Neutral Reward Filtering for Fair Offline-to-Online Diffusion Alignment

Yeon Gyu Han [1 *] **Junah Jung** [2 *] **Dongheon Lee** [3 4 †]

Abstract

Preference optimization (PO) for text-to-image diffusion models can be viewed as an offline-to-online decision-making loop: a reward model learned from offline human preference data guides online adaptation of a generative policy. We show that CLIP-based rewards used in this loop encode perceived demographics in their visual features, so DPO/GRPO can amplify reward bias and narrow who appears in generated images. We propose *Neutral Reward Filtering* (NRF), a reward-side intervention that estimates demographic directions in the reward feature space with a one-time linear probe and removes them by orthogonal projection before rewards are computed. NRF is plug-and-play with both offline pairwise and online group-based PO, and requires no demographic labels during PO optimization beyond the audit stage. Experiments on SDXL and SD1.5 show that NRF restores most of the race/gender entropy lost during alignment while preserving cross-reward quality and prompt fidelity.

1. Introduction

Preference-aligned image generators increasingly use a reward model trained from offline human preference data, then adapt the generator by repeatedly sampling images and optimizing the policy toward higher reward (Wallace et al., 2024; Xue et al., 2025). This is a canonical offline-to-online decision-making pipeline: the learned reward is an offline surrogate for human judgments, while PO performs online adaptation in the generator's action space. Its safety depends not only on reward accuracy for image quality, but also on what social or demographic structure the reward represents.

Figure 1. PO improves reward but can narrow demographic support. NRF filters the reward feature before PO, preserving the alignment gain while recovering demographic coverage.

We study a failure mode in this loop: CLIP-based reward features entangle image quality and perceived demographics. When the generator adapts to such a reward, it learns to increase the probability of demographic groups that the reward model scores more favorably (Radford et al., 2021; Bianchi et al., 2023). On neutral people prompts, standard alignment reduces race entropy by up to roughly half, while diversity-preserving RL methods improve visual variation without recovering demographic representation. This distinction matters: a model can generate diverse poses, backgrounds, and styles while still depicting a narrow population.

Our contribution is a compact reward-side fix. Unlike prompt-embedding or denoising-guidance fairness methods such as ITI-GEN and FairDiffusion, NRF targets the reward signal that drives alignment itself (Zhang et al., 2023; Friedrich et al., 2023). NRF identifies the low-dimensional demographic subspace of the reward feature space, projects it out, and feeds the filtered feature to the original reward head. The PO algorithm itself is unchanged, so the same idea applies to DPO-style offline pair construction and GRPO-style online adaptation. This projection is related to linear representation debiasing and null-space removal methods (Bolukbasi et al., 2016; Ravfogel et al., 2020), but differs in that the filtered representation is used inside the

[*]Equal contribution [†]Corresponding authors. [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. Correspondence to: Dongheon Lee <dhlee.jubilee@gmail.com>.

Accepted at ICML 2026 Workshop on Decision-Making from Offline Datasets to Online Adaptation: Black-Box Optimization to Reinforcement Learning.

reward computation that drives diffusion preference optimization.

2. Methods

Let a learned image reward be $r(x, c) = h(\phi(x), c)$, where ϕ is the reward model's visual encoder and h is its scoring head. We generate a one-time audit set of 10K images from the base model using demographic-neutral prompts and label perceived race/gender with an automatic classifier (Kärkkäinen and Joo, 2021). A linear probe W trained on reward features predicts race with 66.8–74.5% accuracy across HPS-v2, PickScore, ImageReward, and LAION-Aesthetic features (Wu et al., 2023; Kirstain et al., 2023; Xu et al., 2023; Schuhmann et al., 2022), and gender with 77.3–85.1% accuracy; chance is 14.3% and 50%, respectively. A two-layer MLP improves these numbers by only 0.6–2.2%, suggesting that the actionable demographic signal is mostly linear.

Given the probe $W \in \mathbb{R}^{C\times d}$, we compute $W = U\Sigma V^\top$ and take the top k right singular vectors $V_k = [v_1, \ldots, v_k]$. The filtered feature and reward are

$$\tilde{\phi}(x) = (I - V_k V_k^\top)\phi(x), \qquad \tilde{r}(x, c) = h(\tilde{\phi}(x), c). \quad (1)$$

The filtered reward $\tilde{r}$ replaces r in the existing objective. For GRPO, group-relative advantages are computed from $\tilde{r}$; for DPO, preference pairs are ranked by $\tilde{r}$. We use $k = 5$ by default: it reduces race/gender probe accuracy to near chance, while changing the reward feature norm by less than 3% in our SDXL runs. The additional cost is one $d \times k$ projection per generated image, which is negligible compared with denoising and reward-encoder inference.

Feature-level filtering is intentionally earlier than scalar score correction. A scalar baseline that subtracts a per-demographic-group reward mean requires group labels for generated samples during training, is tied to the chosen classifier taxonomy, and cannot remove demographic information that the reward head later mixes with quality. In contrast, NRF removes the corresponding directions before the score is produced; it therefore works with the original PO loss and does not require demographic labels after the one-time audit stage.

Important scope. NRF does not enforce demographic attributes in prompts and does not claim that the uniform distribution is a normative target for every use case. We use uniformity only as a diagnostic reference on deliberately neutral people prompts, with the goal of preventing *additional narrowing caused by alignment*. On prompts that explicitly specify demographics, the text-to-image conditioning path is untouched, so demographic controllability is largely preserved (Table 3).

3. Experiments

3.1. Experimental Setup

We evaluate SDXL (Podell et al., 2024) and SD1.5 (Rombach et al., 2022) with HPS-v2 as the primary training reward. Baselines include the base model, Diffusion-DPO, DanceGRPO, DRIFT, DiverseGRPO, and ITI-GEN combined with GRPO (Zhang et al., 2023; Wallace et al., 2024; Xue et al., 2025; Liu et al., 2025; 2026). All aligned models use LoRA (Hu et al., 2022) and are trained for 2K steps. Evaluation uses 500 neutral prompts spanning occupations, everyday scenes, and generic portraits, with 20 images per prompt. We report reward-based quality (HPS, PickScore, ImageReward), FID (Heusel et al., 2017), Vendi Score for visual diversity (Friedman and Dieng, 2023), and fairness metrics from perceived race/gender labels: entropy (Race-H, Gen-H), demographic parity difference (DPD), and worst-group rate (WGR).

3.2. Main Results

Table 1 shows the central pattern. DanceGRPO reaches the highest HPS score, but reduces Race-H from 1.74 to 0.89 and the worst-group rate from 8% to 1%. DRIFT and DiverseGRPO raise Vendi Score to 0.73 and 0.71, yet their Race-H remains below 1.0. This supports the key diagnosis: visual diversity and demographic fairness are different axes of variation, and optimizing for one does not automatically repair the other.

NRF with GRPO recovers Race-H to 1.67 and Gen-H to 0.96, while keeping HPS within 0.4% of DanceGRPO. Cross-reward metrics remain close as well (PickScore 21.54 vs. 21.63; ImageReward 0.91 vs. 0.93), reducing the risk that the method merely exploits the training reward. On SD1.5, the same trend holds: GRPO drops Race-H from 1.68 to 0.84, while NRF recovers it to 1.62 with a comparable quality gap.

3.3. Diversity and Demographic Fairness

Figure 3 makes the difference between generic diversity and demographic fairness explicit. DRIFT and DiverseGRPO occupy the high-diversity, low-fairness region: they diversify visual modes but leave the reward's demographic preference intact. NRF is not the highest-Vendi method, but it is the only aligned method that simultaneously improves visual diversity over the base model and restores demographic entropy close to the base model. This supports the reward-filtering design choice: fairness should be addressed at the reward feature that drives the policy update, not only through a diversity bonus on generated samples.

Figure 4 further shows that demographic collapse is progressive: entropy begins falling early in training, before

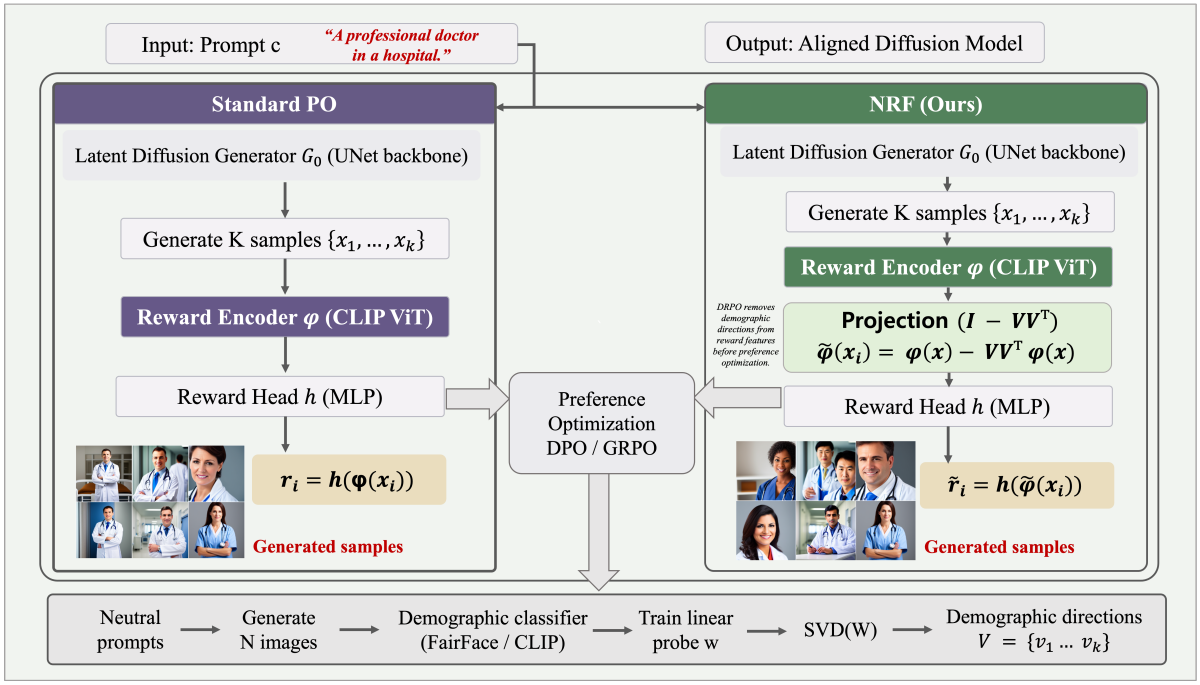

Figure 2. NRF changes only the reward computation in the offline-to-online alignment loop. A one-time preparation stage learns demographic directions in reward-feature space. During PO, generated samples pass through the same reward encoder, but the demographic component is projected out before the reward head and the standard DPO/GRPO update is applied.

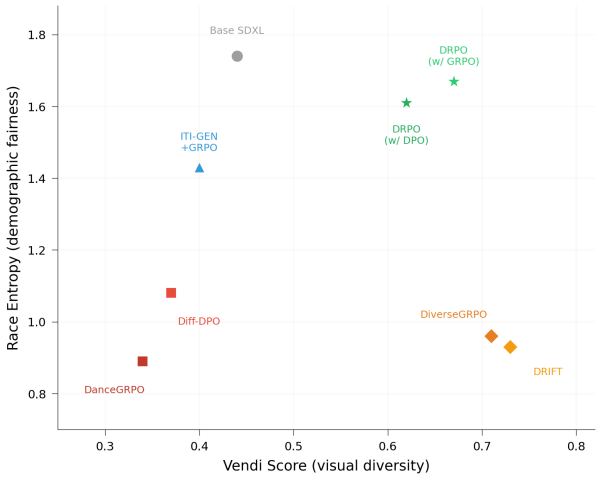

Figure 3. Visual diversity and demographic fairness are separable. Diversity-aware methods move rightward, but remain low in race entropy. NRF moves toward the upper-right region.

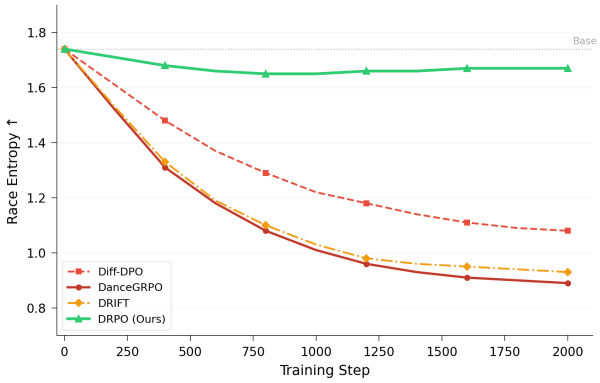

Figure 4. Race entropy during online adaptation. Standard PO and diversity-aware RL progressively narrow demographic coverage; filtering the reward feature prevents the collapse at its source.

large quality gains. This suggests that the demographic component of the reward is an easy shortcut for the policy to exploit. Filtering this component is therefore more direct than adding a generic diversity bonus after the biased reward has already been computed.

3.4. Additional Analysis

Projection dimension. Table 2 shows that increasing k from 1 to 5 steadily restores fairness with minimal quality loss, while $k = 10$ gives little additional fairness and begins to reduce reward quality. This supports using a small demographic subspace rather than aggressively suppressing broad semantic information. The same table also addresses two practical sensitivities: replacing FairFace with CLIP zero-shot labels, or degrading labels to 80% accuracy, still yields a large improvement over unfiltered GRPO.

Table 1. SDXL results. Standard PO improves quality but sharply degrades demographic coverage; diversity methods improve Vendi but not fairness. NRF restores demographic coverage while keeping cross-reward quality close to GRPO.

	Quality and visual diversity					Demographic fairness			
Method	HPS↑	Pick↑	ImgR↑	FID↓	Vendi↑	Race-H↑	Gen-H↑	DPD↓	WGR↑
Base SDXL	25.81	20.14	0.42	44.7	0.44	1.74	0.97	0.18	0.08
Diff-DPO	27.08	21.27	0.81	40.3	0.37	1.08	0.78	0.41	0.02
DanceGRPO	27.42	21.63	0.93	38.9	0.34	0.89	0.69	0.49	0.01
DRIFT	27.18	21.41	0.87	33.1	**0.73**	0.93	0.72	0.46	0.02
DiverseGRPO	27.29	21.52	0.90	33.8	0.71	0.96	0.73	0.44	0.02
ITI-GEN+GRPO	26.84	21.08	0.79	39.2	0.40	1.43	0.88	0.27	0.05
NRF w/ DPO	26.93	21.19	0.78	37.4	0.62	1.61	0.94	0.21	0.07
NRF w/ GRPO	27.31	21.54	0.91	35.9	0.67	**1.67**	**0.96**	**0.19**	**0.07**

Table 2. Ablations on SDXL/GRPO. The default $k = 5$ balances fairness and quality; scalar output-level debiasing is weaker than feature projection.

Variant	Quality↑	Race-H↑
No projection	27.42_{HPS}	0.89
$k = 1$	27.39_{HPS}	1.34
$k = 3$	27.35_{HPS}	1.58
$k = 5$ (default)	27.31_{HPS}	1.67
$k = 10$	27.14_{HPS}	1.69
Output-level means	27.38_{HPS}	1.23
CLIP zero-shot labels	27.33_{HPS}	1.59
80% noisy labels	27.32_{HPS}	1.56

Table 3. Compact robustness checks on SDXL/HPS-v2.

Concern	Check and outcome
Nonlinear residual	After projection: MLP race 16.3% and gender 52.1% (chance: 14.3%, 50%).
Classifier dependence	Race-H gain from GRPO to NRF: FairFace 0.89→1.67, CLIP 0.86→1.59, DeepFace 0.79→1.49.
Prompt fidelity	On 70 demographic-specified prompts, match rate is 76.3% vs. 78.6% for GRPO; HPS is 27.28 vs. 27.39.
Reward/OOD shift	Pearson $\rho(r, \tilde{r}) = 0.963$; top-10% reward-rank preservation is 91.4%.

Validation checks. Potential failure modes are summarized in Table 3. First, a nonlinear probe on filtered HPS-v2 features remains near chance, indicating that projection does not merely hide demographics from a linear classifier. Second, fairness gains are not tied to FairFace alone: CLIP zero-shot and DeepFace (Serengil and Ozpinar, 2021) show the same relative recovery. Third, demographic-specified prompts preserve controllability. Finally, the filtered reward is highly correlated with the original reward, and top-ranked images mostly remain top-ranked, which helps explain why the frozen reward head does not exhibit a severe out-of-distribution failure in our runs.

The Stage-1 prompt set is an important design choice. We use 200 broad neutral prompts spanning occupations, everyday situations, and generic portraits so that the learned subspace is not dominated by a single concept such as "doctor" or "CEO." In deployment, the same audit can be repeated with task-specific neutral prompts, and V_k can be refreshed if the policy or prompt distribution shifts substantially.

Audit-subspace robustness. NRF uses a static audit subspace estimated from broad neutral prompts, so it should not be interpreted as a once-and-for-all demographic filter for every deployment distribution. If downstream prompts shift toward niche domains, historical attire, culturally specific concepts, or specialized occupations, the audit prompts should be refreshed and the filtered reward should be rechecked. There is also a possible trade-off: demographic directions can be entangled with legitimate cultural or semantic cues. For this reason, we treat NRF as a deployment-time mitigation that requires periodic auditing, rather than as a guarantee that all demographic information has been safely removed.

We also checked whether marginal race/gender recovery hides intersectional collapse. On the 14-way race×gender distribution, GRPO reduces entropy from 2.38 to 1.21 and the worst-group rate from 3.1% to 0.3%; NRF recovers entropy to 2.26 and the worst-group rate to 2.7%. Age is harder: a nonlinear probe retains more age signal after linear projection, so age entropy improves only partially. This suggests that iterative null-space removal, adversarial removal, or adaptive nonlinear filters may be useful extensions, but they would require additional checks for reward-rank preservation, prompt fidelity, and stability during online adaptation.

4. Discussion

NRF treats fairness in diffusion alignment as a reward design problem for offline-to-online decision-making. When

the reward model encodes demographics, policy adaptation can amplify this information even if the base model was more balanced. Filtering the reward feature is lightweight, optimizer-agnostic, and compatible with stronger PO methods.

Limitations and Future Work. The current short paper has limitations. The demographic labels are automatic perceived-attribute labels, not identity ground truth, and the categories may not match all cultural contexts. The uniform reference distribution is a diagnostic for neutral prompts, not a universal fairness objective; for concepts whose demographics are intentionally specified, prompt fidelity should take precedence. The projection is static and linear, so it may not fully remove nonlinear or distribution-shifted demographic encodings. We also view hidden-layer OOD analysis and double-blind human evaluation as important future validation, especially for judging quality preservation beyond proxy rewards.

Conclusion. Relative to prompt-based fairness methods, NRF is not meant to decide which groups should appear for every prompt. Its narrower role is to prevent a learned reward from turning offline preference data into an online policy update that systematically drops groups. This makes it complementary to dataset documentation, fairness-aware prompting, and downstream human review.

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).

References

Federico Bianchi, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Myra Cheng, Debora Nozza, Tatsunori Hashimoto, Dan Jurafsky, James Zou, and Aylin Caliskan. Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 1493–1504. Association for Computing Machinery, 2023. doi: 10.1145/3593013.3594095.

Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam T. Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Dan Friedman and Adji Bousso Dieng. The Vendi Score: A diversity evaluation metric for machine learning. *Transactions on Machine Learning Research*, 2023.

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.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Kimmo Kärkkäinen and Jungseock Joo. 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*, pages 1548–1558, 2021.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Henglin Liu, Huijuan Huang, Jing Wang, Chang Liu, Xiu Li, and Xiangyang Ji. DiverseGRPO: Mitigating mode collapse in image generation via diversity-aware GRPO. *arXiv preprint arXiv:2512.21514*, 2025.

Jinmei Liu, Haoru Li, Zhenhong Sun, Chaofeng Chen, Yatao Bian, Bo Wang, Daoyi Dong, Chunlin Chen, and Zhi Wang. Beyond the Dirac Delta: Mitigating diversity collapse in reinforcement fine-tuning for versatile image generation. *arXiv preprint arXiv:2601.12401*, 2026.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations*, 2024.

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. 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, 2021.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7237–7256, 2020.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5b: An open large-scale dataset for training next generation image-text models. In *Advances in Neural Information Processing Systems*, volume 35, pages 25278–25294, 2022.

Sefik Ilkin Serengil and Alper Ozpinar. Hyperextended lightface: A facial attribute analysis framework. In *2021 International Conference on Engineering and Technology*, pages 1–4. IEEE, 2021.

Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8228–8238, 2024.

Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human Preference Score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. ImageReward: Learning and evaluating human preferences for text-to-image generation. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Zeyue Xue, Jie Wu, Yu Gao, Fangyuan Kong, Lingting Zhu, Mengzhao Chen, Zhiheng Liu, Wei Liu, Qiushan Guo, Weilin Huang, and Ping Luo. DanceGRPO: Unleashing GRPO on visual generation. *arXiv preprint arXiv:2505.07818*, 2025.

Cheng Zhang, Xuanbai Chen, Siqi Chai, Chen Henry Wu, Dmitry Lagun, Thabo Beeler, and Fernando De la Torre. ITI-GEN: Inclusive text-to-image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3969–3980, 2023.