DSPO: Direct Score Preference Optimization for Diffusion Model Alignment

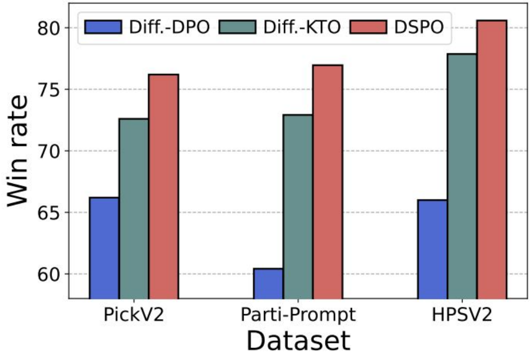
Huaisheng Zhu, Teng Xiao, Vasant Honavar

tl;dr: a score-matching-based perspective for designing empirically effective alignment algorithms for diffusion models

Diffusion-DPO

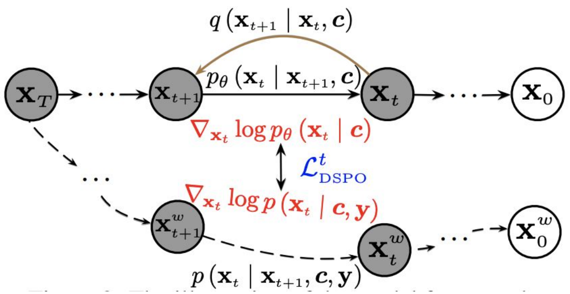
$$-\mathbb{E}\left[\log\sigma\left(\lambda\log\frac{p_\theta\left(\mathbf{x}_t^w\mid\mathbf{x}_{t+1}^w,\boldsymbol{c}\right)}{p_{\mathrm{ref}}\left(\mathbf{x}_t^w\mid\mathbf{x}_{t+1}^w,\boldsymbol{c}\right)}-\lambda\log\frac{p_\theta\left(\mathbf{x}_t^l\mid\mathbf{x}_{t+1}^l,\boldsymbol{c}\right)}{p_{\mathrm{ref}}\left(\mathbf{x}_t^l\mid\mathbf{x}_{t+1}^l,\boldsymbol{c}\right)}\right)\right]$$

Con: It adapts DPO from LLMs, but the upper-bounded loss on diffusion models may limit performance.



Win-rate (vs SD15) for DSPO and preference learning baselines based on Aesthetics reward. "Diff." represents "Diffusion"

DSPO: Score-based Preference Alignment



Human Preference Score Model:

$$\nabla_{\mathbf{x}_t}\log p_\theta\left(\mathbf{x}_t\mid\boldsymbol{c},\mathbf{y}\right)=\nabla_{\mathbf{x}_t}\log p_\theta\left(\mathbf{x}_t\mid\boldsymbol{c}\right)+\nabla_{\mathbf{x}_t}\log p\left(\mathbf{y}\mid\mathbf{x}_t,\boldsymbol{c}\right)$$

$\boldsymbol{c}$ is the conditional variable of human preference for the image.

Direct Score Preference Optimization:

$$\min_\theta\omega(t)\left\|\nabla_{\mathbf{x}_t}\log p_\theta\left(\mathbf{x}_t\mid\boldsymbol{c}\right)-\left(\nabla_{\mathbf{x}_t}\log p\left(\mathbf{x}_t\mid\boldsymbol{c}\right)+\gamma\nabla_{\mathbf{x}_t}\log p\left(\mathbf{y}\mid\mathbf{x}_t,\boldsymbol{c}\right)\right)\right\|_2^2$$

Final Objective:

$$\mathcal{L}^t_{\mathrm{DSPO}}=A(t)\left\|\boldsymbol{\epsilon}_{\theta,t+1}-\boldsymbol{\epsilon}_{t+1}-\lambda\gamma\left(1-\sigma\left(r\left(\boldsymbol{c},\mathbf{x}_t\right)-r\left(\boldsymbol{c},\mathbf{x}_t^l\right)\right)\left(\boldsymbol{\epsilon}_{\theta,t+1}-\boldsymbol{\epsilon}_{\mathrm{ref},t+1}\right)\right)\right\|_2^2$$

$$r\left(\mathbf{x}_t,\boldsymbol{c}\right)=-\left(\left\|\boldsymbol{\epsilon}_\theta\left(\boldsymbol{x}_{t+1},t+1\right)-\boldsymbol{\epsilon}_{t+1}\right\|_2^2-\left\|\boldsymbol{\epsilon}_{\mathrm{ref}}\left(\boldsymbol{x}_{t+1},t+1\right)-\boldsymbol{\epsilon}_{t+1}\right\|_2^2\right)$$

Pro: designed from score matching and achieving good results

How does DSPO perform?

Results: Among all methods, our DSPO achieves the best performance on 3 datasets based on win rate from reward models.

Stable Diffusion 1.5:

Dataset	Method	Pick Score	HPS	Aesthetics	CLIP	Image Reward
PickV2	SFT	70.20	84.20	75.80	61.20	76.40
	Diff.-DPO	71.60	70.20	66.20	58.80	63.60
	Diff.-KTO	71.40	84.40	72.60	60.02	77.00
	DSPO	**73.60**	**84.80**	**76.20**	**61.80**	**78.00**
Parti-Prompt	SFT	64.27	85.72	75.74	54.72	71.38
	Diff.-DPO	61.18	66.48	60.42	**55.45**	62.19
	Diff.-KTO	64.80	86.16	72.92	54.34	71.51
	DSPO	**65.32**	**87.50**	**76.96**	54.86	**71.75**
HPSV2	SFT	79.03	91.97	78.56	60.47	80.78
	Diff.-DPO	76.06	72.13	66.00	58.50	64.22
	Diff.-KTO	79.18	92.15	77.87	59.28	81.96
	DSPO	**79.90**	**92.56**	**80.59**	**61.13**	**82.31**

Stable Diffusion XL:

Dataset	Method	Pick Score	HPS	Aesthetics	CLIP	Image Reward
PickV2	SFT	20.80	40.60	23.20	44.80	34.40
	Diff.-DPO	**75.20**	76.20	54.10	59.40	65.20
	MaPO	54.40	69.60	**68.20**	51.20	61.40
	DSPO	74.00	**80.00**	54.20	**59.60**	**68.60**
Parti-Prompt	SFT	17.03	33.02	27.81	36.58	37.18
	Diff.-DPO	65.44	74.08	56.86	**60.54**	66.85
	MaPO	58.34	66.54	**68.23**	47.43	58.64
	DSPO	**67.46**	**81.80**	57.84	55.02	**73.47**
HPSV2	SFT	18.18	45.28	26.72	39.13	47.22
	Diff.-DPO	70.31	80.81	50.78	**59.31**	68.75
	MaPO	59.62	77.90	**62.31**	50.90	62.09
	DSPO	**72.59**	**83.47**	51.41	57.34	**70.09**

Takeaway: a novel score-based preference alignment algorithm to fine-tune text-to-image diffusion models (see our paper for details).

Paper

Github