

Ranking-based Preference Optimization for Diffusion Models from Implicit User Feedback

Yi-Lun Wu Bo-Kai Ruan Chiang Tseng Hong-Han Shuai

Institute of Electrical and Computer Engineering
National Yang Ming Chiao Tung University

NeurIPS 2025

Introduction

Preference alignment for diffusion models often relies on RL frameworks such as Direct Preference Optimization (DPO) [5, 6]. However, these approaches face two challenges:

- Collecting paired preference data is labor-intensive.
- The sigmoid-based surrogate loss limits learning efficiency in diffusion models.

Inverse Reinforcement Learning (IRL)

To remove the need for paired data, we directly learn preferences from high-quality samples using an inverse RL framework:

$$\mathbb{E}_{\mathbf{c} \sim \mathcal{C}, \bar{\mathbf{x}}_0 \sim \mathcal{D}(\mathbf{c})} \left[\underbrace{r(\bar{\mathbf{x}}_0, \mathbf{c})}_{\text{reward for expert data}} \right] - \mathbb{E}_{\mathbf{c} \sim \mathcal{C}, \mathbf{x}_0 \sim p_{\theta}(\mathbf{x}_0 | \mathbf{c})} \left[\underbrace{r(\mathbf{x}_0, \mathbf{c})}_{\text{reward for policy model}} \right], \quad (1)$$

where \mathcal{D} denotes the expert dataset.

Goal: Learn a reward model that distinguishes expert samples from generated ones.

Policy Optimization

The inner loop of the inverse RL involves reward maximization for the policy model. Following prior works [1, 2, 3, 4] the optimal policy has a closed-form solution:

$$\hat{p}_{\theta}(\mathbf{x}_0|\mathbf{c}) = \frac{1}{Z(\mathbf{c})} p_{\theta_{\text{ref}}}(\mathbf{x}_0|\mathbf{c}) \exp\left(\frac{1}{\beta} r(\mathbf{x}_0, \mathbf{c})\right), \quad (2)$$

where $Z(\mathbf{c})$ is the intractable normalization term.

Goal: Learn a policy model that maximizes rewards to imitate expert demonstration.

Parameterization of the Reward Model

We propose to parameterize the reward model to mirror the closed-form structure of the policy model:

$$r_{\phi}(\mathbf{x}_0, \mathbf{c}) = \beta \log \frac{p_{\phi}(\mathbf{x}_0|\mathbf{c})}{p_{\theta_{\text{ref}}}(\mathbf{x}_0|\mathbf{c})} + \beta \log Z(\mathbf{c}), \quad (3)$$

This ensures an elegant equivalence between the policy and reward models in the inner loop.

$$\hat{p}_{\theta}(\mathbf{x}_0|\mathbf{c}) = \hat{p}_{\phi}(\mathbf{x}_0|\mathbf{c}). \quad (4)$$

Thus, the IRL problem is reduced to reward optimization.

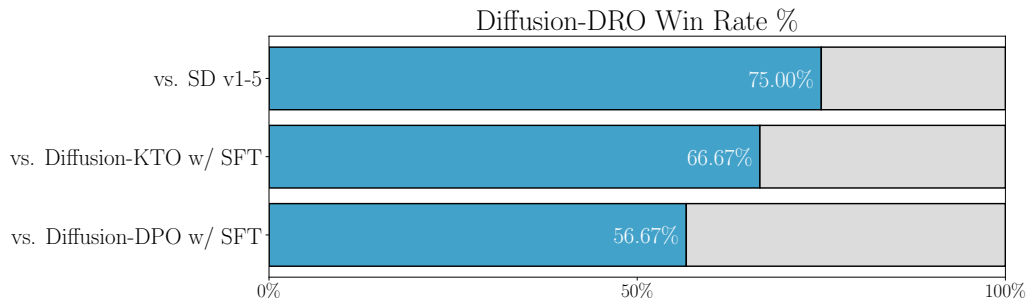
Experiments

Table 1: Automated win rates of Diffusion-DRO compared to baseline methods.

Baseline Method	Pick-a-Pic v2 Test				HPDv2 Benchmark			
	PickScore	Aesthetic Score	CLIP Score	ImageReward	PickScore	Aesthetic Score	CLIP Score	ImageReward
SD v1-5	87.80	85.20	48.40	88.60	90.47	82.91	46.59	87.69
SD v1-5 w/ SFT	71.20	58.00	66.40	57.80	70.62	57.22	64.97	62.03
SPIN-Diffusion	56.20	64.80	58.20	70.60	54.87	62.78	54.78	69.78
Diffusion-SPO	62.80	63.60	71.40	78.00	60.59	67.66	75.78	77.94
Diffusion-SPO w/ SFT	86.60	81.60	42.40	87.20	88.75	80.25	42.69	85.78
Diffusion-DPO	78.40	83.20	41.40	84.20	79.75	79.97	39.09	82.25
Diffusion-DPO w/ SFT	64.00	55.00	59.00	56.20	63.62	56.12	59.91	58.75
Diffusion-KTO	74.20	69.00	42.20	66.60	71.19	71.03	39.81	62.81
Diffusion-KTO w/ SFT	70.20	58.60	64.00	58.60	71.09	56.12	65.31	62.75


User Study

- For each comparison, we collected 2,100 pairwise preference annotations between Diffusion-DRO and each baseline.
- The win rate represents the proportion of cases where participants preferred the results from Diffusion-DRO.



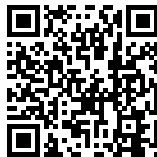
Thank you.



 arXiv



 Code



 Model

References I

- [1] Dongyoung Go, Tomasz Korbak, Germàn Kruszewski, Jos Rozen, Nahyeon Ryu, and Marc Dymetman. Aligning language models with preferences through f -divergence minimization. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202, pages 11546–11583, 23–29 Jul 2023.
- [2] Tomasz Korbak, Hady Elsahar, Germán Kruszewski, and Marc Dymetman. On reinforcement learning and distribution matching for fine-tuning language models with no catastrophic forgetting. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 16203–16220. Curran Associates, Inc., 2022.
- [3] Xue Bin Peng, Aviral Kumar, Grace Zhang, and Sergey Levine. Advantage-weighted regression: Simple and scalable off-policy reinforcement learning, 2019. URL <https://arxiv.org/abs/1910.00177>.
- [4] Jan Peters and Stefan Schaal. Reinforcement learning by reward-weighted regression for operational space control. In *Proceedings of the 24th International Conference on Machine Learning (ICML)*, pages 745–750, 2007.

References II

- [5] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pages 53728–53741. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/a85b405ed65c6477a4fe8302b5e06ce7-Paper-Conference.pdf.
- [6] 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 (CVPR)*, pages 8228–8238, June 2024.