RETHINKING PREFERENCE ALIGNMENT FOR DIFFUSION MODELS WITH CLASSIFIER-FREE GUIDANCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Preference finetuning can improve text-to-image diffusion models while reducing sample diversity and shifting generation away from the pretrained model. We study whether light preference adaptation can supply a reusable guidance direction whose strength is set at inference. Preference-Guided Diffusion (PGD) amplifies the prediction difference between a lightly DPO-tuned model and its reference during sampling. Contrastive PGD (cPGD) constructs this direction from separate generators fitted to preferred and dispreferred images with ordinary denoising losses. Their contrast introduces preference guidance at inference, with a local connection to reweighted DPO. Experiments with Stable Diffusion 1.5 and SDXL demonstrate improved preference scores and favorable reward–diversity and reward–prior-preservation trade-offs relative to the evaluated DPO baseline. In a multiple-selection human study with 55 prompts and 20 participants, PGD, cPGD, and DPO are selected in 45.5%, 33.0%, and 29.5% of decisions, respectively. The learned directions also improve compatible, already aligned checkpoints, while partial-step guidance and model merging reduce inference cost.

1 INTRODUCTION

Text-to-image diffusion models produce high-quality images from natural-language prompts (Ho et al., 2020; Song & Ermon, 2019; Song et al., 2021; Podell et al., 2023; Saharia et al., 2022), but their pretraining objective does not directly encode human preferences for aesthetics or prompt adherence. Preference datasets provide a way to address this gap (Kirstain et al., 2023). Direct preference optimization (DPO) (Rafailov et al., 2023) adapts a model using preferred and dispreferred examples without a separately trained reward model. Its diffusion counterpart, Diffusion-DPO (Wallace et al., 2024), makes this approach practical for image generation.

Preference alignment nevertheless requires more than improving the training objective. Finetuning on a comparatively small preference dataset can move a generator away from useful structure learned during pretraining. In our experiments, prolonged Diffusion-DPO training can concentrate samples in a few modes or degrade image quality. These observations raise a practical question: can useful preference information be learned through light adaptation and amplified during sampling?

A two-dimensional mixture experiment illustrates the distinction. We pretrain a diffusion model on eight Gaussian components, designate four as preferred, and finetune on preference pairs. Continued DPO training can distort the mixture and reduce its mode coverage. By contrast, amplifying the difference between a lightly finetuned predictor and the reference shifts samples toward the preferred components while retaining more of the mixture structure. This observation motivates separating the amount of preference training from the strength of preference guidance.

Classifier-free guidance (CFG) offers a way to separate learning a condition from controlling its influence (Ho & Salimans, 2022). We apply this principle to preference adaptation: a lightly trained model supplies a direction, while the pretrained model remains the base generator. Preference-Guided Diffusion (PGD) amplifies the prediction difference between a DPO-tuned checkpoint and its reference at inference. Further finetuning can change both the magnitude and direction of this residual, whereas guidance scales the residual of a fixed checkpoint. Related inference-time alignment studies use the same form of model-difference guidance (Jin et al., 2025; Cheng et al., 2026).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



Figure 1: Eight-Gaussian experiment. DPO varies training duration at two regularization settings; PGD varies guidance for a fixed checkpoint. Circles and crosses mark preferred and dispreferred components. Every panel spans $[-6, 6]^2$.



Figure 2: Qualitative comparison of the base model, Diffusion-DPO, and PGD. The illustrated DPO outputs exhibit reduced diversity (top) and degraded image quality (bottom).

Here, we examine how light preference training supplies that direction and how its amplification changes the trade-off between preference reward and preservation of the base distribution.

Treating preference status as a condition also suggests fitting preferred and dispreferred distributions directly. We therefore train two generators independently with ordinary denoising losses and use their prediction difference to guide a separate base model, giving contrastive PGD (cPGD). Both branches learn to generate their respective data; their contrast introduces repulsion only during sampling. A local analysis connects the difference of the branch updates to a reweighted DPO direction.

We evaluate PGD and cPGD with SD1.5 and SDXL on three prompt suites. Guidance improves aggregate preference scores and yields favorable reward–diversity and reward–prior-preservation trade-offs relative to the evaluated DPO checkpoint. Human judgments corroborate the preference gains, with the strongest improvement for PGD. The same learned directions improve already aligned checkpoints and transfer to an SDXL-compatible model with a different architecture, showing that their usefulness extends beyond the generator from which they were learned.

In summary, our contributions are:

- We study preference adaptation as a reusable sampling direction, showing how light training followed by amplification can improve preference scores while retaining favorable diversity and base-distribution trade-offs.
- We develop cPGD by independently fitting preferred and dispreferred distributions, and establish a local connection between their contrastive update and reweighted DPO.
- We demonstrate preference gains and transfer on SD1.5 and SDXL, supported by human judgments, and evaluate partial-step guidance and model merging as practical cost–quality trade-offs.

2 Related Work

Alignment with human preferences. Preference alignment adapts generative models using learned rewards or preference labels (Christiano et al., 2017; Stiennon et al., 2020; Ouyang et al., 2022; Rafailov et al., 2023). For diffusion models, reward-based approaches optimize generation through reinforcement learning (Black et al., 2024; Fan et al., 2023) or backpropagation through the sampling process (Clark et al., 2024). Diffusion-DPO (Wallace et al., 2024) instead learns directly from preferred–dispreferred image pairs using a denoising-based approximation to the DPO objective. Subsequent methods revise the preference objective through human-utility modeling (Li et al., 2024), margin-based optimization (Hong et al., 2026), and alternative reward or score formulations (Yang et al., 2024; Zhu et al., 2025).

Guidance and inference-time alignment. Classifier guidance (Dhariwal & Nichol, 2021) and classifier-free guidance (CFG) (Ho & Salimans, 2022) modify diffusion sampling to emphasize a

desired condition. Guidance can also exploit differences between models under the same condition: autoguidance uses a weaker version of the generator to improve sample quality (Karras et al., 2024). For preference alignment, RLG (Jin et al., 2025) combines base and aligned model predictions, including those of DPO-finetuned models, while Diffusion Blend (Cheng et al., 2026) combines diffusion processes to control reward trade-offs and regularization strength. PGD shares this base-plus-residual composition; our study examines light preference finetuning as a way to learn the residual, and evaluates how its amplification affects preference scores, diversity, and preservation of the base distribution.

Positive and negative preference models. Diffusion-NPO (Wang et al., 2025) learns a negatively aligned model by reversing preference labels or rewards and incorporates it into the unconditional or negative branch of CFG. CHATS (Fu et al., 2025) jointly trains preferred and dispreferred generators with a coupled preference objective, then combines their predictions through proxy-prompt guidance. cPGD instead fits each preferred or dispreferred subset with its own ordinary denoising objective, and contrasts the predictions during sampling. This removes the coupled preference loss from training and retains a separate base predictor at inference, so the learned direction can guide compatible base checkpoints.

3 PRELIMINARIES

3.1 DIFFUSION MODELS

For a clean image latent x_0 and prompt c, the forward process produces

$$x_t = \sqrt{\bar{\alpha}_t}\, x_0 + \sqrt{1-\bar{\alpha}_t}\, \epsilon, \qquad \epsilon \sim \mathcal{N}(0, I),$$

where $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$ specifies the noise schedule. A noise predictor ϵ_θ is trained by minimizing

$$L_{\text{den}}(\theta) = \mathbb{E}_{x_0,c,t,\epsilon}\left[\omega_t \left\|\epsilon - \epsilon_\theta(x_t, t, c)\right\|_2^2\right], \tag{1}$$

with a prescribed time weight ω_t, often set to one. Its corresponding score estimate is

$$s_\theta(x_t, t, c) := -\frac{\epsilon_\theta(x_t, t, c)}{\sqrt{1-\bar{\alpha}_t}} \approx \nabla_{x_t} \log p_{\theta,t}(x_t \mid c).$$

Thus noise predictions and scores differ by a known time-dependent factor; the linear guidance rules below can be implemented in either representation at the same noise level.

3.2 DIRECT PREFERENCE OPTIMIZATION

Let $\mathcal{D}$ contain preference triples (x_+, x_-, c), where x_+ is preferred to x_-. For a fixed reference π_{ref}, DPO minimizes (Rafailov et al., 2023)

$$\begin{aligned} L_{\text{DPO}}(\theta) &= -\mathbb{E}_{\mathcal{D}} \log \sigma(\beta m_\theta), \\ m_\theta &= \log \frac{\pi_\theta(x_+ \mid c)}{\pi_{\text{ref}}(x_+ \mid c)} - \log \frac{\pi_\theta(x_- \mid c)}{\pi_{\text{ref}}(x_- \mid c)}. \end{aligned} \tag{2}$$

where σ is the sigmoid and $\beta > 0$ controls regularization toward the reference in the underlying KL-regularized preference objective. The implicit reward $r_\theta(x, c) = \beta \log[\pi_\theta(x \mid c)/\pi_{\text{ref}}(x \mid c)] + C(c)$ gives the Bradley–Terry model (Bradley & Terry, 1952)

$$p_\theta(x_+ \succ x_- \mid c) = \sigma(r_\theta(x_+, c) - r_\theta(x_-, c)) = \sigma(\beta m_\theta). \tag{3}$$

Diffusion-DPO. Image likelihoods are intractable for standard diffusion models. Diffusion-DPO instead uses a variational surrogate expressed through denoising errors (Wallace et al., 2024). For noisy preferred and dispreferred samples x_t^+ and x_t^-, define

$$\delta_\theta^\pm = \|\epsilon^\pm - \epsilon_\theta(x_t^\pm, t, c)\|_2^2 - \|\epsilon^\pm - \epsilon_{\text{ref}}(x_t^\pm, t, c)\|_2^2. \tag{4}$$

The training objective is

$$L_{\text{Diff-DPO}}(\theta) = -\mathbb{E}_{\mathcal{D},t,\epsilon^+,\epsilon^-} \log \sigma\left[-\beta T \omega_t \left(\delta_\theta^+ - \delta_\theta^-\right)\right], \tag{5}$$

where t is sampled uniformly from the T diffusion steps. We use this denoising form directly when deriving the local connection to cPGD.

3.3 CLASSIFIER-FREE GUIDANCE

CFG combines conditional and unconditional predictions (Ho & Salimans, 2022):

$$\widehat{s}_{\mathrm{CFG}} = s_u + \gamma(s_c - s_u), \qquad s_c = s_\theta(x_t, t, c), \quad s_u = s_\theta(x_t, t, \varnothing).$$

The null condition $\varnothing$ is learned by dropping text conditioning during training, and γ controls text guidance. For exact scores at a fixed noise level, this is the score of a density proportional to $p_t(x_t)\, p_t(c \mid x_t)^\gamma$, when normalizable. This fixed-time identity does not in general identify the terminal distribution produced by the guided reverse process. We use the linear composition as a sampling rule and reserve w for the separate preference-guidance strength.

4 METHOD

4.1 PREFERENCE-GUIDED INFERENCE

In the PGD and cPGD sampling equations, each predictor includes text CFG at the fixed scale γ from Sec. 3.3; w independently controls preference guidance. The denoisers in their training objectives remain the underlying model outputs before this CFG combination.

Let s_{ref} be the predictor of a pretrained reference model and s_{DPO} that of a lightly DPO-tuned copy. PGD uses their difference as an inference-time preference direction:

$$\widehat{s}_{\mathrm{PGD}} = s_{\mathrm{ref}} + w\,(s_{\mathrm{DPO}} - s_{\mathrm{ref}})\,. \tag{6}$$

All predictions are evaluated at the same (x_t, t, c), omitted for brevity. At $w = 0$, PGD recovers the reference; at $w = 1$, it recovers the tuned predictor. Larger weights extrapolate along the learned residual. This gives a lightly trained module a sampling influence that can be adjusted without further finetuning.

The same learned residual can steer a compatible target predictor:

$$\widehat{s}_{\mathrm{target}} = s_{\mathrm{target}} + w\,(s_{\mathrm{DPO}} - s_{\mathrm{ref}})\,.$$

The source reference remains inside the residual. Transfer requires compatible latent coordinates, noise levels, conditioning, and prediction parameterizations; the target's parameters remain fixed. Guidance strength controls the empirical trade-off among preference scores, diversity, and deviation from the base distribution.

4.2 CONTRASTIVE PREFERENCE-GUIDED DIFFUSION

PGD learns a preference direction through DPO and controls its strength during sampling. We extend this separation to the treatment of dispreferred images: both branches learn to generate their respective data, and repulsion is introduced by subtracting their predictions at inference. Let $\mathcal{D}_+$ and $\mathcal{D}_-$ be the preferred and dispreferred image–prompt marginals of $\mathcal{D}$. Starting from the same reference parameters θ_0, we independently train

$$L_{\mathrm{pos}}(\theta_+) = \mathbb{E}_{(x_0,c)\sim\mathcal{D}_+,t,\epsilon}\left\|\epsilon - \epsilon_{\theta_+}(x_t, t, c)\right\|_2^2, \tag{7}$$

$$L_{\mathrm{neg}}(\theta_-) = \mathbb{E}_{(x_0,c)\sim\mathcal{D}_-,t,\epsilon}\left\|\epsilon - \epsilon_{\theta_-}(x_t, t, c)\right\|_2^2, \tag{8}$$

where x_t follows the forward process in Sec. 3.1. At each sampling step, contrastive PGD (cPGD) combines predictions at the same (x_t, t, c):

$$\widehat{s}_{\mathrm{cPGD}} = s_{\mathrm{ref}} + w\big(s_{\theta_+} - s_{\theta_-}\big). \tag{9}$$

Both branches fit their respective data distributions with ordinary denoising losses. The difference between their underlying conditional scores estimates the gradient of the log density ratio between these distributions; Appendix D.2 develops this interpretation and its extension to text CFG.

Connection to reweighted Diffusion-DPO. The contrast also has a local connection to preference optimization. Consider a *single* model with parameters θ, and let $\ell_\theta(\xi)$ be the per-pair loss in Eq. (5), where ξ includes the preference pair, timestep, and Gaussian noise. Differentiation gives

$$\begin{aligned} \nabla_\theta \ell_\theta(\xi) &= a_\theta(\xi)\big[\nabla_\theta \delta_\theta^+ - \nabla_\theta \delta_\theta^-\big], \\ a_\theta(\xi) &= \beta T \omega_t\, \sigma\big[\beta T \omega_t (\delta_\theta^+ - \delta_\theta^-)\big]\,. \end{aligned} \tag{10}$$



Figure 3: Representative outputs from the base model, its DPO-tuned counterpart, and PGD.



Figure 4: Qualitative comparison of cPGD with its components. Positive SFT and Negative SFT sample the generators fitted to preferred and dispreferred images, respectively.



Figure 5: Comparison of preference-optimization methods on SDXL. Columns show outputs from the base model (SDXL), DPO, MaPO, NPO, PGD, and cPGD. The examples illustrate differences in prompt adherence and visual appearance.

Because the reference predictions are fixed, the derivatives of $\delta_\theta^{\pm}$ are those of the corresponding denoising losses. Removing the positive scalar a_θ through a detached inverse weight yields

$$\begin{aligned} L_{\mathrm{rw}}(\theta) &= \mathbb{E}_\xi\left[\mathrm{sg}\big(a_\theta(\xi)^{-1}\big)\,\ell_\theta(\xi)\right], \\ \nabla_\theta L_{\mathrm{rw}}(\theta) &= \nabla_\theta L_{\mathrm{pos}}(\theta) - \nabla_\theta L_{\mathrm{neg}}(\theta), \end{aligned} \tag{11}$$

where sg stops gradients through the weight. At a common initialization, subtracting one equal-size gradient-descent update for each branch recovers this single-model update direction. Linearizing the score predictor then connects that update difference to the inference composition in Eq. (9), as shown in Appendix D.1. The connection is local: it does not equate the complete optimization trajectories. The inverse weight is used only for this analysis; cPGD trains its two generators with Eqs. (7)–(8).

Table 1: Win rates of preference optimization methods against the SDXL model on the Pick-a-Pic v2 test set and the Parti-Prompts benchmark. Model checkpoints for other methods are provided by their respective authors. Within each target-model block, the best values are **bold** and the second distinct values are underlined; ties share rank.

Base Model	Inference Strategy	Pick-a-Pic v2 test (424 prompts)						Parti-Prompts (1632 prompts)						Avg. ↑
		PS ↑	HPSv2 ↑	HPSv3 ↑	Aes ↑	CLIP ↑	IR ↑	PS ↑	HPSv2 ↑	HPSv3 ↑	Aes ↑	CLIP ↑	IR ↑	
SDXL	–	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0
	NPO	58.7	59.2	69.1	**52.1**	37.5	53.5	55.0	56.4	62.1	55.1	39.0	51.5	54.1
	PGD	78.8	79.0	73.6	51.9	63.7	69.3	**78.7**	78.4	**78.2**	**68.8**	54.0	69.1	70.3
	cPGD	**80.0**	**80.2**	**77.1**	50.9	**64.9**	**69.8**	75.8	**80.3**	77.8	57.2	**62.0**	**74.0**	**70.8**
DPO-SDXL	–	71.7	77.6	67.9	53.3	61.6	65.8	64.0	70.3	64.0	57.5	58.5	69.7	65.2
	NPO	76.9	81.8	81.4	53.8	57.8	70.8	70.6	78.4	77.5	63.4	56.3	70.8	70.0
	PGD	**83.3**	**85.4**	**85.6**	59.7	**62.3**	**73.6**	**80.8**	**83.9**	**81.7**	**67.6**	57.5	**76.1**	**74.8**
	cPGD	80.9	77.6	84.7	**63.9**	58.7	64.9	73.9	79.2	72.8	59.4	**64.1**	74.4	71.2
MaPO-SDXL	–	55.9	65.3	61.8	68.2	50.2	68.2	52.0	64.4	58.5	72.4	48.2	65.0	60.8
	PGD	**80.4**	**81.6**	**81.6**	**75.7**	51.4	72.2	**78.9**	77.8	**79.6**	**77.8**	53.6	72.7	**73.6**
	cPGD	77.4	78.8	72.9	69.1	**59.4**	**72.4**	72.5	**81.1**	76.9	70.1	**58.2**	**74.4**	71.9
SPO-SDXL	–	89.4	83.0	96.0	81.8	33.3	78.8	87.8	85.5	92.2	88.1	31.7	74.9	76.9
	PGD	92.2	86.1	96.5	**82.1**	42.5	81.4	91.4	87.7	**93.9**	**88.4**	48.0	77.3	80.6
	cPGD	**92.9**	**88.4**	**96.7**	78.8	**53.8**	**84.4**	**92.0**	**90.3**	**93.9**	83.6	**50.4**	**81.3**	**82.2**

Table 2: Win rates of preference optimization methods against the SD1.5 model on the Pick-a-Pic v2 test set and the Parti-Prompts benchmark. Model checkpoints for other methods are provided by their respective authors. Within each target-model block, the best values are **bold** and the second distinct values are underlined; ties share rank.

Base Model	Inference Strategy	Pick-a-Pic v2 test (424 prompts)						Parti-Prompts (1632 prompts)						Avg. ↑
		PS ↑	HPSv2 ↑	HPSv3 ↑	Aes ↑	CLIP ↑	IR ↑	PS ↑	HPSv2 ↑	HPSv3 ↑	Aes ↑	CLIP ↑	IR ↑	
SD1.5	–	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0
	PGD	**78.3**	71.2	67.9	62.3	58.5	63.7	**68.0**	65.0	59.6	58.1	55.0	56.9	63.7
	cPGD	76.9	**71.7**	**71.9**	**63.2**	**59.9**	**72.2**	66.4	**76.9**	**68.9**	**68.1**	**58.4**	**71.0**	**68.8**
DPO-SD1.5	–	76.4	67.7	66.3	65.1	55.9	60.6	67.3	64.8	64.5	62.2	53.6	61.0	63.8
	PGD	**79.2**	70.3	66.3	66.3	59.4	63.2	**74.2**	67.4	65.0	63.1	55.5	62.9	66.1
	cPGD	79.0	**81.8**	**75.5**	**71.7**	**62.3**	**76.2**	73.8	**74.1**	**69.6**	**69.0**	**59.0**	**67.3**	**71.6**
KTO-SD1.5	–	72.6	78.1	76.2	68.6	58.5	75.0	66.6	78.3	71.9	68.8	53.3	71.3	69.9
	PGD	**81.6**	**83.3**	**80.2**	70.3	**61.1**	**77.4**	**72.1**	**80.3**	**72.8**	**72.4**	55.1	**73.8**	**73.4**
	cPGD	76.7	80.2	75.9	**70.5**	60.4	74.3	66.2	77.1	69.5	68.4	**55.9**	72.2	70.6
SPO-SD1.5	–	71.2	63.0	64.9	68.2	38.7	61.1	68.6	61.2	64.2	71.9	37.7	61.6	61.0
	PGD	79.7	70.3	69.6	69.8	44.3	67.9	74.2	66.5	66.2	72.2	46.7	67.0	66.2
	cPGD	**82.3**	**81.8**	**75.5**	**71.2**	**60.6**	**76.2**	**74.8**	**71.9**	**71.3**	**73.9**	**47.5**	**72.9**	**71.7**

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

Data and evaluation prompts. We train primarily on Pick-a-Pic v2 (Kirstain et al., 2023), which contains approximately 900,000 preference pairs from 58,000 unique prompts. We also study HPDv3 (Ma et al., 2025), with 1.08M text–image pairs and 1.17M pairwise annotations, using both the full dataset and a subset selected by image-source provenance. Evaluation uses the Pick-a-Pic v2 test split (424 prompts), HPDv2 (Wu et al., 2023) (400 prompts), and Parti-Prompts (Yu et al., 2022) (1,632 prompts).

Models and comparisons. We use Stable Diffusion v1.5 (SD1.5) (Rombach et al., 2022) and Stable Diffusion XL (SDXL) (Podell et al., 2023). Comparisons include Diffusion-DPO (Wallace et al., 2024), Diffusion-KTO (Li et al., 2024), MaPO (Hong et al., 2026), and Diffusion-NPO (Wang et al., 2025). Preferred-image supervised fine-tuning (SFT-Pref) is also included in the SD1.5 mean-score comparisons in Fig. 21. The tabulated baseline checkpoints are supplied by the respective authors; these comparisons evaluate released systems rather than retraining every method under one common training budget. We additionally apply guidance to SPO (Liang et al., 2024) checkpoints to test compatibility with an online preference-optimization method.

Metrics. We evaluate PickScore (PS) (Kirstain et al., 2023), HPSv2 (Wu et al., 2023), HPSv3 (Ma et al., 2025), ImageReward (IR) (Xu et al., 2024), CLIP score (Radford et al., 2021), and Aesthetics score (Aes) (Schuhmann, 2022). For each metric, the win rate is the percentage of prompts on which a method scores above the original, untuned base model, using one image per prompt with matched sampling settings and seed. The average win rate is an unweighted mean of the reported metrics. We measure within-prompt diversity by pairwise squared distances between CLIP embeddings. For prior preservation, we compute FID between the method's images and images generated by SDXL

under matched prompts and seeds. This FID measures deviation from the generated base distribution, rather than fidelity to real images. Diversity and FID use 25 prompts and 40 seeds per prompt; details appear in Appendix B.

Implementation. We use an effective training batch size of 2,048 image pairs. We use AdamW for SD1.5 and Adafactor for SDXL, with base learning rates of 3×10^{-8} and 5×10^{-9}, respectively. The Diffusion-DPO parameter β is 3,000 for SD1.5 and 5,000 for SDXL. PGD uses a module trained for 2,000 steps; cPGD uses positive and negative generators trained for 500 steps with ordinary denoising losses. Both backbones use 50-step sampling, with text CFG scales of 7.5 for SD1.5 and 5.0 for SDXL.

5.2 Results

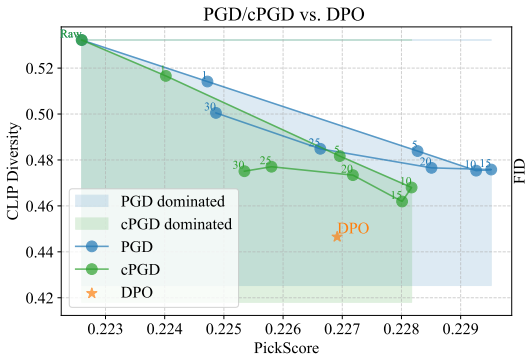


Figure 6: Reward–diversity

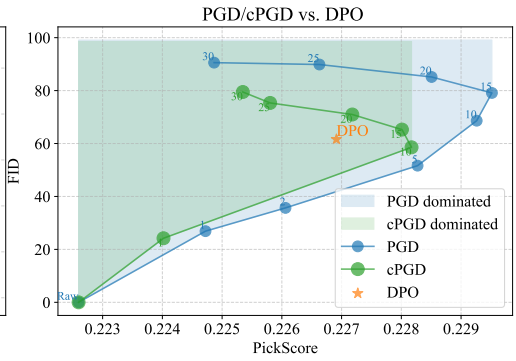


Figure 7: Reward–FID to SDXL

Preference alignment. Tables 1 and 2 report win rates against the untuned backbone. On SDXL, the mean across the two prompt suites and six metrics is 70.3% for PGD and 70.8% for cPGD, compared with 65.2% for released DPO; adding PGD to DPO-SDXL raises it to 74.8%. On untuned SD1.5, cPGD reaches 68.8%, compared with 63.7% for PGD. The relative advantage depends on the target: PGD is stronger on DPO- and MaPO-tuned SDXL, while cPGD leads on SPO-SDXL. Changes in Aes and CLIP score indicate distinct trade-offs across criteria.

Human judgments provide complementary evidence (Appendix G). Across 55 prompts and 20 participants, PGD is selected in 45.5% of decisions, cPGD in 33.0%, and DPO in 29.5%. Participants choose one to three images from six anonymized, randomly ordered alternatives, so these are selection rates rather than pairwise win rates.

Diversity and prior preservation. Figures 6 and 7 trace the effect of the guidance scale on PickScore, CLIP diversity, and FID to the generated SDXL reference. PGD at $w = 5$ and cPGD at $w = 10$ each achieve higher PickScore, greater CLIP diversity, and lower FID than the plotted DPO checkpoint. Increasing the scale further changes this trade-off: higher reward need not preserve the diversity or proximity to the base distribution obtained at a weaker scale. Shading visualizes trade-off envelopes obtained by linearly connecting the evaluated settings.

Transfer to preference-tuned models. Reusing the guidance modules improves aggregate performance on DPO, MaPO, KTO, and SPO targets (Tables 1 and 2). Figure 17 additionally applies SDXL modules to KOALA (Lee et al., 2023), using their shared latent space without updating the target parameters.

Guidance strength and training duration. We vary the guidance scale for SDXL modules trained on Pick-a-Pic v2 for 300, 500, 1,000, and 2,000 steps. Figure 9 shows improvements under moderate amplification and declining scores at larger scales, with a smaller decline for the shorter training runs. The examples in Fig. 8 show corresponding changes in prompt attributes and visual appearance.

Preference-data composition. Table 3 compares training on full HPDv3 with training on a subset selected by image-source provenance. On the subset, cPGD obtains the highest score on all six metrics. On the full dataset, PGD leads on PickScore, HPSv3, and Aes, whereas cPGD leads on

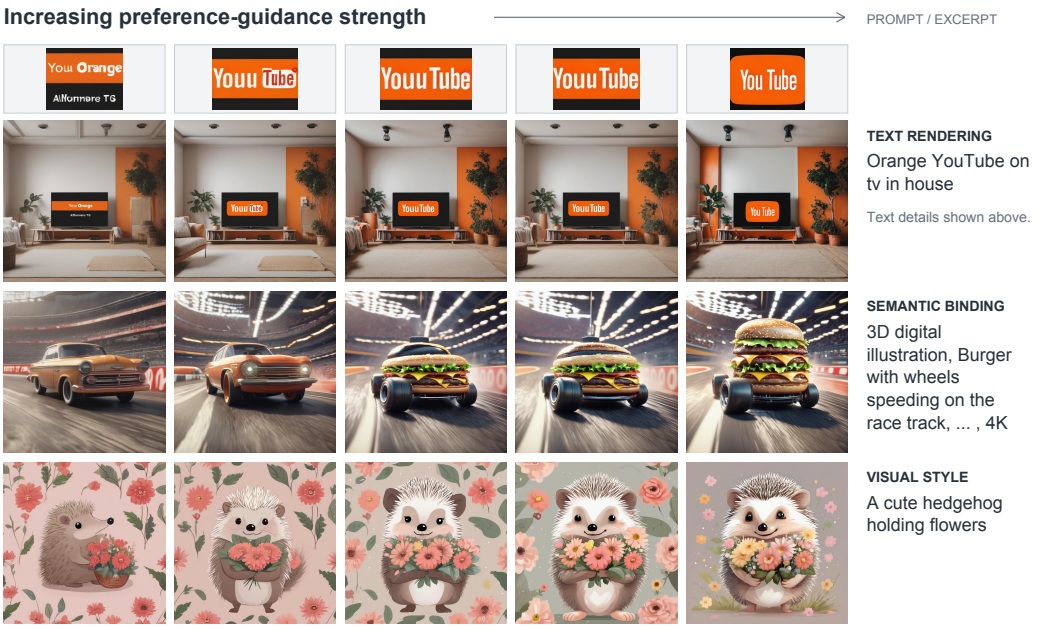


Figure 8: Qualitative effect of preference-guidance strength. Each row uses the same prompt, with guidance weight w increasing from left to right. Rows illustrate text rendering, semantic binding, and visual style; the top insets show text details from the first row. Prompt excerpts appear at right.

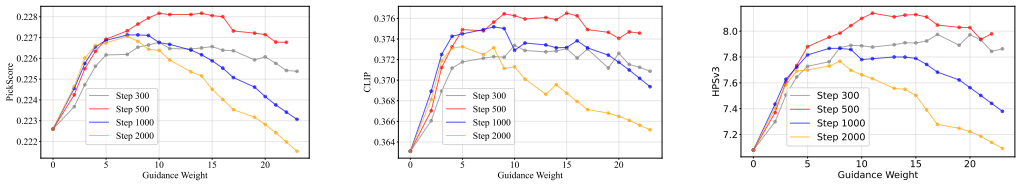


Figure 9: Effect of guidance weight w on automatic metrics (SDXL). Left: PickScore; Middle: CLIP score; Right: HPSv3. "Step" denotes the training steps of the guidance module. Scores initially improve with guidance and can decline at larger scales.

HPSv2, CLIP, and ImageReward. Both variants improve most metrics over DPO in both settings. Because source filtering changes the training examples as well as their composition, these results establish sensitivity to the choice of preference data without isolating image-quality variance as the cause.

Table 3: Alignment performance after training on two HPDv3 configurations. Subset retains pairs whose images belong to the specified source categories; Fullset uses full HPDv3. All methods use SDXL. The best results are bold and the second-best are underlined.

Set	Method	PS ↑	HPSv2 ↑	HPSv3 ↑	Aes ↑	CLIP ↑	IR ↑
Subset	SDXL	0.222	0.277	7.079	6.052	0.363	0.658
	DPO	0.225	0.282	8.032	6.112	0.361	0.823
	PGD	<u>0.226</u>	<u>0.285</u>	<u>9.459</u>	<u>6.203</u>	<u>0.364</u>	<u>0.874</u>
	cPGD	**0.228**	**0.289**	**10.045**	**6.267**	**0.365**	**1.031**
Fullset	SDXL	0.222	0.277	7.079	6.052	0.363	0.658
	DPO	0.226	0.284	8.261	6.165	<u>0.366</u>	0.918
	PGD	**0.229**	<u>0.287</u>	**10.065**	**6.505**	0.364	<u>1.063</u>
	cPGD	<u>0.227</u>	**0.290**	<u>9.243</u>	<u>6.179</u>	**0.373**	**1.143**

6 DISCUSSION

Learning a direction and amplifying it. Training duration and guidance strength affect alignment differently. Further finetuning can change both the magnitude and direction of a model's prediction residual, whereas increasing w scales the residual of a fixed checkpoint. A lightly trained model can therefore supply useful preference information even when that information has only a weak effect on its own samples. The guidance sweeps support this distinction: moderate amplification improves scores, and shorter training runs show a smaller decline at large scales. Amplification still depends on the learned direction and magnifies its errors as well as its useful signal.

The neural tangent kernel view connects this behavior to local features of the reference model (Jacot et al., 2018). Let $f_\theta(u)$ be the sampling prediction at a fixed input $u = (x_t, t, c)$. For a small parameter displacement,

$$f_{\theta_0+\Delta\theta}(u) - f_{\theta_0}(u) \approx J_{\theta_0}(u)\Delta\theta.$$

Here $J_{\theta_0}(u)$ is the parameter Jacobian; its inner products define the tangent kernel. Within this local approximation, preference training selects a change expressed through the reference model's tangent features. Guidance can strengthen the effect of that change without further optimization of the features themselves. For cPGD, the common-initialization gradient-step analysis in Sec. 4.2 relates the initial contrast of these changes to reweighted Diffusion-DPO. As parameter displacements grow, curvature limits this interpretation and the correspondence to a merged model.

Why fit the dispreferred distribution? Here, "dispreferred" denotes the lower-ranked side of a preference pair, rather than an absolute category of low-quality images. The negative generator supplies a reference specific to those preferences. Contrasting it with the positive generator provides relative distributional evidence while both training objectives remain generative. The empirical advantage over PGD varies with the training data and target model: cPGD leads across the HPDv3 subset metrics and on untuned SD1.5, whereas PGD is stronger on DPO- and MaPO-tuned SDXL.

Reusing and compressing guidance. Transfer to already aligned checkpoints shows that the learned direction remains useful beyond its source generator. Explicit guidance preserves this modularity and allows the sampling strength to be adjusted. For deployment with one model, local linearization also motivates extrapolating the parameter updates. The merged checkpoints retain preference gains (Table 4), although curvature can amplify the discrepancy from explicit guidance at large scales, so the merge coefficient is selected separately (Appendix E). Partial-step guidance provides another compromise: guiding 10 of 50 steps gives a 70.2% PickScore win rate at $1.4\times$ baseline sampling time, versus 79.1% at $3.0\times$ for all steps (Table 6). These options make the inference budget a practical choice after preference training.

7 CONCLUSION

Our results support treating preference adaptation as a direction for guiding generation. Lightly trained residuals improve preference scores, expose useful reward–diversity trade-offs, and transfer to compatible generators. cPGD extends this approach by learning both sides of a preference through ordinary denoising and introducing their contrast at inference, with a local connection to reweighted DPO. Together with partial-step guidance and model merging, these findings make preference adaptation reusable across sampling strengths, target checkpoints, and inference budgets.

AI USE STATEMENT

AI assistants were used to support the preparation of this manuscript, primarily for reorganizing LaTeX source code, improving writing clarity, and refining the presentation of the text. The authors reviewed all AI-assisted modifications and verified the correctness of the equations, experimental results, tables, and technical claims. The authors remain solely responsible for the final manuscript, including its content, conclusions, and scientific integrity.

REFERENCES

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/14f75513f0f1ca01de1e826b52e6b840-Abstract-Conference.html.

Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

Min Cheng, Fatemeh Doudi, Dileep Kalathil, Mohammad Ghavamzadeh, and Panganamala R. Kumar. Diffusion blend: Inference-time multi-preference alignment for diffusion models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/54093759b45e92c1f68ccfc655820464-Abstract-Conference.html.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2818–2829, 2023.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Kevin Clark, Paul Vicol, Kevin Swersky, and David J. Fleet. Directly fine-tuning diffusion models on differentiable rewards. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/13e8be77982beb73d7ed0bbf122f9f3c-Abstract-Conference.html.

Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Carles Domingo-Enrich, Michal Drozdzal, Brian Karrer, and Ricky T. Q. Chen. Adjoint matching: Fine-tuning flow and diffusion generative models with memoryless stochastic optimal control, 2025. URL https://arxiv.org/abs/2409.08861.

Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. DPOK: Reinforcement learning for fine-tuning text-to-image diffusion models. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/fc65fab891d83433bd3c8d966edde311-Abstract-Conference.html.

Minghao Fu, Guo-Hua Wang, Liangfu Cao, Qing-Guo Chen, Zhao Xu, Weihua Luo, and Kaifu Zhang. CHATS: Combining human-aligned optimization and test-time sampling for text-to-image generation. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 17852–17874. PMLR, 2025. URL https://proceedings.mlr.press/v267/fu25g.html.

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 (NeurIPS)*, volume 30, 2017.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *Advances in Neural Information Processing Systems (NeurIPS) Workshop on Deep Generative Models*, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 2020.

Jiwoo Hong, Sayak Paul, Noah Lee, Kashif Rasul, James Thorne, and Jongheon Jeong. Margin-aware preference optimization for aligning diffusion models without reference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 4744–4752, 2026. doi: 10.1609/aaai.v40i6.42476. URL https://ojs.aaai.org/index.php/AAAI/article/view/42476.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Luozhijie Jin, Zijie Qiu, Jie Liu, Zijie Diao, Lifeng Qiao, Ning Ding, Alex Lamb, and Xipeng Qiu. Inference-time alignment control for diffusion models with reinforcement learning guidance, 2025. URL https://arxiv.org/abs/2508.21016.

Tero Karras, Miika Aittala, Tuomas Kynkäänniemi, Jaakko Lehtinen, Timo Aila, and Samuli Laine. Guiding a diffusion model with a bad version of itself. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/5ee7ed60a7e8169012224dec5fe0d27f-Abstract-Conference.html.

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. *Advances in Neural Information Processing Systems*, 36:36652–36663, 2023.

Youngwan Lee, Kwanyong Park, Yoorhim Cho, Yong-Ju Lee, and Sung Ju Hwang. Koala: Empirical lessons toward memory-efficient and fast diffusion models for text-to-image synthesis, 2023.

Shufan Li, Konstantinos Kallidromitis, Akash Gokul, Yusuke Kato, and Kazuki Kozuka. Aligning diffusion models by optimizing human utility. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Zhanhao Liang, Yuhui Yuan, Shuyang Gu, Bohan Chen, Tiankai Hang, Mingxi Cheng, Ji Li, and Liang Zheng. Aesthetic post-training diffusion models from generic preferences with step-by-step preference optimization. *arXiv preprint arXiv:2406.04314*, 2024.

Zhen Liu, Tim Z. Xiao, Weiyang Liu, Yoshua Bengio, and Dinghuai Zhang. Efficient diversity-preserving diffusion alignment via gradient-informed gflownets. In *International Conference on Learning Representations*, 2025.

Yuhang Ma, Yunhao Shui, Xiaoshi Wu, Keqiang Sun, and Hongsheng Li. Hpsv3: Towards wide-spectrum human preference score, 2025. URL https://arxiv.org/abs/2508.03789.

Chenlin Meng, Yang Song, Jiaming Song, Jiaming Wu, Prafulla Dhariwal, Alexey Nichol, Xinyang Chen, Jascha Sohl-Dickstein, and Stefano Ermon. On distillation of guided diffusion models. *arXiv preprint arXiv:2306.05544*, 2023.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

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, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

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 *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/a85b405ed65c6477a4fe8302b5e06ce7-Abstract-Conference.html.

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*, pp. 10684–10695, 2022.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Lit, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Raphael Gontijo-Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations*, 2022.

Christoph Schuhmann. Laion-aesthetics. https://laion.ai/blog/laion-aesthetics/, 2022. Accessed: 2023 - 11- 10.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=PxTIG12RRHS.

Yang Song, Chenlin Meng, and Stefano Ermon. Consistency models. In *International Conference on Machine Learning*, 2023.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. Learning to summarize with human feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

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*, pp. 8228–8238, 2024.

Fu-Yun Wang, Yunhao Shui, Jingtan Piao, Keqiang Sun, and Hongsheng Li. Diffusion-NPO: Negative preference optimization for better preference aligned generation of diffusion models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/b20852ac0a083262ecc98b49bca43086-Abstract-Conference.html.

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, 2023. URL https://arxiv.org/abs/2306.09341.

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. *Advances in Neural Information Processing Systems*, 36, 2024.

Shentao Yang, Tianqi Chen, and Mingyuan Zhou. A dense reward view on aligning text-to-image diffusion with preference. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 55998–56032, 2024.

Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *Trans. Mach. Learn. Res.*, 2022.

Huaisheng Zhu, Teng Xiao, and Vasant G Honavar. DSPO: Direct score preference optimization for diffusion model alignment. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=xyfb9HHvMe.