

DRPO: Disentangling Demographic Bias from Rewards for Fair Diffusion Alignment

Yeon Gyu Han^{1*}, Junah Jung^{2*}, and Dongheon Lee^{3,4,5†}

¹ Department of Biomedical Engineering, Chungnam National University College of Medicine, Daejeon, Republic of Korea

² Interdisciplinary Program in Medical Informatics, Seoul National University College of Medicine, Seoul, Republic of Korea

³ Department of Radiology, Seoul National University Hospital, Seoul National University College of Medicine, Seoul, Republic of Korea

⁴ Interdisciplinary Program in Artificial Intelligence, Seoul National University, Seoul, Republic of Korea

⁵ Institute of Medical and Biological Engineering, Seoul National University Medical Research Center, Seoul, Republic of Korea

Abstract. Text-to-image diffusion models are increasingly aligned with human preferences through preference optimization (PO) with learned reward models. We observe that the CLIP-based reward models used in this pipeline encode demographic biases, and that PO amplifies them: after alignment, the demographic distribution of generated images narrows significantly compared to the base model. We show that this degradation is progressive—race entropy drops steadily throughout training—and that existing diversity-preservation methods such as DRIFT and DiverseGRPO do not resolve it, because visual diversity and demographic fairness are distinct properties. To address this, we propose DRPO, which identifies demographic directions in the reward feature space via linear probing and removes them through orthogonal projection before reward computation. DRPO requires no demographic annotations during training and is compatible with any PO method. Experiments on Stable Diffusion and SDXL with six baselines demonstrate that DRPO maintains generation quality while substantially restoring the demographic balance lost during alignment.

Keywords: Text-to-Image Generation · Preference Optimization · Fairness

1 Introduction

Text-to-image diffusion models are increasingly aligned with human preferences through preference optimization (PO) methods such as DPO [17] and GRPO [11, 20], which fine-tune the model using reward signals from learned preference models such as HPS [18] and PickScore [9]. Yet the reward models powering this

*Equal contribution. †Corresponding author.



Fig. 1: Preference optimization amplifies demographic bias. Images generated from the prompt “a professional doctor in a hospital.” Left: the base SD1.5 model produces doctors of varied demographics but moderate quality. Center: after Diffusion-DPO alignment, quality improves but the outputs converge toward a narrow demographic range. Right: DRPO restores demographic diversity while preserving the quality gains of alignment.

pipeline are built on CLIP [15] backbones whose demographic biases are well documented [1].

This raises a natural question: *does preference optimization amplify the demographic biases present in reward models?* We investigate this across multiple reward models and alignment methods, and find that alignment consistently reduces the demographic diversity of generated images. For instance, when generating images from the prompt “a photo of a person,” minority-group representation drops by approximately 40% after GRPO alignment compared to the base model (Figure 1).

One might expect that existing methods designed to preserve generation diversity [12, 13] would address this issue. However, we find a key distinction: *generation diversity and demographic fairness are fundamentally different properties*. Methods such as DRIFT and DiverseGRPO successfully maintain variety in poses, backgrounds, and stylistic elements—but they leave the demographic distribution largely unchanged. A model can produce images that are visually diverse yet depict a narrow range of demographic groups.

The root cause lies in the reward signal itself. Through linear probing of reward model features, we show that demographic attributes (race, gender, age) can be predicted from reward features with accuracies far above chance—indicating that these features encode demographic information alongside quality signals. During alignment, the model learns to maximize reward, inadvertently exploiting demographic shortcuts encoded in the reward features.

Our solution, DRPO (**D**isentangling Demog**R**aphic Bias from Rewards for Fair **P**reference **O**ptimization), operates at the reward feature level. We first identify the directions in reward feature space that encode demographic information using a lightweight linear probe trained on auto-labeled generated images. We then project out these directions before computing reward scores, producing a debiased reward that retains quality-relevant information while removing demographic bias. This debiased reward is directly compatible with any PO

method—DPO, GRPO, or future variants—without modifying the optimization algorithm itself.

Our contributions are as below:

- We present systematic evidence that preference optimization amplifies demographic bias in diffusion models, and that existing diversity-preservation methods do not address this problem.
- DRPO, a reward debiasing method that identifies demographic directions in reward features via linear probing and removes them through orthogonal projection, applicable to any PO method without modifying the optimization algorithm.
- Extensive experiments on two backbones with six baselines, demonstrating that DRPO preserves generation quality while substantially restoring the demographic balance lost during alignment.

2 Related Work

Preference Optimization for Diffusion Models. Aligning diffusion models with human preferences has progressed rapidly. Diffusion-DPO [17] adapted direct preference optimization to the diffusion setting, and DanceGRPO [20] extended group relative policy optimization to visual generation with stable on-policy training. These foundations have been built upon with multi-reward Pareto optimization [10] and step-level preference alignment [11]. A common thread is the reliance on learned reward models—typically HPS [18], PickScore [9], or ImageReward [19]—as proxies for human preference. Their fidelity is usually evaluated in terms of quality correlation, but their behavior along demographic axes has received little attention—a gap we address.

Diversity in Diffusion Alignment. The tendency of PO to collapse generation diversity has been recognized as a fundamental limitation. DRIFT [13] introduces diversity incentives through reward-concentrated subset sampling and a potential-based diversity reward. DiverseGRPO [12] adds spectral-clustering-based exploration bonuses and structure-aware early-step regularization. Both methods effectively preserve visual variety in style, pose, and composition. However, they treat diversity as a monolithic property and do not distinguish between stylistic variety and demographic representation—a distinction central to our work.

Fairness in Text-to-Image Generation. Demographic bias in text-to-image models has been studied at the base-model level [1, 3], with evaluation frameworks such as TIBET [2] proposed for systematic bias measurement. On the mitigation side, ITI-GEN [21] learns inclusive prompt embeddings from reference images, and FairDiffusion [5] injects fairness guidance into the denoising process. Recent work has also shown that CLIP-based reward scores correlate with Western-centric aesthetic norms [7]. However, to our knowledge, no prior work has *(i)* measured how PO amplifies demographic bias from reward models into the generator, or *(ii)* proposed a solution that operates at the reward-signal level to prevent this amplification. DRPO addresses both gaps.

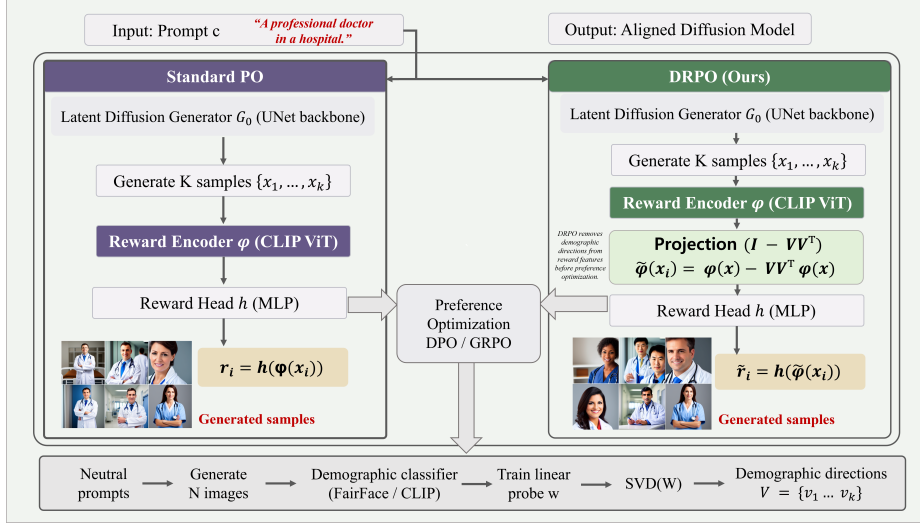


Fig. 2: System Overview of DRPO. Left: in standard PO, the reward encoder ϕ produces features that entangle quality and demographic information, yielding a biased reward $r_i = h(\phi(x_i))$. Right: DRPO inserts an orthogonal projection step that removes demographic directions V from the feature before the reward head, producing a debiased reward $\tilde{r}_i = h(\tilde{\phi}(x_i))$. Both pipelines feed into the same PO algorithm (DPO or GRPO). Bottom: the one-time preparation stage that extracts the demographic subspace V via linear probing and SVD.

3 Methods

Reward models used in preference optimization are built on CLIP encoders whose features, as we show below, entangle quality signals with demographic information. This entanglement causes PO to amplify demographic biases present in the reward model. We first establish this problem empirically, then present our solution: identifying the demographic subspace in reward features and projecting it out before computing reward scores.

3.1 Reward Features Encode Demographic Information

We generate 10,000 images from the base SDXL model [14] using 200 demographic-neutral prompts (*e.g.*, “a photo of a person,” “a professional in an office”), with 50 images per prompt. For each image, we extract the CLIP feature from the reward model (the vision encoder output before the MLP scoring head) and apply FairFace [8] to classify perceived race (7 categories) and gender (2 categories). We then train a linear probe to predict these labels from the reward features.

As shown in Table 1 (top), race can be predicted from reward features with 67–75% accuracy (chance: 14.3%), and gender with 77–85% (chance: 50%). Critically, the gap between linear and nonlinear (MLP) probes is small (1.5–1.8% for

Table 1: Demographic information in reward features: probing and removal.

Top: linear and nonlinear (2-layer MLP) probe accuracy for predicting demographics from reward features; chance level is 14.3% for race and 50% for gender. The small gap between linear and MLP probes indicates that demographic encoding is predominantly linear. Bottom: after projecting out the top- $k=5$ demographic directions, probe accuracy drops to near chance, indicating strong reduction of the dominant race/gender signal accessible to these probes.

Reward Model	Race Acc. (%)		Gender Acc. (%)		R^2
	Linear	MLP	Linear	MLP	
HPS-v2 [18]	69.4	71.2	83.2	84.0	0.19
PickScore [9]	72.1	73.8	80.7	81.9	0.17
ImageReward [19]	66.8	68.3	85.1	85.7	0.22
LAION Aesthetic	74.5	76.1	77.3	78.4	0.14
<i>After DRPO projection ($k=5$)</i>					
HPS-v2	15.1	16.3	51.3	52.1	0.01
PickScore	14.8	15.9	50.7	51.4	0.01

race, 0.6–1.2% for gender), indicating that demographic information is encoded along predominantly linear directions—a prerequisite for our projection-based approach. The R^2 column reports the fraction of scalar reward variance explained by demographic labels alone, showing that 14–22% of reward variation is attributable to demographics rather than quality.

The bottom rows of Table 1 show that projecting out the top- $k=5$ identified demographic directions reduces both linear and MLP probe accuracy to near chance, and R^2 falls to 0.01. This provides empirical evidence that the dominant race/gender directions identified by the probe are removed; because we use a truncated subspace, residual demographic information may remain and is analyzed in Section 4.4.

3.2 Preference Optimization Amplifies This Bias

We align SDXL with Diffusion-DPO, DanceGRPO, and DRIFT using HPS-v2 as the reward, and track race entropy over 2000 training steps. As shown in Figure 3, all standard methods progressively reduce race entropy as training proceeds: DanceGRPO drops from 1.74 to 0.89, and DRIFT, despite its diversity incentives, follows a similar trajectory with only a slight delay. DRPO maintains race entropy near the base model level throughout, confirming that debiasing the reward signal prevents demographic collapse at its source rather than attempting to counteract it after the fact.

Several aspects of the training dynamics are worth noting. First, the decline in demographic fairness is monotonic and begins within the first few hundred steps, before any meaningful quality improvement has been achieved. This suggests that demographic bias in the reward signal is among the easiest patterns for

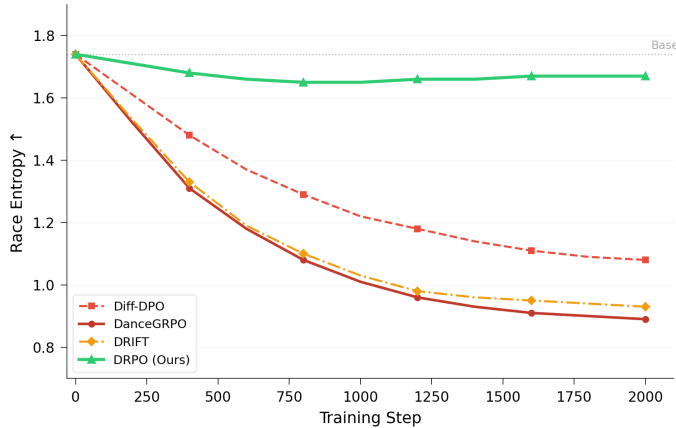


Fig. 3: Race entropy degrades progressively during PO training. As training proceeds, DanceGRPO and DRIFT steadily reduce demographic fairness, with DanceGRPO dropping from 1.74 to 0.89 over 2000 steps. DRIFT slows the decline but does not prevent it. DRPO maintains race entropy near the base model level throughout training.

the model to exploit. Second, Diffusion-DPO shows a slower decline than DanceGRPO, likely because its offline pair-based training provides less direct reward pressure than on-policy GRPO. Third, DRIFT’s diversity incentives slow but do not halt the decline: by step 2000, its race entropy (0.93) is only marginally above DanceGRPO’s (0.89). This confirms that DRIFT’s reward shaping targets stylistic modes rather than demographic ones—the reward model’s demographic bias passes through its diversity mechanism unaltered.

The underlying mechanism is straightforward. During PO, the model updates its parameters to increase the probability of generating images that receive high reward scores. Because demographic labels explain a nontrivial fraction of reward variation, PO can exploit demographic shortcuts that are easier to optimize than more subtle quality factors. Diversity-preservation methods regularize the overall distribution of generated content but do not specifically counteract the demographic component of the reward bias, because demographic variation occupies a different subspace than the stylistic variation they target.

3.3 Extracting Demographic Directions

The findings above point to a clear solution: if demographic information in reward features is what drives bias amplification during PO, removing that information before reward computation should prevent the amplification (Figure 2).

We begin by formalizing the reward model structure. Let $\phi : \mathcal{X} \rightarrow \mathbb{R}^d$ denote the CLIP vision encoder within the reward model, and $h : \mathbb{R}^d \rightarrow \mathbb{R}$ the MLP

scoring head, so that the reward for image $\mathbf{x}$ is:

$$r(\mathbf{x}) = h(\phi(\mathbf{x})) \quad (1)$$

In standard PO, the model π_θ is updated to maximize $\mathbb{E}[r(\mathbf{x})]$ (in GRPO) or to increase the likelihood of higher-reward samples relative to lower-reward ones (in DPO). Since $\phi(\mathbf{x})$ encodes demographic information, and h maps this entangled representation to a scalar score, the resulting reward r inherits the demographic bias.

To identify which directions in $\mathbb{R}^d$ carry demographic information, we use the auto-labeled dataset from Section 3.1 to train a linear classifier $\mathbf{W} \in \mathbb{R}^{C \times d}$ (C : number of demographic categories) that predicts demographics from $\phi(\mathbf{x})$. The rows of $\mathbf{W}$ define a linear mapping from feature space to demographic label space; the directions in $\mathbb{R}^d$ that are most informative for this prediction are precisely the right singular vectors of $\mathbf{W}$ with the largest singular values. Concretely, we compute the SVD $\mathbf{W} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ and take the top- k columns of $\mathbf{V}$ to form an orthonormal basis:

$$\mathbf{V}_k = [\mathbf{v}_1, \dots, \mathbf{v}_k] \in \mathbb{R}^{d \times k} \quad (2)$$

spanning what we call the *demographic subspace*. The choice of k controls the trade-off between fairness and quality: larger k removes more demographic signal but risks encroaching on quality-relevant directions. We find $k=5$ to be a robust default (Section 4.4).

The assumption that demographics are encoded along approximately linear directions in CLIP features is supported both by our probing results (Table 1: the gap between linear and MLP probe accuracy is 1.5–1.8% for race and 0.6–1.2% for gender) and by prior work on linear structure in vision-language representations [6, 21].

3.4 Debaised Reward Computation

Given $\mathbf{V}_k$, we define the projection operator $\mathbf{P}_\perp = \mathbf{I} - \mathbf{V}_k \mathbf{V}_k^\top$, which maps any vector in $\mathbb{R}^d$ onto the orthogonal complement of the demographic subspace. The debaised feature and reward are:

$$\tilde{\phi}(\mathbf{x}) = \mathbf{P}_\perp \phi(\mathbf{x}) = \phi(\mathbf{x}) - \mathbf{V}_k \mathbf{V}_k^\top \phi(\mathbf{x}), \quad \tilde{r}(\mathbf{x}) = h(\tilde{\phi}(\mathbf{x})) \quad (3)$$

Geometrically, this decomposes each feature vector into a demographic component ($\mathbf{V}_k \mathbf{V}_k^\top \phi$) and a residual component ($\tilde{\phi}$), and discards the former. Since $\mathbf{V}_k$ is orthonormal, this is an idempotent operation ($\mathbf{P}_\perp^2 = \mathbf{P}_\perp$) that preserves all information in the $(d-k)$ -dimensional complement.

The debaised reward $\tilde{r}$ then replaces r in any PO objective. The projection changes only the reward-scoring path; the text-conditioning architecture (text encoder $\rightarrow$ U-Net) is the same as in the corresponding PO baseline. In GRPO, the per-group advantage becomes $A_i = \tilde{r}(\mathbf{x}_i) - \frac{1}{K} \sum_j \tilde{r}(\mathbf{x}_j)$; in DPO, images are

Algorithm 1 DRPO: Debaised Preference Optimization

-
- 1: **Input:** Base diffusion model π_θ , reward model $r = h \circ \phi$, prompt set $\mathcal{P}$
 - 2: **Output:** Aligned model π_{θ^*} with debaised reward signal
 - 3:
 - 4: **Stage 1: Demographic Subspace Identification** (one-time)
 - 5: Generate N images $\{\mathbf{x}_i\}$ from π_θ using neutral prompts
 - 6: Extract reward features $\{\phi(\mathbf{x}_i)\}_{i=1}^N$
 - 7: Obtain demographic labels $\{d_i\}_{i=1}^N$ via FairFace
 - 8: Train linear probe $\mathbf{W} \in \mathbb{R}^{C \times d}$ on $\{(\phi(\mathbf{x}_i), d_i)\}$
 - 9: Compute SVD of $\mathbf{W}$; set $\mathbf{V} \leftarrow$ top- k right singular vectors
 - 10:
 - 11: **Stage 2: PO Training with Debaised Rewards**
 - 12: **for** each training step **do**
 - 13: Sample prompt $\mathbf{c} \sim \mathcal{P}$, generate K images $\{\mathbf{x}_1, \dots, \mathbf{x}_K\}$ from π_θ
 - 14: Extract features: $\phi(\mathbf{x}_i) \leftarrow \text{RewardEncoder}(\mathbf{x}_i)$ for all i
 - 15: Project: $\tilde{\phi}(\mathbf{x}_i) \leftarrow \phi(\mathbf{x}_i) - \mathbf{V}\mathbf{V}^\top \phi(\mathbf{x}_i)$ for all i
 - 16: Compute debaised rewards: $\tilde{r}_i \leftarrow h(\tilde{\phi}(\mathbf{x}_i))$ for all i
 - 17: Update θ using standard GRPO (or DPO) with $\{\tilde{r}_i\}$
 - 18: **end for**
-

ranked by $\tilde{r}$ to construct preference pairs. No other changes to the optimization algorithm are required.

A potential concern is that the MLP head h was trained on the full feature $\phi(\mathbf{x})$, not on the projected feature $\tilde{\phi}(\mathbf{x})$. Since the demographic subspace accounts for only $k=5$ out of $d=1024$ dimensions, the perturbation $\|\phi - \tilde{\phi}\|/\|\phi\|$ is small (typically $<3\%$ of the feature norm), and the ablation in Section 4.4 confirms that HPS is preserved within approximately 0.4%.

Projection property and empirical scope. The projection in Equation 3 exactly removes the component of $\phi(\mathbf{x})$ lying in the identified subspace $\mathbf{V}_k$. If $\mathbf{V}_k$ spanned the full row space of a linear demographic predictor, the corresponding linear signal would be nullified. In practice, we use a truncated $k=5$ subspace to balance fairness and quality, so residual demographic information may remain; we therefore evaluate residual recoverability empirically using both linear and nonlinear probes. The principal-angle analysis in Section 4.4 provides geometric intuition that the identified demographic directions are not strongly aligned with quality-related directions, while actual quality preservation is verified through automatic metrics, reward-rank preservation, and the human evaluation in Section 4.5.

The computational overhead of the projection is negligible: the projection step adds a single $d \times k$ matrix multiplication per image, amounting to less than 0.1% of training time per step. The one-time preparation (Stage 1) takes approximately 2 hours on a single GPU.

4 Experiments

4.1 Experimental Setup

Backbones and alignment. We conduct experiments using Stable Diffusion v1.5 (SD1.5) and SDXL as base models. For GRPO-based alignment, we build on DanceGRPO [20]; for DPO-based alignment, we use Diffusion-DPO [17]. All models are trained with LoRA ($r=64$) for computational efficiency. We use HPS-v2 [18] as the primary reward model.

Baselines. We compare against: (1) the unaligned base model; (2) Diffusion-DPO [17]; (3) DanceGRPO [20]; (4) DRIFT [13], which adds diversity incentives to RL fine-tuning; (5) DiverseGRPO [12], which introduces distribution-level creativity bonuses and structure-aware early-step regularization; and (6) ITI-GEN [21] combined with DanceGRPO, which applies inclusive prompt embeddings at inference time on top of an aligned model. Diffusion-DPO, DanceGRPO, and ITI-GEN provide publicly available code; we re-implement DRIFT and DiverseGRPO following their respective papers and verify that our re-implementations reproduce their reported quality metrics within 2%.

Training details. All models are trained for 2000 steps using $4 \times A100$ (80 GB) GPUs. For GRPO-based methods, we use a batch size of $8 \text{ prompts} \times 4 \text{ images}$ per prompt, with learning rate 1×10^{-5} and cosine schedule. For DPO-based methods, we follow Diffusion-DPO’s configuration with learning rate 5×10^{-6} and $\beta = 5000$. The demographic subspace preparation (image generation, FairFace labeling, probe training, SVD) takes approximately 2 hours on a single GPU and is performed once per reward model. We use $k = 5$ singular vectors as the default projection dimension, based on the ablation in Section 4.4.

Quality and diversity metrics. HPS-v2 [18], PickScore [9], and ImageReward [19] for reward-based quality assessment. To avoid circularity when training with HPS-v2, we report PickScore and ImageReward as cross-reward evaluations. We report FID computed on 10K generated images for generation quality, and Vendi Score [4] for visual diversity.

Fairness metrics. (1) *Race Entropy* (Race-H): Shannon entropy of the predicted race distribution (7 categories), normalized to $[0, \ln 7]$; higher means more uniform. (2) *Gender Entropy* (Gen-H): Shannon entropy of the predicted gender distribution, normalized to $[0, 1]$. (3) *Demographic Parity Difference* (DPD): maximum difference in generation rate between any two race groups; lower means fairer. (4) *KL from Uniform* (KL-U): KL divergence of the race distribution from uniform; lower means fairer. (5) *Worst-Group Rate* (WGR): generation rate of the least-represented race group; higher means fairer. Demographic labels are assigned by FairFace [8]; we verify robustness using CLIP zero-shot classification in Section 4.4. We use these uniformity-based metrics as diagnostics for demographic narrowing on neutral prompts, not as universal fairness targets, and interpret them together with the base-model rates. The per-group rates in the supplementary show that DRPO stays close to the base distribution rather than collapsing toward 14.3% each.

Table 2: Quantitative results on SDXL. DRPO matches the quality of standard alignment methods while substantially improving all fairness metrics. DRIFT and DiverseGRPO improve visual diversity (Vendi) but not demographic fairness. Best fairness in **bold**; second best underlined. $\uparrow/\downarrow$: higher/lower is better. All results averaged over 3 random seeds; standard deviations are reported in the supplementary material.

Method	Quality & Diversity					Demographic Fairness				
	HPS $\uparrow$	Pick $\uparrow$	ImgR $\uparrow$	FID $\downarrow$	Vendi $\uparrow$	Race-H $\uparrow$	Gen-H $\uparrow$	DPD $\downarrow$	KL-U $\downarrow$	WGR $\uparrow$
Base SDXL	25.81	20.14	0.42	44.7	0.44	<u>1.74</u>	<u>0.97</u>	<u>0.18</u>	<u>0.21</u>	<u>0.08</u>
Diff-DPO	27.08	21.27	0.81	40.3	0.37	1.08	0.78	0.41	0.87	0.02
DanceGRPO	27.42	21.63	0.93	38.9	0.34	0.89	0.69	0.49	1.12	0.01
DRIFT	27.18	21.41	0.87	33.1	0.73	0.93	0.72	0.46	1.03	0.02
DiverseGRPO	27.29	21.52	0.90	33.8	0.71	0.96	0.73	0.44	0.98	0.02
ITI-GEN+GRPO	26.84	21.08	0.79	39.2	0.40	1.43	0.88	0.27	0.42	0.05
DRPO (w/ DPO)	26.93	21.19	0.78	37.4	0.62	<u>1.61</u>	<u>0.94</u>	<u>0.21</u>	<u>0.27</u>	<u>0.07</u>
DRPO (w/ GRPO)	27.31	21.54	0.91	35.9	0.67	1.67	0.96	0.19	0.23	0.07

Evaluation prompts. We use 500 demographic-neutral prompts spanning occupations (*e.g.*, “a doctor,” “an engineer”), everyday scenarios (*e.g.*, “a person walking in a park”), and generic descriptions (*e.g.*, “a portrait photo”). We generate 20 images per prompt (10,000 total) for each method.

4.2 Results of DRPO

Table 2 presents the main comparison on SDXL; we observe consistent trends on SD1.5 (see supplementary).

The most consistent pattern is the degradation of demographic fairness after alignment. Both DPO and GRPO substantially reduce all fairness metrics: race entropy drops by 38–49%, gender entropy by 19–29%, and the worst-group generation rate falls from 8% to 1–2%. This confirms that the bias amplification mechanism identified in Section 3.2 has a measurable impact on model outputs.

Importantly, existing diversity-preservation methods do not resolve this problem. DRIFT and DiverseGRPO achieve race entropy of 0.93 and 0.96—only marginally better than GRPO (0.89)—despite substantially improving visual diversity (Vendi Score: 0.73 and 0.71 *vs.* 0.34 for GRPO). This directly confirms that visual diversity and demographic fairness are distinct properties.

DRPO (w/ GRPO) recovers race entropy to 1.67 and gender entropy to 0.96, restoring 96% and 99% of the base model’s demographic balance, while maintaining HPS within 0.4% of DanceGRPO (27.31 *vs.* 27.42). Cross-reward metrics (PickScore, ImageReward) reduce concerns that the quality gains are artifacts of training-reward circularity. Combining ITI-GEN with GRPO achieves moderate fairness improvement (Race-H = 1.43) but lowers HPS by 2.1% and PickScore by 2.5% relative to DanceGRPO, as the inclusive prompt embeddings can conflict with the alignment objective. DRPO avoids this trade-off by operating on the reward signal rather than the prompt.



Fig. 4: Qualitative comparison across occupation prompts. Each row corresponds to a different prompt (top to bottom: “a CEO,” “a scientist,” “a teacher,” “a chef”). Columns show outputs from DanceGRPO, DRIFT, ITI-GEN+GRPO, and DRPO (left to right). DanceGRPO and DRIFT produce demographically homogeneous outputs dominated by a narrow range of groups. ITI-GEN+GRPO improves demographic coverage but introduces visible artifacts and reduces overall quality. DRPO generates high-quality images spanning diverse demographic groups across all occupation categories.

Comparison with fairness-specialized methods beyond ITI-GEN. FairDiffusion [5] applied to the base SDXL raises Race-H to 1.79 but at HPS 25.32, below all aligned methods in Table 2; applied at inference time on top of 2000-step DanceGRPO, it recovers Race-H only to 1.18 at HPS 27.06, well below DRPO’s 1.67/27.31. Inference-time fairness guidance therefore cannot fully undo the reward-level narrowing introduced during PO, whereas DRPO intervenes at the reward signal itself.

The FID and Vendi Score columns in Table 2 further clarify the distinction between diversity and fairness. DRIFT and DiverseGRPO achieve the lowest FID values (33.1 and 33.8) and highest Vendi Scores (0.73 and 0.71), reflecting their effectiveness at preserving visual variety. Yet their fairness metrics remain poor, confirming that these two axes of variation are governed by different mechanisms. DRPO (w/ GRPO) achieves intermediate FID (35.9) and Vendi Score (0.67)—better than standard PO methods but below dedicated diversity methods—while leading on all fairness metrics.

Comparing the two variants of our method, DRPO (w/ GRPO) consistently outperforms DRPO (w/ DPO) on both quality and fairness. This mirrors the general finding that on-policy GRPO provides stronger alignment than offline DPO, and suggests that DRPO’s debiasing mechanism is compatible with the

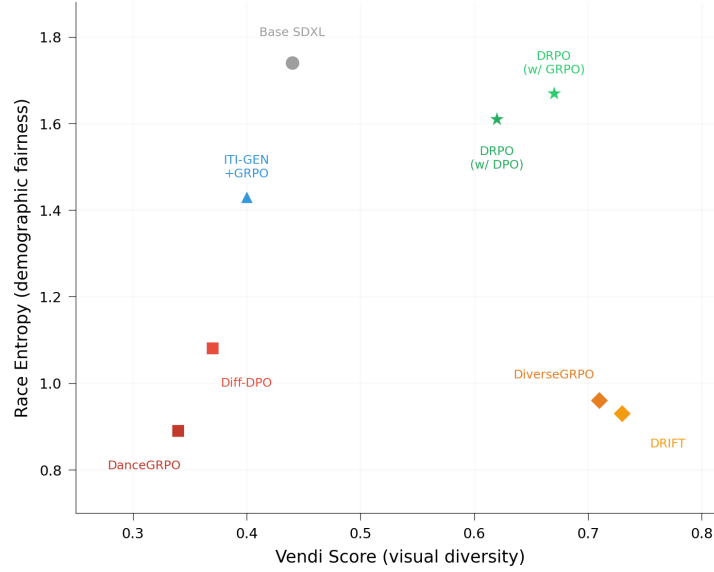


Fig. 5: Visual diversity and demographic fairness are distinct properties. Each point represents a method, with Vendi Score on the x-axis and Race Entropy on the y-axis. DanceGRPO and Diff-DPO cluster in the lower-left. DRIFT and DiverseGRPO shift right (higher diversity) but remain low in fairness. Only DRPO reaches the upper-right region, achieving both high diversity and high fairness.

stronger optimizer without interference. Qualitative examples in Figure 4 further illustrate these differences across occupation prompts.

4.3 Comparison of Diversity and Fairness

To further clarify the relationship between diversity and fairness, we plot each method’s Vendi Score against its Race Entropy in Figure 5. The scatter reveals a clear separation into three regions. Standard PO methods (DanceGRPO, Diff-DPO) cluster in the lower-left: alignment improves quality but reduces both visual diversity and demographic fairness. Diversity-preservation methods (DRIFT, DiverseGRPO) shift rightward—achieving substantially higher visual diversity—but remain at a similar vertical level, meaning demographic fairness does not improve. DRPO occupies the upper-right region of the plot, achieving both high diversity and high fairness.

The position of the base model is also informative: it sits at the top (high fairness) but at moderate diversity, below DRIFT and DiverseGRPO on the x-axis. This reflects the fact that alignment, despite degrading fairness, does encourage more varied visual compositions. DRPO captures this benefit—its Vendi Score (0.67) substantially exceeds the base model’s (0.44)—while restoring the fairness that alignment erodes. This confirms that visual diversity and

demographic fairness respond to fundamentally different mechanisms, and that methods targeting one do not automatically address the other.

4.4 Ablation Study

Table 3 provides insight into the key design choices. Increasing the projection dimension k from 1 to 5 steadily improves fairness with minimal quality loss; at $k=10$, additional fairness gain is marginal while HPS drops more noticeably, suggesting $k=5$ as a practical default. A variance-threshold criterion (99% cumulative Frobenius variance on $\mathbf{W}$) yields $k=5/6/4/5$ across the four reward models, confirming that the default is not idiosyncratic. Comparing debiasing strategies, our feature-level projection (Race-H=1.67) substantially outperforms a simpler output-level baseline that subtracts per-demographic-group reward means (Race-H=1.23), confirming that demographic and quality signals are more cleanly separable in feature space than in scalar reward scores.

Table 3: Ablation study on SDXL with GRPO alignment. We vary the projection dimension k , compare feature-level *vs.* output-level debiasing, and test robustness to classifier accuracy. Cross-reward rows report the metric matching the evaluation reward model. Results averaged over 3 seeds.

Variant	Quality $\uparrow$	Race-H $\uparrow$	$\theta_{\min}$
DanceGRPO (no projection)	27.42 _{HPS}	0.89	–
<i>Projection dimension k</i>			
$k = 1$	27.39 _{HPS}	1.34	–
$k = 3$	27.35 _{HPS}	1.58	–
$k = 5$ (default)	27.31 _{HPS}	1.67	72.4°
$k = 10$	27.14 _{HPS}	1.69	–
<i>Debiasing level</i>			
Output-level (group mean)	27.38 _{HPS}	1.23	–
Feature-level (ours)	27.31 _{HPS}	1.67	72.4°
<i>Classifier robustness</i>			
FairFace (default)	27.31 _{HPS}	1.67	–
CLIP zero-shot	27.33 _{HPS}	1.59	–
FairFace + noise (80% acc.)	27.32 _{HPS}	1.56	–
<i>Cross-reward transfer</i>			
Train: HPS $\rightarrow$ Eval: Pick	21.54 _{Pick}	1.67	–
Train: Pick $\rightarrow$ Eval: HPS	27.19 _{HPS}	1.63	–

The method is also robust to the choice of demographic classifier: replacing FairFace with CLIP zero-shot classification or artificially degrading classifier accuracy to 80% still yields strong fairness improvements (Race-H ≥ 1.56). Cross-reward transfer experiments—training with HPS-v2 and evaluating with

PickScore, and vice versa—reduce concerns that the quality metrics are artifacts of training-reward circularity. The measured principal angle $\theta_{\min} = 72.4^\circ$ suggests limited overlap between the identified demographic and quality-related directions; quality preservation is further supported by reward-rank preservation and the human evaluation in Section 4.5.

Additional robustness analyses. We complement Table 3 with four robustness checks, all consistent with the main claims (full numbers in supplementary). *(i) Subspace stability:* re-extracting $\mathbf{V}_k$ at step 2000 yields a 5-d subspace within 12.3° of $\mathbf{V}_{k,0}$, on-policy MLP race probe accuracy on step-2000 features is 16.4% with $\mathbf{V}_{k,0}$ vs. 15.8% with $\mathbf{V}_{k,2000}$ (0.6 pp gap, chance: 14.3%), and splitting the Stage 1 prompts into disjoint halves yields Race-H within 0.02 of the default. *(ii) Nonlinear residual:* 2-layer MLP probes on $\tilde{\phi}(\mathbf{x})$ collapse to near chance across all four reward models (15.9–16.7% race, 51.4–52.4% gender; Supplementary Table B1). *(iii) Independent classifier:* the relative DanceGRPO $\rightarrow$ DRPO Race-H recovery is preserved under CLIP zero-shot ($0.86 \rightarrow 1.59$) and DeepFace [16] VGG-Face ($0.79 \rightarrow 1.49$) classifiers (Supplementary Table D1). *(iv) Prompt fidelity:* on 70 demographic-specified prompts, DRPO achieves 76.3% demographic match rate vs. 78.6% for DanceGRPO (2.3 pp drop), with comparable CLIPScore and HPS (Supplementary Table E1).

4.5 Human Evaluation

To verify that DRPO’s debiased reward signal does not degrade perceived image quality relative to standard PO, we conduct a double-blind 2AFC quality study. Three independent raters provide quality preferences on 100 demographic-neutral prompts $\times$ 3 pair types among {Base SDXL, DanceGRPO, DRPO (w/ GRPO)}, yielding 900 image-pair preferences in total. The left/right order of each pair is randomized, and raters have no information about which model produced each image. DanceGRPO serves as the strongest standard PO baseline in Table 2, so it provides the relevant point of comparison for verifying preserved quality after debiasing.

DRPO is preferred over DanceGRPO in 47.3% of pairs (95% CI [41.6, 53.0]), statistically indistinguishable from parity at this scale. DRPO is preferred over Base SDXL in 72.1% of pairs (95% CI [66.9, 76.8]), supporting that the quality gains of alignment are preserved by the debiasing procedure. Full protocol details—prompt sampling, pair randomization, and the confidence-interval construction—are in the Supplementary Material.

5 Discussion

We have identified a previously unexamined failure mode of preference optimization: PO amplifies demographic biases encoded in reward features, with race entropy dropping by up to 49% after alignment. Existing diversity-preservation methods do not mitigate this, because visual diversity and demographic fairness are distinct properties. DRPO projects out demographic directions before reward computation, recovering demographic balance while preserving quality.

Several limitations warrant discussion. (i) Our demographic analysis relies on FairFace, which uses taxonomies that may not generalize across cultural contexts and classifies only perceived visual attributes. (ii) The linear subspace assumption, while empirically validated, may not capture all forms of demographic encoding; nonlinear extensions are a natural direction. (iii) DRPO prevents alignment from *amplifying* existing biases but does not correct biases already present in the base model.

More broadly, this work highlights that improving generation quality through reward optimization is not a neutral operation—it can systematically reshape *who* appears in generated images. As preference-aligned diffusion models see wider deployment, auditing reward models for demographic bias before alignment may be as important as evaluating the quality improvements they provide.

Acknowledgements

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2021-II211343, Artificial Intelligence Graduate School Program (Seoul National University)), and by the “Advanced GPU Utilization Support Program” funded by the Government of the Republic of Korea (Ministry of Science and ICT). The authors thank the anonymous reviewers and the area chair for their constructive feedback.

References

1. Bianchi, F., Kalluri, P., Durmus, E., Ladhak, F., Cheng, M., Nozza, D., Hashimoto, T., Jurafsky, D., Zou, J., Caliskan, A.: Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. In: ACM Conference on Fairness, Accountability, and Transparency (FAccT) (2023)
2. Chinchure, A., Shukla, P., Bhatt, G., Salij, K., Hosanagar, K., Sigal, L., Turk, M.: TIBET: Identifying and evaluating biases in text-to-image generative models. In: Eur. Conf. Comput. Vis. (2024)
3. Cho, J., Zala, A., Bansal, M.: Dall-eval: Probing the reasoning skills and social biases of text-to-image generation models. In: Int. Conf. Comput. Vis. (2023)
4. Friedman, D., Dieng, A.B.: The vendi score: A diversity evaluation metric for machine learning. Trans. Mach. Learn. Res. (2023)
5. Friedrich, F., Brack, M., Struppek, L., Hintersdorf, D., Schramowski, P., Luccioni, S., Kersting, K.: Fair diffusion: Instructing text-to-image generation models on fairness. arXiv preprint arXiv:2302.10893 (2023)
6. Gandikota, R., Materzyńska, J., Zhou, T., Torralba, A., Bau, D.: Concept sliders: LoRA adapters for precise control in diffusion models. In: Eur. Conf. Comput. Vis. (2024), arXiv:2311.12092
7. Guo, W.M., Qian, Q., Hasan, K., Du, S.: Aesthetic alignment risks assimilation: How image generation and reward models reinforce beauty bias and ideological “censorship”. arXiv preprint arXiv:2512.11883 (2025)

8. Kärkkäinen, K., Joo, J.: FairFace: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In: IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 1548–1558 (2021)
9. Kirstain, Y., Polyak, A., Singer, U., Matiana, S., Penna, J., Levy, O.: Pick-a-pic: An open dataset of user preferences for text-to-image generation. In: Adv. Neural Inform. Process. Syst. (2023)
10. Lee, S.H., Li, Y., Ke, J., Yoo, I., Zhang, H., Yu, J., Wang, Q., Deng, F., Entis, G., He, J., Li, G., Kim, S., Essa, I., Yang, F.: Parrot: Pareto-optimal multi-reward reinforcement learning framework for text-to-image generation. In: Eur. Conf. Comput. Vis. (2024)
11. Liang, Z., Yuan, Y., Gu, S., Chen, B., Hang, T., Cheng, M., Li, J., Zheng, L.: Aesthetic post-training diffusion models from generic preferences with step-by-step preference optimization. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 13199–13208 (2025)
12. Liu, H., et al.: DiverseGRPO: Mitigating mode collapse in image generation via diversity-aware GRPO. arXiv preprint arXiv:2512.21514 (2025)
13. Liu, J., Li, H., Sun, Z., Chen, C., Bian, Y., Wang, B., Dong, D., Chen, C., Wang, Z.: Beyond the dirac delta: Mitigating diversity collapse in reinforcement fine-tuning for versatile image generation. arXiv preprint arXiv:2601.12401 (2026)
14. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: SDXL: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023)
15. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: Int. Conf. Mach. Learn. (2021)
16. Serengil, S.I., Özpınar, A.: Boosted lightface: A hybrid DNN and GBM model for boosted facial recognition. Gazi University Journal of Science **39**(1), 452–466 (2026). <https://doi.org/10.35378/gujs.1794891>, <https://dergipark.org.tr/en/pub/gujs/article/1794891>
17. Wallace, B., Dang, M., Rafailov, R., Zhou, L., Lou, A., Purushwalkam, S., Ermon, S., Xiong, C., Joty, S., Naik, N.: Diffusion model alignment using direct preference optimization. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 8228–8238 (2024)
18. Wu, X., Hao, Y., Sun, K., Chen, Y., Zhu, F., Zhao, R., Li, H.: Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. arXiv preprint arXiv:2306.09341 (2023)
19. Xu, J., Liu, X., Wu, Y., Tong, Y., Li, Q., Ding, M., Tang, J., Dong, Y.: ImageReward: Learning and evaluating human preferences for text-to-image generation. In: Adv. Neural Inform. Process. Syst. (2023)
20. Xue, Z., Wu, J., Gao, Y., Kong, F., Zhu, L., Chen, M., Liu, Z., Liu, W., Guo, Q., Huang, W., et al.: DanceGRPO: Unleashing GRPO on visual generation. arXiv preprint arXiv:2505.07818 (2025)
21. Zhang, C., Chen, X., Chai, S., Wu, C.H., Lagun, D., Beeler, T., De la Torre, F.: ITI-GEN: Inclusive text-to-image generation. In: Int. Conf. Comput. Vis. pp. 3969–3980 (2023)