Self-Supervised Direct Preference Optimization for Text-to-Image Diffusion Models

Liang Peng · Boxi Wu · Haoran Cheng · Yibo Zhao · Xiaofei He

Fabu Inc. , Zhejiang University

Motivation & Abstract

Data requirements	SFT	Self-DPO	DPO
Text Caption	✓	✓	✓
Image	✓	✓	✓
Extra Preference Image	–	–	✓

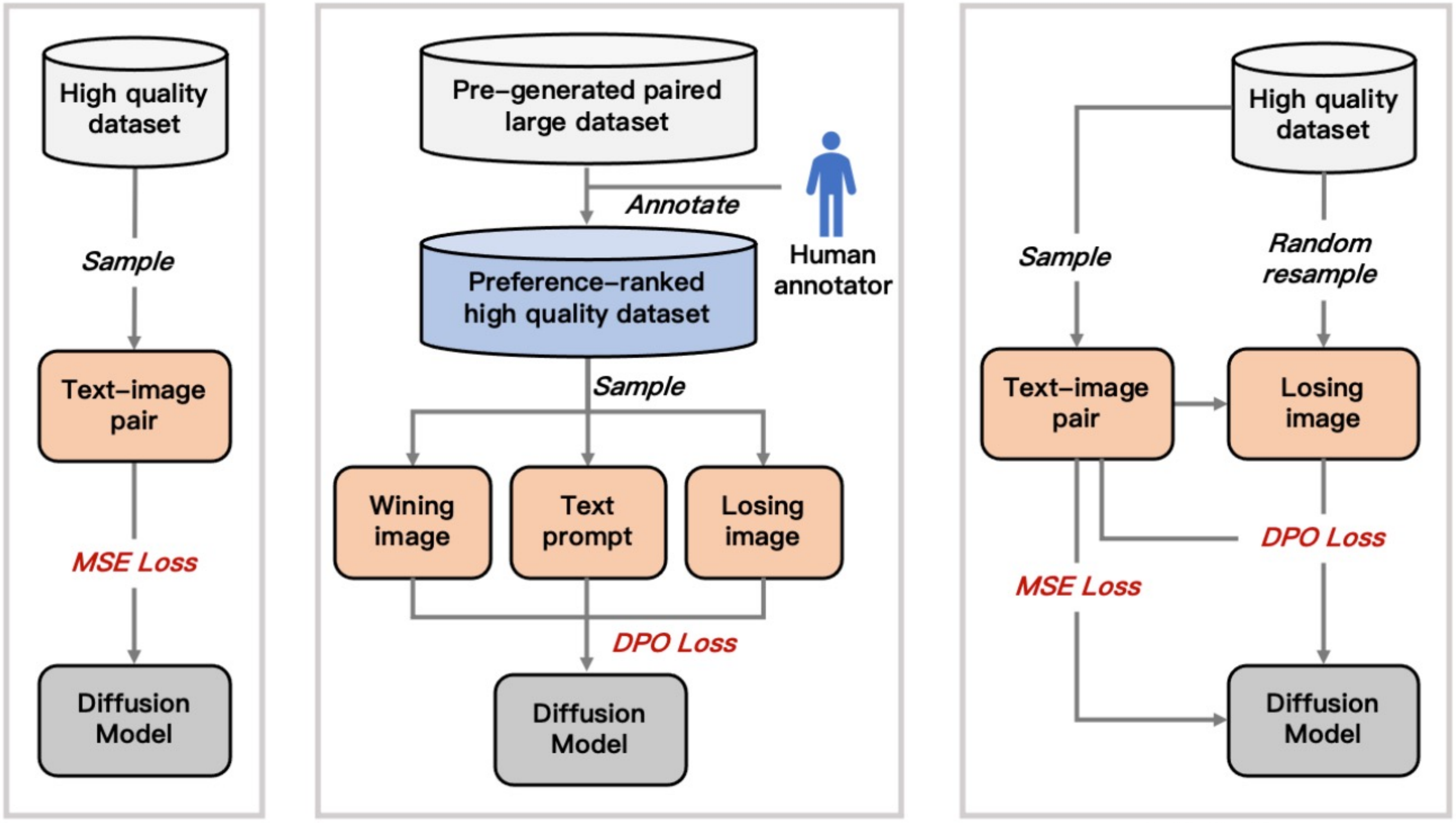
Motivations & Solutions:

- ☐ DPO has shown strong results in aligning text-to-image models with user intent.
- ☐ The practical adoption of DPO is hindered by a costly and rigid training pipeline: it first requires the offline generation of a large set of image candidates, followed by extensive human annotation in the form of pairwise ranking.
- ☐ Instead of requiring external preference data, our method constructs training pairs dynamically using self-supervised image transformations.
- ☐ We identifies a "winning image” that satisfies human-aligned quality criteria, then generate a corresponding "losing image" by intentionally degrading the winner, either through visual-quality reductions or text–image misalignment.

Contributions:

- Self-DPO alleviates the need for costly pre-generated images and human ranking efforts, enables scalable and dynamic training, and introduces greater diversity into the preference supervision signal.
- Self-DPO not only matches but surpasses conventional DPO in both qualitative and quantitative metrics. We attribute it to the greater diversity of preferences sampled during training.

Method



SFT *DPO* *Self–DPO*

Overview:

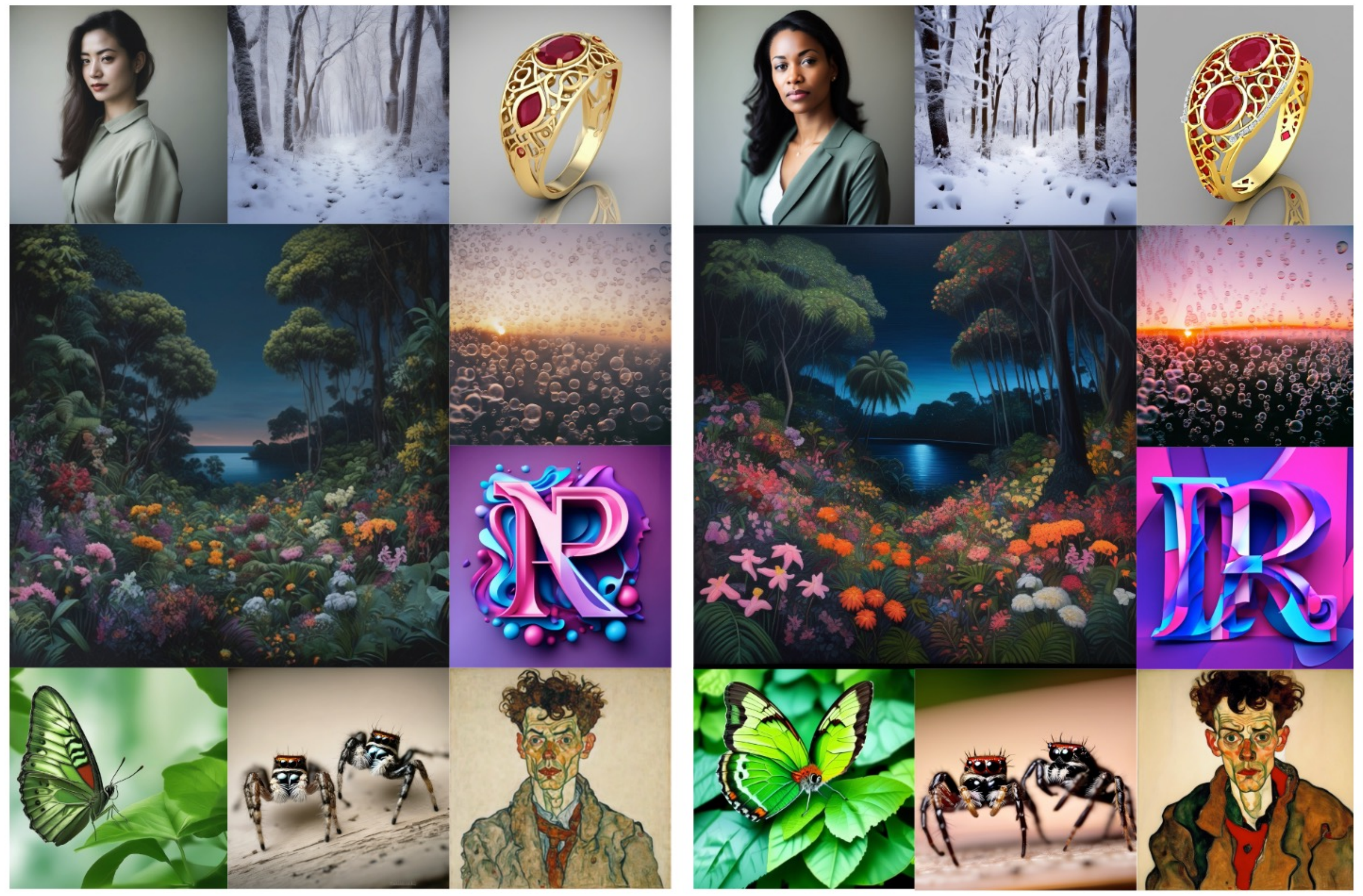
We generate the "losing" images self-supervisedly, enabling direct preference optimization without extra collecting and ranking steps. This lightweight procedure eliminates the substantial overhead of conventional DPO while retaining the same data requirements as standard SFT.

$$\mathcal{L}_{\text{Self-DPO}} = -\log\sigma\Big(C\Big(\big(\|\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t^w,t)-\boldsymbol{\epsilon}^w\|_2^2 - \|\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t^{sl},t)-\boldsymbol{\epsilon}^{sl}\|_2^2\big) - \big(\|\boldsymbol{\epsilon}_{\text{ref}}(\boldsymbol{x}_t^w,t)-\boldsymbol{\epsilon}^w\|_2^2 - \|\boldsymbol{\epsilon}_{\text{ref}}(\boldsymbol{x}_t^{sl},t)-\boldsymbol{\epsilon}^{sl}\|_2^2\big)\Big)\Big),$$

$$\textit{where} \quad \boldsymbol{x}^{sl} = \textbf{Downgrade}(\boldsymbol{x}^w)$$

The downgrade operation can be simply performed by randomly selecting images from the training dataset. For each self-generated image pair, the winning sample closely aligns with the prompt, whereas the losing sample fails to correspond to the description.

Experiments & Results



DPO Self-DPO

Datasets	Methods		SD1.5 P.S.	SD1.5 Aes.	SD1.5 CLIP	SD1.5 HPS	SD1.5 I.R.	SDXL P.S.	SDXL Aes.	SDXL CLIP	SDXL HPS	SDXL I.R.
Pick-a-Pic V2	Base	Avg score	20.57	53.15	32.58	26.17	-14.81	22.10	**60.01**	35.86	26.83	50.62
	SFT		21.10	**56.35**	33.75	27.03	45.03	21.48	57.84	35.67	26.67	30.89
	DPO		20.91	54.07	33.19	26.46	4.13	**22.57**	59.93	37.30	27.30	81.14
	Self-DPO		**21.23**	**56.35**	**34.79**	**27.33**	**71.00**	22.34	59.97	**37.53**	**27.89**	**103.96**
	SFT	Win rate	75.00	77.20	60.40	90.20	80.00	19.40	31.80	47.00	44.60	42.4
	DPO		73.80	60.00	60.00	71.80	61.00	**72.60**	47.20	**63.00**	79.80	69.8
	Self-DPO		**78.60**	**77.80**	**68.40**	**94.20**	**85.20**	60.80	**50.80**	62.40	**93.80**	**79.2**
PartiPrompts	Base	Avg score	21.39	53.13	33.21	26.79	1.48	22.63	57.69	35.77	27.33	69.78
	SFT		21.75	**55.31**	33.93	27.57	50.75	22.02	56.41	35.31	27.13	47.29
	DPO		21.61	53.58	33.88	26.98	21.43	**22.90**	57.85	36.95	27.73	103.36
	Self-DPO		**21.84**	55.09	**35.11**	**27.84**	**75.66**	22.79	**58.69**	**37.00**	**28.30**	**117.50**
	SFT	Win rate	67.28	**70.89**	53.43	85.42	73.35	21.38	38.11	45.10	43.75	40.93
	DPO		67.10	57.17	56.74	61.83	63.05	**63.42**	53.62	**62.32**	73.10	68.44
	Self-DPO		**69.85**	68.50	**63.24**	**89.40**	**81.00**	56.19	**60.48**	60.17	**92.16**	**76.84**
HPD V2	Base	Avg score	20.84	54.32	33.96	26.84	-11.79	22.78	61.34	37.68	27.68	78.27
	SFT		21.57	**57.41**	35.26	27.89	57.74	22.24	60.08	37.39	27.76	66.62
	DPO		21.30	55.80	34.68	27.22	13.24	**23.18**	**61.35**	**38.45**	28.14	102.74
	Self-DPO		**21.58**	57.10	**36.30**	**28.11**	**76.13**	22.98	61.30	38.35	**28.77**	**110.67**
	SFT	Win rate	**79.53**	**75.31**	59.34	90.10	81.16	23.47	37.28	46.63	58.22	45.81
	DPO		75.72	66.28	57.56	72.43	64.69	**72.66**	**50.28**	**58.69**	80.56	69.78
	Self-DPO		**79.53**	74.03	**68.47**	**92.49**	**85.19**	58.78	48.06	55.65	**94.97**	**72.50**