Debiased DPO for Diffusion Models

Khiem Pham* Cornell University
Quang Nguyen* Rutgers University
Tung Nguyen* Cornell Tech
Jingsen Zhu Cornell University

Michele Santacatterina New York University
Dimitris Metaxas Rutgers University
Ramin Zabih Cornell Tech

Abstract

Direct Preference Optimization (DPO) enables offline alignment of diffusion models from preference pairs without explicit reward modeling, but scaling DPO requires large amounts of expensive human preference data. In this paper, we consider a semi-supervised setting in which only a small subset of preference pairs is annotated by humans, while the remaining, much larger pool is labeled with inexpensive synthetic feedback, e.g., vision-language models scoring image pairs or self-training from the current model. However, such synthetic supervision is not a substitute for human judgment: it can be systematically misaligned, so naively mixing human and synthetic preferences can lead to misaligned model behavior. We introduce DeDPO, a debiased objective that integrates a doubly robust estimator from causal inference into DPO to correct synthetic-label misalignment while preserving the simplicity of DPO. Experiments show improved label efficiency and robustness to synthetic-labeler quality, allowing DeDPO to outperform standard DPO under the same human-label budget and to approach fully human-supervised performance despite using four times fewer human-labeled examples.

1 Introduction

Text-to-image diffusion models have demonstrated impressive capabilities in generating images from natural language prompts [13, 8, 40]. Trained at scale on large image–text corpora [42], these models generalize broadly, but their pretraining objectives do not directly capture subjective, user-valued criteria such as fine-grained prompt adherence, aesthetic preference, or style consistency. As a result, generated images can be plausible yet misaligned with what humans actually prefer.

Preference-based alignment addresses this gap by fine-tuning models using pairwise comparisons. In language models, RLHF [36, 16] learns a reward model from human preferences and optimizes the policy via reinforcement learning [35, 43]. Similar ideas have been applied to image generation [3, 9]. However, training and maintaining a reward model, as well as performing RL optimization, increases both system complexity and computational cost. Direct Preference Optimization (DPO) [38] simplifies this pipeline by directly optimizing a preference objective from paired comparisons without explicitly learning a reward model. Diffusion-DPO [49] extends DPO to diffusion models and has been adopted in a growing set of generative tasks, including image inpainting [45], video generation [27, 53], and physically grounded 3D generation [24]. Its simplicity and offline nature make DPO-style training an appealing default for diffusion alignment.

Despite this appeal, **DPO-based methods inherit a key bottleneck: collecting large-scale, high-quality human preference labels is expensive and time-consuming**. This limitation motivates the use of *synthetic* feedback, such as pairwise preferences generated by vision-language models, reward models, or self-training, to reduce reliance on costly human annotation [12, 22]. However, synthetic

*Equal contribution. Correspondence: {dkp45,tn375}@cornell.edu and quang.ng@rutgers.edu.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).



Figure 1: **Mismatch between synthetic and human preferences on Pick-a-Pic v2 [20].** Qwen2.5-VL tends to favor semantic coherence and style consistency, but often underweights photorealism-related cues such as *"8k"* or *"hyper-detailed"*. Human raters place more weight on photorealistic fidelity, even when some style details are missed. This gap motivates DeDPO, which uses a small set of human labels to correct synthetic preference bias.

preferences are not a drop-in replacement for human judgments because they can be systematically misaligned. As Fig. 1 illustrates, this misalignment can vary across prompts, domains, and sampling settings. In practice, synthetic labels are often unavoidable when aligning large diffusion models under a limited human-label budget. The key question is therefore not whether to combine human and synthetic preferences, but how to combine them. Treating synthetic labels as ground truth and training DPO on both human and synthetic pairs can degrade alignment when the synthetic signal diverges from human preferences.

To address this problem, we propose **DeDPO**, a debiased semi-supervised DPO objective for diffusion alignment. The main idea is simple: use synthetic preferences for coverage, but correct their systematic errors with a small set of high-quality human labels instead of treating them as ground truth. Concretely, DeDPO augments standard DPO with a doubly robust correction term while preserving DPO's practical advantages: fully offline training, no separate reward model, and compatibility with existing diffusion pipelines. Our contributions are threefold. **First**, we formulate diffusion alignment with mixed human and synthetic preferences as a debiased semi-supervised learning problem and derive a simple objective that corrects synthetic-label bias using labeled data. **Second**, building on the doubly robust semi-supervised estimator of [61] and the orthogonal statistical learning framework of [10], we instantiate this theory for Diffusion-DPO: we show that the resulting estimator is unbiased and doubly robust under standard assumptions, which helps explain why naive mixing can fail when synthetic preferences are misaligned. We do not claim a new general doubly robust estimator. **Third**, across two datasets, different backbones, and several synthetic labelers, we show that DeDPO improves over matched DPO baselines, with the largest gains when synthetic labels are more misaligned with humans, and approaches the fully human-supervised reference point while using only **25% of the human annotations**. Throughout the paper, we use *misalignment* and *bias* interchangeably, except when referring to human bias. Here, statistical bias means a systematic divergence between a predictor's distribution and a target distribution, rather than simple random noise. In our setting, this includes both a model–human gap (e.g., Qwen2.5-VL preferring aesthetics in Fig. 1) and a human–prompt gap (e.g., humans favoring photorealism over strict prompt compliance). We use *alignment* specifically to refer to reducing the model–human gap.

2 Related Work

Alignment with synthetic preferences. To reduce the cost of human comparisons, many diffusion-alignment methods rely on *synthetic* preference labels. A common approach trains reward models such as PickScore and HPSv2 on large human-preference datasets and then uses them to annotate image pairs at scale [20, 52, 49, 17]. Another line of work uses self-training or pseudo-labeling from the current diffusion model itself [23, 21, 28]. Both settings can suffer from *double-dipping*, where the same model or human-preference source is reused for both labeling and training. We instead study a label-scarce regime with a small human-labeled set and a larger unlabeled pool labeled by

pretrained external VLMs such as Qwen2.5-VL and SmolVLM [48, 33]. Because these VLM judges are imperfect and systematically biased relative to humans, naively mixing their labels with scarce human data can misguide DPO. Our method addresses this with a principled debiasing mechanism anchored by the small human-labeled subset.

Robust alignment with noisy labels. Synthetic preferences can disagree with human preferences, so robustness to label error is important. Existing work mainly follows two directions: methods based on explicit noise models [4, 7, 34] and methods based on distributionally robust optimization (DRO) [56, 51, 19, 32]. However, these approaches are primarily designed for random or worst-case label corruption. Our setting is different: the main problem is *systematic bias* in synthetic preferences, and we also have access to a small set of high-quality human labels. This reference set lets us *debias* synthetic supervision directly, rather than only smoothing labels or optimizing for worst-case robustness. Unlike robust reward-modeling approaches in RLHF [57, 5, 11, 59, 37], our method stays reward-model-free and focuses on biased synthetic preferences in diffusion alignment.

Doubly Robust / Debiased estimators. Doubly robust (DR) and debiased estimators were developed to handle biased or missing supervision [39, 41, 2]. Their key idea is to combine two estimators of the same target so that the final estimator remains reliable if either one is accurate [6, 18, 10]. This makes them well suited to our setting, where synthetic preferences provide broad but imperfect supervision and a small human-labeled set provides a high-quality reference. Closest prior work applies DR ideas to semi-supervised classification [61], while recent alignment work such as Xu et al. [54] studies related debiasing ideas for language models. **Our setting differs in three ways**: we study *pairwise preference optimization* rather than classification, we focus on diffusion alignment rather than language generation, and we use *external* synthetic labelers instead of relying only on self-training. Compared with self-training methods [23, 21, 28], external supervision avoids confirmation bias and better matches the sample-splitting conditions behind DR theory. Compared with Xu et al. [54], our method remains fully offline and integrates debiasing directly into Diffusion DPO.

3 Preliminaries

Text-to-image diffusion model. Diffusion models [47, 15, 46, 8, 40] generate images by learning to reverse a gradual noising process applied to image latents. Formally, they model the transformation from a standard Gaussian noise map $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \boldsymbol{I})$ to a sample from the target data distribution via a denoising diffusion process. The forward (noising) process perturbs a clean image latent $\boldsymbol{x}_0$ across multiple timesteps t as:

$$\boldsymbol{x}_t = \alpha_t \boldsymbol{x}_0 + \sigma_t \boldsymbol{\epsilon}, \tag{1}$$

where α_t and σ_t are time-dependent scaling coefficients satisfying $\alpha_0 = \sigma_T = 1$ and $\alpha_T = \sigma_0 = 0$. The reverse process learns to recover $\boldsymbol{x}_0$ by denoising $\boldsymbol{x}_t$ using a noise estimator $\boldsymbol{\epsilon}_\theta$, trained to predict the noise $\boldsymbol{\epsilon}$ via:

$$\min_\theta \mathbb{E}_{t\sim\mathcal{U}(0,T),\boldsymbol{\epsilon}\sim\mathcal{N}(0,\boldsymbol{I})} \left\|\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t) - \boldsymbol{\epsilon}\right\|_2^2. \tag{2}$$

In the text-to-image setting, models are further conditioned on text embeddings $\mathbf{c}$, modifying the noise predictor to $\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t, \mathbf{c})$. This conditioning enables the generation to be guided by textual prompts. To strengthen alignment between the image and the text, many models incorporate classifier-free guidance (CFG) [8, 14], which blends conditional and unconditional predictions using a guidance scale γ. The guided output is computed as:

$$\hat{\boldsymbol{\epsilon}}_\theta(\boldsymbol{x}_t, t, \mathbf{c}) = \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t) + \gamma\left(\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t, \mathbf{c}) - \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)\right), \tag{3}$$

where $\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t, \mathbf{c})$ denotes the prediction conditioned on the text, and $\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)$ denotes the unconditional prediction. CFG enhances both the fidelity and semantic relevance of generated images without relying on an external classifier.

DPO for text-to-image diffusion models. Direct Preference Optimization (DPO) [38] offers an efficient alternative to reward modeling for aligning generative models with human preferences. Diffusion-DPO [49] extends this approach to text-to-image diffusion models by adapting the DPO objective to the denoising process. Since diffusion models generate images through a sequence of latent states over T timesteps, directly computing the full trajectory likelihood is intractable. To address this, Diffusion-DPO proposes a surrogate loss based on denoising score comparisons at

individual timesteps. For a given text prompt $\boldsymbol{c}$ and an image pair $(\boldsymbol{x}_t^w, \boldsymbol{x}_t^l)$ annotated by human preference $\boldsymbol{x}_t^w \succ \boldsymbol{x}_t^l$, with Gaussian noise samples $(\boldsymbol{\epsilon}^w, \boldsymbol{\epsilon}^l) \sim \mathcal{N}(0, \boldsymbol{I})$, the preference loss is approximated as:

$$\begin{aligned} L_{\text{DPO}}(\theta) &= \mathbb{E}_{t,\boldsymbol{c},\boldsymbol{x}_t^w,\boldsymbol{x}_t^l}\left[-\log \mathcal{G}_\theta(\boldsymbol{c}, \boldsymbol{x}_t^l, \boldsymbol{x}_t^w)\right], \\ \mathcal{G}_\theta(\boldsymbol{c}, \boldsymbol{x}_t^l, \boldsymbol{x}_t^w) &= \sigma\Big(-\beta\big(\|\boldsymbol{\epsilon}^w - \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t^w, t, c)\|_2^2 - \|\boldsymbol{\epsilon}^w - \boldsymbol{\epsilon}_{\text{ref}}(\boldsymbol{x}_t^w, t, c)\|_2^2 \\ &\qquad - \|\boldsymbol{\epsilon}^l - \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t^l, t, c)\|_2^2 + \|\boldsymbol{\epsilon}^l - \boldsymbol{\epsilon}_{\text{ref}}(\boldsymbol{x}_t^l, t, c)\|_2^2\big)\Big). \end{aligned} \tag{4}$$

where $\sigma(\cdot)$ is the sigmoid function, $\boldsymbol{x}_t^w$ and $\boldsymbol{x}_t^l$ are the noise-perturbed images at timestep t (Eq. (1)), $\boldsymbol{\epsilon}_\theta$ is the trainable noise predictor and $\boldsymbol{\epsilon}_{\text{ref}}$ is a frozen reference model. This formulation encourages the model to improve denoising quality on preferred samples relative to less preferred ones, effectively aligning generation with preference signals without relying on reward modeling.

4 Method

4.1 Problem setting.

We consider a semi-supervised setting for text-to-image diffusion alignment via DPO, where only a small set of prompts is annotated with human preference pairs alongside a larger pool of synthetically labeled data. Our objective is to leverage limited human supervision together with abundant but noisy synthetic labels to achieve performance competitive with fully human-annotated training.

DPO as binary classification. We first re-interpret DPO in Eq. (4) as an equivalent binary classification problem. For an *unordered* image pair $(\boldsymbol{x}^0, \boldsymbol{x}^1)$ with text prompt $\boldsymbol{c}$, we denote $\boldsymbol{z}$ as the preference label over the pair, where $\boldsymbol{z} = 1$ denotes preference on $\boldsymbol{x}^1$ and $\boldsymbol{z} = 0$ otherwise. With this label, we can rewrite the loss in Eq. (4) as a binary cross-entropy loss

$$\begin{aligned} L_{\text{DPO}}(\theta) &= \mathbb{E}_{t,\boldsymbol{c},\boldsymbol{x}_t^0,\boldsymbol{x}_t^1}\left[\mathcal{L}\big(\mathcal{G}_\theta(\boldsymbol{c}, \boldsymbol{x}_t^0, \boldsymbol{x}_t^1), \boldsymbol{z}\big)\right], \\ \mathcal{L}(a, b) &= -b \log a - (1 - b)\log(1 - a). \end{aligned} \tag{5}$$

which is a binary classification loss with predictive model $\hat{\boldsymbol{z}} = \mathcal{G}_\theta(\boldsymbol{c}, \boldsymbol{x}_t^0, \boldsymbol{x}_t^1)$ and target label $\boldsymbol{z}$.

Notes on symbols. We drop the subscript t for brevity. For example, $\boldsymbol{x}_t^0$ is written as $\boldsymbol{x}^0$. We also drop the subscripts of the expectation operator $\mathbb{E}$. $\mathbb{E}_n$ denotes the average over n samples (equivalent to the operator $\frac{1}{n}\sum_{i=1}^n \cdot$). $\boldsymbol{y}$ denotes the input triplet $\boldsymbol{y} = (\boldsymbol{c}, \boldsymbol{x}^0, \boldsymbol{x}^1)$. Therefore, Eq. (5) can be written as:

$$L_{\text{DPO}}(\theta) = \mathbb{E}_n \mathcal{L}(\mathcal{G}_\theta(\boldsymbol{y}), \boldsymbol{z}) \tag{6}$$

We will slightly overload $\mathcal{G}$ to mean any mapping from the input $\boldsymbol{y}$ to the label prediction $\hat{\boldsymbol{z}}$, reserving $\mathcal{G}_\theta$ for the model's implicit margin (Eq. (4)). Additionally, we note that the expectation is also taken over the denoising time t.

Naive pseudo-labeling is biased. In our semi-supervised setting, let $\mathcal{D}_l = \{(\boldsymbol{y}_i, z_i)\}_{i=1}^{n_l}$ denote labeled pairs with human preferences, and $\mathcal{D}_u = \{\boldsymbol{y}_j\}_{j=1}^{n_u}$ denote unlabeled pairs. For unlabeled data, we obtain synthetic preferences $\hat{z} = \hat{\mathcal{G}}(\boldsymbol{y})$ from a cheap annotator $\hat{\mathcal{G}}$ (e.g., a pretrained VLM). A natural baseline is to *impute* the missing human label on $\mathcal{D}_u$ with $\hat{z}$, and optimize the empirical risk:

$$L_{\text{OR}}(\theta) = \mathbb{E}_{n_l}\mathcal{L}(\mathcal{G}_\theta(\boldsymbol{y}_l), z_l) + \mathbb{E}_{n_u}\mathcal{L}(\mathcal{G}_\theta(\boldsymbol{y}_u), \hat{\mathcal{G}}(\boldsymbol{y}_u)). \tag{7}$$

This objective corresponds to standard pseudo-label training: it treats $\hat{z}$ as if it were ground-truth supervision and simply mixes labeled and pseudo-labeled data. However, this approach is generally biased as an estimator of the true DPO risk. When $\hat{\mathcal{G}}$ is biased, $\hat{z}$ deviates from the true human preference z in a *directional* way, not merely as random noise. As a result, minimizing Eq. (7) can drive θ toward the synthetic labeler's preferences rather than human preferences.

4.2 DeDPO loss.

Our **key contribution** in the paper is to use debiased learning: use the unbiased signal from the small labeled set to alleviate bias from the large pseudo-labeled set. Concretely, the algorithm works by

training on synthetic labels for *all* samples, and adds a correction term on $\mathcal{D}_l$ that replaces pseudo labels by true labels:

$$L_{\text{DeDPO}}(\theta) = \mathbb{E}_{n_l+n_u}\Big[\mathcal{L}\big(\mathcal{G}_\theta(\boldsymbol{y}), \hat{\mathcal{G}}(\boldsymbol{y})\big)\Big] + \mathbb{E}_{n_l}\Big[\mathcal{L}\big(\mathcal{G}_\theta(\boldsymbol{y}_l), z_l\big) - \mathcal{L}\big(\mathcal{G}_\theta(\boldsymbol{y}_l), \hat{\mathcal{G}}(\boldsymbol{y}_l)\big)\Big]. \tag{8}$$

In the following section, we demonstrate the properties that make DeDPO an effective loss function for alignment: (i) Eq. (8) is *unbiased* regardless of the synthetic label quality (ii) the estimation error of the learned logit depends only *weakly* on synthetic-label error.

4.3 Properties of DeDPO.

4.3.1 DeDPO is unbiased

A key property of Eq. (8) is that the labeled-set correction term exactly cancels the bias introduced by training on synthetic labels. As a result, Eq. (8) is an *unbiased* estimator of the population DPO risk, regardless of the quality of the synthetic labeler $\hat{\mathcal{G}}$.

Proposition 1. *The expectation of our proposed debiased loss in Eq.* (8) *is equal to the expectation of the original DPO loss in Eq.* (4)*, i.e., L_{DeDPO} is unbiased:*

$$\mathbb{E}[L_{DeDPO}(\theta)] = \mathbb{E}[L_{DPO}(\theta)] \tag{9}$$

Proposition 1 guarantees that using large amounts of synthetic supervision does not introduce asymptotic bias. As a sanity check, if $\hat{\mathcal{G}}(\boldsymbol{y}) = z$ is perfect, the correction term vanishes and Eq. (8) reduces to training on all $n_l + n_u$ samples. If $\hat{\mathcal{G}}$ is low-quality, the correction term anchors learning to the human-labeled subset while still allowing the unlabeled pool to provide coverage. The proof of proposition 1 is the same in [61] but it only shows the average effect of the debiased loss on the synthetic labels. Following, we explain its effect on individual labels. First, we combine the 3 loss terms of Eq. (8) into one with a new target label:

Proposition 2. *Our debiased loss is characterized by using pure pseudo-labels for unlabeled data and applying an amplified correction from pseudo-labels toward ground-truth labels for labeled data.*

$$L_{DeDPO}(\theta) = \mathbb{E}_{n_l+n_u}\Big[\mathcal{L}\big(\mathcal{G}_\theta(\boldsymbol{y}), \hat{\boldsymbol{w}}\big)\Big],$$
$$\hat{\boldsymbol{w}} = \begin{cases} \hat{\mathcal{G}}(\boldsymbol{y}), & \text{if } \boldsymbol{y} \text{ is unlabeled}, \\ \hat{\mathcal{G}}(\boldsymbol{y}) + \dfrac{n_l+n_u}{n_l}\big(\mathbf{z} - \hat{\mathcal{G}}(\boldsymbol{y})\big), & \text{if } \boldsymbol{y} \text{ is labeled with } \mathbf{z}. \end{cases} \tag{10}$$

Please refer to the supplementary for the proof. Intuitively, for unlabeled data points, our loss simply uses the pseudo-labels as the supervision signal. For labeled data points, the loss introduces a correction term $\mathbf{z} - \hat{\mathcal{G}}(\boldsymbol{y})$ that moves the pseudo-label toward the true label, amplified by the importance weight $\frac{n_l+n_u}{n_l} \geq 1$, which reflects the local density ratio. When the pseudo-label is already correct, this correction term vanishes and the loss reduces to standard supervised learning. When the pseudo-label is incorrect, the amplified correction strengthens the influence of the human label on training. Through shared model parameters, corrections on labeled examples can also influence predictions on unlabeled examples.

4.3.2 Convergence of DeDPO

We now discuss a second theoretical property of our debiased loss, namely how the convergence rate depends on the number of samples and the quality of the synthetic preference model $\hat{\mathcal{G}}$. Let $f_\theta(\boldsymbol{y}) = \operatorname{logit} \mathcal{G}_\theta(\boldsymbol{y})$, $\hat{f} = f_{\hat{\theta}}$, and let f^* minimize the population DPO risk over the logit class. Write $\mathcal{G}^*(\boldsymbol{y}) = \mathbb{E}[\boldsymbol{z} \mid \boldsymbol{y}]$. Although we use a fixed labeling probability $p = \frac{n_l}{n_l+n_u}$, we state the theorem more generally for probabilities that may depend on $\boldsymbol{y}$ and require estimation. With $\boldsymbol{r}$ indicating an observed label, the inverse weight in proposition 2 becomes $\gamma^*(\boldsymbol{y}) = 1/\Pr(\boldsymbol{r} = 1 \mid \boldsymbol{y})$, estimated by $\hat{\gamma}(\boldsymbol{y})$. Our setting corresponds to $\hat{\gamma} = \gamma^* = 1/p = \frac{n_l+n_u}{n_l}$.

Theorem 1 (Informal)**.** *Suppose the synthetic labeler $\hat{\mathcal{G}}$ and any estimate $\hat{\gamma}$ are independent of the training sample. Under bounded logits and inverse weights, a logit class of finite VC-subgraph dimension, and the sampling and optimality conditions stated in the appendix, the empirical DeDPO minimizer satisfies, with high probability,*

$$\|\hat{f} - f^*\|_2^2 \le \widetilde{O}\Big(\frac{1}{n_l + n_u}\Big) + O\Big(\|\hat{\gamma} - \gamma^*\|_4^2 \|\hat{\mathcal{G}} - \mathcal{G}^*\|_4^2\Big). \tag{11}$$

The notation $\widetilde{O}$ suppresses logarithmic factors. The full statement and proof follow the framework of [10] and are given in the appendix. The theorem has two useful consequences.

Corollary 1 (Known or estimated propensity)**.** *Under the assumptions of theorem 1, with $n = n_l + n_u$, $\|\hat{f} - f^*\|_2^2 \le \widetilde{O}(1/n)$ with high probability if either:*

1. *$\hat{\gamma} = \gamma^*$, with no accuracy requirement on $\hat{\mathcal{G}}$; or*
2. *with high probability, $\|\hat{\gamma} - \gamma^*\|_4 \|\hat{\mathcal{G}} - \mathcal{G}^*\|_4 = O(n^{-1/2})$, as when both nuisance errors are $O(n^{-1/4})$.*

The first case applies directly to our random labeling design. The nuisance term vanishes, so persistent synthetic-label bias does not prevent convergence to the population optimum. No convergence of the synthetic preference model is required.

For estimated propensity, fourth-order dependence is recovered when the propensity error is bounded by a constant times the synthetic preference error. The nuisance term is then $O(\|\hat{\mathcal{G}} - \mathcal{G}^*\|_4^4)$, so the synthetic preference model only needs to converge at the slower rate $n^{-1/4}$ for the learned logit to retain the statistical rate $\widetilde{O}(1/n)$.

Notably, the independence assumption is violated by self-training on the same preference labels but is satisfied by a fixed pretrained external labeler. This "double-dipping" can have practical consequences: if the synthetic preference model is fitted on the same labeled data used for correction, the correction in proposition 2 can overfit toward zero and provide little corrective signal. *Therefore, debiasing may be less effective with self-training than with an external model*; see Fig. 3a and the accompanying discussion.

5 Experiments

5.1 Experimental setup

Datasets. We use Pick-a-Pic v2 [20] and HPDv2 [52] for training. For each dataset, we construct a high-quality 5K subset by ranking preference pairs with PickScore [20] and retaining the top examples. This step removes many ambiguous or low-quality pairs, making the scarce human-labeled subset a cleaner and more label-efficient anchor for debiasing. To verify that this curation does not drive our conclusions, Sec. C.1 repeats the main comparison on a uniformly sampled, unfiltered 5K subset. From this 5K pool, we sample a small human-labeled subset and a larger unlabeled subset, then replace the labels of the unlabeled portion with synthetic preferences. Unless otherwise stated, we use a 25%/75% split (1.25K human-labeled pairs and 3.75K pseudo-labeled pairs) and Qwen2.5-VL as the synthetic labeler, yielding the PaP-5K and HPD-5K settings reported in Table 1. We use PartiPrompt [60] for evaluation, which contains 1632 test prompts.

Notes on notations. A two-number suffix denotes the human-labeled and unlabeled split (e.g., DeDPO+25-75 uses 25% human labels and 75% synthetic/unlabeled data), while a single number denotes fully supervised training (e.g., DPO+25 uses 25% human labels only). Unless noted otherwise, DeDPO and DPO refer to the 25-75 setting.

Metrics. We evaluate preference alignment using three primary metrics: PickScore [20], HPSv3 [31], and ImageReward [55]. For **PickScore** and **HPSv3**, we report the mean score over the 1632 PartiPrompt test prompts. For ImageReward, we report **ImageReward@Top-1(%)**, defined as the percentage of prompts for which a model is ranked first among the compared methods.

Table 1: Main results on SD1.5 and SDXL trained on PaP-5K and HPD-5K, derived from PickaPicv2 [58] and HPDv2 [52] (Section 5.1). #L / #U: number of labeled / unlabeled pairs; #U=0 denotes full supervision. PickScore and HPSv3 are reported as mean ± std over 15 seeds.

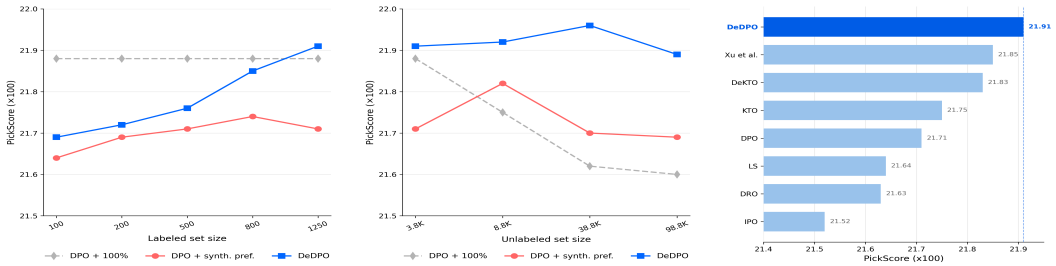
Train set	Model	Method	#L	#U	Pickscore [20] (↑)	HPSv3 [31] (↑)	ImageReward@Top-1(%) [55] (↑)
PaP-5K	SD1.5	SFT	1250	0	21.64 ± 0.011	4.81 ± 0.074	15.32
		DPO+25	1250	0	21.76 ± 0.011	4.99 ± 0.073	15.99
		DPO+25-75	1250	3750	21.71 ± 0.017	5.07 ± 0.074	19.42
		DeDPO+25-75	1250	3750	$\mathbf{21.91 \pm 0.011}$	$\mathbf{5.19 \pm 0.071}$	**22.37**
		DPO+100	5000	0	21.88 ± 0.012	5.15 ± 0.072	26.90
	SDXL	SFT	1250	0	22.01 ± 0.013	5.24 ± 0.079	17.59
		DPO+25	1250	0	22.57 ± 0.016	6.53 ± 0.066	18.69
		DPO+25-75	1250	3750	22.61 ± 0.014	6.71 ± 0.061	19.61
		DeDPO+25-75	1250	3750	$\mathbf{22.83 \pm 0.013}$	$\mathbf{6.99 \pm 0.070}$	**21.94**
		DPO+100	5000	0	22.84 ± 0.015	7.01 ± 0.069	22.18
HPD-5K	SD1.5	SFT	1250	0	21.48 ± 0.011	4.65 ± 0.074	15.81
		DPO+25	1250	0	21.61 ± 0.011	4.75 ± 0.073	18.22
		DPO+25-75	1250	3750	21.69 ± 0.017	4.76 ± 0.074	20.21
		DeDPO+25-75	1250	3750	$\mathbf{21.71 \pm 0.011}$	$\mathbf{4.79 \pm 0.071}$	**22.26**
		DPO+100	5000	0	21.61 ± 0.012	4.79 ± 0.072	22.50
	SDXL	SFT	1250	0	21.60 ± 0.013	5.11 ± 0.079	15.83
		DPO+25	1250	0	22.48 ± 0.016	6.28 ± 0.066	17.82
		DPO+25-75	1250	3750	22.52 ± 0.014	6.38 ± 0.062	20.12
		DeDPO+25-75	1250	3750	$\mathbf{22.55 \pm 0.013}$	$\mathbf{6.46 \pm 0.070}$	**22.81**
		DPO+100	5000	0	22.53 ± 0.015	6.49 ± 0.069	23.42

Baselines. We compare **DeDPO** against three primary baselines, each targeting a specific question: the effect of preference optimization, synthetic supervision, and debiasing. **SFT** is a non-preference-optimized reference for measuring the gain from DPO-style alignment over standard supervised fine-tuning. **DPO+25** uses the same human-label budget as DeDPO but no synthetic labels, testing whether synthetic preferences add useful supervision. **DPO+25-75** uses the same human/synthetic split as DeDPO but treats synthetic labels as ground truth, serving as the direct baseline for isolating the effect of our debiasing correction. We additionally report **DPO+100** as a full-human-supervision reference point, not as a primary baseline DeDPO is expected to outperform, since it uses four times more human annotations than **DeDPO+25-75**. It measures whether DeDPO matches DPO+100 on automated metrics while using substantially fewer human labels. We evaluate all methods on **SD1.5** and **SDXL** using the metrics described above.

Implementation details. (i) Training. For SD1.5, we fine-tune for 1,000 steps using AdamW [29] with a learning rate of 1×10^{-7}, a piecewise-constant schedule with 10 warmup steps, and a global batch size of 128 across 2 GPUs with 8 gradient accumulation steps. For SDXL, we train for 100 steps using AdaFactor [44] with a learning rate of 2×10^{-8} and 5 warmup steps under the same schedule. **(ii) Inference.** At test time, we generate images using 20 denoising steps with DPMSolver++ [30] and a classifier-free guidance scale of 7.5, applied uniformly across all methods and model scales. All results in Tab. 1 are averaged over 15 random seeds. Each training mini-batch contains equal numbers of labeled and unlabeled pairs, which sets the correction weight to $w_t = 2$ (Sec. B.1).

5.2 Experimental results

Comparison to DPO [49]. Tab. 1 shows that DeDPO improves over the two direct baselines, DPO+25 and DPO+25-75, across both datasets and both backbones. These are the key comparisons because DPO+25 matches the human-label budget, while DPO+25-75 matches the same human/synthetic split without debiasing. The clearest gains appear on PaP-5K, where synthetic labels are less aligned with human preferences. Relative to DPO+25-75, DeDPO improves PickScore by $+0.20/+0.22$, HPSv3 by $+0.12/+0.28$, and ImageReward@Top-1 by $+2.95/+2.33$ points for SD1.5/SDXL. These gains are larger than the corresponding run-to-run variation shown in Table 1. It also exceeds DPO+25 in every PaP-5K setting, suggesting that synthetic preferences are most useful when combined with explicit debiasing rather than naively treated as ground truth. On HPD-5K, the gains over DPO+25-75 are smaller: $+0.02/+0.03$ in PickScore, $+0.03/+0.08$ in HPSv3, and $+2.05/+2.69$ points in ImageReward@Top-1 for SD1.5/SDXL. For PickScore and HPSv3, these smaller differences should be interpreted as matching within error bars or improving within seed variance rather than



(a) PickScore under sparse human supervision: fixed 5K total datapoints with varying labeled set size. Gray dashed line shows fully supervised DPO on 5K human labels.

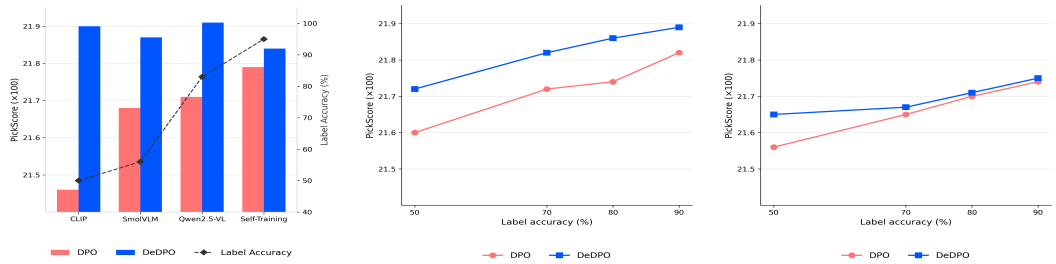
(b) Effect of scaling synthetic preferences with fixed 1.25K human labels. Gray dashed line shows fully supervised DPO on 5K–100K human labels.

(c) Comparison of DeDPO with other robust methods. 1.25K human labels and 3.75K Qwen2.5-VL labels.

Figure 2: Performance under label-scarce regimes. **(a)** DeDPO is numerically above DPO across the plotted human-label budgets (100–1250 pairs). **(b)** DeDPO shows a smaller drop as the synthetic pool scales, with the gray dashed line showing fully supervised DPO on 5K–100K PaP human labels as a reference. **(c)** DeDPO exceeds several compared baselines in this setting.

as clear statistical separations. This smaller margin is expected because synthetic labels on HPD-5K agree more closely with human preferences, about 93% versus 83% on PaP-5K, leaving less bias for the correction term to remove. Compared with SFT, all preference-optimized methods are substantially stronger on automated metrics, confirming that preference optimization is itself important for alignment. DeDPO also matches DPO+100 on some automated metrics and remains close on others while using only 25% of the human supervision. Measured as the fraction of the DPO+25-75→DPO+100 gap that is closed, DeDPO recovers 96–100% of the PickScore gap on SD1.5 and SDXL (Sec. C.3). The advantage over DPO+25-75 also persists without PickScore-based curation (Sec. C.1), on a DiT backbone (PixArt-Σ, Sec. C.2), and at a 50/50 human/synthetic split (Sec. C.4).

Comparison to other robust methods. We compare DeDPO against LS [34], IPO [1], DRO [51], Xu et al. [54], and KTO [25]. LS and DRO address noisy preferences through label smoothing and worst-case distributional optimization, respectively. In our setting, however, the main issue is *systematic synthetic bias*: the model labeler makes errors in a structured, non-random way. As a result, LS and DRO are mismatched to this problem and underperform DeDPO in Fig. 2c. We also test DeKTO, a debiased variant of the point-wise KTO loss, and find that it improves over KTO numerically (21.83 vs. 21.75), suggesting that the same debiasing idea may extend beyond DPO-style objectives. Xu et al. is the numerically closest competitor (21.85 vs. 21.91), but requires more than 2× the training time due to online generation. Overall, Fig. 2c shows that DeDPO achieves the strongest reported PickScore among the compared methods in this setting.



(a) VLM Labeler

(b) Systematic rater bias

(c) Random Preference Flip

Figure 3: **Labeler sensitivity under three types of corruption.** (PickScore ×100, SD1.5, PaP-5K). Label accuracy denotes the percentage of uncorrupted synthetic labels. DeDPO+25-75 is numerically above DPO+25-75 across the plotted labeler types and corruption levels, with the largest visible gaps under structured, directional bias.

Human evaluation.

We conduct a blind pairwise human evaluation on HPSv2 prompts [52] from four categories: *anime*, *photo*, *paintings*, and *concept-arts*. For each comparison, approximately 20 raters judge around 800 image pairs on *visual quality*, *general preference*, and *prompt following* using a five-point scale (-2 to $+2$), with model identities hidden and left–right order balanced. Against the matched DPO+25-75 baseline (Tab. 2), raters significantly prefer DeDPO+25-75 for visual quality (57.3%) and general preference (54.1%). Prompt following is directionally positive (53.4%) but has a high tie rate (49.8%) and is not significant. To anchor these results, we add comparisons against the full-human reference DPO+100. Despite using only 25% of the human labels, DeDPO+25-75 remains close to DPO+100: general preference modestly favors DPO+100, while the confidence intervals include 50% for visual quality and prompt following. DPO+100 itself is preferred over DPO+25-75 on all three dimensions. Human judgments thus corroborate the automatic metrics.

Table 2: Blind pairwise human evaluation on HPSv2 prompts. Each entry is the tie-excluded win rate (%) of the first method, with a Wilson 95% confidence interval.

Comparison	Visual quality	General preference	Prompt following
DeDPO+25-75 vs. DPO+25-75	57.3 [53.7, 60.8]	54.1 [50.4, 57.7]	53.4 [49.0, 57.8]
DPO+100 vs. DeDPO+25-75	53.7 [49.4, 57.8]	54.0 [50.1, 57.9]	51.0 [44.3, 57.5]
DPO+100 vs. DPO+25-75	57.3 [53.4, 61.1]	56.3 [51.5, 60.9]	58.6 [53.0, 63.7]

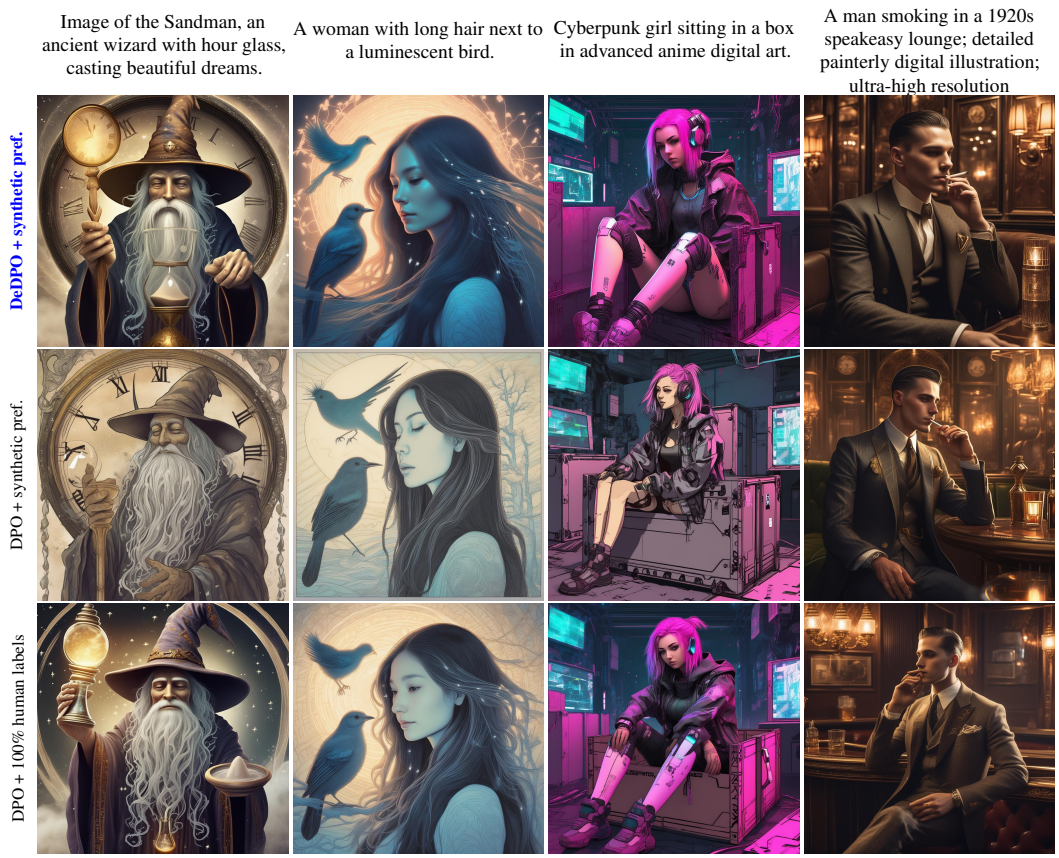


Figure 4: **Qualitative comparison** of DeDPO+25-75, DPO+100, and naively mixed DPO+25-75. DeDPO produces images closer to the fully supervised DPO+100 reference, while avoiding the stylized bias often introduced by naive mixing. Across examples, DeDPO is generally more photorealistic and better aligned with human preferences, consistent with Fig. 1. This suggests that DeDPO narrows the gap to fully human-supervised alignment using only 25% of the human labels.

Qualitative results. Building on Fig. 1, which highlights disagreement between Qwen2.5-VL and human preferences, our qualitative evaluation in Fig. 4 suggests that DeDPO reduces the tendency toward illustrative and stylized generations induced by the synthetic labeler. When trained on a mixed-label dataset comprising 25% human annotations and 75% synthetic labels from Qwen2.5-VL, DeDPO often produces more photorealistic images than DPO+25-75. These qualitative observations are consistent with the automated-metric trends, though they do not replace a human preference study.

5.3 Analysis

Labeler sensitivity. We study three kinds of sensitivity in the 25–75 setting: **(i) labeler sensitivity** across model-based labelers of varying quality (CLIP, SmolVLM, Self-Training, and Qwen2.5-VL), **(ii) sensitivity to systematic synthetic bias** favoring animated over photorealistic images, and **(iii) sensitivity to random label flipping**. Label accuracy denotes agreement between synthetic and human labels on the unlabeled set; for Self-Training, the reported 95% accuracy is measured at the final DPO step. As shown in Fig. 3, DeDPO is numerically above DPO across the plotted corruption types and accuracy levels, with the largest visible gaps under model-based labelers and systematic bias. Notably, DeDPO with the weakest labeler (CLIP, 50% accuracy) remains close to DeDPO with the strongest one (Qwen2.5-VL, 83%), suggesting reduced sensitivity to labeler quality. Self-Training [23, 28] gives the strongest non-debiased baseline (21.79) but benefits least from debiasing (21.84), consistent with the double-dipping effect in theorem 1. Under random flipping, both methods degrade smoothly, but DeDPO remains numerically above DPO, with the largest gap at 50% accuracy (21.65 vs. 21.56).

Performance under extreme sparsity. In the extremely sparse regime (100–1250 human preference pairs out of 5K total), DeDPO is numerically above DPO at every tested label budget (Fig. 2a), with the remaining pairs pseudo-labeled by Qwen2.5-VL. This directly reflects our motivation: curating reliable human preferences is expensive, so the key question is how much alignment benefit can be extracted from a very small labeled set. As more high-quality labels are added, the gap between DeDPO and DPO increases in Fig. 2a, a pattern consistent with DeDPO extracting more value from additional human annotations while continuing to leverage synthetic supervision.

Effect of scaling synthetic preferences. We fix 1250 human-labeled pairs and progressively enlarge the unlabeled pool with additional PickScore-sorted Pick-a-Pic data. As shown in Fig. 2b, DeDPO attains the highest plotted PickScore across the tested unlabeled-data budgets: it improves slightly up to 38.8K unlabeled samples and degrades less at 98.8K than DPO+25-75, which peaks earlier at 8.8K and then drops more noticeably. Even fully supervised DPO+100 deteriorates as more low-quality pairs are added, consistent with [58]. These trends are consistent with DeDPO making better use of larger pseudo-labeled pools while keeping training anchored to the human-labeled subset.

6 Conclusion

In this work, we address the scalability bottleneck of preference optimization by proposing DeDPO. Motivated by the prohibitive cost of human supervision, DeDPO integrates a debiased estimation method from causal inference to robustly leverage large volumes of inexpensive yet misaligned synthetic preference labels. Empirical results across multiple diffusion models and synthetic annotators show that DeDPO improves over naive synthetic training, with the largest gains when synthetic labels are more misaligned with human preferences, and achieves performance that remains competitive with fully human-supervised DPO while using substantially fewer human labels. Overall, these results suggest that properly debiased synthetic feedback can serve as a practical and cost-effective complement to scarce human supervision for scalable human–AI alignment.

Acknowledgments and Disclosure of Funding

The authors received no third-party funding or support for this work. The authors declare no competing financial interests.

References

[1] Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Remi Munos, Mark Rowland, Michal Valko, and Daniele Calandriello. A general theoretical paradigm to understand learning from human preferences. In *International Conference on Artificial Intelligence and Statistics*, pages 4447–4455. PMLR, 2024.

[2] Heejung Bang and James M. Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005. doi: 10.1111/j.1541-0420.2005.00377.x.

[3] Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. *arXiv preprint arXiv:2305.13301*, 2023.

[4] Alexander Bukharin, Ilgee Hong, Haoming Jiang, Zichong Li, Qingru Zhang, Zixuan Zhang, and Tuo Zhao. Robust reinforcement learning from corrupted human feedback. *Advances in Neural Information Processing Systems*, 37:124093–124113, 2024.

[5] Souradip Chakraborty, Jiahao Qiu, Hui Yuan, Alec Koppel, Furong Huang, Dinesh Manocha, Amrit Singh Bedi, and Mengdi Wang. Maxmin-rlhf: Alignment with diverse human preferences. *arXiv preprint arXiv:2402.08925*, 2024.

[6] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters, 2018.

[7] Sayak Ray Chowdhury, Anush Kini, and Nagarajan Natarajan. Provably robust dpo: Aligning language models with noisy feedback. *arXiv preprint arXiv:2403.00409*, 2024.

[8] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

[9] 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. *Advances in Neural Information Processing Systems*, 36:79858–79885, 2023.

[10] Dylan J Foster and Vasilis Syrgkanis. Orthogonal statistical learning. *The Annals of Statistics*, 51(3):879–908, 2023.

[11] Jiayi Fu, Xuandong Zhao, Chengyuan Yao, Heng Wang, Qi Han, and Yanghua Xiao. Reward shaping to mitigate reward hacking in rlhf. *arXiv preprint arXiv:2502.18770*, 2025.

[12] Shangmin Guo, Biao Zhang, Tianlin Liu, Tianqi Liu, Misha Khalman, Felipe Llinares, Alexandre Rame, Thomas Mesnard, Yao Zhao, Bilal Piot, et al. Direct language model alignment from online ai feedback. *arXiv preprint arXiv:2402.04792*, 2024.

[13] Cheng-Ju Ho, Chen-Hsuan Tai, Yen-Yu Lin, Ming-Hsuan Yang, and Yi-Hsuan Tsai. Diffusion-ss3d: Diffusion model for semi-supervised 3d object detection. *Advances in Neural Information Processing Systems*, 36:49100–49112, 2023.

[14] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

[15] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[16] Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

[17] Shyamgopal Karthik, Huseyin Coskun, Zeynep Akata, Sergey Tulyakov, Jian Ren, and Anil Kag. Scalable ranked preference optimization for text-to-image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18399–18410, 2025.

[18] Edward H Kennedy. Semiparametric doubly robust targeted double machine learning: a review. *Handbook of statistical methods for precision medicine*, pages 207–236, 2024.

[19] Cheol Woo Kim, Shresth Verma, Mauricio Tec, and Milind Tambe. Lightweight robust direct preference optimization. *arXiv preprint arXiv:2510.23590*, 2025.

[20] 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*, 2023.

[21] Aishik Konwer, Zhijian Yang, Erhan Bas, Cao Xiao, Prateek Prasanna, Parminder Bhatia, and Taha Kass-Hout. Enhancing sam with efficient prompting and preference optimization for semi-supervised medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20990–21000, 2025.

[22] Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and Sushant Prakash. Rlaif vs. rlhf: Scaling reinforcement learning from human feedback with ai feedback. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.

[23] Seonggyun Lee, Sungjun Lim, Seojin Park, Soeun Cheon, and Kyungwoo Song. Semi-supervised preference optimization with limited feedback. *arXiv preprint arXiv:2511.00040*, 2025.

[24] Ruining Li, Chuanxia Zheng, Christian Rupprecht, and Andrea Vedaldi. Dso: Aligning 3d generators with simulation feedback for physical soundness. In *ICCV*, 2025.

[25] Shufan Li, Konstantinos Kallidromitis, Akash Gokul, Yusuke Kato, and Kazuki Kozuka. Aligning diffusion models by optimizing human utility. *Advances in Neural Information Processing Systems*, 37:24897–24925, 2024.

[26] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[27] Jie Liu, Gongye Liu, Jiajun Liang, Ziyang Yuan, Xiaokun Liu, Mingwu Zheng, Xiele Wu, Qiulin Wang, Menghan Xia, Xintao Wang, et al. Improving video generation with human feedback. *arXiv preprint arXiv:2501.13918*, 2025.

[28] Xinxin Liu, Ming Li, Zonglin Lyu, Yuzhang Shang, and Chen Chen. Learning from noisy preferences: A semi-supervised learning approach to direct preference optimization. In *The Fourteenth International Conference on Learning Representations*, 2026.

[29] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[30] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. *Machine Intelligence Research*, pages 1–22, 2025.

[31] Yuhang Ma, Xiaoshi Wu, Keqiang Sun, and Hongsheng Li. Hpsv3: Towards wide-spectrum human preference score. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15086–15095, 2025.

[32] Debmalya Mandal, Paulius Sasnauskas, and Goran Radanovic. Distributionally robust reinforcement learning with human feedback. *arXiv preprint arXiv:2503.00539*, 2025.

[33] Andrés Marafioti, Orr Zohar, Miquel Farré, Merve Noyan, Elie Bakouch, Pedro Cuenca, Cyril Zakka, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, et al. Smolvlm: Redefining small and efficient multimodal models. *arXiv preprint arXiv:2504.05299*, 2025.

[34] Eric Mitchell. A note on dpo with noisy preferences & relationship to ipo. `https://ericmitchell.ai/cdpo.pdf`, November 2023. Version 1.1.

[35] Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*, pages 1928–1937. PmLR, 2016.

[36] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[37] Vishakh Padmakumar, Chuanyang Jin, Hannah Rose Kirk, and He He. Beyond the binary: Capturing diverse preferences with reward regularization. *arXiv preprint arXiv:2412.03822*, 2024.

[38] 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. *Advances in neural information processing systems*, 36:53728–53741, 2023.

[39] James M. Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866, 1994. doi: 10.1080/01621459.1994.10476818.

[40] 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.

[41] Daniel O. Scharfstein, Andrea Rotnitzky, and James M. Robins. Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, 94(448):1096–1120, 1999. doi: 10.1080/01621459.1999.10473862.

[42] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.

[43] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[44] Noam Shazeer and Mitchell Stern. Adafactor: Adaptive learning rates with sublinear memory cost. In *International Conference on Machine Learning*, pages 4596–4604. PMLR, 2018.

[45] Yutao Shen, Junkun Yuan, Toru Aonishi, Hideki Nakayama, and Yue Ma. Follow-your-preference: Towards preference-aligned image inpainting. *arXiv preprint arXiv:2509.23082*, 2025.

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

[47] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

[48] Qwen Team. Qwen2.5-vl, January 2025. URL `https://qwenlm.github.io/blog/qwen2.5-vl/`.

[49] 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.

[50] Zijie J Wang, Evan Montoya, David Munechika, Haoyang Yang, Benjamin Hoover, and Duen Horng Chau. Diffusiondb: A large-scale prompt gallery dataset for text-to-image generative models. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 893–911, 2023.

[51] Junkang Wu, Yuexiang Xie, Zhengyi Yang, Jiancan Wu, Jiawei Chen, Jinyang Gao, Bolin Ding, Xiang Wang, and Xiangnan He. Towards robust alignment of language models: Distributionally robustifying direct preference optimization. *arXiv preprint arXiv:2407.07880*, 2024.

[52] Xiaoshi Wu, Keqiang Sun, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score: Better aligning text-to-image models with human preference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2096–2105, 2023.

[53] Ziyi Wu, Anil Kag, Ivan Skorokhodov, Willi Menapace, Ashkan Mirzaei, Igor Gilitschenski, Sergey Tulyakov, and Aliaksandr Siarohin. Densedpo: Fine-grained temporal preference optimization for video diffusion models. *arXiv preprint arXiv:2506.03517*, 2025.

[54] Erhan Xu, Kai Ye, Hongyi Zhou, Luhan Zhu, Francesco Quinzan, and Chengchun Shi. Doubly robust alignment for large language models. *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. to appear, also available as arXiv:2506.01183.

[55] 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*, 36:15903–15935, 2023.

[56] Zaiyan Xu, Sushil Vemuri, Kishan Panaganti, Dileep Kalathil, Rahul Jain, and Deepak Ramachandran. Robust llm alignment via distributionally robust direct preference optimization. *arXiv preprint arXiv:2502.01930*, 2025.

[57] Yuzi Yan, Xingzhou Lou, Jialian Li, Yiping Zhang, Jian Xie, Chao Yu, Yu Wang, Dong Yan, and Yuan Shen. Reward-robust rlhf in llms. *arXiv preprint arXiv:2409.15360*, 2024.

[58] Yongjin Yang, Sihyeon Kim, Hojung Jung, Sangmin Bae, SangMook Kim, Se-Young Yun, and Kimin Lee. Automated filtering of human feedback data for aligning text-to-image diffusion models. *arXiv preprint arXiv:2410.10166*, 2024.

[59] Kai Ye, Hongyi Zhou, Jin Zhu, Francesco Quinzan, and Chengchun Shi. Robust reinforcement learning from human feedback for large language models fine-tuning. *arXiv preprint arXiv:2504.03784*, 2025.

[60] 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. *arXiv preprint arXiv:2206.10789*, 2(3):5, 2022.

[61] Banghua Zhu, Mingyu Ding, Philip Jacobson, Ming Wu, Wei Zhan, Michael Jordan, and Jiantao Jiao. Doubly-robust self-training. *Advances in Neural Information Processing Systems*, 36: 41413–41431, 2023.

Supplementary Material

A Theoretical Proofs

Proof of proposition 1.

Proof. Simply reorder the terms in Eq. (8) as follows:

$$\begin{aligned} L_{\text{DeDPO}}(\theta) = \; & \mathbb{E}_{n_l} \mathcal{L}(\mathcal{G}_\theta(\boldsymbol{y}_l), \mathbf{z}_l) \\ & + \left[\mathbb{E}_{n_l+n_u} \mathcal{L}(\mathcal{G}_\theta(\boldsymbol{y}), \hat{\mathcal{G}}(\boldsymbol{y})) - \mathbb{E}_{n_l} \mathcal{L}(\mathcal{G}_\theta(\boldsymbol{y}_l), \hat{\mathcal{G}}(\boldsymbol{y}_l)) \right] \end{aligned} \tag{12}$$

Then, the expectation of the first term is the population-level DPO loss Eq. (6), and the expectation of the second term is 0 because both are unbiased estimators of $\mathbb{E}[\mathcal{L}(\mathcal{G}_\theta(\boldsymbol{y}), \hat{\mathcal{G}}(\boldsymbol{y}))]$. □

This shows that we can safely use synthetic preference labels while preserving an unbiased loss, leading to bias-free learning. Moreover, when the synthetic labels are perfect, i.e., $\hat{\mathcal{G}}(\boldsymbol{y}) = \mathbf{z}$, the debiased loss becomes $\mathbb{E}_{n_l+n_u} \mathcal{L}(\mathcal{G}_\theta(\boldsymbol{y}), \hat{\mathcal{G}}(\boldsymbol{y}))$, thereby using all of the data. When the synthetic labels are completely wrong, the loss asymptotically reduces to $\mathbb{E}_{n_l} \mathcal{L}(\mathcal{G}_\theta(\boldsymbol{y}_l), \mathbf{z}_l)$, i.e., only the labeled data is used.

Proof of proposition 2.

Proof. First, we rewrite the second term of Eq. (8):

$$\begin{aligned} & L_{\text{DeDPO}}(\theta) = \mathbb{E}_{n_l+n_u} \mathcal{L}(\mathcal{G}_\theta(\boldsymbol{y}), \hat{\mathcal{G}}(\boldsymbol{y})) \\ & + \mathbb{E}_{n_l} (\mathcal{L}(\mathcal{G}_\theta(\boldsymbol{y}_l), \boldsymbol{z}_l) - \mathcal{L}(\mathcal{G}_\theta(\boldsymbol{y}_l), \hat{\mathcal{G}}(\boldsymbol{y}_l))) \\ & = \mathbb{E}_{n_l+n_u} \mathcal{L}(\mathcal{G}_\theta(\boldsymbol{y}), \hat{\mathcal{G}}(\boldsymbol{y})) \\ & + \mathbb{E}_{n_l+n_u} \frac{n_l + n_u}{n_l} \boldsymbol{r} \left(\mathcal{L}(\mathcal{G}_\theta(\boldsymbol{y}), \boldsymbol{z}) - \mathcal{L}(\mathcal{G}_\theta(\boldsymbol{y}), \hat{\mathcal{G}}(\boldsymbol{y})) \right) \end{aligned} \tag{13}$$

where $\boldsymbol{r}$ is a binary variable equal to 1 if $\boldsymbol{y}$ is labeled and 0 otherwise. Second, we note that the binary cross-entropy loss is linear in its second argument, i.e., the target label. Therefore, since all terms now share the same prediction $\mathcal{G}_\theta(\boldsymbol{y})$, we can add, subtract, and rescale the targets with weight $\frac{n_l+n_u}{n_l} \boldsymbol{r}$ to obtain:

$$\begin{aligned} & L_{\text{DeDPO}}(\theta) = \\ & \mathbb{E}_{n_l+n_u} \mathcal{L} \left(\mathcal{G}_\theta(\boldsymbol{y}), \hat{\mathcal{G}}(\boldsymbol{y}) + \frac{n_l + n_u}{n_l} \boldsymbol{r} (\boldsymbol{z} - \hat{\mathcal{G}}(\boldsymbol{y})) \right) \\ & = \mathbb{E}_{n_l+n_u} \mathcal{L} \left(\mathcal{G}_\theta(\boldsymbol{y}), \hat{\boldsymbol{w}} \right) \end{aligned} \tag{14}$$

for $\hat{\boldsymbol{w}}$ defined in the original claim. □

A.1 Finite-sample convergence of DeDPO

Notation. Let $f_\theta(\boldsymbol{y})$ be the implicit preference logit of the diffusion model with parameters θ. Recall that $\boldsymbol{y} = (\boldsymbol{c}, \boldsymbol{x}^0, \boldsymbol{x}^1)$ is the input and $\boldsymbol{z}$ is the preference label. With $\mathcal{G}_\theta$ denoting the implicit preference classifier, we write

$$f_\theta(\boldsymbol{y}) = \operatorname{logit} \mathcal{G}_\theta(\boldsymbol{y}) = \log \frac{\mathcal{G}_\theta(\boldsymbol{y})}{1 - \mathcal{G}_\theta(\boldsymbol{y})}.$$

In terms of the denoising model, the same logit is

$$\begin{aligned} f_\theta(\boldsymbol{y}) = -\beta \Big[& \|\boldsymbol{\epsilon}^1 - \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t^1, t, \boldsymbol{c})\|_2^2 - \|\boldsymbol{\epsilon}^1 - \boldsymbol{\epsilon}_{\text{ref}}(\boldsymbol{x}_t^1, t, \boldsymbol{c})\|_2^2 \\ & - \|\boldsymbol{\epsilon}^0 - \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t^0, t, \boldsymbol{c})\|_2^2 + \|\boldsymbol{\epsilon}^0 - \boldsymbol{\epsilon}_{\text{ref}}(\boldsymbol{x}_t^0, t, \boldsymbol{c})\|_2^2 \Big]. \end{aligned}$$

The timestep and noise variables are implicit in $\boldsymbol{y}$ and in the expectations. Let $\mathcal{F} = \{f_\theta : \theta \in \Theta\}$ be the class of such logit functions, and write $\hat{f} = f_{\hat{\theta}}$ for the fitted logit function.

Write $\hat{g} = \hat{\mathcal{G}}$ for the pretrained synthetic preference model and define

$$g^*(\boldsymbol{y}) := \mathcal{G}^*(\boldsymbol{y}) = \mathbb{E}[\boldsymbol{z} \mid \boldsymbol{y}].$$

Write $R(f) = \mathbb{E}[\ell(f(\boldsymbol{y}), \boldsymbol{z})]$ for the population risk and let $f^* \in \arg\min_{f \in \mathcal{F}} R(f)$.

We use the indicator $\boldsymbol{r} \in \{0, 1\}$, where $\boldsymbol{r} = 1$ means the human label is observed. Define the true inverse labeling probability and its estimated version by

$$\gamma^*(\boldsymbol{y}) := \frac{1}{\Pr(\boldsymbol{r} = 1 \mid \boldsymbol{y})}, \qquad \hat{\gamma}(\boldsymbol{y}) := \frac{1}{\widehat{\Pr}(\boldsymbol{r} = 1 \mid \boldsymbol{y})}.$$

For constant labeling probability p, the true inverse weight reduces to $\gamma^*(\boldsymbol{y}) = 1/p$.

Using the augmented target from proposition 2, write $w(g, \gamma) = g(\boldsymbol{y}) + \boldsymbol{r}\gamma(\boldsymbol{y})(\boldsymbol{z} - g(\boldsymbol{y}))$ and $L(f, g, \gamma) = \mathbb{E}[\ell(f(\boldsymbol{y}), w(g, \gamma))]$ for the DeDPO population loss. Here the inverse weight $\gamma(\boldsymbol{y})$ may depend on $\boldsymbol{y}$. Recall the logistic loss expressed in terms of a logit a and target b,

$$\ell(a, b) = -b \log \sigma(a) - (1 - b) \log(1 - \sigma(a)) = \log(1 + e^a) - ba. \tag{15}$$

The directional derivative of the population loss at f in the direction h is

$$D_f L(f, g, \gamma)[h] := \frac{d}{dt} L(f + th, g, \gamma)\Big|_{t=0}. \tag{16}$$

Empirical loss and population optima. Write $\mathbb{E}_n[U] = n^{-1} \sum_{i=1}^n U(\boldsymbol{y}_i, \boldsymbol{z}_i, \boldsymbol{r}_i)$ for the empirical average, with $n = n_l + n_u$.

$$\hat{w} = w(\hat{g}, \hat{\gamma}), \qquad \widehat{L}(f) = \mathbb{E}_n[\ell(f(\boldsymbol{y}), \hat{w})].$$

Let

$$\hat{f} \in \arg\min_{f \in \mathcal{F}} \widehat{L}(f), \qquad \bar{f} \in \arg\min_{f \in \mathcal{F}} L(f, \hat{g}, \hat{\gamma}).$$

Assumption 1 (First-order optimality)**.** *The population minimizers $\bar{f}$ and f^* satisfy*

$$\begin{aligned} D_f L(\bar{f}, \hat{g}, \hat{\gamma})[f - \bar{f}] \geq 0, \\ D_f L(f^*, g^*, \gamma^*)[f - f^*] \geq 0 \end{aligned} \qquad \textit{for all } f \in \mathcal{F}.$$

This condition holds, for example, when $\mathcal{F}$ is convex in logit space. For a nonconvex class it is an additional assumption on the population optima.

Theorem 2 (Finite-sample convergence with nuisance estimates)**.** *Suppose $(\boldsymbol{y}_i, \boldsymbol{z}_i, \boldsymbol{r}_i)_{i=1}^n$ are i.i.d., $\boldsymbol{z}_i, \boldsymbol{r}_i \in \{0, 1\}$, and $\boldsymbol{r} \perp \boldsymbol{z} \mid \boldsymbol{y}$. Assume that, for some fixed $\Gamma \geq 1$,*

$$\Gamma^{-1} \leq \Pr(\boldsymbol{r} = 1 \mid \boldsymbol{y}) \leq 1, \qquad 0 \leq \hat{g}(\boldsymbol{y}) \leq 1, \qquad 1 \leq \hat{\gamma}(\boldsymbol{y}) \leq \Gamma$$

almost surely, and that $(\hat{g}, \hat{\gamma})$ are fixed or estimated independently of these n observations. Let $\mathcal{F}$ be a fixed class of logit functions taking values in $[-M, M]$, for $M > 0$, with VC-subgraph dimension at most d. Suppose the minimizers exist, the relevant empirical suprema and minimizers are measurable, and assumption 1 holds. Put

$$\lambda = \sigma(M)(1 - \sigma(M)), \qquad K = (d + 1) \log(en) + \log(2/\delta), \qquad 0 < \delta < 1.$$

Then, conditionally on the fitted nuisance functions, with probability at least $1 - \delta$,

$$\boxed{\|\hat{f} - f^*\|_2^2 \leq C_{M,\lambda,\Gamma} \frac{K}{n} + \frac{2}{\lambda^2} \|\hat{g} - g^*\|_4^2 \|\hat{\gamma} - \gamma^*\|_4^2.} \tag{17}$$

The constant is independent of n, d, δ.

We first establish the curvature property used in both parts of the proof. The empirical-process step uses the concentration results of orthogonal statistical learning [10]. This logit-error guarantee complements the analysis of the loss gradient at the population optimum in [61].

Lemma 1 (Strong convexity)**.** *Fix g, γ, and assume that the augmented target $w = w(g, \gamma)$ satisfies $\mathbb{E}|w| < \infty$. For any two logit functions f, f' taking values in $[-M, M]$,*

$$D_f L(f, g, \gamma)[f - f'] - D_f L(f', g, \gamma)[f - f'] \geq \lambda \|f - f'\|_2^2, \tag{18}$$

where $\lambda = \sigma(M)(1 - \sigma(M)) > 0$.

Proof. With the augmented target w held fixed, the first and second derivatives of the pointwise loss are

$$\ell'(a, w) = \sigma(a) - w \tag{19}$$

and

$$\ell''(a, w) = \sigma'(a) = \sigma(a)(1 - \sigma(a)). \tag{20}$$

The second derivative is independent of w. On $[-M, M]$, it is bounded below by λ.

Let $h = f - f'$. The chain rule gives

$$\frac{d}{dt}\ell(f(\boldsymbol{y}) + th(\boldsymbol{y}), w) = \ell'(f(\boldsymbol{y}) + th(\boldsymbol{y}), w)\, h(\boldsymbol{y}).$$

Since h is bounded and $\mathbb{E}|w| < \infty$, we can differentiate under the expectation. Evaluating at $t = 0$ yields

$$D_f L(f, g, \gamma)[h] = \mathbb{E}[(\sigma(f(\boldsymbol{y})) - w)h(\boldsymbol{y})].$$

Taking the difference of the derivatives at f and f', in the same direction h, cancels the augmented target,

$$D_f L(f, g, \gamma)[h] - D_f L(f', g, \gamma)[h] = \mathbb{E}[(\sigma(f(\boldsymbol{y})) - \sigma(f'(\boldsymbol{y})))h(\boldsymbol{y})].$$

By the mean value theorem, for each $\boldsymbol{y}$ there is a value $\xi_{\boldsymbol{y}}$ between $f(\boldsymbol{y})$ and $f'(\boldsymbol{y})$ such that

$$\sigma(f(\boldsymbol{y})) - \sigma(f'(\boldsymbol{y})) = \sigma'(\xi_{\boldsymbol{y}})(f(\boldsymbol{y}) - f'(\boldsymbol{y})).$$

Since $\xi_{\boldsymbol{y}} \in [-M, M]$, we have $\sigma'(\xi_{\boldsymbol{y}}) \geq \lambda$, so

$$(\sigma(f(\boldsymbol{y})) - \sigma(f'(\boldsymbol{y})))h(\boldsymbol{y}) = \sigma'(\xi_{\boldsymbol{y}})h(\boldsymbol{y})^2 \geq \lambda h(\boldsymbol{y})^2.$$

Taking expectations gives

$$\begin{aligned} &D_f L(f, g, \gamma)[h] - D_f L(f', g, \gamma)[h] \\ &\quad \geq \lambda \mathbb{E}[h(\boldsymbol{y})^2] = \lambda \|f - f'\|_2^2, \end{aligned} \tag{21}$$

which proves the lemma. □

Proof of theorem 2. Condition on $(\hat{g}, \hat{\gamma})$ throughout the proof. Independence allows us to treat these functions as fixed while the observations remain i.i.d.

Population error. Let $h = \bar{f} - f^*$. Applying lemma 1 with the same nuisance functions at both logit functions gives

$$D_f L(\bar{f}, \hat{g}, \hat{\gamma})[h] - D_f L(f^*, \hat{g}, \hat{\gamma})[h] \geq \lambda \|\bar{f} - f^*\|_2^2. \tag{22}$$

Taking $f = f^*$ in assumption 1 and using $h = \bar{f} - f^*$ gives

$$D_f L(\bar{f}, \hat{g}, \hat{\gamma})[h] \leq 0.$$

Substituting this inequality into Eq. (22) yields

$$\begin{aligned} \lambda \|\bar{f} - f^*\|_2^2 &\leq -D_f L(f^*, \hat{g}, \hat{\gamma})[h] \\ &= -\mathbb{E}\big[\big(\sigma(f^*(\boldsymbol{y})) - w(\hat{g}, \hat{\gamma})\big)h(\boldsymbol{y})\big]. \end{aligned} \tag{23}$$

Cancellation of the first-order nuisance errors. Write $\hat{w} = w(\hat{g}, \hat{\gamma})$ and $w^* = w(g^*, \gamma^*)$. With the true nuisance functions, $\mathbb{E}[w^* \mid \boldsymbol{y}] = g^*(\boldsymbol{y})$, so $L(f, g^*, \gamma^*) = R(f)$. Taking $f = \bar{f}$ in the second inequality of assumption 1 gives

$$D_f L(f^*, g^*, \gamma^*)[h] = \mathbb{E}[(\sigma(f^*(\boldsymbol{y})) - w^*)h(\boldsymbol{y})] \geq 0.$$

Adding this nonnegative term to the right-hand side of Eq. (23) cancels the sigmoid terms,

$$\begin{aligned} \lambda\|\bar{f} - f^*\|_2^2 &\leq -D_f L(f^*, \hat{g}, \hat{\gamma})[h] + D_f L(f^*, g^*, \gamma^*)[h] \\ &= \mathbb{E}[(\hat{w} - w^*)h(\boldsymbol{y})]. \end{aligned} \tag{24}$$

Define the nuisance errors $e_g = \hat{g} - g^*$ and $e_\gamma = \hat{\gamma} - \gamma^*$. Expanding the target difference gives

$$\hat{w} - w^* = \boldsymbol{r}e_\gamma(\boldsymbol{z} - g^*) + e_g(1 - \boldsymbol{r}\gamma^*) - \boldsymbol{r}e_\gamma e_g, \tag{25}$$

where all functions on the right are evaluated at $\boldsymbol{y}$. The definitions of g^*, γ^* and the assumption $\boldsymbol{r} \perp \boldsymbol{z} \mid \boldsymbol{y}$ give

$$\mathbb{E}[\boldsymbol{r}(\boldsymbol{z} - g^*) \mid \boldsymbol{y}] = 0, \qquad \mathbb{E}[1 - \boldsymbol{r}\gamma^* \mid \boldsymbol{y}] = 0.$$

Since e_g, e_γ, and h depend only on $\boldsymbol{y}$, the first two terms of Eq. (25) vanish after multiplication by $h(\boldsymbol{y})$ and averaging. Hence

$$\mathbb{E}[(\hat{w} - w^*)h(\boldsymbol{y})] = -\mathbb{E}[\boldsymbol{r}e_\gamma(\boldsymbol{y})e_g(\boldsymbol{y})h(\boldsymbol{y})],$$

and Eq. (24) implies

$$\lambda\|\bar{f} - f^*\|_2^2 \leq \left|\mathbb{E}[\boldsymbol{r}e_\gamma(\boldsymbol{y})e_g(\boldsymbol{y})h(\boldsymbol{y})]\right|. \tag{26}$$

Thus only the product of the two nuisance errors remains.

Finally, since $0 \leq \boldsymbol{r} \leq 1$, Hölder's inequality with exponents $4, 4, 2$ gives

$$\begin{aligned} \lambda\|\bar{f} - f^*\|_2^2 &\leq \mathbb{E}[|e_\gamma(\boldsymbol{y})e_g(\boldsymbol{y})h(\boldsymbol{y})|] \\ &\leq \|\hat{g} - g^*\|_4\|\hat{\gamma} - \gamma^*\|_4\|\bar{f} - f^*\|_2. \end{aligned} \tag{27}$$

Dividing when $\|\bar{f} - f^*\|_2 > 0$ and squaring gives the following bound, which is immediate when the norm is zero:

$$\|\bar{f} - f^*\|_2^2 \leq \frac{1}{\lambda^2}\|\hat{g} - g^*\|_4^2\|\hat{\gamma} - \gamma^*\|_4^2. \tag{28}$$

Empirical error. Abbreviate $L(f) = L(f, \hat{g}, \hat{\gamma})$. For each $f \in \mathcal{F}$ define the pointwise loss difference

$$U_f(\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{r}) = \ell(f(\boldsymbol{y}), \hat{w}) - \ell(\bar{f}(\boldsymbol{y}), \hat{w}), \qquad \hat{w} = w(\hat{g}, \hat{\gamma}).$$

Thus $\mathbb{E}[U_f] = L(f) - L(\bar{f})$ and $\mathbb{E}_n[U_f] = \widehat{L}(f) - \widehat{L}(\bar{f})$. Empirical optimality gives $\mathbb{E}_n[U_{\hat{f}}] \leq 0$, so

$$L(\hat{f}) - L(\bar{f}) \leq (\mathbb{E} - \mathbb{E}_n)U_{\hat{f}}. \tag{29}$$

Let $A = \Gamma$. The bounds on $\hat{g}$ and $\hat{\gamma}$ imply $1 - A \leq \hat{w} \leq A$. Hence

$$|\ell'(a, \hat{w})| = |\sigma(a) - \hat{w}| \leq A, \qquad a \in [-M, M].$$

The mean value theorem and the definition of variance give

$$\begin{aligned} |U_f| &\leq A|f(\boldsymbol{y}) - \bar{f}(\boldsymbol{y})| \leq 2MA, \\ \mathrm{Var}(U_f) &\leq \mathbb{E}[U_f^2] \leq A^2\|f - \bar{f}\|_2^2. \end{aligned} \tag{30}$$

We apply the uniform concentration bound in [10, Appendix K, Lemma 14]. We show how the VC-subgraph assumption supplies its complexity condition. Put $B = \max\{1, M\}$ and normalize the logit class by B. For every probability measure Q, the VC-subgraph covering bound gives

$$\log N(\varepsilon, \mathcal{F}/B, L_2(Q)) \leq C(d + 1)\log(C/\varepsilon), \qquad 0 < \varepsilon < 1,$$

where N is the minimum number of $L_2(Q)$ balls of radius ε covering the class. Translation by $\bar{f}/B$ preserves covering numbers. Passing to

$$\mathcal{H} = \{t(f - \bar{f})/B : f \in \mathcal{F},\ 0 \leq t \leq 1\}$$

adds only a term of order $\log(C/\varepsilon)$ to their logarithms. An additional normalization by 2 bounds $\mathcal{H}$ by 1 and changes only universal constants in the cited entropy calculations. In particular, the entropy integral for a neighborhood of radius $0 < s < 1$ is bounded by

$$\int_{s^2/8}^{s} \sqrt{\frac{C(d+1)\log(C/\varepsilon)}{n}}\, d\varepsilon \le Cs\sqrt{\frac{(d+1)\log(C/s)}{n}}.$$

Consequently, $\tau = C_0\sqrt{K/n}$, with C_0 sufficiently large, satisfies the critical-radius and confidence requirements in the cited lemma. This is the VC entropy calculation of [10, Appendix H.1.1, Lemmas 8–10], applied to the normalized class. When this choice is larger than a fixed constant, the desired theorem bound follows directly from $\|\hat{f} - f^*\|_2^2 \le 4M^2$, after enlarging its constant.

On the normalized class, the loss $(a, \hat{w}) \mapsto \ell(Ba, \hat{w})$ is AB-Lipschitz. Lemma 14 therefore yields, with probability at least $1 - \delta$, simultaneously for all $f \in \mathcal{F}$,

$$\begin{aligned} |(\mathbb{E}_n - \mathbb{E})U_f| &\le CAB\tau\left(\frac{\|f - \bar{f}\|_2}{B} + \tau\right) \\ &\le CA\left[\|f - \bar{f}\|_2\sqrt{\frac{K}{n}} + B\frac{K}{n}\right]. \end{aligned} \tag{31}$$

Conditioning makes $\bar{f}$ a fixed comparator for this application. The uniform event permits substituting the data-dependent $\hat{f}$.

For completeness, the curvature lower bound also gives a loss-gap bound. For $h = f - \bar{f}$, Taylor's integral formula gives

$$\begin{aligned} L(f) - L(\bar{f}) &= D_f L(\bar{f})[h] + \int_0^1 (1-t)\mathbb{E}[\sigma'(\bar{f}(\boldsymbol{y}) + th(\boldsymbol{y}))h(\boldsymbol{y})^2]\, dt \\ &\ge \frac{\lambda}{2}\|f - \bar{f}\|_2^2. \end{aligned}$$

The first-order term is nonnegative by assumption 1; the line segment between bounded logits stays in $[-M, M]$. Evaluating the loss along that segment does not require the intermediate functions to belong to $\mathcal{F}$. Combine this inequality with Eqs. (29) and (31) and set $s = \|\hat{f} - \bar{f}\|_2$ to obtain

$$\frac{\lambda}{2}s^2 \le CAs\sqrt{\frac{K}{n}} + CAB\frac{K}{n}.$$

Young's inequality gives

$$CAs\sqrt{\frac{K}{n}} \le \frac{\lambda}{4}s^2 + \frac{C^2A^2}{\lambda}\frac{K}{n}.$$

Subtracting the quadratic term and dividing by $\lambda/4$ yields

$$\|\hat{f} - \bar{f}\|_2^2 \le 4\left(\frac{C^2A^2}{\lambda^2} + \frac{CAB}{\lambda}\right)\frac{K}{n} = C_{M,\lambda,\Gamma}\frac{K}{n}. \tag{32}$$

Combining the two errors. Finally,

$$\begin{aligned} \|\hat{f} - f^*\|_2^2 &\le 2\|\hat{f} - \bar{f}\|_2^2 + 2\|\bar{f} - f^*\|_2^2 \\ &\le C_{M,\lambda,\Gamma}\frac{K}{n} + \frac{2}{\lambda^2}\|\hat{g} - g^*\|_4^2\|\hat{\gamma} - \gamma^*\|_4^2, \end{aligned}$$

where the statistical constant has been enlarged to absorb the factor 2. This proves the conditional result; averaging the conditional failure probability gives the same guarantee unconditionally. □

Proof of corollary 1.

Proof. When $\hat{\gamma} = \gamma^*$, Eq. (17) gives $\|\hat{f} - f^*\|_2^2 \le C_{M,\lambda,\Gamma}K/n$ with probability at least $1 - \delta$. For estimated propensity, let $\mathcal{E}$ be an event determined by the fitted nuisances such that $\Pr(\mathcal{E}) \ge 1 - \eta$ and, on $\mathcal{E}$,

$$\|\hat{\gamma} - \gamma^*\|_4\|\hat{g} - g^*\|_4 \le bn^{-1/2}.$$

Conditional on any such fitted nuisances, the theorem gives

$$\|\hat{f} - f^*\|_2^2 \le C_{M,\lambda,\Gamma} \frac{K}{n} + \frac{2b^2}{\lambda^2 n}$$

with probability at least $1 - \delta$. A union bound with $\mathcal{E}^c$ gives joint probability at least $1 - \delta - \eta$. □

Fourth-order nuisance dependence. If, for a constant C_0,

$$\|\hat{\gamma} - \gamma^*\|_4 \le C_0 \|\hat{g} - g^*\|_4,$$

then the product term satisfies

$$\|\hat{\gamma} - \gamma^*\|_4^2 \|\hat{g} - g^*\|_4^2 \le C_0^2 \|\hat{g} - g^*\|_4^4.$$

This condition is needed only for the fourth-order specialization; the theorem itself imposes no convergence rate on either nuisance.

The theorem is a finite-sample guarantee for the empirical logit minimizer. With a fixed function-class dimension, its statistical term is of order $1/n$ up to logarithmic factors.

B Experiment Setup

B.1 Implementation of DeDPO

To make the method concrete, Fig. 5 shows a minimal implementation of DeDPO. In practice, DeDPO uses exactly the same logits and binary cross-entropy components as standard Diffusion-DPO, and differs only by an additional correction term applied on the labeled subset. Consequently, moving from the original DPO objective to our debiased objective requires only a small code change and introduces essentially no additional algorithmic complexity. We include this implementation sketch to complement the theory and to emphasize that the practical gains of DeDPO do not depend on a complicated training pipeline.

Balanced mini-batches and loss weighting. Let B_L and B_U denote the numbers of labeled and unlabeled pairs in a mini-batch, with total size $B = B_L + B_U$, and let $q_L = B_L/B$ be its labeled sampling fraction. This fraction is distinct from the dataset fraction $p = n_l/(n_l + n_u)$. The correction weight is $w_t = 1/q_L$. Writing ℓ_L, $\ell_{\hat{L}}$, and $\ell_{\hat{U}}$ for the mean human-label loss on labeled pairs and the mean pseudo-label losses on labeled and unlabeled pairs, respectively, the mini-batch loss is

$$\widehat{L}_{\text{DeDPO}} = q_L \ell_{\hat{L}} + (1 - q_L)\ell_{\hat{U}} + q_L w_t (\ell_L - \ell_{\hat{L}}).$$

The code computes the component losses with `reduction="sum"` and divides the combined DeDPO loss by B, yielding this same expression. In all experiments, we use equal numbers of labeled and unlabeled pairs in each mini-batch, regardless of the pool sizes (n_l, n_u). We obtain this balance by undersampling the larger pool or oversampling the smaller pool, giving $B_L = B_U$, $q_L = 1/2$, and $w_t = 2$. The loss then reduces to $\frac{1}{2}[\ell_{\hat{L}} + \ell_{\hat{U}} + 2(\ell_L - \ell_{\hat{L}})]$. Balanced batches ensure that the human correction is represented in every update. Each reuse of a pair independently resamples the diffusion timestep and Gaussian noise. Under the random split used in our experiments, labeled and unlabeled pairs have the same population distribution, so this rebalancing preserves the population target of Eq. (8).

B.2 Datasets

Pick-a-Pic v2. Pick-a-Pic [20] is collected from interactions in a public text-to-image web application: a user writes a prompt, is shown two generated images, and selects a preferred image or a tie; the rejected image is then replaced and the process continues. The original paper reports more than 500k examples and 35k distinct prompts in the release used there, and also notes that the newer v2 release contains more than one million examples. The images are generated from multiple backbones and classifier-free guidance settings, and the data collection process is centered on real, intrinsically motivated users rather than paid crowd workers. This provenance is both the main strength and the main caveat of PaP. On the positive side, it offers strong ecological validity, broad prompt diversity, and direct human pairwise supervision. On the other hand, it naturally reflects the preferences of a

```
# model_diff    : policy MSE difference (image_1 - image_0), shape (B,)
# ref_diff      : reference MSE difference (image_1 - image_0), shape (B,)
# y_true_L      : ground truth preferences for labeled subset, shape (B_L,)
# y_pseudo      : external model pseudo-labels for entire batch, shape (B,)
# beta          : KL regularization coefficient
# B_L, B_U      : labeled and unlabeled mini-batch sizes
# wt            : inverse labeled sampling fraction, (B_L + B_U) / B_L

# Calculate Diffusion DPO logits
logits = -0.5 * beta * (model_diff - ref_diff)

# Split batches into Labeled (L) and Unlabeled (U) subsets
logits_L, logits_U = logits[:B_L], logits[B_L:]
y_pseudo_L, y_pseudo_U = y_pseudo[:B_L], y_pseudo[B_L:]

# Component BCE losses (summed over each subset)
loss_L = F.binary_cross_entropy_with_logits(
    logits_L, y_true_L, reduction="sum")
loss_pseudo_L = F.binary_cross_entropy_with_logits(
    logits_L, y_pseudo_L, reduction="sum")
loss_pseudo_U = F.binary_cross_entropy_with_logits(
    logits_U, y_pseudo_U, reduction="sum")

# Standard Semi-Supervised Diffusion DPO
SemiSup_DPO_loss = loss_L / B_L + loss_pseudo_U / B_U

# Debiased Diffusion DPO
Debiased_DPO_loss = (
    loss_pseudo_L + loss_pseudo_U
    + wt * (loss_L - loss_pseudo_L)
) / (B_L + B_U)
```

Figure 5: Pseudocode for Debiased Diffusion DPO, shown alongside the standard semi-supervised objective for comparison. The code allows arbitrary positive batch sizes B_L, B_U, with $w_t = (B_L + B_U)/B_L$. Our experiments use $B_L = B_U$ and $w_t = 2$; see Sec. B.1.

self-selected online creator community and the biases induced by the app interaction loop itself, so its prompt distribution, stylistic preferences, and ambiguity patterns need not match those of more balanced benchmark-style datasets.

HPDv2. HPDv2 [52] was introduced as a larger and cleaner human-preference dataset designed to mitigate several known biases in earlier resources. The paper reports 798,090 preference choices over 433,760 image pairs, with prompts and images deliberately drawn from diverse sources to improve coverage and reduce dataset bias. The official repository provides additional context: prompts are sourced from DiffusionDB [50] and MSCOCO captions [26]; DiffusionDB prompts are first cleaned with ChatGPT to remove biased function words; and human annotators then rank images generated by different text-to-image models for the same prompt. Relative to Pick-a-Pic, this makes HPDv2 a stronger clean-label source, but not a bias-free one. Its prompt cleaning and curated benchmark design improve stability and breadth, but they also make the resulting distribution more controlled and less representative of unconstrained real-user prompting. In that sense, HPDv2 should be viewed as less noisy and more curated, rather than as perfectly unbiased.

On PickScore-based data curation. A potential concern is that using PickScore for training-data selection may introduce circularity when PickScore is also reported as one of the evaluation metrics. We clarify that PickScore is used only as a coarse curation signal before training, not as supervision for the preference objective and not as part of the test-time evaluation pipeline. Specifically, we use PickScore to rank candidate preference pairs and select a higher-quality 5K pool from each source dataset. This filtering step is intended to remove highly ambiguous or low-quality pairs so that the small human-labeled subset provides a cleaner anchor for debiasing. After this pool is formed, the actual training labels are either the original human preference labels or synthetic pairwise

judgments produced by Qwen2.5-VL; PickScore is never used to assign pseudo-labels or to decide the preferred image in a training pair. This distinction avoids direct circularity. The evaluation prompts are drawn from the held-out PartiPrompt benchmark, and each method is evaluated on newly generated images, not on the curated training pairs. Thus, PickScore-based sorting does not expose the model to the evaluation prompts or evaluation images. Moreover, the same curated pools are used for all competing methods, including **DPO+25**, **DPO+25-75**, and DeDPO. Therefore, any effect of PickScore-based selection is shared across baselines and cannot by itself explain the relative improvement of DeDPO over the direct semi-supervised baseline. The comparison between DeDPO and **DPO+25-75** is especially controlled: both methods use the same human/synthetic split and differ only in whether the debiasing correction is applied.

B.3 Evaluation metrics.

We evaluate text-to-image preference alignment using **PickScore**, **HPSv3**, and **ImageReward**, instead of relying on older CLIPScore- or aesthetic-based metrics. Although CLIPScore and aesthetic predictors were widely used in earlier text-to-image work, they are only indirect proxies for human preference. CLIPScore primarily measures image–text representation similarity and can miss failures in realism, visual quality, or fine-grained composition, while aesthetic predictors are image-only models and therefore cannot determine whether an otherwise attractive image actually follows the prompt. More recent preference-modeling work suggests that evaluators trained on human judgments better align with human preferences than these earlier proxies. Accordingly, we use PickScore [20], which is trained from real user pairwise preferences; HPSv3 [31], a recent broad-coverage human preference metric intended to better capture differences across modern image generators and quality regimes; and ImageReward [55], a general-purpose reward model trained from expert human comparisons. For PickScore and HPSv3, we report the mean score over all PartiPrompt test prompts. For ImageReward, we report **ImageReward@Top-1(%)**, defined as the percentage of prompts for which a method attains the highest ImageReward score among all compared methods. This metric is best interpreted as a prompt-level win rate: rather than averaging reward magnitudes across prompts, it measures how often a method is ranked first on a per-prompt basis. We include it as a complementary relative metric because it is less sensitive to score calibration across prompts while still capturing how consistently a model wins against the baselines on a common evaluation set. Therefore, improvements in ImageReward@Top-1 indicate that gains are distributed broadly across prompts, rather than being driven by a small number of especially favorable examples.

C Additional Results

C.1 Training without PickScore-based curation

PickScore is used to rank candidate training pairs for PaP-5K and HPD-5K and is also one of our evaluation metrics. To rule out an interaction between this curation rule and the method, we remove PickScore ranking entirely and sample a uniformly random 5K subset of Pick-a-Pic v2. Tab. 3 shows that DeDPO retains its advantage over the matched DPO+25-75 baseline on both a U-Net (SD1.5) and a DiT (PixArt-Σ at 512 resolution). The gain is largest on ImageReward@Top-1, which is independent of PickScore ($+5.52$ and $+6.44$ points), so the improvement cannot be attributed solely to PickScore-based curation.

Table 3: Results on a uniformly sampled, unfiltered 5K subset of Pick-a-Pic v2 (no PickScore ranking).

Model	Method	PickScore (↑)	ImageReward@Top-1 (%) (↑)
SD1.5	DPO+25	21.67	19.11
	DPO+25-75	21.69	23.13
	DeDPO+25-75	**21.75**	**28.65**
	DPO+100	21.75	29.11
PixArt-Σ-512	DPO+25	22.34	18.99
	DPO+25-75	22.57	22.23
	DeDPO+25-75	**22.58**	**28.67**
	DPO+100	22.67	30.11

C.2 DiT backbone

All main experiments use U-Net backbones (SD1.5 and SDXL). To test whether the effect is specific to U-Nets, we train PixArt-Σ, a DiT-based text-to-image model, on PaP-5K at 512 and 1024 resolution. As shown in Tab. 4, DeDPO again improves over DPO+25-75 and approaches the DPO+100 full-human reference at both resolutions, e.g., at 512 resolution, PickScore 22.72 vs. 22.66 and ImageReward@Top-1 29.78 vs. 20.75.

Table 4: Results with a DiT backbone (PixArt-Σ) trained on PaP-5K.

	PixArt-Σ-512		**PixArt-Σ-1024**	
Method	**PickScore** ($\uparrow$)	**IR@Top-1 (%)** ($\uparrow$)	**PickScore** ($\uparrow$)	**IR@Top-1 (%)** ($\uparrow$)
DPO+25	22.60	19.30	22.81	20.71
DPO+25-75	22.66	20.75	22.84	24.02
DeDPO+25-75	**22.72**	**29.78**	**22.87**	**26.71**
DPO+100	22.75	30.17	22.89	28.56

C.3 Fraction of the full-human gap recovered

Raw differences in PickScore or ImageReward are metric-specific and do not indicate how much of the distance to full human supervision has been closed. For example, on PaP-5K with SD1.5, the entire PickScore improvement from SFT to DPO+100 is only 0.24 points ($21.64 \rightarrow 21.88$). We therefore report the gap recovered by DeDPO, $\frac{\text{DeDPO}-\text{DPO+25-75}}{\text{DPO+100}-\text{DPO+25-75}}$, capped at 100%. Tab. 5 shows that DeDPO recovers 96–100% of the PickScore gap on SD1.5 and SDXL and a majority of the gap in most other cells. The lowest value (ImageReward@Top-1 on PaP-5K with SD1.5, 39%) occurs where the full-human headroom is largest; we report all settings rather than selecting favorable ones.

Table 5: Fraction of the DPO+25-75$\rightarrow$DPO+100 gap recovered by DeDPO+25-75 (capped at 100%). SD1.5 and SDXL rows use the 15-seed means of Tab. 1; PixArt-Σ rows use the single-run results of Tab. 4. †DPO+100 does not exceed DPO+25-75 in this cell, so the gap is not positive; DeDPO scores above both.

Setting	**PickScore gap**	**ImageReward@Top-1 gap**
PaP-5K / SD1.5	100%	39%
PaP-5K / SDXL	96%	91%
HPD-5K / SD1.5	100%†	90%
HPD-5K / SDXL	100%	82%
PaP-5K / PixArt-Σ-512	67%	96%
PaP-5K / PixArt-Σ-1024	60%	59%

C.4 Balanced 50/50 human/synthetic split

Our main setting targets the label-scarce regime, and Figs. 2a and 2b already sweep human/synthetic ratios from roughly 1/99 to 25/75. To test a larger human fraction, we also train on a 50/50 split on SD1.5 and PixArt-Σ-512 (Tab. 6). DeDPO still outperforms naive mixing on both backbones, with PickScore gains of $+0.15$ and $+0.07$ and a head-to-head ImageReward@Top-1 win rate of 64.47% and 61.15%. This is consistent with the mechanism of the correction, whose benefit scales with the amount of biased synthetic supervision in the pool and vanishes only as the human fraction approaches 100% or as the labeler becomes unbiased (proposition 1).

C.5 Behavior with strong synthetic labelers

By proposition 1, if the synthetic labels equal the human labels pointwise, $\hat{\mathcal{G}}(\boldsymbol{y}) = z$, the correction term vanishes and DeDPO reduces to standard training on all $n_l + n_u$ pairs, so the expected benefit shrinks as the labeler approaches the human target. This matches our observations: the gains are smaller on HPD-5K (about 93% Qwen–human agreement) than on PaP-5K (about 83%). Because these two agreement levels come from different datasets, Fig. 3 provides the controlled comparison on

Table 6: Results with a 50/50 human/synthetic split. ImageReward@Top-1 is computed head-to-head between the two methods of each backbone.

Model	**Method**	**PickScore** (↑)	**ImageReward@Top-1 (%)** (↑)
SD1.5	DPO+50-50	21.70	35.53
	DeDPO+50-50	**21.85**	**64.47**
PixArt-Σ-512	DPO+50-50	22.64	38.85
	DeDPO+50-50	**22.71**	**61.15**

a single dataset (PaP-5K), varying labeler quality from roughly 50% to 90% agreement; DeDPO stays above DPO across this range, with the largest gap under the most biased labelers. A strong but imperfect labeler can still make a small set of systematic errors, which is exactly what the correction targets. We do not claim a finite-sample no-harm guarantee: estimating the correction can add variance, and the theoretical statement is unbiasedness in expectation together with exact reduction to standard training when the labels are perfect. Empirically, performance remains stable in the high-agreement regime, e.g., the small but non-negative HPD-5K margins in Tab. 1.

D Limitations

Our work has several limitations. First, although DeDPO consistently improves over matched semi-supervised DPO baselines, its gains still depend on the quality and structure of the synthetic labeler. When synthetic preferences are only weakly aligned with human judgment, or when their errors are highly instance-specific rather than systematic, the correction induced by the small labeled subset may be less effective. Second, our strongest empirical results are obtained in the label-scarce offline setting studied here, with fixed pretrained synthetic annotators and limited backbone diversity. We do not claim that the same gains will transfer unchanged to all diffusion architectures, prompt distributions, or online preference-optimization pipelines. Third, the theoretical guarantees rely on assumptions such as sample splitting, smoothness, and sufficiently well-behaved function classes; these assumptions clarify the mechanism behind debiasing, but they do not perfectly capture every practical training regime, especially self-training settings that violate independence between the synthetic labeler and the optimized model. Finally, our evaluation relies primarily on automatic preference metrics and qualitative inspection. Although we use stronger human-preference-aligned evaluators than older proxy metrics, these remain imperfect surrogates for large-scale human studies. A useful direction for future work is therefore to extend DeDPO to broader architectures and feedback sources, and to validate it with more direct human evaluation under diverse prompting conditions.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We already discuss the limitaton in the appendix.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: We provide the full proof in the appendix

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: See the experimental setup.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

Justification: The full code will be released upon acceptance. We already provide the pseudo code in appendix.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: See the experimental setup

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: See the main results

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: See the experimental setup

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conform with NeurIPS CoE.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

    Justification: No disinformation or privacy considerations

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [No]

    Justification: The paper proposed an algorithm which is safe.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [No]

    Justification: We use public dataset.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification:

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: See the discussion in human evaluation.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification:

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification:

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.