq 1)$$

Setting $\frac{dV}{d\alpha} = 0$ yields

$$\frac{\alpha}{\sqrt{1-\alpha^2}} = \frac{\|\vec{e}_{\mathcal{A}_\parallel}\|}{\|\vec{e}_{\mathcal{A}_\perp}\|} = \frac{\|\vec{e}_{\mathcal{A}_\parallel}\|}{\sqrt{1 - \|\vec{e}_{\mathcal{A}_\parallel}\|^2}}$$

The unique solution to this equation is $\alpha = \|\vec{e}_{\mathcal{A}_\parallel}\|$. To examine the sign of $\dfrac{dV}{d\alpha}$ around this point, define the function $h(x) = \dfrac{x}{\sqrt{1-x^2}}$ on $[0, 1)$. Since

$$h'(x) = \frac{1}{(1-x^2)^{3/2}} > 0 \quad \text{for } x \in [0, 1),$$

$h(x)$ is strictly increasing on $[0, 1)$. Consequently,

$$\alpha > \|\vec{e}_{\mathcal{A}_\parallel}\| \;\Rightarrow\; h(\alpha) > h(\|\vec{e}_{\mathcal{A}_\parallel}\|) \Rightarrow \; \frac{\alpha}{\sqrt{1-\alpha^2}} > \frac{\|\vec{e}_{\mathcal{A}_\parallel}\|}{\sqrt{1 - \|\vec{e}_{\mathcal{A}_\parallel}\|^2}} \;\Rightarrow\; \frac{dV}{d\alpha} > 0,$$

and conversely,

$$\alpha < \|\vec{e}_{\mathcal{A}_\parallel}\| \;\Rightarrow\; h(\alpha) < h(\|\vec{e}_{\mathcal{A}_\parallel}\|) \Rightarrow \; \frac{\alpha}{\sqrt{1-\alpha^2}} < \frac{\|\vec{e}_{\mathcal{A}_\parallel}\|}{\sqrt{1 - \|\vec{e}_{\mathcal{A}_\parallel}\|^2}} \;\Rightarrow\; \frac{dV}{d\alpha} < 0,$$

Since $w_1, w_2 \geq 0$ and $w_1 + w_2 = 1$, when $\alpha > \|\vec{e}_{\mathcal{A}_\parallel}\|$, it follows that

$$\frac{dF}{d\alpha} = w_1 + w_2 \frac{dV}{d\alpha} > 0$$

Hence, $F(\alpha)$ is strictly increasing on $(\|\vec{e}_{\mathcal{A}_\parallel}\|, 1)$ and cannot achieve an optimum in this interval. Next, we examine the boundary case $\alpha = 1$. At $\alpha = 1$, we have $F(1) = w_1 + w_2(1 - \|\vec{e}_{\mathcal{A}_\parallel}\|)$. At $\alpha = \|\vec{e}_{\mathcal{A}_\parallel}\|$, we have $F(\|\vec{e}_{\mathcal{A}_\parallel}\|) = w_1\|\vec{e}_{\mathcal{A}_\parallel}\|$. The difference between these two values is

$$F(1) - F(\|\vec{e}_{\mathcal{A}_\parallel}\|) = w_1(1 - \|\vec{e}_{\mathcal{A}_\perp}\|) + w_2(1 - \|\vec{e}_{\mathcal{A}_\parallel}\|) \geq 0,$$

which confirms that $F(1) \geq F(\|\vec{e}_{\mathcal{A}_\parallel}\|)$. Thus, we can conclude that all optimal solutions must satisfy $0 \leq \alpha \leq ||\vec{e}_{\mathcal{A}_\parallel}||$.

Within this feasible range, we have $\frac{dL}{d\alpha} = 1 > 0$ and $\frac{dV}{d\alpha} \leq 0$, implying that $L(\alpha)$ is strictly increasing while $V(\alpha)$ is strictly decreasing on $[0, \|\vec{e}_{\mathcal{A}_\parallel}\|]$. For any two feasible points $\alpha_1 < \alpha_2 \leq \|\vec{e}_{\mathcal{A}_\parallel}\|$, it follows that $L(\alpha_1) < L(\alpha_2)$ and $V(\alpha_1) > V(\alpha_2)$, indicating that reducing leakage necessarily increasing self-utility loss. Each α in this interval thus yields a distinct non-dominated pair $(L(\alpha), V(\alpha))$. Therefore, the set $\alpha \in [0, \|\vec{e}_{\mathcal{A}_\parallel}\|]$ fully characterizes the Pareto front.

□

Proof of Theorem 1. For any scalars (x, y) and weights $(w_1, w_2) \in \Delta := \{w_1, w_2 \geq 0,\ w_1 + w_2 = 1\}$, we have

$$\sup_{(w_1, w_2) \in \Delta} \big(w_1 x + w_2 y\big) = \max\{x, y\}.$$

Therefore, the **Problem (3)** yields the following minimax formulation:

$$\min_{0 \leq \alpha \leq \|\vec{e}_{\mathcal{A}_\parallel}\|} \; \sup_{\substack{w_1, w_2 \geq 0 \\ w_1 + w_2 = 1}} \Big\{ w_1 \, L(\alpha) + w_2 \, V(\alpha) \Big\} = \min_{0 \leq \alpha \leq \|\vec{e}_{\mathcal{A}_\parallel}\|} \max\{L(\alpha), \, V(\alpha)\}$$

Since the two loss functions operate on different numerical scales over $\alpha \in [0, ||\vec{e}_{\mathcal{A}_\parallel}||]$, we normalize each to the interval $[0, 1]$:

$$\tilde{L}(\alpha) = \frac{\alpha}{\|\vec{e}_{\mathcal{A}_\parallel}\|}, \qquad \tilde{V}(\alpha) = \frac{1 - \left(\sqrt{1-\alpha^2}\,\|\vec{e}_{\mathcal{A}_\perp}\| + \alpha\|\vec{e}_{\mathcal{A}_\parallel}\|\right)}{1 - \|\vec{e}_{\mathcal{A}_\perp}\|}.$$

This normalization ensures that $\tilde{L}(0) = 0$, $\tilde{L}(||\vec{e}_{\mathcal{A}_\parallel}||) = 1$, and $\tilde{V}(0) = 1$, $\tilde{V}(||\vec{e}_{\mathcal{A}_\parallel}||) = 0$. Hence, **Problem (3)** becomes

$$\min_{0 \le \alpha \le \|\vec{e}_{\mathcal{A}_\parallel}\|} \max\{\tilde{L}(\alpha),\, \tilde{V}(\alpha)\}$$

By **Lemma 2**, over the interval $[0, ||\vec{e}_{\mathcal{A}_\parallel}||]$, $\tilde{L}(\alpha)$ is strictly increasing while $\tilde{V}(\alpha)$ is strictly decreasing. Therefore, $\max\{\tilde{L}, \tilde{V}\}$ is minimized *uniquely* at the point where they intersect:

$$\tilde{L}(\alpha^\star) = \tilde{V}(\alpha^\star) \quad \Longleftrightarrow \quad \frac{\alpha^\star}{\|\vec{e}_{\mathcal{A}_\parallel}\|} = \frac{1 - \left(\sqrt{1-(\alpha^\star)^2}\,\|\vec{e}_{\mathcal{A}_\perp}\| + \alpha^\star\|\vec{e}_{\mathcal{A}_\parallel}\|\right)}{1 - \|\vec{e}_{\mathcal{A}_\perp}\|}.$$

Rearranging, we obtain

$$\|\vec{e}_{\mathcal{A}_\perp}\|\sqrt{1-(\alpha^\star)^2} = 1 - \alpha^\star\Big(\|\vec{e}_{\mathcal{A}_\parallel}\| + \frac{1 - \|\vec{e}_{\mathcal{A}_\perp}\|}{\|\vec{e}_{\mathcal{A}_\parallel}\|}\Big).$$

For this equality to hold, the right-hand side must be positive, which provides an upper bound on $\alpha^\star$:

$$\alpha^\star < \frac{1}{\|\vec{e}_{\mathcal{A}_\parallel}\| + \dfrac{1 - \|\vec{e}_{\mathcal{A}_\perp}\|}{\|\vec{e}_{\mathcal{A}_\parallel}\|}}. \tag{C2}$$

Squaring both sides (as both are nonnegative) yields

$$\|\vec{e}_{\mathcal{A}_\perp}\|^2(1-(\alpha^\star)^2) = \Big(1 - \alpha^\star\Big(\|\vec{e}_{\mathcal{A}_\parallel}\| + \frac{1 - \|\vec{e}_{\mathcal{A}_\perp}\|}{\|\vec{e}_{\mathcal{A}_\parallel}\|}\Big)\Big)^2.$$

Expanding and rearranging leads to a quadratic in $\alpha^\star$:

$$Q(\alpha^*) := \Big(\Big(\|\vec{e}_{\mathcal{A}_\parallel}\| + \frac{1 - \|\vec{e}_{\mathcal{A}_\perp}\|}{\|\vec{e}_{\mathcal{A}_\parallel}\|}\Big)^2 + \|\vec{e}_{\mathcal{A}_\perp}\|^2\Big)(\alpha^\star)^2 - 2\Big(\|\vec{e}_{\mathcal{A}_\parallel}\| + \frac{1 - \|\vec{e}_{\mathcal{A}_\perp}\|}{\|\vec{e}_{\mathcal{A}_\parallel}\|}\Big)\alpha^\star + \|\vec{e}_{\mathcal{A}_\parallel}\|^2 = 0.$$

The discriminant is

$$\begin{aligned}
\Delta &= 4\Big[\big(\|\vec{e}_{\mathcal{A}_\parallel}\| + \frac{1 - \|\vec{e}_{\mathcal{A}_\perp}\|}{\|\vec{e}_{\mathcal{A}_\parallel}\|}\big)^2 - \Big(\Big(\|\vec{e}_{\mathcal{A}_\parallel}\| + \frac{1 - \|\vec{e}_{\mathcal{A}_\perp}\|}{\|\vec{e}_{\mathcal{A}_\parallel}\|}\Big)^2 + \|\vec{e}_{\mathcal{A}_\perp}\|^2\Big)\|\vec{e}_{\mathcal{A}_\parallel}\|^2\Big] \\
&= 4\Big[\big(1 - \|\vec{e}_{\mathcal{A}_\parallel}\|^2\big)\big(\|\vec{e}_{\mathcal{A}_\parallel}\| + \frac{1 - \|\vec{e}_{\mathcal{A}_\perp}\|}{\|\vec{e}_{\mathcal{A}_\parallel}\|}\big)^2 - \|\vec{e}_{\mathcal{A}_\perp}\|^2\|\vec{e}_{\mathcal{A}_\parallel}\|^2\Big] \\
&= 4\,\|\vec{e}_{\mathcal{A}_\perp}\|^2\left[\Big(\|\vec{e}_{\mathcal{A}_\parallel}\| + \frac{1 - \|\vec{e}_{\mathcal{A}_\perp}\|}{\|\vec{e}_{\mathcal{A}_\parallel}\|}\Big)^2 - \|\vec{e}_{\mathcal{A}_\parallel}\|^2\right] \\
&= 4\,\|\vec{e}_{\mathcal{A}_\perp}\|^2\left(\frac{(1 - \|\vec{e}_{\mathcal{A}_\perp}\|)^2}{\|\vec{e}_{\mathcal{A}_\parallel}\|^2} + 2\big(1 - \|\vec{e}_{\mathcal{A}_\perp}\|\big)\right) \\
&> 0
\end{aligned}$$

Thus, there exist two real roots:

$$\alpha_\pm = \frac{\Big(\|\vec{e}_{\mathcal{A}_\parallel}\| + \dfrac{1 - \|\vec{e}_{\mathcal{A}_\perp}\|}{\|\vec{e}_{\mathcal{A}_\parallel}\|}\Big) \;\pm\; \|\vec{e}_{\mathcal{A}_\perp}\|\sqrt{\Big(\|\vec{e}_{\mathcal{A}_\parallel}\| + \dfrac{1 - \|\vec{e}_{\mathcal{A}_\perp}\|}{\|\vec{e}_{\mathcal{A}_\parallel}\|}\Big)^2 - \|\vec{e}_{\mathcal{A}_\parallel}\|^2}}{\Big(\|\vec{e}_{\mathcal{A}_\parallel}\| + \dfrac{1 - \|\vec{e}_{\mathcal{A}_\perp}\|}{\|\vec{e}_{\mathcal{A}_\parallel}\|}\Big)^2 + \|\vec{e}_{\mathcal{A}_\perp}\|^2}$$

Since the solution is unique, we need to verify which root satisfies the feasibility conditions.

Let

$$E := \|\vec{e}_{\mathcal{A}_\parallel}\| + \frac{1 - \|\vec{e}_{\mathcal{A}_\perp}\|}{\|\vec{e}_{\mathcal{A}_\parallel}\|}, \qquad \text{so that} \quad \alpha_+ = \frac{E \;+\; \|\vec{e}_{\mathcal{A}_\perp}\|\sqrt{E^2 - \|\vec{e}_{\mathcal{A}_\parallel}\|^2}}{E^2 + \|\vec{e}_{\mathcal{A}_\perp}\|^2}.$$

The feasibility conditions are:

$$\text{(C1)} \quad 0 < \alpha^* < \|\vec{e}_{\mathcal{A}_\parallel}\|, \qquad \text{(C2)} \quad \alpha^* < \frac{1}{E}.$$

We show that α_+ violates at least one of these conditions.

Case A: $E \geq 1$**.** Consider

$$\alpha_+ - \frac{1}{E} = \frac{E + \|\vec{e}_{\mathcal{A}_\perp}\|\sqrt{E^2 - \|\vec{e}_{\mathcal{A}_\parallel}\|^2}}{E^2 + \|\vec{e}_{\mathcal{A}_\perp}\|^2} - \frac{1}{E} = \frac{\|\vec{e}_{\mathcal{A}_\perp}\|\big(\, E\sqrt{E^2 - \|\vec{e}_{\mathcal{A}_\parallel}\|^2} - \|\vec{e}_{\mathcal{A}_\perp}\|\,\big)}{E(E^2 + \|\vec{e}_{\mathcal{A}_\perp}\|^2)}.$$

The denominator is positive. For the numerator, since $E \geq 1$, we have

$$\big(E\sqrt{E^2 - \|\vec{e}_{\mathcal{A}_\parallel}\|^2}\big)^2 - \|\vec{e}_{\mathcal{A}_\perp}\|^2 = (E^2 - \|\vec{e}_{\mathcal{A}_\parallel}\|^2)E^2 - \|\vec{e}_{\mathcal{A}_\perp}\|^2 = (E^2 - 1)\big(E^2 + \|\vec{e}_{\mathcal{A}_\perp}\|^2\big) \geq 0,$$

Hence,

$$\big(E\sqrt{E^2 - \|\vec{e}_{\mathcal{A}_\parallel}\|^2}\big)^2 - \|\vec{e}_{\mathcal{A}_\perp}\|^2 = \Big(E\sqrt{E^2 - \|\vec{e}_{\mathcal{A}_\parallel}\|^2} - \|\vec{e}_{\mathcal{A}_\perp}\|\Big)\Big(E\sqrt{E^2 - \|\vec{e}_{\mathcal{A}_\parallel}\|^2} + \|\vec{e}_{\mathcal{A}_\perp}\|\Big) \geq 0.$$

Since $E\sqrt{E^2 - \|\vec{e}_{\mathcal{A}_\parallel}\|^2} + \|\vec{e}_{\mathcal{A}_\perp}\| \geq 0$, we have

$$E\sqrt{E^2 - \|\vec{e}_{\mathcal{A}_\parallel}\|^2} - \|\vec{e}_{\mathcal{A}_\perp}\| \geq 0 \Longrightarrow \alpha_+ \;-\; \frac{1}{E} \;\geq\; 0.$$

Thus, α_+ violates (C2) when $E \geq 1$, making it infeasible.

Case B: $E < 1$**.** From $E = ||\vec{e}_{\mathcal{A}_\parallel}|| + \dfrac{1 - ||\vec{e}_{\mathcal{A}_\perp}||}{||\vec{e}_{\mathcal{A}_\parallel}||} < 1$ with $||\vec{e}_{\mathcal{A}_\parallel}||^2 + ||\vec{e}_{\mathcal{A}_\perp}||^2 = 1$, we deduce

$$1 - \|\vec{e}_{\mathcal{A}_\perp}\| < \|\vec{e}_{\mathcal{A}_\parallel}\|(1 - \|\vec{e}_{\mathcal{A}_\parallel}\|) \;\leq\; \tfrac{1}{4} \quad \Longrightarrow \quad \|\vec{e}_{\mathcal{A}_\perp}\| \geq \tfrac{3}{4} \;>\; \tfrac{1}{2}.$$

Evaluating $Q(\alpha^*)$ at $\alpha^* = ||\vec{e}_{\mathcal{A}_\parallel}||$ gives

$$\begin{aligned}
Q\big(\|\vec{e}_{\mathcal{A}_\parallel}\|\big) &= \Big(\Big(\|\vec{e}_{\mathcal{A}_\parallel}\| + \frac{1 - \|\vec{e}_{\mathcal{A}_\perp}\|}{\|\vec{e}_{\mathcal{A}_\parallel}\|}\Big)^2 + \|\vec{e}_{\mathcal{A}_\perp}\|^2\Big)\|\vec{e}_{\mathcal{A}_\parallel}\|^2 - 2\Big(\|\vec{e}_{\mathcal{A}_\parallel}\| + \frac{1 - \|\vec{e}_{\mathcal{A}_\perp}\|}{\|\vec{e}_{\mathcal{A}_\parallel}\|}\Big)\|\vec{e}_{\mathcal{A}_\parallel}\| + \|\vec{e}_{\mathcal{A}_\parallel}\|^2 \\
&= \Big(\|\vec{e}_{\mathcal{A}_\parallel}\| + \frac{1 - \|\vec{e}_{\mathcal{A}_\perp}\|}{\|\vec{e}_{\mathcal{A}_\parallel}\|}\Big)^2\|\vec{e}_{\mathcal{A}_\parallel}\|^2 + \|\vec{e}_{\mathcal{A}_\perp}\|^2\|\vec{e}_{\mathcal{A}_\parallel}\|^2 - 2\Big(\|\vec{e}_{\mathcal{A}_\parallel}\|^2 + 1 - \|\vec{e}_{\mathcal{A}_\perp}\|\Big) + \|\vec{e}_{\mathcal{A}_\parallel}\|^2 \\
&= \Big[\|\vec{e}_{\mathcal{A}_\parallel}\|^4 + 2(1 - \|\vec{e}_{\mathcal{A}_\perp}\|)\|\vec{e}_{\mathcal{A}_\parallel}\|^2 + (1 - \|\vec{e}_{\mathcal{A}_\perp}\|)^2\Big] + \|\vec{e}_{\mathcal{A}_\perp}\|^2\|\vec{e}_{\mathcal{A}_\parallel}\|^2 - 2\Big(\|\vec{e}_{\mathcal{A}_\parallel}\|^2 + 1 - \|\vec{e}_{\mathcal{A}_\perp}\|\Big) + \|\vec{e}_{\mathcal{A}_\parallel}\|^2 \\
&= \|\vec{e}_{\mathcal{A}_\parallel}\|^4 + \Big(2(1 - \|\vec{e}_{\mathcal{A}_\perp}\|) + \|\vec{e}_{\mathcal{A}_\perp}\|^2 - 1\Big)\|\vec{e}_{\mathcal{A}_\parallel}\|^2 + \Big((1 - \|\vec{e}_{\mathcal{A}_\perp}\|)^2 - 2(1 - \|\vec{e}_{\mathcal{A}_\perp}\|)\Big) \\
&= \|\vec{e}_{\mathcal{A}_\parallel}\|^4 + \big(\|\vec{e}_{\mathcal{A}_\perp}\|^2 - 2\|\vec{e}_{\mathcal{A}_\perp}\| + 1\big)\|\vec{e}_{\mathcal{A}_\parallel}\|^2 + \big(\|\vec{e}_{\mathcal{A}_\perp}\|^2 - 1\big) \\
&= \|\vec{e}_{\mathcal{A}_\parallel}\|^4 + \big((1 - \|\vec{e}_{\mathcal{A}_\parallel}\|^2) - 2\|\vec{e}_{\mathcal{A}_\perp}\| + 1\big)\|\vec{e}_{\mathcal{A}_\parallel}\|^2 + \big((1 - \|\vec{e}_{\mathcal{A}_\parallel}\|^2) - 1\big) \\
&= \|\vec{e}_{\mathcal{A}_\parallel}\|^4 + \big(2 - 2\|\vec{e}_{\mathcal{A}_\perp}\| - \|\vec{e}_{\mathcal{A}_\parallel}\|^2\big)\|\vec{e}_{\mathcal{A}_\parallel}\|^2 - \|\vec{e}_{\mathcal{A}_\parallel}\|^2 \\
&= \|\vec{e}_{\mathcal{A}_\parallel}\|^4 + \big(2 - 2\|\vec{e}_{\mathcal{A}_\perp}\|\big)\|\vec{e}_{\mathcal{A}_\parallel}\|^2 - \|\vec{e}_{\mathcal{A}_\parallel}\|^4 - \|\vec{e}_{\mathcal{A}_\parallel}\|^2 \\
&= \|\vec{e}_{\mathcal{A}_\parallel}\|^2\big(1 - 2\|\vec{e}_{\mathcal{A}_\perp}\|\big) \\
&< 0 \qquad (\text{since } \|\vec{e}_{\mathcal{A}_\perp}\| > \tfrac{1}{2}),
\end{aligned}$$

Because the quadratic coefficient $E^2 + ||\vec{e}_{\mathcal{A}_\perp}||^2 > 0$, one root lies below $||\vec{e}_{\mathcal{A}_\parallel}||$ and the other lies strictly above it. The larger root corresponds to α_+, so

$$\alpha_+ \;>\; \|\vec{e}_{\mathcal{A}_\parallel}\|,$$

which violates (C1). Therefore, α_+ is infeasible also when $E < 1$.

Combining **Cases A** and **B**, the α_+ can never simultaneously satisfies both (C1) and (C2).

Therefore,

$$\alpha^\star := \alpha_- = \frac{\left(\|\vec{e}_{\mathcal{A}_\parallel}\| + \frac{1-\|\vec{e}_{\mathcal{A}_\perp}\|}{\|\vec{e}_{\mathcal{A}_\parallel}\|}\right) - \|\vec{e}_{\mathcal{A}_\perp}\| \sqrt{\left(\|\vec{e}_{\mathcal{A}_\parallel}\| + \frac{1-\|\vec{e}_{\mathcal{A}_\perp}\|}{\|\vec{e}_{\mathcal{A}_\parallel}\|}\right)^2 - \|\vec{e}_{\mathcal{A}_\parallel}\|^2}}{\left(\|\vec{e}_{\mathcal{A}_\parallel}\| + \frac{1-\|\vec{e}_{\mathcal{A}_\perp}\|}{\|\vec{e}_{\mathcal{A}_\parallel}\|}\right)^2 + \|\vec{e}_{\mathcal{A}_\perp}\|^2}.$$

The optimal debiased embedding is

$$\vec{u}^\star = \sqrt{1-(\alpha^\star)^2}\,\frac{\vec{e}_{\mathcal{A}_\perp}}{\|\vec{e}_{\mathcal{A}_\perp}\|} \;+\; \alpha^\star\,\frac{\vec{e}_{\mathcal{A}_\parallel}}{\|\vec{e}_{\mathcal{A}_\parallel}\|}\,.$$

The attribute leakage and self-utility loss at $\vec{u}^\star$ are given by

$$L(\alpha^\star) = \alpha^\star \qquad \text{and} \qquad \tilde{L}(\alpha^\star) = \tilde{V}(\alpha^\star) \implies \frac{\alpha^\star}{\|\vec{e}_{\mathcal{A}_\parallel}\|} = \frac{V(\alpha^\star)}{1-\|\vec{e}_{\mathcal{A}_\perp}\|} \implies V(\alpha^\star) = (1-\|\vec{e}_{\mathcal{A}_\perp}\|)\,\frac{\alpha^\star}{\|\vec{e}_{\mathcal{A}_\parallel}\|}\,.$$

Finally, by **Proposition 1**, we have a tighter upper bound on the cross-utility loss as:

$$\ell_{\text{cross}} \le \sqrt{2(1-\|\vec{e}^{(I)}_{\mathcal{A}_\perp}\|)\,\frac{\alpha^\star}{\|\vec{e}^{(I)}_{\mathcal{A}_\parallel}\|}} + \sqrt{2(1-\|\vec{e}^{(T)}_{\mathcal{A}_\perp}\|)\,\frac{\alpha^\star}{\|\vec{e}^{(T)}_{\mathcal{A}_\parallel}\|}} \le \sqrt{2\,(1-\|\vec{e}^{(I)}_{\mathcal{A}_\perp}\|)} + \sqrt{2\,(1-\|\vec{e}^{(T)}_{\mathcal{A}_\perp}\|)}$$

□

B. Details of LLM

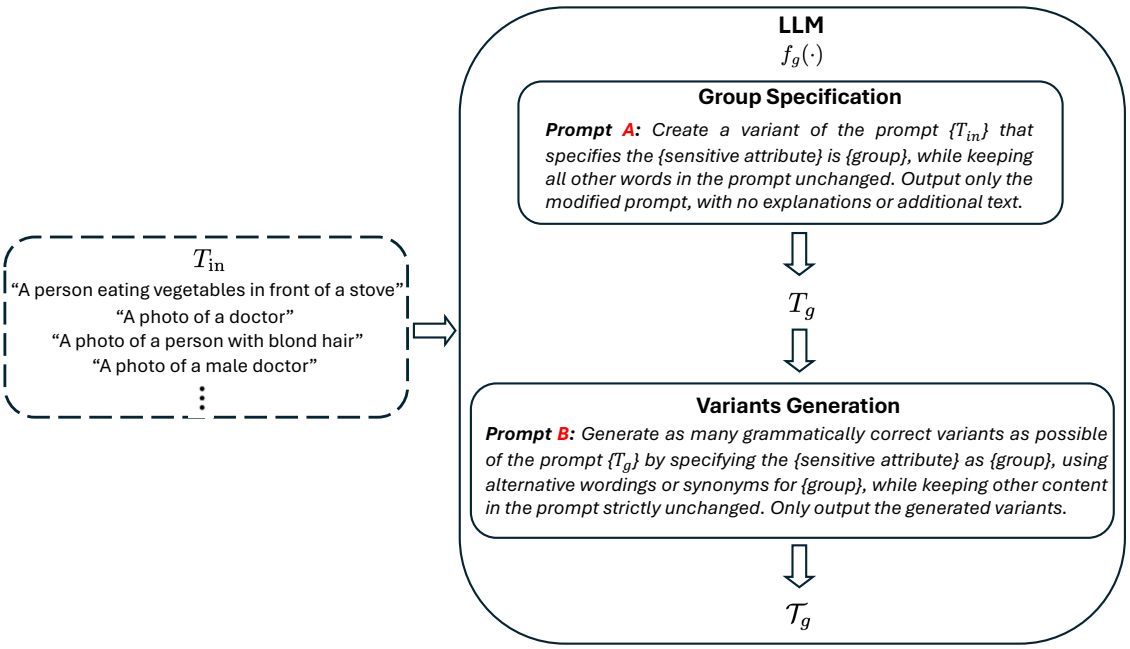

Figure 1. An illustration of the prompts we use in an LLM to insert the group specification and generate the variants for each group g.

As shown in Fig. 1, this module is template agnostic. The input prompt can take any form, whether neutral or group-specific, and all follow the same process. Once a sensitive attribute and its groups are defined, an LLM $f_g(\cdot)$ is used for each group to generate prompts. For each group, we collect the group-specific prompt T_g and its variants $\mathcal{T}_g$, which are then fed into the text encoder of a VLM as described in Section 4.1 of the main paper. As shown in Fig. 2 and Fig. 3, we provide several examples of the LLM inputs and outputs.

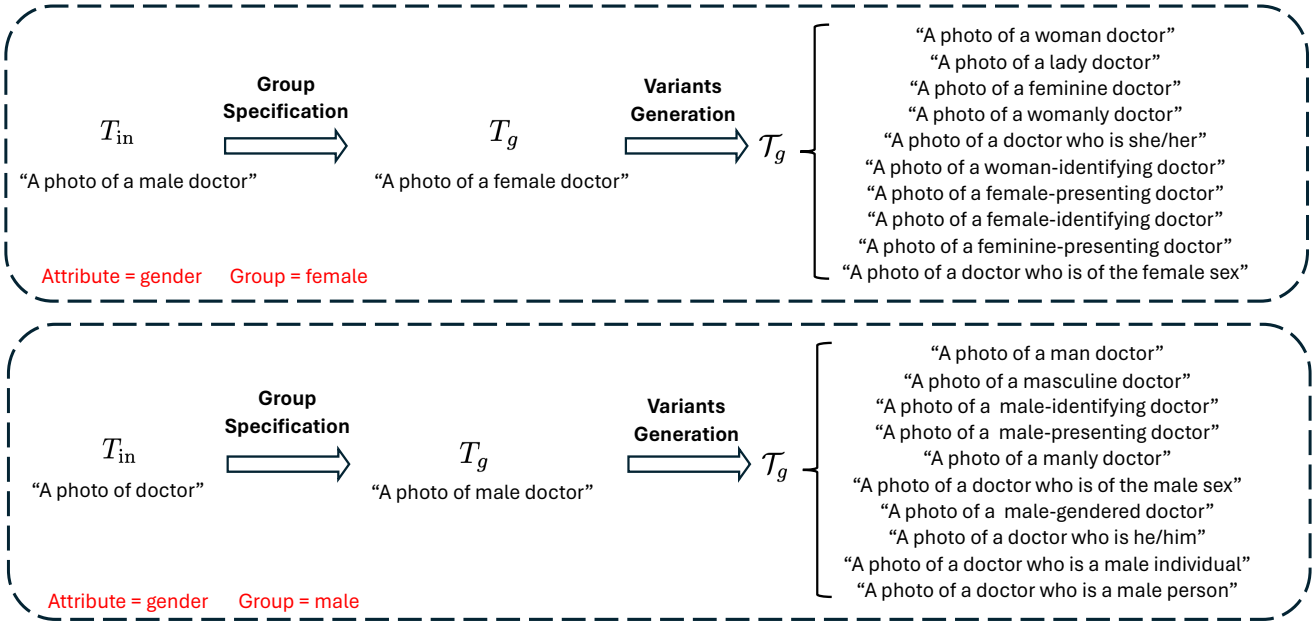

Figure 2. An illustration of the LLM output when the input is group-specific/neutral. The sensitive attribute is gender, and the selected group is female/male.

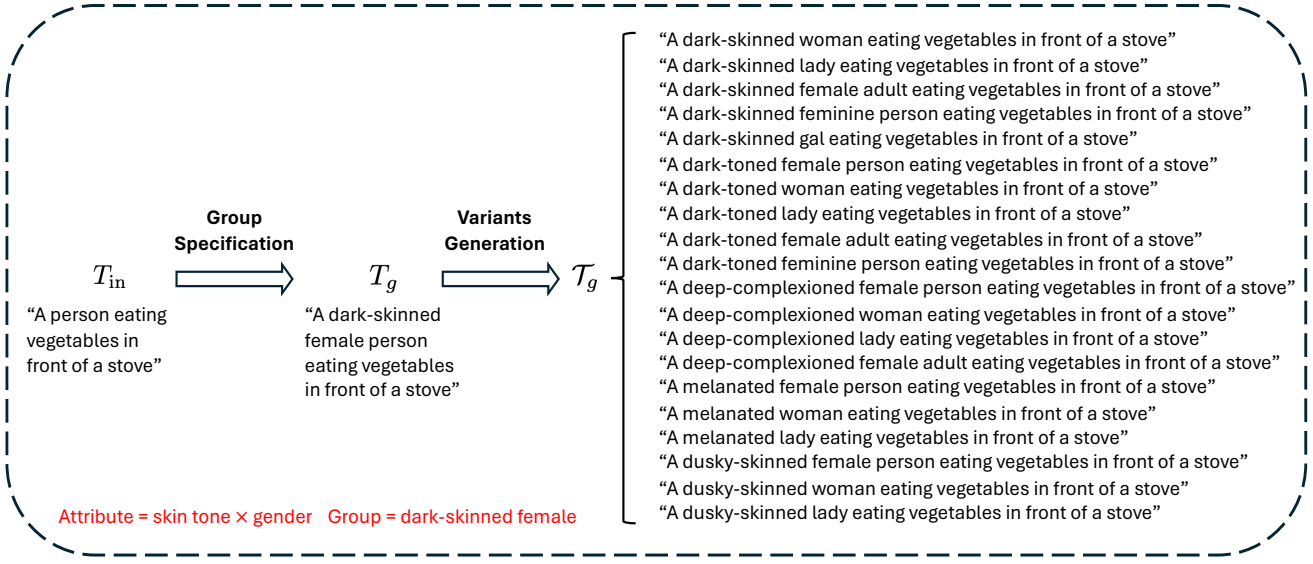

Figure 3. An illustration of the LLM output when the input is neutral. The sensitive attribute is the intersection of skin tone and gender, and the selected group is dark-skinned female.

C. Prompts in Text-to-Image Generation

We use the following 34 occupations: *doctor, aerospace engineer, computer programmer, electrical engineer, scientist, artist, designer, musician, painter, photographer, singer, writer, architect, civil engineer, engineer, software developer, childcare worker, coach, dentist, clerk, housekeeper, massage therapist, nurse, psychologist, social worker, teacher, professor, CEO, skateboarder, surfer, baseball player, football player, soccer player, tennis player*.

The occupation list is adapted from [2, 3] with several modifications for clarity and redundancy reduction. Specifically, we remove *maid* as it is inherently gendered, and delete *dental assistant* and *dental hygienist* due to their similarity to *dentist*. We also exclude *programmer* since *computer programmer* is already included, and remove *therapist* as it overlaps with *massage therapist*. Additionally, *author* is omitted because *writer* serves the same purpose, and *social assistant* is excluded due to the presence of *social worker*.

We then use these 34 occupations to create prompts following the template "a photo of a/an occupation." As shown in Fig. 4, this results in 34 neutral prompts for fairness evaluation. We also construct 34 group-explicit prompts to evaluate utility. For each occupation, the group-explicit prompt specifies the gender group that the original model is biased toward. This choice is intentional, as many debiasing methods tend to overcorrect and fail to generate images for the originally dominant group once debiased. For example, the original SD model tends to generate male images for the occupation *doctor*, while the debiased model balances the gender distribution but often struggles to generate male images when explicitly prompted, as it has been adjusted to produce more female representations.

Fairness

A photo of a doctor	A photo of a coach
A photo of an aerospace engineer	A photo of a dentist
A photo of a computer programmer	A photo of a clerk
A photo of an electrical engineer	A photo of a housekeeper
A photo of a scientist	A photo of a massage therapist
A photo of an artist	A photo of a nurse
A photo of a designer	A photo of a psychologist
A photo of a musician	A photo of a social worker
A photo of a painter	A photo of a teacher
A photo of a photographer	A photo of a professor
A photo of a singer	A photo of a CEO
A photo of a writer	A photo of a skateboarder
A photo of an architect	A photo of a surfer
A photo of a civil engineer	A photo of a baseball player
A photo of an engineer	A photo of a football player
A photo of a software developer	A photo of a soccer player
A photo of a childcare worker	A photo of a tennis player

Utility

A photo of a male doctor	A photo of a male coach
A photo of a male aerospace engineer	A photo of a male dentist
A photo of a male computer programmer	A photo of a female clerk
A photo of a male electrical engineer	A photo of a female housekeeper
A photo of a male scientist	A photo of a female massage therapist
A photo of a female artist	A photo of a female nurse
A photo of a female designer	A photo of a female psychologist
A photo of a male musician	A photo of a female social worker
A photo of a female painter	A photo of a female teacher
A photo of a female photographer	A photo of a male professor
A photo of a female singer	A photo of a male CEO
A photo of a male writer	A photo of a male skateboarder
A photo of a male architect	A photo of a male surfer
A photo of a male civil engineer	A photo of a male baseball player
A photo of a male engineer	A photo of a male football player
A photo of a male software developer	A photo of a male soccer player
A photo of a female childcare worker	A photo of a female tennis player

Figure 4. An illustration of neutral prompts used for fairness evaluation and group-explicit prompts used for utility evaluation in the text-to-image generation task.

D. Evaluations of Text-to-Image Generation

For the **fairness evaluation**, we use 34 neutral prompts. For each prompt, we generate 100 images using five models: the original SD, Orth-Proj, Orth-Cali, SFID, and ours, resulting in a total of **17,000 generated images**. For quantitative evaluation, we query BLIP-2 with the question: "What is the perceived gender of the person in this image?" and compute $\overline{\text{SP}}_{\mathcal{T}_o}$ based on BLIP-2's annotations. For qualitative evaluation, we select the top-5 prompts $\mathcal{T}_o^5 \subset \mathcal{T}_o$ showing the highest SP_t scores under the original SD model: *"A photo of a doctor," "A photo of a CEO," "A photo of a computer programmer," "A photo of a designer," and "A photo of a civil engineer."* Each prompt has 100 generated images per model, totaling **2,500 images** for manual inspection. Three independent annotators answer the same question asked of BLIP-2 with three options: *male*, *female*, or *unsure*. Images labeled as *unsure* or with disagreement among annotators are regenerated until unanimous agreement is reached. As a result, all analyzed images have unanimously agreed perceived gender labels (male or female).

For the **utility evaluation**, we use 34 group-explicit prompts. Again, 100 images are generated per prompt for each of the five models, resulting in another **17,000 images**. For quantitative evaluation, we compute the CLIP score to measure image–text alignment and report the average score across all 34 prompts. For qualitative evaluation, we select the top-5 prompts with the lowest CLIP scores under the original SD model: *"A photo of a male electrical engineer," "A photo of a female photographer," "A photo of a male professor," "A photo of a male dentist," and "A photo of a male football player."*

Each prompt produces 100 images per model, totaling **2,500 images** for manual inspection. The same three annotators identify images where the perceived gender does not match the one specified in the prompt. The number of mismatched images is denoted as N_I, and the *generation accuracy* is computed as $\text{Acc}^{\text{G}} = (N_G - N_I)/N_G$, where $N_G = 500$ is the total number of images per prompt. Images labeled as *unsure* or with non-unanimous annotations are regenerated until full agreement is reached, ensuring that all final images have unanimously agreed perceived gender labels.

Ethical Statement. We do not collect any personal information from the annotators, and all annotations are performed exclusively on AI-generated images. No personal data of real human subjects is involved in this study.

E. Datasets

CelebA. The CelebA is a binary facial attribute public dataset widely used as a benchmark for fairness evaluation in computer vision. It consists of facial attribute images, and we use its original validation and test sets, totaling 39,829 images for both utility and fairness evaluation.

FACET. The FACET dataset is a publicly available benchmark created by Meta AI for evaluating fairness in computer vision. It contains approximately 32,000 images, each annotated with various attributes, including perceived gender, age, and skin tone. The original dataset provides bounding boxes for each person. By cropping each individual from the original images, we obtain 49,550 single-person images. For our experiments, we use perceived gender as the sensitive attribute. To maintain consistency with previous debiasing studies [1, 5], we exclude samples labeled as non-binary or unsure. The dataset also includes 52 occupation classes, which serve as the classification labels. To ensure statistically reliable results, we remove classes with fewer than 1000 samples, resulting in 26,834 images across 18 occupation categories. The retained occupations are: *lawman, laborer, boatman, basketball_player, tennis_player, backpacker, speaker, soldier, farmer, guard, dancer, singer, ballplayer, soccer_player, repairman, guitarist, seller, motorcyclist.*

Flickr30K. Flickr30K is a public image–caption dataset containing 31,783 real-world images, each paired with up to five human-written descriptive sentences, resulting in a total of 158,915 image–text pairs. We first use YOLOv8 [4] to select images containing only one person. For these selected images, we retain captions that include gender-explicit words from the list *["man", "woman", "boy", "girl", "gentleman", "guy", "lady", "female", "male"]*. Captions containing any of *["man", "boy", "gentleman", "guy", "male"]* are labeled as *male*, while those with *["woman", "girl", "lady", "female"]* are labeled as *female*. After filtering, we obtain 3,079 image–text pairs for the retrieval task. To evaluate fairness, we further modify the captions by replacing gender-explicit words with the neutral term *"person"*, and these neutralized captions are used as text queries.

COCO2017. The COCO2017 dataset is one of the most widely used image–caption datasets in computer vision, containing five human-written captions per image. We use a subset of the original training and validation sets provided by [6], which includes explicit annotations for perceived gender and skin tone. These annotations were collected by multiple annotators using a more rigorous and reliable labeling protocol. In total, 28,316 images were annotated. We further exclude samples where either the perceived gender or skin tone is labeled as unsure, resulting in 1,368 image–text pairs. Similar to Flickr30K, to evaluate fairness, we modify the captions by replacing gender-explicit words with the neutral term *"person"*, and use these neutralized captions as text queries.

F. Results of Other VLMs

Table 1. Experimental results of CLIP (ResNet50) on the zero-shot image classification task.

Datasets	CelebA			FACET		
Evaluation Metric	F1 ↑	Δ_{EO}^{Avg} (G×A) ↓	Δ_{EO}^{Max} (G×A) ↓	Macro F1 ↑	Δ_{EO}^{Avg} (G) ↓	Δ_{EO}^{Max} (G) ↓
Baseline CLIP (ResNet50)	60.7±0.5	15.3±0.3	30.0±0.4	59.2±0.2	9.3±0.6	48.9±0.1
✔ 🔥 **I&T SFID**	56.4±0.2	14.6±0.5	24.9±0.3	[50.8]±0.6	9.4±0.4	47.2±0.2
✔ 🔥 **I&T FairerCLIP**	[56.6]±0.3	14.2±0.1	**[21.4]**±0.4	50.7±0.5	8.9±0.6	47.2±0.2
✘ 🔥 **I&T PRISM**	56.3±0.1	14.7±0.4	[21.8]±0.3	50.5±0.6	**[8.4]**±0.5	47.6±0.2
✘ ❄ **I&T PRISM-mini**	56.0±0.6	13.7±0.3	24.9±0.1	49.2±0.4	8.9±0.5	46.6±0.2
✘ ❄ **I&T RoboShot**	55.8±0.4	**[10.9]**±0.2	**[21.4]**±0.3	50.7±0.1	9.0±0.6	**[45.6]**±0.5
✘ ❄ **T Orth-Proj**	56.4±0.3	14.6±0.5	24.4±0.2	47.8±0.6	8.8±0.4	47.9±0.1
✘ ❄ **T Orth-Cali**	55.6±0.2	14.5±0.4	24.5±0.3	49.6±0.6	8.9±0.1	47.6±0.5
✘ ❄ **I&T Ours**	**[58.5]**±0.4	[11.5]±0.3	[21.8]±0.2	**[58.6]**±0.1	[8.5]±0.2	[45.8]±0.5

Table 2. Experimental results of CLIP (ResNet50) on the text-to-image retrieval task.

Datasets	COCO2017			Flickr30K		
Evaluation Metric	R@5 ↑	R@10 ↑	MS@1000 (G×ST) ↓	R@5 ↑	R@10 ↑	MS@1000 (G) ↓
Baseline CLIP (ResNet50)	71.4±0.5	81.8±0.1	13.3±0.3	78.5±0.6	85.9±0.4	18.6±0.5
✔ 🔥 **I&T SFID**	[66.6]±0.2	76.3±0.3	12.7±0.4	74.9±0.6	79.6±0.2	15.2±0.3
✔ 🔥 **I&T PromptArray**	66.5±0.5	[76.5]±0.4	12.2±0.1	74.9±0.3	[81.6]±0.3	15.4±0.5
✔ 🔥 **I&T FairerCLIP**	65.8±0.6	74.8±0.3	11.8±0.5	75.2±0.2	[81.6]±0.4	16.0±0.3
✔ ❄ **I&T CLIP-clip**	65.6±0.1	74.9±0.5	**[10.0]**±0.6	[75.3]±0.3	80.8±0.4	**[13.2]**±0.5
✘ ❄ **T Orth-Proj**	64.2±0.2	72.5±0.6	11.2±0.4	74.9±0.5	78.7±0.2	14.1±0.6
✘ ❄ **T Orth-Cali**	64.7±0.4	74.6±0.3	11.0±0.5	74.9±0.2	80.4±0.6	16.1±0.3
✘ ❄ **I&T Ours**	**[68.3]**±0.5	**[78.4]**±0.2	[10.1]±0.6	**[76.6]**±0.4	**[85.1]**±0.3	[13.6]±0.1

Table 3. Experimental results of BLIP on the zero-shot image classification task.

Datasets	CelebA			FACET		
Evaluation Metric	F1 ↑	Δ_{EO}^{Avg} (G×A) ↓	Δ_{EO}^{Max} (G×A) ↓	Macro F1 ↑	Δ_{EO}^{Avg} (G) ↓	Δ_{EO}^{Max} (G) ↓
Baseline BLIP	46.9±0.6	14.7±0.4	29.2±0.6	68.4±0.3	8.8±0.5	48.0±0.4
✔ 🔥 **I&T SFID**	[52.5]±0.2	**[12.1]**±0.1	26.4±0.3	60.6±0.1	8.2±0.2	47.1±0.5
✔ 🔥 **I&T FairerCLIP**	52.1±0.4	13.4±0.6	**[22.6]**±0.4	[61.0]±0.3	8.9±0.6	46.5±0.3
✘ 🔥 **I&T PRISM**	[5]±0.5	13.8±0.2	26.1±0.5	60.3±0.4	**[6.8]**±0.3	46.0±0.5
✘ ❄ **I&T PRISM-mini**	49.5±0.4	14.0±0.5	25.1±0.1	60.0±0.2	7.4±0.1	46.1±0.4
✘ ❄ **I&T RoboShot**	52.1±0.1	14.7±0.1	25.1±0.3	60.3±0.5	7.6±0.6	**[44.4]**±0.1
✘ ❄ **T Orth-Proj**	48.2±0.3	14.6±0.2	27.0±0.2	58.3±0.2	8.5±0.2	46.3±0.2
✘ ❄ **T Orth-Cali**	49.6±0.6	14.4±0.1	27.5±0.6	59.5±0.3	8.0±0.6	46.0±0.5
✘ ❄ **I&T Ours**	**[54.8]**±0.4	[12.2]±0.5	[23.2]±0.3	**[68.4]**±0.3	[7.3]±0.4	[44.7]±0.3

Table 4. Experimental results of BLIP on the text-to-image retrieval task.

Datasets	COCO2017			Flickr30K		
Evaluation Metric	R@5 ↑	R@10 ↑	MS@1000 (G×ST) ↓	R@5 ↑	R@10 ↑	MS@1000 (G) ↓
Baseline BLIP	93.9±0.5	97.2±0.1	15.1±0.6	93.9±0.4	96.8±0.2	18.3±0.6
✔ 🔥 **I&T SFID**	[90.6]±0.3	95.9±0.4	12.5±0.1	86.1±0.6	89.9±0.5	13.5±0.1
✔ 🔥 **I&T PromptArray**	90.4±0.2	[96.0]±0.3	12.9±0.4	86.4±0.3	91.5±0.4	12.8±0.5
✔ 🔥 **I&T FairerCLIP**	90.0±0.6	94.7±0.2	13.3±0.3	[91.0]±0.2	[91.7]±0.3	12.0±0.4
✔ ❄ **I&T CLIP-clip**	90.3±0.4	94.5±0.6	**[11.2]**±0.5	[91.0]±0.1	90.8±0.1	**[8.1]**±0.2
✘ ❄ **T Orth-Proj**	90.8±0.5	92.8±0.4	13.5±0.2	90.7±0.6	88.1±0.6	8.8±0.3
✘ ❄ **T Orth-Cali**	90.3±0.1	94.1±0.5	13.7±0.5	90.7±0.3	89.9±0.3	9.0±0.1
✘ ❄ **I&T Ours**	**[92.2]**±0.4	**[96.7]**±0.2	[12.0]±0.3	**[93.3]**±0.5	**[96.5]**±0.6	[8.4]±0.5

G. More Illustrative Examples

Figure 5. Illustrative examples for $o =$ "designer". We randomly sample ten generated images for each method. Male-looking samples are marked in red and numbered. On the left, a more balanced ratio of female- and male-looking samples indicates lower bias, while on the right, fewer male-looking samples reflect better preservation of self-utility.

Figure 6. Illustrative examples for $o =$ "computer programmer". We randomly sample ten generated images for each method. Female-looking samples are marked in red and numbered. On the left, a more balanced ratio of female- and male-looking samples indicates lower bias, while on the right, fewer female-looking samples reflect better preservation of self-utility.

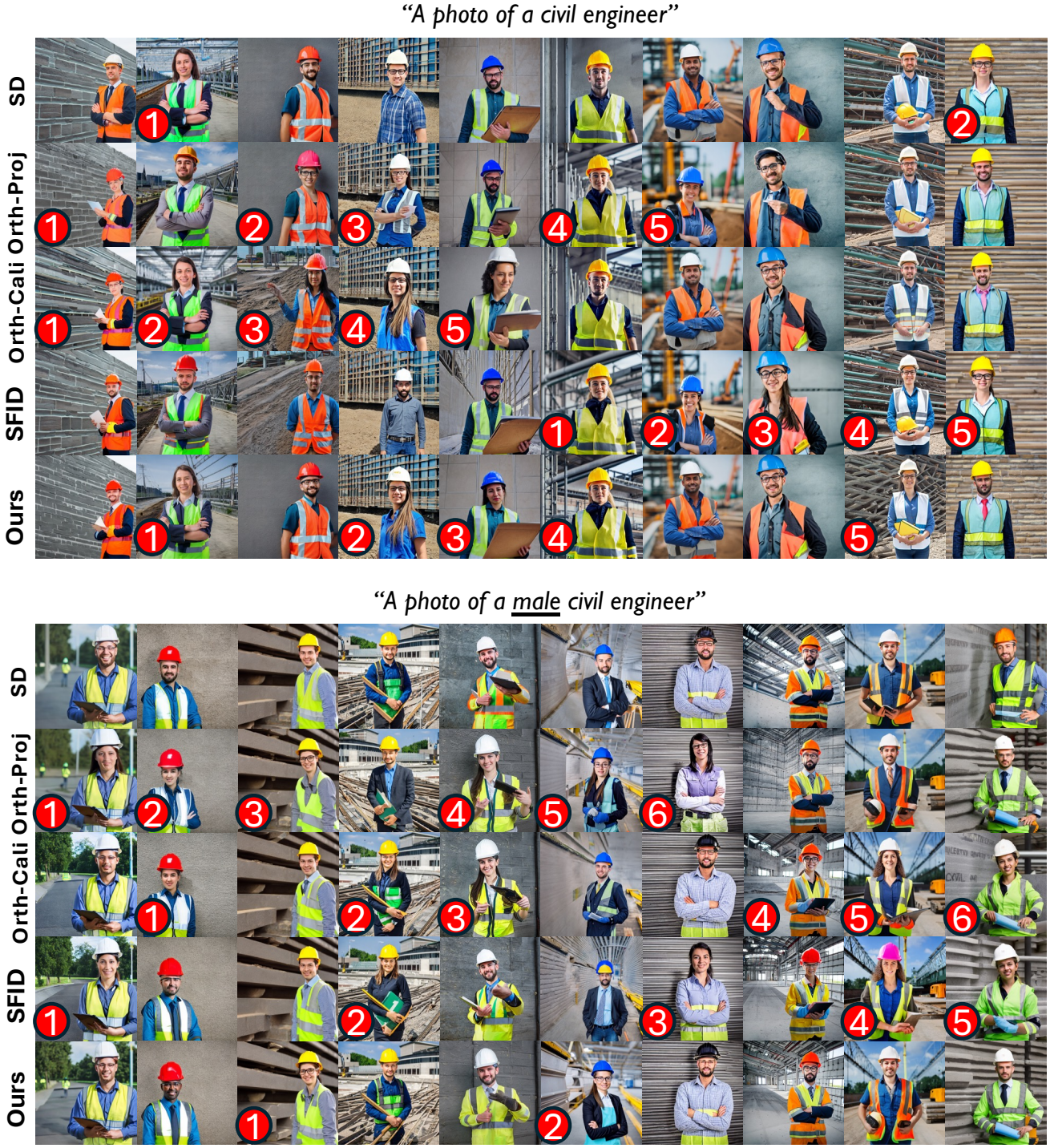

Figure 7. Illustrative examples for o = "civil engineer". We randomly sample ten generated images for each method. Female-looking samples are marked in red and numbered. On the left, a more balanced ratio of female- and male-looking samples indicates lower bias, while on the right, fewer female-looking samples reflect better preservation of self-utility.

Figure 8. Illustrative examples for $o =$ "CEO". We randomly sample ten generated images for each method. Female-looking samples are marked in red and numbered. On the left, a more balanced ratio of female- and male-looking samples indicates lower bias, while on the right, fewer female-looking samples reflect better preservation of self-utility.

References

[1] Ching-Yao Chuang, Varun Jampani, Yuanzhen Li, Antonio Torralba, and Stefanie Jegelka. Debiasing vision-language models via biased prompts. *arXiv preprint arXiv:2302.00070*, 2023. 9

[2] Felix Friedrich, Manuel Brack, Lukas Struppek, Dominik Hintersdorf, Patrick Schramowski, Sasha Luccioni, and Kristian Kersting. Fair diffusion: Instructing text-to-image generation models on fairness. *arXiv preprint arXiv:2302.10893*, 2023. 8

[3] Yusuke Hirota, Min-Hung Chen, Chien-Yi Wang, Yuta Nakashima, Yu-Chiang Frank Wang, and Ryo Hachiuma. SANER: Annotation-free societal attribute neutralizer for debiasing CLIP. In *The Thirteenth International Conference on Learning Representations*, 2025. 8

[4] Glenn Jocher, Ayush Chaurasia, Alex Stoken, Jirka Borovec, NanoCode012, Yonghye Kwon, Kalen Michael, TaoXie, Jiacong Fang, imyhxy, Lorna, Yifu Zeng, Colin Wong, Abhiram V, Diego Montes, Zhiqiang Wang, Cristi Fati, Jebastin Nadar, Laughing, and Mrinal Jain. ultralytics/yolov5: v7.0 - yolov5 sota realtime instance segmentation (v7.0), 2022. 9

[5] Hoin Jung, Taeuk Jang, and Xiaoqian Wang. A unified debiasing approach for vision-language models across modalities and tasks. *Advances in Neural Information Processing Systems*, 37:21034–21058, 2024. 9

[6] Dora Zhao, Angelina Wang, and Olga Russakovsky. Understanding and evaluating racial biases in image captioning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 14830–14840, 2021. 9