Boost Your Human Image Generation Model via Direct Preference Optimization

Sanghyeon Na* Yonggyu Kim* Hyunjoon Lee†
Kakao
{orca.ai, arthur.a, malfo.y}@kakaocorp.com

Figure 1. **Top:** HG-DPO generates high-quality human images that encompass a wide range of actions, appearances, group sizes, and backgrounds. **Bottom left:** This is because HG-DPO improves the base model to generate images with more realistic anatomical features and poses, while also better aligning with the prompt (red text in the prompt). **Bottom right:** The benefits of HG-DPO transfer to personalized text-to-image tasks without additional training, generating high-quality images with the identity of concept image.

Abstract

Human image generation is a key focus in image synthesis due to its broad applications, but even slight inaccuracies in anatomy, pose, or details can compromise realism. To address these challenges, we explore Direct Preference Optimization (DPO), which trains models to generate preferred (winning) images while diverging from non-preferred (losing) ones. However, conventional DPO methods use generated images as winning images, limiting realism. To overcome this limitation, we propose an enhanced DPO approach that incorporates high-quality real images as win-

ning images, encouraging outputs to resemble real images rather than generated ones. However, implementing this concept is not a trivial task. Therefore, our approach, ***HG-DPO*** *(**H**uman image **G**eneration through **DPO**), employs a novel curriculum learning framework that gradually improves the output of the model toward greater realism, making training more feasible. Furthermore, HG-DPO effectively adapts to personalized text-to-image tasks, generating high-quality and identity-specific images, which highlights the practical value of our approach.*

1. Introduction

Human image generation is a key focus in generative modeling due to its broad applications in entertainment and social media. Despite advances in text-to-image generation [23, 55, 58, 63, 64, 67] using diffusion models [16, 27, 73], it remains difficult to generate realistic human images because even slight inaccuracies in anatomy, pose, or fine details can create artifacts that reduce realism. Conventional approaches, which rely on supervised fine-tuning of diffusion models using high-quality images, often struggle to achieve the desired realism. Our objective is to build upon this fine-tuned model as a base, enhancing it to produce realistic human images as shown in Figure 1.

In response, we explore Direct Preference Optimization (DPO) [62, 78], which trains models on pairs of preferred (winning) and non-preferred (losing) images, guiding outputs toward preferred characteristics while avoiding non-preferred ones. This approach suits complex tasks like human image generation by leveraging the contrasts between winning and losing pairs. However, Diffusion-DPO [78] and its variants [12, 17, 21, 24, 28, 39, 86, 88] struggle to achieve high realism in human image generation because they use *generated* images as winning images, limiting output quality to the suboptimal level of generated images. To overcome this, we propose an enhanced DPO method with a novel preference structure that uses *real* images as winning images, while treating generated images as losing images.

Notably, this preference structure integrates the training mechanism of Generative Adversarial Networks (GANs) [22], which has proven highly effective in guiding human image generation toward greater realism [30–32, 53], into diffusion models. In GANs, a discriminator assesses how closely a generated image resembles real images and penalizes deviations, thereby guiding the outputs to be more similar to real images. Similarly, HG-DPO penalizes outputs that resemble generated (losing) images, while encouraging outputs more similar to real (winning) images. Both methods guide outputs to resemble real images rather than generated ones, enhancing realism. However, unlike GANs, HG-DPO implements it through DPO framework by defining a preference structure between real and generated images.

Based on this concept, our initial experiments used a naive approach, applying DPO with real images as winning and generated images as losing. However, this approach fell short, likely due to the significant domain gap between real and generated images. For example, real images typically have more realistic compositions, poses, and intricate details. This gap can make a single-stage training difficult.

To bridge this gap, we integrate curriculum learning [3] into the DPO framework, gradually training the model from easy to hard tasks. As shown in Figure 2, HG-DPO training consists of three stages: easy, normal, and hard. To create tasks of varying difficulty at each stage, each stage utilizes a dataset constructed in a different manner. As a result, in the easy stage, the model learns basic human preferences, focusing on undistorted anatomy and poses, and better image-text alignment. The normal stage enhances visual quality by capturing more realistic compositions and poses. The hard stage refines fine details to match real images, enabling high-fidelity outputs. This gradual progression allows the model to produce images that closely resemble real ones. Notably, unlike existing DPO datasets [34, 82], our datasets are constructed without costly human feedback.

Furthermore, HG-DPO can improve personalized text-to-image (PT2I) [7, 8, 48, 65, 69, 84], by generating higher-quality images tailored to specific identities, as shown in Figure 1, without requiring additional training. This versatility highlights its practical value for creative and social media applications.

In summary, our contributions are as follows:

(i) We propose a novel DPO approach, HG-DPO, to generate high-quality human images. Unlike existing DPO methods, our approach uses real images as winning images. This can be viewed as introducing the training mechanism of GANs into diffusion models.

(ii) However, implementing our approach is challenging. To address this, we present a three-stage curriculum learning pipeline that enables the model to generate realistic images through gradual improvement.

(iii) Unlike existing DPO datasets, our proposed methods for constructing DPO datasets does not require costly human feedback, making it more efficient.

(iv) We demonstrate that HG-DPO effectively adapts to PT2I tasks without additional training, highlighting the practical value of our work.

2. Related Work

Aligning diffusion models with human preferences. Direct Preference Optimization (DPO) offers an improvement over Reinforcement Learning from Human Feedback

* Equal contribution.
† Corresponding author.

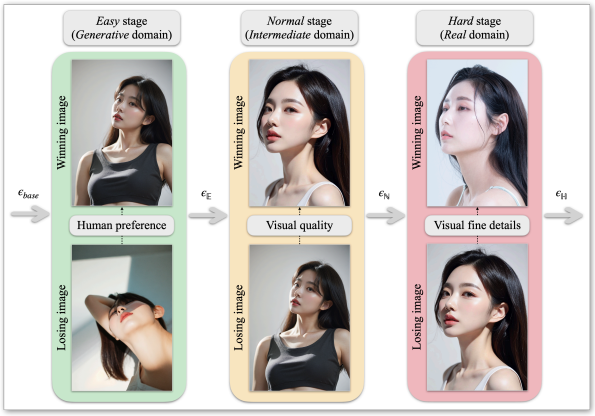

Figure 2. **Three-stage training of HG-DPO.** It progressively enhances the model's human image generation capabilities.

(RLHF) [56], which is widely used to align large language models (LLMs) to human preferences [9, 10, 13, 36, 42–44, 71, 83, 92], by directly using human preference dataset without requiring a reward model. Building on the success of DPO in the field of LLMs, Diffusion-DPO [78] extends DPO to diffusion models.Since then, several methods [12, 17, 21, 24, 28, 39, 86, 88] have demonstrated improved performance over Diffusion-DPO. However, these approaches primarily focus on enhancing training techniques while utilizing the Pick-a-Pic dataset [34], which contains only generated images. On the other hand, DPOK [18], D3PO [87], and AlignProp [59] propose online learning methods based on policy optimization, DPO, and reward model gradients, respectively. However, these approaches also do not incorporate real images into the training process. In contrast, we enable the use of real images in the DPO dataset by integrating curriculum learning into DPO. Among existing methods, Curriculum-DPO [12] is the most closely related to ours, as it also incorporates curriculum learning into DPO. However, its performance gains are limited because it relies exclusively on generated images.

Curriculum Learning. Curriculum learning [3] trains models by first exposing them to simpler data or tasks, then progressively introducing more complex ones. It has proven effective in fields such as computer vision [4, 37, 60, 70, 74–76, 90], natural language processing [2, 35, 40, 45, 66, 89, 91, 93], and reinforcement learning [19, 20, 47, 49, 51, 52, 54]. We employ curriculum learning to ease the transition of model's outputs from the generative to real domain, a shift challenging to achieve through single-stage training.

3. HG-DPO

Given a human-specific paired image-text dataset $\mathcal{D}_{\text{real}} = (\mathcal{P}, \mathcal{X}_{\text{real}})$, where $\mathcal{P}$ is a set of prompts and $\mathcal{X}_{\text{real}}$ is a set of real human images, our objective is to design a model that can generate images with a level of realism similar to that of $\mathcal{X}_{\text{real}}$. We begin by establishing a base model ϵ_{base} by supervised fine-tuning a backbone model, which is a latent diffusion model [64] with $\mathcal{D}_{\text{real}}$. Despite being fine-tuned on high-quality real images, ϵ_{base} often generates low-quality human images as shown in Figure 5.

Therefore, we refine ϵ_{base} to improve its human image generation using HG-DPO, which consists of three progressively challenging DPO stages: easy, normal, and hard (Figure 2). In each stage, the model faces increasingly difficult objectives. The difficulty of each stage is determined by the domain of winning images, the target images that the model aims to generate. Specifically, the easy stage uses generated images as winning images. Since there is no domain gap - the model is trained on the same type of images it generates - this stage represents an easier task. In contrast, the hard stage uses real images as winning images. The model must learn from a real domain, which differs significantly from its generative domain, making it a more difficult task. For losing images, we use the winning images from the previous stage, except in the easy stage, which has no prior stage. Through each stage, the model is progressively refined to generate images that more closely resemble real images. Here, we employ LoRA [29] for training.

3.1. Easy Stage

In the easy stage, we refine ϵ_{base} into $\epsilon_{\mathbb{E}}$ to generate images more likely preferred by humans. To achieve this, we create pairs of winning and losing images, where the winning images exhibit better anatomy, pose, and prompt alignment than the losing images as shown in the green box in Figure 2 and Figure 3.

Image pool generation. The first step is to generate images using a prompt set $\mathcal{P} = \{p^i\}_{i=1}^{D}$ where p^i is a prompt and D is the size of $\mathcal{D}_{\text{real}}$. Instead of generating exactly two images for winning and losing, we create an *image pool* with N distinct images per prompt. For a prompt p^i, the image pool $\mathcal{X}^i_{\text{gen}}$ of size N is defined as:

$$\mathcal{X}^i_{\text{gen}} = \{x^{i_j}_{\text{gen}}\}_{j=1}^{N} \text{ where } x^{i_j}_{\text{gen}} = \mathcal{G}(\epsilon_{base}, p^i, r^{i_j}). \quad (1)$$

Here, $\mathcal{G}$ is a text-to-image sampler with a random seed r^{i_j} used to generate $x^{i_j}_{\text{gen}}$. To generate N different images, we employ N different random seeds $\{r^{i_j}\}_{j=1}^{N}$.

Selection of winning and losing images. We then score the images using a human preference estimator f [34]:

$$\mathcal{S}^i_{\text{gen}} = \{s^{i_j}_{\text{gen}}\}_{j=1}^{N} \text{ where } s^{i_j}_{\text{gen}} = f(x^{i_j}_{\text{gen}}, p^i), \quad (2)$$

where $s^{i_j}_{\text{gen}}$ is a preference score of $x^{i_j}_{\text{gen}}$ considering p^i. Unlike existing DPO datasets [34, 82] that rely on costly human feedback, we use a more efficient AI-based method.

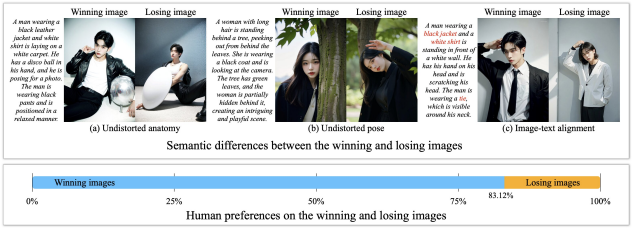

Figure 3. **DPO Dataset for the easy stage.** In the upper figure, $\mathcal{D}_{\mathbb{E}}$, constructed with AI rather than human feedback, shows winning images with superior features over losing images. A user study in the lower figure confirms this outcome.

We select the images with the highest and lowest preference scores from the image pool $\mathcal{X}^i_{\text{gen}}$ as the winning and losing images, $x^{\mathbf{w},i}_{\mathbb{E}}$ and $x^{\mathbf{l},i}_{\mathbb{E}}$, respectively. It ensures clear semantic superiority of the winning image, consistent with human preferences (Figure 3). Formally, this is defined as:

$$(x^{\mathbf{w},i}_{\mathbb{E}},\ x^{\mathbf{l},i}_{\mathbb{E}}) = (\mathcal{X}^i_{\text{gen}}[j^w_{\mathbb{E}}],\ \mathcal{X}^i_{\text{gen}}[j^l_{\mathbb{E}}]) \quad (3)$$
$$\text{where } (j^w_{\mathbb{E}},\ j^l_{\mathbb{E}}) = (\underset{j^w_{\mathbb{E}}\in\{1,\dots,N\}}{\text{argmax}}\ S^i_{\text{gen}},\ \underset{j^l_{\mathbb{E}}\in\{1,\dots,N\}}{\text{argmin}}\ S^i_{\text{gen}}).$$

By assigning winning and losing images for all prompts in $\mathcal{P}$, we complete the dataset: $\mathcal{D}_{\mathbb{E}} = \{(p^i, x^{\mathbf{w},i}_{\mathbb{E}}, x^{\mathbf{l},i}_{\mathbb{E}})\}^D_{i=1}$.

Statistics matching loss. With $\mathcal{D}_{\mathbb{E}}$, we can apply the objective function of Diffusion-DPO ($\mathcal{L}_{D\text{-}DPO}$) to update ϵ_{base} to $\epsilon_{\mathbb{E}}$. However, $\epsilon_{\mathbb{E}}$ trained this way produces color shift artifacts, making images unnatural (see $N > 2$ in Figure 8). This issue arises from the statistics of latents sampled by $\epsilon_{\mathbb{E}}$ diverging from those sampled by ϵ_{base}. To address this, we design a statistics matching loss to prevent this divergence. Let l^t be a noisy latent of a winning image generated by the forward diffusion process at timestep t during training, and let $\epsilon^\theta_{\mathbb{E}}$ be the model being updated in the easy stage. From l^t, we sample l^{t-1}_θ and l^{t-1}_{base}, which are latents sampled using $\epsilon^\theta_{\mathbb{E}}$ and ϵ_{base}, respectively. The statistics matching loss is defined as:

$$\mathcal{L}_{stat} = \mathbb{E}_{\mathcal{D}_{\mathbb{E}},\ t\sim\mathcal{U}(0,\ T)}\left[\ ||\ \mu(l^{t-1}_\theta) - \mu(l^{t-1}_{base})\ ||^2_2\ \right] \quad (4)$$

where μ calculates the channel-wise mean. $\mathcal{L}_{stat}$ only matches the mean because it sufficiently resolves the color shift artifacts. During the easy stage, ϵ_{base} is updated using the combined objective: $\mathcal{L} = \mathcal{L}_{D\text{-}DPO} + \lambda_{stat}\mathcal{L}_{stat}$. $\mathcal{L}_{stat}$ is applied only in the easy stage, since color shift artifacts are not observed in the normal and hard stages.

3.2. Normal Stage

While $\epsilon_{\mathbb{E}}$ generates human images aligned with human preferences, they lack the realistic compositions and poses of real images. To address this, we refine $\epsilon_{\mathbb{E}}$ into $\epsilon_{\mathbb{N}}$ through the normal stage, improving visual quality (*e.g.*, realistic composition and pose). Instead of immediately using real images as winning images after the easy stage, we introduce the intermediate domain for winning images as shown in the yellow box in Figure 2. The normal stage facilitates training by providing an intermediate task before progressing from the easy to the hard stage.

Intermediate domain. The intermediate domain has mixed characteristics of generated and real domains. Specifically, we produce the intermediate images through *Stochastic Differential Reconstruction (SDRecon)*, which is similar with SDEdit [50], where noise is added to a real image to generate a noisy image, which is then reconstructed back into a real image using ϵ_{base}. Here, we can use $\epsilon_{\mathbb{E}}$, but by using ϵ_{base} instead, we eliminate the need to wait for $\epsilon_{\mathbb{E}}$ to be fully trained in order to construct the intermediate domain. This process retains certain features of the real image (*e.g.*, composition and pose), while other features, like texture and fine details, resemble those of the generative domain. For example, in Figure 2, the intermediate image (the winning image in the yellow box) maintains the pose of the real image (the winning image in the red box), but it lacks the fine details present in the real image (*e.g.*, texture) resulting in a smoother and more synthetic appearance.

We perform SDRecon by varying diffusion timesteps to control noise magnitude, creating multi-level intermediate domains. It allows us to adaptively select the most appropriate intermediate domain during training. Formally, for each prompt $p^i \in \mathcal{P}$ and paired real image $x^i_{\text{real}} \in \mathcal{X}_{\text{real}}$, we use a set of T timesteps, $\mathcal{T} = \{t_1, t_2, \dots, t_T\}$, to generate a set of T intermediate images:

$$\mathcal{X}^i_{\text{int}} = \{x^{i_t}_{\text{int}}\}_{t\in\mathcal{T}} \text{ where } x^{i_t}_{\text{int}} = \mathcal{R}(\epsilon_{base}, p^i, x^i_{\text{real}}, t) \quad (5)$$

where $\mathcal{R}$ is the proposed SDRecon operator. When $t = t_T$, strong noise produces an intermediate image closest to the generative domain, while $t = t_1$ results in an intermediate image closest to the real domain.

Selection of winning and losing images. We score each image in $\mathcal{X}^i_{\text{int}}$ to select the optimal winning image:

$$\mathcal{S}^i_{\text{int}} = \{s^{i_t}_{\text{int}}\}_{t\in\mathcal{T}},\ \text{where } s^{i_t}_{\text{int}} = f(x^{i_t}_{\text{int}}, p^i). \quad (6)$$

Then, we designate the image with the highest score in $\mathcal{X}^i_{\text{int}}$ as the winning image $x^{\mathbf{w},i}_{\mathbb{N}}$, while the winning image from the easy stage is used as the losing image $x^{\mathbf{l},i}_{\mathbb{N}}$. Here, we consider only mid-level images from $\mathcal{X}^i_{\text{int}}$ as candidates for winning images. Formally, this process is defined as

$$(x^{\mathbf{w},i}_{\mathbb{N}},\ x^{\mathbf{l},i}_{\mathbb{N}}) = (\mathcal{X}^i_{\text{int}}[j^w_{\mathbb{N}}],\ \mathcal{X}^i_{\text{gen}}[j^l_{\mathbb{N}}]) \quad (7)$$
$$\text{where } (j^w_{\mathbb{N}},\ j^l_{\mathbb{N}}) = (\underset{j^w_{\mathbb{N}}\in\hat{\mathcal{T}}}{\text{argmax}}\ S^i_{\text{int}},\ \underset{j^l_{\mathbb{N}}\in\{1,\dots,N\}}{\text{argmax}}\ S^i_{\text{gen}}).$$

Model	P-Score (↑)	HPS (↑)	I-Reward (↑)	AES (↑)	CLIP (↑)	FID (↓)	CI-Q (↑)	CI-S (↑)	ATHEC (↑)
HPD v2 [82]	21.7855	0.2829	-0.0002	6.1132	29.99	35.98	0.9015	0.9592	19.22
Pick-a-Pic v2 [34]	21.7433	0.2831	0.0268	6.1315	30.05	37.89	0.8801	0.9464	19.00
Diffusion-DPO [78]	17.9314	0.2425	-1.8873	5.1183	24.03	112.67	0.8198	0.9438	**36.30**
NCP-DPO [21]	18.7679	0.2560	-1.6644	5.5309	24.37	96.72	0.7164	0.8702	18.81
MAPO [28]	20.4401	0.2707	-0.3613	5.4477	28.35	59.36	0.7117	0.8343	32.68
Curriculum-DPO [12]	22.4381	0.2869	0.6532	6.1925	31.50	35.35	0.8886	0.9561	23.36
AlignProp [59]	**23.0202**	0.2854	0.1989	**6.2773**	29.67	49.92	0.8599	0.9661	17.05
HG-DPO (Ours)	22.6043	**0.2872**	**0.7568**	6.1785	**31.57**	**29.41**	**0.9343**	**0.9858**	29.41

Table 1. **Quantitative comparison with the previous methods.** HG-DPO achieves superior performance over the existing methods.

Figure 4. **Qualitative comparison with the previous methods.** HG-DPO generates high-quality human images with more realistic compositions and poses, providing superior text alignment compared to the prior methods.

Here, $\hat{\mathcal{T}} = \{t \mid t_r \leq t \leq t_g\}$ where $t_1 < t_r$ and $t_g < t_T$. Also, note that $x_{\mathbb{N}}^{\mathbf{l},i} = x_{\mathbb{E}}^{\mathbf{w},i}$. By applying this process to all prompts in $\mathcal{P}$, we obtain a set of triplets $\{p^i, x_{\mathbb{N}}^{\mathbf{w},i}, x_{\mathbb{N}}^{\mathbf{l},i}\}_{i=1}^{D}$.

Filtering. Instead of using all D triplets, we filter them to retain only those containing images of sufficiently high quality, which are expected to benefit training. Specifically, a triplet is included in the normal stage dataset $\mathcal{D}_{\mathbb{N}}$ only when the score of $x_{\mathbb{N}}^{\mathbf{w},i}$ meets or exceeds that of $x_{\mathbb{N}}^{\mathbf{l},i}$:

$$\mathcal{D}_{\mathbb{N}} = \{(p^i, x_{\mathbb{N}}^{\mathbf{w},i}, x_{\mathbb{N}}^{\mathbf{l},i})\}_{i \in \mathcal{K}} \quad \text{where } \mathcal{K} = \{k | \mathcal{S}_{\text{int}}^k[j_{\mathbb{N}}^w] \geq \mathcal{S}_{\text{gen}}^k[j_{\mathbb{N}}^l]\}. \tag{8}$$

During the normal stage, we obtain $\epsilon_{\mathbb{N}}$ by updating $\epsilon_{\mathbb{E}}$ using $\mathcal{L}_{D\text{-}DPO}$ with the dataset $\mathcal{D}_{\mathbb{N}}$.

3.3. Hard Stage

The objective of the hard stage is to enhance $\epsilon_{\mathbb{N}}$ enabling it to generate images that closely resemble real images, including visual fine details. To achieve this goal, as shown in the red box of Figure 2, the hard stage is designed to leverage *real* images as winning images. However, we use images from the intermediate domain t_1, which are nearly indistinguishable from real images, as this approach has yielded better results in our experiments. Formally, the dataset for the hard stage, denoted as $\mathcal{D}_{\mathbb{H}} = \{(p^i, x_{\mathbb{H}}^{\mathbf{w},i}, x_{\mathbb{H}}^{\mathbf{l},i})\}_{i \in \mathcal{K}}$ where $\mathcal{K}$ is the same as $\mathcal{K}$ in Eq. (8), is constructed with winning and losing images defined as

$$(x_{\mathbb{H}}^{\mathbf{w},i}, x_{\mathbb{H}}^{\mathbf{l},i}) = (\mathcal{X}_{\text{int}}^i[t_1], x_{\mathbb{N}}^{\mathbf{w},i}). \tag{9}$$

Using the dataset $\mathcal{D}_{\mathbb{H}}$, we train $\epsilon_{\mathbb{N}}$ via $\mathcal{L}_{D\text{-}DPO}$ to obtain $\epsilon_{\mathbb{H}}$.

3.4. Training the Text Encoder

To improve image-text alignment, we train the text encoder separately from the U-Net, following TexForce [6]. We use both $\epsilon_{\mathbb{H}}$ and the enhanced text encoder at inference. The text encoder is trained only up to the easy stage, as the goal is to improve image-text alignment rather than visual quality.

4. Experimental Settings

In this section, we describe our experimental settings. Additional details not covered in this section are available in the Appendices.

Datasets. We constructed an internal dataset consisting of approximately 300k high-quality human images. From this dataset, 5k images were randomly selected for testing, while the remaining images were utilized for training. Captions for the training and test images were generated using LLaVA [41] and Qwen2-VL [80], respectively. For evaluation, we generate images using these 5k test prompts.

Model	P-Score (↑)	HPS (↑)	I-Reward (↑)	AES (↑)	CLIP (↑)	FID (↓)	CI-Q (↑)	CI-S (↑)	ATHEC (↑)
Base (ϵ_{base})	21.7364	0.2819	-0.0665	6.1061	29.72	37.34	0.9058	0.9573	18.73
Naive	17.9314	0.2425	-1.8873	5.1183	24.03	112.67	0.8198	0.9438	**36.30**
Easy ($\epsilon_{\mathbb{E}}$)	22.5384	**0.2878**	0.7146	6.1775	31.56	36.00	0.9057	0.9547	19.58
Normal ($\epsilon_{\mathbb{N}}$)	22.5422	0.2865	0.6515	6.1637	31.45	**26.05**	0.9302	0.9778	25.47
Hard ($\epsilon_{\mathbb{H}}$)	22.4698	0.2867	0.5791	**6.1955**	31.15	28.66	**0.9365**	**0.9859**	30.08
Hard ($\epsilon_{\mathbb{H}}$) + TE (HG-DPO)	**22.6043**	0.2872	**0.7568**	6.1785	**31.57**	29.41	0.9343	0.9858	29.41
E2E training	21.3244	0.2773	-0.0487	6.0892	29.50	57.10	0.7962	0.7862	9.46
Hard w/o easy	19.8417	0.2687	-0.8262	5.7305	27.36	72.34	0.8801	0.9631	19.72
Hard w/o normal	22.2541	0.2849	0.3932	6.1990	30.30	32.24	0.9233	0.9577	21.55
Base (ϵ_{base}) + SFT	21.732	0.2808	0.0140	6.0948	30.67	34.05	0.8860	0.9531	22.09
Easy ($\epsilon_{\mathbb{E}}$) + SFT	21.858	0.2818	0.1788	6.0569	31.20	35.23	0.8802	0.9498	23.10
Normal ($\epsilon_{\mathbb{N}}$) + SFT	21.9077	0.2818	0.1840	6.0691	31.08	35.12	0.8882	0.9534	23.29

Table 2. **Quantitative analysis of the HG-DPO pipeline.** The row labeled *Base (ϵ_{base})* shows the base model's performance, while *Naive* to *Hard ($\epsilon_{\mathbb{H}}$) + TE* illustrate model progress through curriculum stages, ending with the final model, **HG-DPO**. Subsequent rows examine the importance of each curriculum stage. *E2E training* refers to a model trained end-to-end using the combined training datasets from all three stages, instead of the proposed three-stage training. *Hard w/o easy* and *Hard w/o normal* exclude the easy and normal stages from the training pipeline, respectively. The last three rows indicate models with supervised fine-tuning (*SFT*) using winning images of the hard stage after each curriculum step.

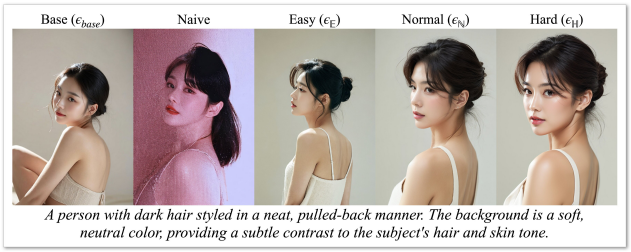

Figure 5. **Qualitative progress.** ϵ_{base} evolves as it progresses through each stage of the HG-DPO pipeline up to the hard stage.

Figure 6. **Qualitative results by the enhanced text encoder.** With the improved text encoder, Hard ($\epsilon_{\mathbb{H}}$) + TE can achieve the enhanced image-text alignment, retaining the image quality of $\epsilon_{\mathbb{H}}$.

Figure 7. **Qualitative results illustrating the importance of each stage.** To generate high-quality images like the one labeled as Hard ($\epsilon_{\mathbb{H}}$), each stage of the HG-DPO pipeline is essential.

Metrics. To assess prompt-aware human preferences, we use PickScore (P-Score) [34], HPS-v2 (HPS) [82], and ImageReward (I-Reward) [85]. For prompt-independent preferences, we use the AestheticScore (AES) estimator [68]. For image-text alignment, we employ CLIP [61]. We apply FID [25] to measure the distance between the generated and real distributions using 5k test images. We also use CLIP-IQA [79] to evaluate image quality (CI-Q) and sharpness (CI-S), and ATHEC [57] for additional sharpness assessment. CI-S uses pretrained CLIP, while ATHEC relies on the standard deviation of the Laplacian of image pixels. To quantify the color shift artifacts, we convert RGB images to HSV and calculate the circular mean of hue. We measure color shift severity of the target model by comparing the mean hue difference between the target model and the base model. Finally, for identity similarity in PT2I tasks, we compute feature distances using ArcFace [14] and VGGFace [5]. Higher values indicate better performance for all metrics except FID, hue distance, ArcFace, and VGGFace.

Baselines. We compare HG-DPO with several existing methods. To show that publicly available datasets may not yield the best results, we include baselines trained on the HPD v2 [82] and Pick-a-Pic v2 [34] datasets. To demonstrate that the naive approach described in Section 1 is suboptimal even with improved objectives, we introduce baselines trained with NCP-DPO [21] and MAPO [28], which have outperformed Diffusion-DPO [78]. We compare HG-DPO with Curriculum-DPO [12], which applies curriculum learning within DPO, but uses only generated images. To clearly demonstrate the effectiveness of incorporating real images, Curriculum-DPO is trained on our effective image pool dataset (Eqs. (1) and (2)). Finally, we introduce AlignProp [59] as an online learning baseline using PickScore [34] as a reward model to highlight the advantage of our approach over the online learning approach.

5. Analysis on HG-DPO

We quantitatively and qualitatively analyze HG-DPO to demonstrate its effectiveness. We retain the notations in-

Model	P-Score (↑)	HPS (↑)	I-Reward (↑)	AES (↑)	CLIP (↑)	FID (↓)	CI-Q (↑)	CI-S (↑)	ATHEC (↑)	Hue (↓)
Base (ϵ_{base})	21.7364	0.2819	-0.0665	6.1061	29.72	37.34	**0.9058**	**0.9573**	18.73	-
$N = 2$	22.1939	0.2854	0.3610	6.1408	30.66	**34.44**	0.8887	0.9472	18.96	**10.24**
$N > 2$	**22.5688**	0.2872	**0.7830**	**6.2544**	31.50	37.29	0.8879	0.9471	**27.20**	98.54
$N > 2 + \beta \uparrow$	22.2506	0.2864	0.5435	6.1129	31.30	36.00	0.8416	0.9141	19.17	23.77
$N > 2 + \mathcal{L}_{stat}$ ($\epsilon_{\mathbb{E}}$)	22.5384	**0.2878**	0.7146	6.1775	**31.56**	36.00	0.9057	0.9547	19.58	27.94

Table 3. **Quantitative analysis of the easy stage.** For $\mathcal{D}_{\mathbb{E}}$, $N = 2$ generates exactly two images per prompt, while $N > 2$ builds an image pool as defined in Eq. (1). $N > 2 + \beta \uparrow$ and $N > 2 + \mathcal{L}_{stat}$ add regularization to address the color shift artifacts in $N > 2$. Specifically, $N > 2 + \beta \uparrow$ applies a higher β, which is a strength of the original regularization in $\mathcal{L}_{D\text{-}DPO}$, and $N > 2 + \mathcal{L}_{stat}$ integrates $\mathcal{L}_{stat}$ (Eq. (4)). $N > 2 + \mathcal{L}_{stat}$, which is highlighted in blue, is our proposed training configuration for the easy stage.

troduced in Section 3. In Tables 1, 2, 3, and 4, **bold** text and underlined text indicate the best and second-best results, respectively. The row corresponding to our final model, HG-DPO, or the proposed training configuration is highlighted in blue in each table. Additional results not included in this section including a user study are provided in the Appendices.

5.1. Comparison with the Previous Methods

In this section, we compare HG-DPO with existing methods discussed in Section 4. As shown in Figure 4, HG-DPO produces more natural poses (first row) and better text alignment (second row) than models trained on the HPD v2 [82] and Pick-a-Pic v2 [34] datasets. This highlights the limitations of relying solely on public datasets for human image generation, supported by the quantitative results in Table 1.

We also evaluate baselines trained with the naive single-stage approach, where real images are treated as winning images and generated ones as losing images, using various objective functions (Diffusion-DPO, NCP-DPO, and MAPO). However, Figure 4 shows that this naive approach fails to deliver satisfactory results, regardless of the objective function. In contrast, HG-DPO, leveraging curriculum learning, generates high-quality human images. Table 1 further confirms the superiority of our curriculum-based approach over naive methods.

Additionally, we compare HG-DPO to Curriculum-DPO, which also employs curriculum learning but is exclusively trained on generated images. In Figure 4, Curriculum-DPO fails to achieve realistic compositions, poses, and fine details compared to HG-DPO, demonstrating the benefits of including real images in training. This is consistent with the superior realism (FID and CI-Q) and sharpness (CI-S, and ATHEC) scores of HG-DPO compared to those of Curriculum-DPO in Table 1.

Finally, Figure 4 shows that AlignProp produces images with a distinct aesthetic style, deviating from the model's original style. This likely stems from optimizing only PickScore, which may cause catastrophic forgetting and mode collapse [11, 38, 78]. As a result, while AlignProp achieves higher P-Score and AES in Table 1, it scores lower on other metrics compared to HG-DPO.

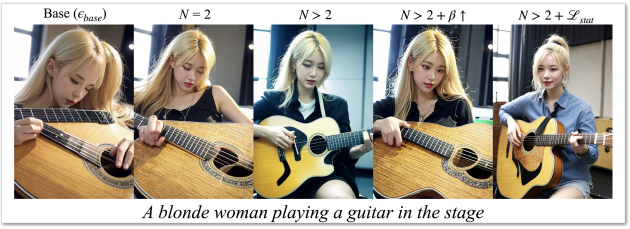

Figure 8. **Qualitative analysis of the easy stage.** In the easy stage, only the model trained with both $\mathcal{D}_{\mathbb{E}}$ and $\mathcal{L}_{stat}$, namely $N > 2 + \mathcal{L}_{stat}$, produces images without distortions in pose and color.

5.2. Progress through the HG-DPO Pipeline

We evaluate the impact of each stage in the HG-DPO pipeline. First, the naive single-stage approach, which treats real images as winning and images generated from their captions as losing, performs worse than ϵ_{base} (Figure 5 and Table 2). This underscores the challenge of using real images as winning images. Additionally, the model labeled as Naive in Table 2 corresponds to Diffusion-DPO in Table 1.

To address this, we first train ϵ_{base} to obtain $\epsilon_{\mathbb{E}}$ through the easy stage. In Table 2, $\epsilon_{\mathbb{E}}$ shows significant improvements over ϵ_{base} in human preference metrics (P-Score, HPS, I-Reward, and AES) and image-text alignment (CLIP). Figure 5 also shows that $\epsilon_{\mathbb{E}}$ produces better anatomical features than ϵ_{base}. These gains stem from the superiority of the winning images in $\mathcal{D}_{\mathbb{E}}$ over the losing images (Figure 3). However, FID and CI-Q, which measure image realism, show no significant improvement since the winning images are generated rather than real.

In the normal stage, $\epsilon_{\mathbb{N}}$ uses intermediate domains to produce more realistic images than $\epsilon_{\mathbb{E}}$. Figure 5 shows that $\epsilon_{\mathbb{N}}$ generates more realistic composition, pose, and facial quality than $\epsilon_{\mathbb{E}}$, consistent with FID and CI-Q improvements in Table 2. As realism improves, image sharpness also increases, reflected in CI-S and ATHEC scores.

Through the hard stage, $\epsilon_{\mathbb{H}}$ achieves even greater realism and sharpness than $\epsilon_{\mathbb{N}}$ (Figure 5), supported by higher CI-Q, CI-S, and ATHEC in Table 2. However, $\epsilon_{\mathbb{H}}$ shows a slight FID degradation compared to $\epsilon_{\mathbb{N}}$, though it still surpasses ϵ_{base} and $\epsilon_{\mathbb{E}}$. Given that $\epsilon_{\mathbb{H}}$ achieves higher CI-Q, CI-S, and ATHEC scores than $\epsilon_{\mathbb{N}}$, this likely reflects reduced image

Model	P-Score (↑)	HPS (↑)	I-Reward (↑)	AES (↑)	CLIP (↑)	FID (↓)	CI-Q (↑)	CI-S (↑)	ATHEC (↑)	ArcFace (↓)	VGGFace (↓)
IB [69]	21.6847	0.2807	-0.1045	6.0697	29.72	39.61	0.9034	0.9427	18.02	0.2662	72.11
HG-DPO + IB	**22.5674**	**0.2855**	**0.6864**	**6.1321**	**31.24**	**29.30**	**0.9279**	**0.9806**	**26.79**	**0.2586**	**71.45**

Table 4. **Quantitative results on PT2I.** HG-DPO brings its improvements to PT2I generation while preserving the identity injection capability of pre-trained PT2I model, InstantBooth (IB) [69]. The qualitative results are reported in the Appendices.

diversity. However, our goal prioritizes image quality over diversity. Also, $\epsilon_{\mathbb{H}}$ shows lower image-text alignment than $\epsilon_{\mathbb{E}}$ and $\epsilon_{\mathbb{N}}$, though it remains better than ϵ_{base}. This may result from the absence of the prompt-aware preference estimator (PickScore) in the hard stage, which was used in the easy and normal stages to select winning images (Eqs. (3) and (7)). To address this, we first improve visual quality by training the U-Net through the hard stage, accepting some image-text alignment degradation, and then refine the text encoder to restore image-text alignment.

Our final model, **HG-DPO** (Hard ($\epsilon_{\mathbb{H}}$) + TE), integrates text encoder enhancements, improving image-text alignment over $\epsilon_{\mathbb{H}}$ (Table 2 and Figure 6). This likely boosts prompt-aware human preference metrics (P-Score, HPS, and I-Reward) in Table 2. HG-DPO also preserves the high realism and sharpness of $\epsilon_{\mathbb{H}}$, as shown by FID, CI-Q, CI-S, and ATHEC scores. Overall, HG-DPO significantly outperforms ϵ_{base} across all metrics, aligning with Figure 1.

5.3. Necessity of the Each Proposed Stage

We demonstrate that each stage in the proposed HG-DPO pipeline is essential for achieving optimal results.

Firstly, the model labeled as E2E training in Table 2 and Figure 7 is trained in an end-to-end manner by combining training datasets from all three stages, but this approach leads to suboptimal results, showing the effectiveness of our three-stage curriculum learning.

The model labeled as Hard w/o easy, which is trained only on the normal and hard stages, also produces a noticeably degraded results as shown in Table 2 and Figure 7, underscoring the importance of the easy stage. In contrast, Hard w/o normal, trained only on the easy and hard stages, generates a more natural image than Hard w/o easy, as illustrated in Figure 7. However, when it comes to fine detail, $\epsilon_{\mathbb{H}}$, which includes the normal stage, produces a more realistic image than Hard w/o normal. Table 2 supports this finding, where $\epsilon_{\mathbb{H}}$ achieves higher scores in FID, CI-Q, CI-S, and ATHEC, indicating the normal stage's crucial role to achieve optimal results.

In addition, compared to the result for $\epsilon_{\mathbb{H}}$, the results for Base (ϵ_{base}) + SFT, Easy ($\epsilon_{\mathbb{E}}$) + SFT, and Normal ($\epsilon_{\mathbb{N}}$) + SFT in Table 2 and Figure 7 reveal that supervised fine-tuning (SFT) with the hard-stage winning images is suboptimal when applied to models trained on earlier stages. Furthermore, these three models generate similar images, suggesting that SFT, regardless of the starting point (ϵ_{base}, $\epsilon_{\mathbb{E}}$, or $\epsilon_{\mathbb{N}}$), may undesirably lead to forgetting knowledge previously learned through each stage.

5.4. Additional Analysis on the Easy Stage

The significantly degraded results of Hard w/o Easy in Table 2 and Figure 7 highlight the importance of the easy stage. To explore this further, we conduct an additional analysis of the easy stage.

In Table 3, we observe that the model labeled as $N = 2$ underperforms in human preference metrics and image-text alignment compared to configurations the model labeled as $N > 2$, highlighting the effectiveness of employing the image pool. This observation is further supported by Figure 8, where $N = 2$ results in an image with a distorted pose, similar to that produced by ϵ_{base}, whereas $N > 2$ generates an image with an undistorted pose.

However, Figure 8 shows that for $N > 2$, unnatural color shift artifacts occur, corroborated by the highest hue distance in Table 3. To mitigate these artifacts, increasing the weight of the regularization term from the original Diffusion-DPO objective significantly reduces the hue distance, as seen in the $N > 2 + \beta \uparrow$ results. However, this adjustment leads to a noticeable decline in human preference metrics and image-text alignment in Table 3. In contrast, our proposed statistics matching loss reduces the hue distance while maintaining strong performance in human preference metrics and image-text alignment, as evidenced by the results of $N > 2 + \mathcal{L}_{stat}$ in Table 3 and Figure 8.

5.5. Personalized T2I with HG-DPO

To improve PT2I generation, we adapt pre-trained HG-DPO LoRA layers to the pre-trained InstantBooth (IB) [69]. As shown in Table 4, HG-DPO + IB outperforms IB in human preferences, image-text alignment, image realism, and image sharpness while preserving identity injection, as indicated by similar ArcFace and VGGFace distances.

6. Conclusion

HG-DPO represents a significant advancement in human image generation by integrating real images and curriculum learning into the DPO framework. By gradually training the model from achieving basic anatomical accuracy to complex details of real images, HG-DPO narrows the realism gap between generated and real images. Furthermore, HG-DPO is adaptable to personalized T2I generation, consistently improving image quality. This adaptability makes it a valuable tool for creative applications and social media.

Appendices

A. Additional Results of HG-DPO

In this section, we present additional qualitative and quantitative results of HG-DPO to demonstrate the effectiveness of HG-DPO.

A.1. Text-to-Image Generation

As demonstrated in Figure 9, HG-DPO successfully generates high-quality human images with diverse actions, appearances, group sizes, and backgrounds. This is made possible by HG-DPO's effective enhancement of the base model, as demonstrated by extensive experimental results in our manuscript and Figure 10.

As a result, in Table 5, HG-DPO outperforms other existing methods. Table 5 is similar to Table 1 in the manuscript but differs in two key aspects: it includes additional baselines, DPOK [18] and D3PO [87], and uses 10 random seeds instead of a single one. To train DPOK and D3PO, we use our training prompt set $\mathcal{P}$ and PickScore [34] as the reward model. While D3PO originally uses human feedback, we follow the authors' setup by using the reward model instead. The results in Table 6, which converts Table 5 to samplewise win rates, further highlight the effectiveness of HG-DPO.

Furthermore, HG-DPO significantly outperforms the base model and the previous approaches in the user study, as shown in Figure 11. In the user study, we evaluated a selected subset of the baselines introduced in Section 4 of our manuscript against HG-DPO. Specifically, since the model trained with HPD [82] yields results similar to the model trained with Pick-a-Pic [34] (see Figure 4 in our manuscript), we compared HG-DPO exclusively with the model trained using Pick-a-Pic [34], which is widely used in DPO-related studies. Furthermore, we excluded Diffusion-DPO [78], NCP-DPO [21], and MAPO [28] from the user study because these models often failed to generate images reliably and exhibited severe artifacts (see Figure 4 in our manuscript).

A.2. Personalized Text-to-Image Generation

HG-DPO significantly improves personalized text-to-image (PT2I) generation. As shown in Figure 12, this allows the generation of high-quality images that accurately reflect specific identities. Notably, these improvements are achieved without compromising the identity injection capability of the base PT2I model.

B. Additional Analysis on the Easy Stage

In this section, we present additional experimental results and further analysis of the easy stage.

B.1. Effectiveness of the Easy Stage

In the easy stage, we refine the base model to generate images that align more closely with human preferences as shown in Figure 13. Specifically, the model is improved to produce images with undistorted poses and anatomies and stronger alignment with the given prompts.

B.2. Image Pool with AI Feedback

In our manuscript, we propose a method for selecting winning and losing images from the image pool using AI feedback (PickScore [34]). This method assumes that a larger PickScore difference between the winning and losing images indicates greater semantic differences, which are crucial for enhancing the model through DPO and align better with actual human preferences. As shown in Figure 14, comparing the image with the highest PickScore to the image with the l-th highest PickScore reveals that the semantic differences between the two images (e.g., anatomy, pose, and text-image alignment) become more pronounced as l increases. By choosing the images with the highest and the 20th highest PickScores as the winning and losing images, respectively, we accentuate the semantic differences between them, better reflecting human preferences.

B.3. Statistics Matching Loss

In this section, we further analyze the statistics matching loss.

B.3.1. Hypothesis test

Here, we validate the hypothesis underlying the statistics matching loss, $\mathcal{L}_{stat}$. Let us denote the model obtained by training ϵ_{base} through the easy stage without $\mathcal{L}_{stat}$ as $\hat{\epsilon_{\mathbb{E}}}$. $\hat{\epsilon_{\mathbb{E}}}$ is a model that suffers from the color shift artifacts. As explained in our manuscript, we hypothesize that the cause of the color shift artifacts is the divergence between the latent statistics sampled by $\hat{\epsilon_{\mathbb{E}}}$ and those of ϵ_{base} during inference. $\mathcal{L}_{stat}$ is designed to prevent such divergence based on this assumption.

To verify our hypothesis more directly, we design an inference-time statistics matching approach called *latent adaptive normalization* (*LAN*). If the gaps in the channel-wise statistics of the latents during inference cause the color shift artifacts, then eliminating those gaps should resolve those artifacts.

Let $\hat{h}_{\mathbb{E}}^{t-1}$ and h_{base}^{t-1} denote the latents sampled from the same random noise using $\hat{\epsilon_{\mathbb{E}}}$ and ϵ_{base} at inference time with timestep t, respectively. Formally, we define

$$\hat{h}_{\mathbb{E}}^{t-1} = \psi(h_{\mathbb{E}}^{t}, p, t, \hat{\epsilon_{\mathbb{E}}}) \tag{10}$$

$$h_{base}^{t-1} = \psi(h_{base}^{t}, p, t, \epsilon_{base}) \tag{11}$$

where ψ denotes a inference-time latent sampler and p denotes an inference prompt. Then, we define LAN as fol-

Figure 9. **Qualitative results of HG-DPO.** HG-DPO is capable of effectively generating high-quality human images that encompass a wide range of actions, appearances, group sizes, and backgrounds.

Model	P-Score (↑)	HPS (↑)	I-Reward (↑)	AES (↑)	CLIP (↑)	FID (↓)	CI-Q (↑)	CI-S (↑)	ATHEC (↑)
HPD v2	21.7211	0.2821	-0.1353	6.0928	29.71	39.53	0.8856	0.9507	17.45
Pick-a-Pic v2	21.6778	0.2821	-0.1352	6.0999	29.72	40.85	0.8614	0.9383	17.43
Diffusion-DPO	18.0731	0.2408	-1.9616	5.0637	23.49	160.11	0.6638	0.8715	**40.31**
NCP-DPO	17.4631	0.2327	-2.0222	4.7983	21.53	184.81	0.6342	0.8236	12.09
MAPO	20.3971	0.2692	-0.5150	5.4260	28.22	63.33	0.6459	0.7566	30.71
Curriculum-DPO	22.4298	0.2868	0.5823	6.1874	31.43	37.02	0.8857	0.9528	21.63
AlignProp	**22.8933**	0.2843	0.0693	**6.2670**	29.50	53.87	0.8534	0.9609	15.67
DPOK	21.6709	0.2809	-0.2344	6.0998	29.25	41.52	0.8756	0.9332	15.68
D3PO	21.6905	0.2810	-0.1914	6.0764	29.59	41.26	0.8902	0.9508	17.41
HG-DPO (Ours)	22.5781	**0.2871**	**0.7384**	6.1758	**31.53**	**30.91**	**0.9327**	**0.9852**	28.28

Table 5. **Quantitative comparison with the previous methods.** HG-DPO achieves superior performance over the existing methods across nearly all evaluation metrics. **Bold** text and underlined text indicate the best and second-best results, respectively. The row corresponding to our final model, HG-DPO, is highlighted in blue. For a more accurate comparison, we evaluate using 10 random seeds.

Model	P-Score (↑)	HPS (↑)	I-Reward (↑)	AES (↑)	CLIP (↑)	CI-Q (↑)	CI-S (↑)	ATHEC (↑)
vs HPD v2	85.13 %	76.44 %	82.25 %	62.17 %	73.15 %	82.31 %	86.64 %	93.75 %
vs Pick-a-Pic v2	86.03 %	76.14 %	82.08 %	61.06 %	72.64 %	89.63 %	90.87 %	93.79 %
vs Diffusion-DPO	99.97 %	99.96 %	99.67 %	96.91 %	96.48 %	96.74 %	88.17 %	27.24 %
vs NCP-DPO	99.97 %	99.85 %	99.78 %	95.04 %	98.95 %	99.61 %	96.94 %	97.02 %
vs MAPO	97.92 %	98.35 %	88.51 %	96.26 %	84.78 %	98.89 %	98.10 %	42.07 %
vs Curriculum-DPO	60.85 %	51.86 %	57.10 %	49.51 %	50.21 %	84.35 %	88.66 %	82.74 %
vs AlignProp	33.82 %	62.12 %	75.80 %	37.73 %	74.02 %	95.35 %	85.45 %	97.98 %
vs DPOK	86.19 %	80.71 %	84.06 %	61.80 %	77.67 %	85.67 %	91.82 %	95.91 %
vs D3PO	85.90 %	81.70 %	83.69 %	64.16 %	74.82 %	81.09 %	87.46 %	92.90 %

Table 6. **Samplewise win rates (%) of HG-DPO against the previous methods.** HG-DPO achieves superior performance over the existing methods across nearly all evaluation metrics. This table converts Table 5 into win rates, which means that these results are also calculated using 10 random seeds.

Figure 10. **Qualitative enhancements in text-to-image generation through HG-DPO.** HG-DPO improves the base model's capability to generate human images with more realistic poses and anatomies that align more accurately with the given prompt.

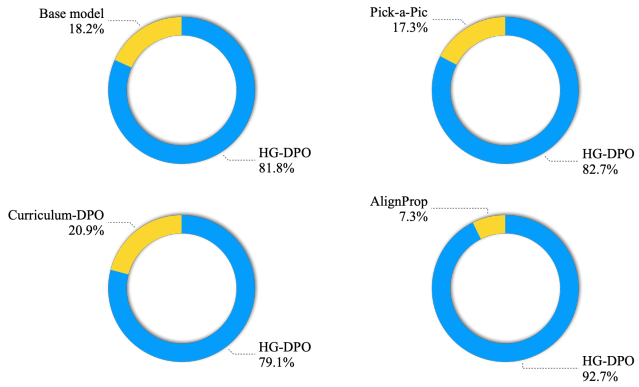

Figure 11. **User studies comparing HG-DPO and baselines.** HG-DPO demonstrates superior performance compared to the base model and previous approaches in human evaluation. Participants were tasked with choosing the image that exhibited higher realism and better alignment with the given prompt from the outputs of the two models. The detailed process for conducting the user study is described in Section F.5.

Figure 12. **Qualitative advancements achieved through in personalized text-to-image (PT2I) generation through HG-DPO.** HG-DPO improves the base model's capability to generate human images with more realistic poses and anatomies that align more accurately with the given prompt, and these improvements extend to PT2I generation. As a result, we can produce high-quality images that accurately reflect the identity of the concept image shown in the bottom left.

lows:

$$h_{\mathbb{E}}^{t-1} = \left(\frac{\hat{h}_{\mathbb{E}}^{t-1} - \mu(\hat{h}_{\mathbb{E}}^{t-1})}{\sigma(\hat{h}_{\mathbb{E}}^{t-1})} \right) \sigma(h_{base}^{t-1}) + \mu(h_{base}^{t-1}) \quad (12)$$

where μ and σ calculate the channel-wise mean and standard deviation from the input, respectively. $h_{\mathbb{E}}^{t-1}$ is used in Eq. (10) of the supplementary material at the next inference timestep.

Table 7 reveals that $N > 2$ ($\hat{\epsilon_{\mathbb{E}}}$) + LAN significantly reduces the hue distance compared to $N > 2$ ($\hat{\epsilon_{\mathbb{E}}}$). Furthermore, $N > 2$ ($\hat{\epsilon_{\mathbb{E}}}$) + LAN achieves comparable performance to $N > 2$ ($\hat{\epsilon_{\mathbb{E}}}$) in human preference metrics (P-Score, HPS, I-Reward, and AES) and image-text alignment (CLIP). These findings validate LAN's effectiveness in addressing the color shift artifacts and support the hypothesis underlying the design of $\mathcal{L}_{stat}$.

However, because LAN requires additional sampling from ϵ_{base} during inference, it incurs higher computational costs during inference compared to $N > 2 + \mathcal{L}_{stat}$. For this reason, we propose $\mathcal{L}_{stat}$ as a more computationally efficient solution to mitigate the color shift artifacts.

B.3.2. What causes the color shift artifacts?

The color shift artifacts arise from the deviation of the channel-wise statistics of latents sampled using $\hat{\epsilon_{\mathbb{E}}}$ from those sampled using ϵ_{base}, as demonstrated by the effectiveness of LAN in the previous paragraph. Here, to find the cause of this deviation, we further analyze the winning and losing images used in the easy stage. Specifically, we calculate the cosine distance of channel-wise statistics of

Figure 13. **Qualitative advancements achieved through the easy stage.** We enhance the base model through the easy stage to generate images that better align with human preferences. Specifically, the model is improved to produce images with undistorted poses and anatomies and stronger alignment with the given prompts.

encoded latents of winning and losing images. In Table 8, the results reveal that the cosine distance between the latents' means for the winning and losing images is *0.2035*, while the cosine distance for their standard deviations is *0.005*. Since DPO trains the model to learn the differences between winning and losing images, it can be inferred that the differences in the channel-wise **mean** values of latents present in the dataset are also learned by the model. This can encourage the model to shift the mean of the sampled latents far from that of the losing image and close to that of the winning image.

B.3.3. Why is it sufficient to match only the mean?

$\mathcal{L}_{stat}$ mitigates the color shift by preventing the aforementioned mean shift through the mean matching loss. Interestingly, as reported in the previous paragraph, we can observe that the cosine distance of standard deviation between the latents of winning and losing images is close to zero. We believe this is why matching only the mean in $\mathcal{L}_{stat}$ is sufficient to prevent the color shift artifacts.

B.3.4. Importance of the statistics matching loss

As illustrated in Figure 15, the absence of $\mathcal{L}_{stat}$ results in generated images appearing unnatural due to the color shift artifacts. Incorporating $\mathcal{L}_{stat}$ effectively eliminates these artifacts, producing noticeably more natural images.

C. Additional Analysis on the Normal Stage

In this section, we present additional experimental results and further analysis of the normal stage.

C.1. Effectiveness of the Normal Stage

We further explore the role of the normal stage, which refines $\epsilon_{\mathbb{E}}$, derived from the easy stage, to produce $\epsilon_{\mathbb{N}}$. While the easy stage enables $\epsilon_{\mathbb{E}}$ to generate images aligned with human preferences resulting in undistorted anatomical features and poses, they still fall short of achieving the realism found in real human portrait images. For example, as shown in Figure 16, although the poses are largely free from distortion, they still appear somewhat unnatural compared to those in real photographs. The normal stage enhances $\epsilon_{\mathbb{E}}$ by improving its ability to generate compositions and poses that are not only distortion-free but also realistic, closely mirroring those found in the real dataset. Figure 16 illustrates that $\epsilon_{\mathbb{N}}$ achieves significantly more realistic compositions and poses, derived from real human portrait images, than $\epsilon_{\mathbb{E}}$.

C.2. Intermediate Domains

In the normal stage, we introduce intermediate domains for winning images. Figure 17 illustrates the outcomes of the *SDRecon* operation used to create these intermediate domains, along with the winning images employed during the normal stage.

C.2.1. Intermediate domains with SDRecon

As shown in Figure 17, we use 10 intermediate domains, labeled from t_1 to t_T. While t_1 is nearly identical to a real image, t_T resembles a generated image, retaining little of the real image's original features. As the transition progresses from t_1 to t_T, the characteristics of the real image gradually fade, increasingly resembling those of a generated image.

Figure 14. **Visualization of the image pool.** This figure shows the image pool with the size of 20 for the prompt in the leftmost column. The column labeled as 1st contains images with the highest PickScore, while the column labeled as 20th contains images with the 20th highest PickScore, i.e., the lowest PickScore, in the image pool. By selecting the image with the highest PickScore from this image pool as the winning image and the image with the 20th highest PickScore as the losing image, we magnify the semantic differences between the winning and losing images.

Model	P-Score (↑)	HPS (↑)	I-Reward (↑)	AES (↑)	CLIP (↑)	FID (↓)	CI-Q (↑)	CI-S (↑)	ATHEC (↑)	Hue (↓)
Base (ϵ_{base})	21.7364	0.2819	-0.0665	6.1061	29.72	37.34	<u>0.9058</u>	**0.9573**	18.73	-
$N = 2$	22.1939	0.2854	0.3610	6.1408	30.66	**34.44**	0.8887	0.9472	18.96	**10.24**
$N > 2$ ($\hat{\epsilon_{\mathbb{E}}}$)	<u>22.5688</u>	0.2872	**0.7830**	**6.2544**	31.50	37.29	0.8879	0.9471	**27.20**	98.54
$N > 2$ ($\hat{\epsilon_{\mathbb{E}}}$) + $\beta \uparrow$	22.2506	0.2864	0.5435	6.1129	31.30	<u>36.00</u>	0.8416	0.9141	19.17	23.77
$N > 2$ ($\hat{\epsilon_{\mathbb{E}}}$) + LAN	**22.6474**	**0.2885**	<u>0.7677</u>	<u>6.1899</u>	**31.60**	37.08	**0.9086**	0.9521	18.65	<u>16.13</u>
$N > 2 + \mathcal{L}_{stat}$ ($\epsilon_{\mathbb{E}}$)	22.5384	<u>0.2878</u>	0.7146	6.1775	<u>31.56</u>	<u>36.00</u>	0.9057	<u>0.9547</u>	<u>19.58</u>	27.94

Table 7. **Quantitative analysis of the easy stage.** For $\mathcal{D}_{\mathbb{E}}$, $N = 2$ generates exactly two images per prompt, while $N > 2$ builds an image pool. $N > 2 + \beta \uparrow$, $N > 2+$ LAN, and $N > 2 + \mathcal{L}_{stat}$ add regularization to address the color shift artifacts in $N > 2$. Specifically, $N > 2 + \beta \uparrow$ applies a higher β, which is a strength of the original regularization in $\mathcal{L}_{D\text{-}DPO}$, $N > 2+$ LAN applies latent adaptive normalization (Section B.3), and $N > 2 + \mathcal{L}_{stat}$ integrates $\mathcal{L}_{stat}$. **Bold** text and <u>underlined</u> text indicate the best and second-best results, respectively. The row corresponding to the proposed training configuration in the easy stage is highlighted in blue.

	Mean	Standard deviation
Cosine distance	0.2035	0.0005

Table 8. **Difference of channel-wise statistics between winning and losing images.** Cosine distance of channel-wise statistics of encoded latents of winning and losing images. For the encoding, we use the encoder of VAE [33] used in HG-DPO.

Specifically, fine-detail information is lost first, followed by the loss of pose information.

C.2.2. Winning images from the intermediate domains

As depicted in Figure 17, we select four intermediate domains, t_4 through t_7, as candidates for the winning images in the normal stage. This is because the our qualitative analysis reveals that these domains generally retain the realistic pose of the real image while exhibiting fine details resembling those of generated images. Among these candidates, the image with the highest PickScore [34] is chosen as the winning image.

D. Additional Analysis on the Hard Stage

In this section, we present additional experimental results and further analysis of the hard stage.

D.1. Effectiveness of the Hard Stage

We investigate the impact of the hard stage, which refines $\epsilon_{\mathbb{N}}$, obtained from the normal stage, to produce $\epsilon_{\mathbb{H}}$. While $\epsilon_{\mathbb{N}}$ achieves realistic composition and poses during the normal stage, it struggles to generate fine details. For instance, as shown in Figures 18, 19, 20, and 21, $\epsilon_{\mathbb{N}}$ 1) fails to accurately depict fine facial features such as eyes and lips, 2) requires better shading, and 3) suffers from image blurriness. Although these details may seem minor, they play a crucial role in achieving overall image realism. The hard stage addresses these limitations by enhancing $\epsilon_{\mathbb{N}}$, resulting in $\epsilon_{\mathbb{H}}$, which excels in generating realistic fine details. Figures 18, 19, 20, and 21 illustrate that $\epsilon_{\mathbb{H}}$ effectively produces fine details that $\epsilon_{\mathbb{N}}$ cannot, significantly improving image realism. As shown in Figure 22, in a user study comparing $\epsilon_{\mathbb{N}}$ and $\epsilon_{\mathbb{H}}$, $\epsilon_{\mathbb{H}}$ is rated higher, further demonstrating its effectiveness.

Figure 15. **Qualitative enhancements achieved with the statistics matching loss.** The statistics matching loss effectively removes the color shift artifacts, leading to the generation of significantly more natural images.

Figure 16. **Qualitative advancements achieved through the normal stage.** $\epsilon_{\mathbb{N}}$, derived by refining $\epsilon_{\mathbb{E}}$ through the normal stage, generates images with more **realistic compositions and poses** compared to $\epsilon_{\mathbb{E}}$.

D.2. Winning Images of the Hard Stage

In the hard stage, we employ images from the intermediate domain t_1 as winning images instead of real images. As illustrated in Figure 17, images from the intermediate domain t_1 are visually nearly indistinguishable from real human portrait images, making this approach effectively comparable to using real images directly as winning images. This choice is motivated by the observation that, while real images and intermediate domain t_1 images appear almost identical to the human eye, utilizing intermediate domain images leads to slightly better quantitative performance. Specifically, as demonstrated in Table 9, the model trained with intermediate domain t_1 images achieves results similar to those trained with real images, with a slight improvement in CI-Q scores.

D.3. Effectiveness of the Enhanced Text Encoder

We train the text encoder during the easy stage to enhance image-text alignment and employ it alongside $\epsilon_{\mathbb{H}}$, derived from the hard stage, during inference. As shown in Figure 23, the enhanced text encoder effectively improves image-text alignment without compromising the image quality achieved by $\epsilon_{\mathbb{H}}$.

E. Limitations

Through a three-stage training pipeline, HG-DPO enhances the base model to generate not only realistic anatomical features and poses but also fine details with greater realism.

Figure 17. **Visualization of the intermediate domains.** The images labeled t_1 to t_T are reconstructed from real images using the *SDRecon* operation. The image labeled generated image is produced via text-to-image generation based on the caption of the real image. As the labels progress toward t_T, SDRecon applies increasingly stronger noise to the real image, causing it to lose more of its original characteristics and resemble the generated image more closely. For the normal stage, we select four intermediate domains, t_4 to t_7, as candidates for winning images, because they maintain the realistic pose of the real image while adopting the fine details typical of the generated image. The image with the highest PickScore among these candidates is chosen as the winning image.

Figure 18. **Qualitative advancements achieved through the hard stage.** $\epsilon_{\mathbb{H}}$, derived by refining $\epsilon_{\mathbb{N}}$ through the hard stage, generates finer details, especially more realistic depictions of the **eyes**, compared to $\epsilon_{\mathbb{N}}$ as shown in the red box.

Figure 19. **Qualitative advancements achieved through the hard stage.** $\epsilon_{\mathbb{H}}$, derived by refining $\epsilon_{\mathbb{N}}$ through the hard stage, generates finer details, especially more realistic depictions of the **gaze**, compared to $\epsilon_{\mathbb{N}}$ as shown in the red box.

Figure 20. **Qualitative advancements achieved through the hard stage.** $\epsilon_{\mathbb{H}}$, derived by refining $\epsilon_{\mathbb{N}}$ through the hard stage, generates finer details, especially more realistic depictions of the **lips**, compared to $\epsilon_{\mathbb{N}}$ as shown in the red box.

Figure 21. **Qualitative advancements achieved through the hard stage.** $\epsilon_{\mathbb{H}}$, derived by refining $\epsilon_{\mathbb{N}}$ through the hard stage, generates **sharper** images with improved fine details, particularly exhibiting more vivid and realistic **shading**, compared to $\epsilon_{\mathbb{N}}$.

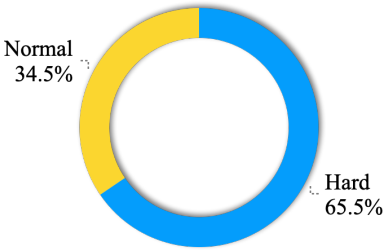

Figure 22. **User study comparing a model trained up to the normal stage ($\epsilon_{\mathbb{N}}$) with one trained through the hard stage ($\epsilon_{\mathbb{H}}$).** Participants were tasked with choosing the image that exhibited higher realism and better alignment with the given prompt from the outputs of the two models. The model trained through the hard stage achieves higher human evaluation scores due to its ability to generate finer details with greater realism compared to the model trained only up to the normal stage.

Despite these improvements, HG-DPO does not address the generation of realistic fingers. As shown in Figure 24, HG-DPO produces an image with sharper and more realistic fine details compared to the base model. However, the generated fingers remain notably unrealistic.

F. Implementation Details

In this section, we provide implementation details on training and inference.

F.1. Details on Supervised Fine-Tuning

First, we introduce the method for obtaining ϵ_{base} through supervised fine-tuning.

Text-to-image dataset. We collected approximately 300k high-quality human images. Each image has a resolution of 704×1024. We use `LLaVa` [41] to generate text prompts for all the collected images for training. This text-to-image dataset corresponds to $\mathcal{D}_{\text{real}}$ in our manuscript.

Furthermore, we use Qwen2-VL [81] for visual question answering to analyze distribution of this dataset, which includes 40.7% male and 59.3% female, and 24.45% child, 2.82% teenager, 41.00% youth, 31.61% adult, and 0.12% elderly. While the proportions of teenagers and elderly appear small, images in these groups may have been reasonably classified into adjacent categories (*e.g.*, teenagers as child/youth, elderly as adult).

Architecture. We employ Stable Diffusion 1.5 (SD1.5) [64], which is pre-trained with large text-to-image datasets, as our backbone model. More specifically, we use majicmix-v7 [1], a fine-tuned model of SD1.5 specialized in human generation. We further fine-tune this backbone model with $\mathcal{D}_{\text{real}}$, to obtain our base model, ϵ_{base}.

Figure 23. **Qualitative advancements achieved through the text encoder enhancement.** By training the text encoder through the easy stage and incorporating it with $\epsilon_{\mathbb{H}}$ during inference, we achieve improved image-text alignment compared to using $\epsilon_{\mathbb{H}}$ alone. Moreover, the use of the enhanced text encoder does not compromise the image quality produced by $\epsilon_{\mathbb{H}}$.

Loss function. For fine-tuning, we use the noise prediction loss [27]. Also, we use DDPM noise scheduler [27] for the forward diffusion process during training.

F.2. Details on HG-DPO Training

In this section, we provide details on how to improve ϵ_{base} using HG-DPO.

Model	P-Score (↑)	HPS (↑)	I-Reward (↑)	AES (↑)	CLIP (↑)	FID (↓)	CI-Q (↑)	CI-S (↑)	ATHEC (↑)
Real	**22.4773**	0.2857	0.5388	6.1953	30.99	**28.56**	0.9298	**0.9885**	29.13
Intermediate t_1	22.4698	**0.2867**	**0.5791**	**6.1955**	**31.15**	28.66	**0.9365**	0.9859	**30.08**

Table 9. **Quantitative results based on the type of images used as winning images in the hard stage.** The row labeled *Real* displays the results for the model trained with real images as winning images, while the row labeled *Intermediate* t_1 shows the results for the model trained using images from the intermediate domain t_1 as winning images. **Bold** text indicates the best results. The row corresponding to the proposed training configuration in the hard stage is highlighted in blue.

F.2.1. Architecture

U-Net. Instead of training the all parameters of ϵ_{base} through HG-DPO, we attach LoRA [29] layers to the all linear layers in the attention modules and only train them. We set LoRA rank as 8.

Text encoder. When training the text encoder, we also attach LoRA [29] layers to the all linear layers in the attention modules and only train them. For the text encoder, we set LoRA rank as 64.

F.2.2. Loss function

DPO loss. We adopt the objective function of Diffusion-DPO ($\mathcal{L}_{D\text{-}DPO}$) [78] with $\beta = 2500$. For $\mathcal{L}_{D\text{-}DPO}$, we use DDPM noise scheduler [27] for the forward diffusion process.

Statistics matching loss. For the statistics matching loss ($\mathcal{L}_{stat}$), we set $\lambda_{stat} = 10000$. Also, for the latent sampling in $\mathcal{L}_{stat}$, we use DDPM sampler [27]. We tried DDIM sampler [72], but there was no significant difference. In addition, classifier-free guidance [26] is not used during the latent sampling in $\mathcal{L}_{stat}$.

F.2.3. Optimization

For the optimization, we set the local batch size to four, which corresponds to the total batch size to 16 because we used four NVIDIA A100 GPUs. As an optimizer, we use the 8-bit Adam optimizer [15] with β_1 and β_2 of the Adam optimizer to 0.9 and 0.999, respectively, and the learning rate to $1e-5$. Additionally, we utilize mixed precision for efficient training. For the easy, normal, and hard stages, we update the model for 300k, 20k, and 20k steps, respectively.

F.2.4. Dataset

Image pool. For the image pool generation, we simply use the prompt set from $\mathcal{D}_{\text{real}}$. Furthermore, as shown in Figure 14, we generate 20 images per prompt for the image pool, which corresponds to $N = 20$ in our manuscript.

Intermediate domains. For the intermediate domains, we introduce 10 intermediate domains from t_1 to $t_{T=10}$ as shown in Figure 17. These 10 domains are generated by evenly dividing the diffusion timesteps from 1 to 1000 into 10 intervals. Specifically, we set $t_1 = 100$, $t_2 = 200$, ..., $t_T = 1000$. Then, we set $t_r = t_4$ and $t_g = t_7$ for candidates of winning images as shown in Figure 17.

Figure 24. **Qualitative results illustrating the limitations of HG-DPO.** While HG-DPO significantly improves the base model in generating more realistic human images, it still struggles to accurately synthesize fingers.

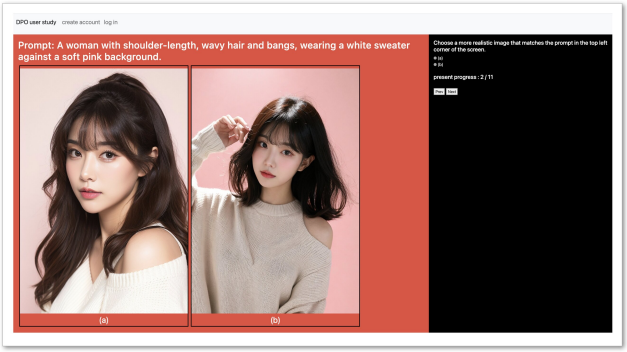

Figure 25. **User study interface.** We conduct the user study by providing a prompt and two images, asking users to choose the one that appeared more realistic considering the given prompt.

F.3. Adaptation to Personalized T2I model

To adapt HG-DPO to the personalized T2I model, we firstly trained InstantBooth [69] using ϵ_{base} as the backbone. After training InstantBooth, we can seamlessly adapt the pre-trained HG-DPO LoRA layers to InstantBooth because they share the same backbone, ϵ_{base}.

F.4. Details on Image Sampling

Sampling method. *DPMSolverMultistepScheduler* [46] in `diffusers` [77] is used with the step size of 50 for sampling the images, using classifier-free guidance [26] with the guidance scale of 5.0.

LoRA configuration. In addition, the LoRA weight of 0.5 is applied to both the U-Net and the text encoder. The LoRA layers in the text encoder are specifically trained to improve image-text alignment rather than visual quality, so they are applied only to a subset of inference timesteps near the noise. Specifically, the text encoder's LoRA layers are activated during inference timesteps 900 to 1000. Additionally, as $\epsilon_{\mathbb{H}}$ focuses on enhancing visual fine details, its LoRA layers are applied solely to the upsampling blocks of the U-Net, while the remaining U-Net blocks are frozen. This approach is chosen because qualitative analysis suggested that applying $\epsilon_{\mathbb{H}}$'s LoRA layers to all U-Net blocks reduces image diversity. This method allows for improved image quality while preserving diversity as much as possible.

F.5. Details on User Study

In Figures 11 and 22, we present the results of user studies. Each participant was tasked with selecting one of two images that best aligned with the given prompt and appeared more realistic. Here, these two images are generated by the models being compared. Evaluations were conducted using a web-based user interface, as illustrated in Figure 25.

G. Broader Impacts

We recognize the potential negative societal impacts of our work. Since our method can generate high-quality human images, it could be misused to create malicious fake images, especially when combined with personalized T2I models. It can cause significant harm to specific individuals. However, our work can also have positive impacts on society when used beneficially, such as in the entertainment or film industries. For instance, users can create desired high-quality profile pictures using text input. It highlights the beneficial uses of our work.

References

[1] majicmix realistic. https://civitai.com/models/43331/majicmix-realistic. 22

[2] Siqi Bao, Huang He, Fan Wang, Hua Wu, Haifeng Wang, Wenquan Wu, Zhen Guo, Zhibin Liu, and Xinchao Xu. Plato-2: Towards building an open-domain chatbot via curriculum learning. *arXiv preprint arXiv:2006.16779*, 2020. 3

[3] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48, 2009. 2, 3

[4] Barış Büyüktaş, Çiğdem Eroğlu Erdem, and Tanju Erdem. Curriculum learning for face recognition. In *2020 28th European Signal Processing Conference (EUSIPCO)*, pages 650–654. IEEE, 2021. 3

[5] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 67–74. IEEE, 2018. 6

[6] Chaofeng Chen, Annan Wang, Haoning Wu, Liang Liao, Wenxiu Sun, Qiong Yan, and Weisi Lin. Enhancing diffusion models with text-encoder reinforcement learning. *arXiv preprint arXiv:2311.15657*, 2023. 5

[7] Hong Chen, Yipeng Zhang, Simin Wu, Xin Wang, Xuguang Duan, Yuwei Zhou, and Wenwu Zhu. Disenbooth: Identity-preserving disentangled tuning for subject-driven text-to-image generation. In *The Twelfth International Conference on Learning Representations*, 2023. 2

[8] Xi Chen, Lianghua Huang, Yu Liu, Yujun Shen, Deli Zhao, and Hengshuang Zhao. Anydoor: Zero-shot object-level image customization. *arXiv preprint arXiv:2307.09481*, 2023. 2

[9] Zixiang Chen, Yihe Deng, Huizhuo Yuan, Kaixuan Ji, and Quanquan Gu. Self-play fine-tuning converts weak language models to strong language models. *arXiv preprint arXiv:2401.01335*, 2024. 3

[10] Pengyu Cheng, Yifan Yang, Jian Li, Yong Dai, and Nan Du. Adversarial preference optimization. *arXiv preprint arXiv:2311.08045*, 2023. 3

[11] Kevin Clark, Paul Vicol, Kevin Swersky, and David J Fleet. Directly fine-tuning diffusion models on differentiable rewards. *arXiv preprint arXiv:2309.17400*, 2023. 7

[12] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, Nicu Sebe, and Mubarak Shah. Curriculum direct preference optimization for diffusion and consistency models. *arXiv preprint arXiv:2405.13637*, 2024. 2, 3, 5, 6

[13] Josef Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang. Safe rlhf: Safe reinforcement learning from human feedback. *arXiv preprint arXiv:2310.12773*, 2023. 3

[14] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019. 6

[15] Tim Dettmers, Mike Lewis, Sam Shleifer, and Luke Zettlemoyer. 8-bit optimizers via block-wise quantization. *arXiv preprint arXiv:2110.02861*, 2021. 23

[16] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 2

[17] Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. Kto: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*, 2024. 2, 3

[18] 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. 3, 9

[19] Meng Fang, Tianyi Zhou, Yali Du, Lei Han, and Zhengyou Zhang. Curriculum-guided hindsight experience replay. *Advances in neural information processing systems*, 32, 2019. 3

[20] Carlos Florensa, David Held, Markus Wulfmeier, Michael Zhang, and Pieter Abbeel. Reverse curriculum generation for reinforcement learning. In *Conference on robot learning*, pages 482–495. PMLR, 2017. 3

[21] Alexander Gambashidze, Anton Kulikov, Yuriy Sosnin, and Ilya Makarov. Aligning diffusion models with noise-conditioned perception. *arXiv preprint arXiv:2406.17636*, 2024. 2, 3, 5, 6, 9

[22] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 2

[23] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10696–10706, 2022. 2

[24] Yi Gu, Zhendong Wang, Yueqin Yin, Yujia Xie, and Mingyuan Zhou. Diffusion-rpo: Aligning diffusion models through relative preference optimization. *arXiv preprint arXiv:2406.06382*, 2024. 2, 3

[25] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 6

[26] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 23, 24

[27] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2, 22, 23

[28] Jiwoo Hong, Sayak Paul, Noah Lee, Kashif Rasul, James Thorne, and Jongheon Jeong. Margin-aware preference optimization for aligning diffusion models without reference. *arXiv preprint arXiv:2406.06424*, 2024. 2, 3, 5, 6, 9

[29] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen.

Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 3, 23

[30] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 2

[31] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020.

[32] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *Advances in neural information processing systems*, 34:852–863, 2021. 2

[33] Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 15

[34] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 3, 5, 6, 7, 9, 15

[35] Tom Kocmi and Ondrej Bojar. Curriculum learning and minibatch bucketing in neural machine translation. *arXiv preprint arXiv:1707.09533*, 2017. 3

[36] Tomasz Korbak, Kejian Shi, Angelica Chen, Rasika Vinayak Bhalerao, Christopher Buckley, Jason Phang, Samuel R Bowman, and Ethan Perez. Pretraining language models with human preferences. In *International Conference on Machine Learning*, pages 17506–17533. PMLR, 2023. 3

[37] M Pawan Kumar, Haithem Turki, Dan Preston, and Daphne Koller. Learning specific-class segmentation from diverse data. In *2011 International conference on computer vision*, pages 1800–1807. IEEE, 2011. 3

[38] Yanyu Li, Xian Liu, Anil Kag, Ju Hu, Yerlan Idelbayev, Dhritiman Sagar, Yanzhi Wang, Sergey Tulyakov, and Jian Ren. Textcraftor: Your text encoder can be image quality controller. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7985–7995, 2024. 7

[39] Zhanhao Liang, Yuhui Yuan, Shuyang Gu, Bohan Chen, Tiankai Hang, Ji Li, and Liang Zheng. Step-aware preference optimization: Aligning preference with denoising performance at each step. *arXiv preprint arXiv:2406.04314*, 2024. 2, 3

[40] Cao Liu, Shizhu He, Kang Liu, Jun Zhao, et al. Curriculum learning for natural answer generation. In *IJCAI*, pages 4223–4229, 2018. 3

[41] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 5, 22

[42] Tianqi Liu, Yao Zhao, Rishabh Joshi, Misha Khalman, Mohammad Saleh, Peter J Liu, and Jialu Liu. Statistical rejection sampling improves preference optimization. *arXiv preprint arXiv:2309.06657*, 2023. 3

[43] Tianqi Liu, Zhen Qin, Junru Wu, Jiaming Shen, Misha Khalman, Rishabh Joshi, Yao Zhao, Mohammad Saleh, Simon Baumgartner, Jialu Liu, et al. Lipo: Listwise preference optimization through learning-to-rank. *arXiv preprint arXiv:2402.01878*, 2024.

[44] Wenhao Liu, Xiaohua Wang, Muling Wu, Tianlong Li, Changze Lv, Zixuan Ling, Jianhao Zhu, Cenyuan Zhang, Xiaoqing Zheng, and Xuanjing Huang. Aligning large language models with human preferences through representation engineering. *arXiv preprint arXiv:2312.15997*, 2023. 3

[45] Xuebo Liu, Houtim Lai, Derek F Wong, and Lidia S Chao. Norm-based curriculum learning for neural machine translation. *arXiv preprint arXiv:2006.02014*, 2020. 3

[46] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *arXiv preprint arXiv:2206.00927*, 2022. 24

[47] Sha Luo, Hamidreza Kasaei, and Lambert Schomaker. Accelerating reinforcement learning for reaching using continuous curriculum learning. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2020. 3

[48] Jian Ma, Junhao Liang, Chen Chen, and Haonan Lu. Subject-diffusion: Open domain personalized text-to-image generation without test-time fine-tuning. *arXiv preprint arXiv:2307.11410*, 2023. 2

[49] Binyamin Manela and Armin Biess. Curriculum learning with hindsight experience replay for sequential object manipulation tasks. *Neural Networks*, 145:260–270, 2022. 3

[50] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021. 4

[51] Nicola Milano and Stefano Nolfi. Automated curriculum learning for embodied agents a neuroevolutionary approach. *Scientific reports*, 11(1):8985, 2021. 3

[52] Adithyavairavan Murali, Lerrel Pinto, Dhiraj Gandhi, and Abhinav Gupta. Cassl: Curriculum accelerated self-supervised learning. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6453–6460. IEEE, 2018. 3

[53] Sanghyeon Na. Mfim: Megapixel facial identity manipulation. In *European Conference on Computer Vision*, pages 143–159. Springer, 2022. 2

[54] Sanmit Narvekar, Jivko Sinapov, Matteo Leonetti, and Peter Stone. Source task creation for curriculum learning. In *Proceedings of the 2016 international conference on autonomous agents & multiagent systems*, pages 566–574, 2016. 3

[55] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 2

[56] 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. 3

[57] Yilang Peng. Athec: A python library for computational aesthetic analysis of visual media in social science research. *Computational Communication Research*, Forthcoming. 6

[58] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 2

[59] Mihir Prabhudesai, Anirudh Goyal, Deepak Pathak, and Katerina Fragkiadaki. Aligning text-to-image diffusion models with reward backpropagation. *arXiv preprint arXiv:2310.03739*, 2023. 3, 5, 6

[60] Wei Qin, Zhenzhen Hu, Xueliang Liu, Weijie Fu, Jun He, and Richang Hong. The balanced loss curriculum learning. *IEEE Access*, 8:25990–26001, 2020. 3

[61] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 6

[62] 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, 2024. 2

[63] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022. 2

[64] 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. 2, 3, 22

[65] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023. 2

[66] Mrinmaya Sachan and Eric Xing. Easy questions first? a case study on curriculum learning for question answering. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 453–463, 2016. 3

[67] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022. 2

[68] Christoph Schuhmann. Laion-aesthetics. https://laion.ai/blog/laion-aesthetics/, 2022. 6

[69] Jing Shi, Wei Xiong, Zhe Lin, and Hyun Joon Jung. Instantbooth: Personalized text-to-image generation without test-time finetuning. *arXiv preprint arXiv:2304.03411*, 2023. 2, 8, 24

[70] Miaojing Shi and Vittorio Ferrari. Weakly supervised object localization using size estimates. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, pages 105–121. Springer, 2016. 3

[71] Feifan Song, Bowen Yu, Minghao Li, Haiyang Yu, Fei Huang, Yongbin Li, and Houfeng Wang. Preference ranking optimization for human alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 18990–18998, 2024. 3

[72] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 23

[73] 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. 2

[74] Petru Soviany, Claudiu Ardei, Radu Tudor Ionescu, and Marius Leordeanu. Image difficulty curriculum for generative adversarial networks (cugan). In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 3463–3472, 2020. 3

[75] Petru Soviany, Radu Tudor Ionescu, Paolo Rota, and Nicu Sebe. Curriculum self-paced learning for cross-domain object detection. *Computer Vision and Image Understanding*, 204:103166, 2021.

[76] Ye Tang, Yu-Bin Yang, and Yang Gao. Self-paced dictionary learning for image classification. In *Proceedings of the 20th ACM international conference on Multimedia*, pages 833–836, 2012. 3

[77] Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, Dhruv Nair, Sayak Paul, William Berman, Yiyi Xu, Steven Liu, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers, 2022. 24

[78] 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. *arXiv preprint arXiv:2311.12908*, 2023. 2, 3, 5, 6, 7, 9, 23

[79] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *AAAI*, 2023. 6

[80] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 5

[81] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 22

[82] Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score

v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023. 2, 3, 5, 6, 7, 9

[83] Zeqiu Wu, Yushi Hu, Weijia Shi, Nouha Dziri, Alane Suhr, Prithviraj Ammanabrolu, Noah A Smith, Mari Ostendorf, and Hannaneh Hajishirzi. Fine-grained human feedback gives better rewards for language model training. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[84] Guangxuan Xiao, Tianwei Yin, William T Freeman, Frédo Durand, and Song Han. Fastcomposer: Tuning-free multi-subject image generation with localized attention. *arXiv preprint arXiv:2305.10431*, 2023. 2

[85] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36, 2024. 6

[86] Kai Yang, Jian Tao, Jiafei Lyu, Chunjiang Ge, Jiaxin Chen, Qimai Li, Weihan Shen, Xiaolong Zhu, and Xiu Li. Using human feedback to fine-tune diffusion models without any reward model. *arXiv preprint arXiv:2311.13231*, 2023. 2, 3

[87] Kai Yang, Jian Tao, Jiafei Lyu, Chunjiang Ge, Jiaxin Chen, Weihan Shen, Xiaolong Zhu, and Xiu Li. Using human feedback to fine-tune diffusion models without any reward model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8941–8951, 2024. 3, 9

[88] Huizhuo Yuan, Zixiang Chen, Kaixuan Ji, and Quanquan Gu. Self-play fine-tuning of diffusion models for text-to-image generation. *arXiv preprint arXiv:2402.10210*, 2024. 2, 3

[89] Runzhe Zhan, Xuebo Liu, Derek F Wong, and Lidia S Chao. Meta-curriculum learning for domain adaptation in neural machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 14310–14318, 2021. 3

[90] Bowen Zhang, Yidong Wang, Wenxin Hou, Hao Wu, Jindong Wang, Manabu Okumura, and Takahiro Shinozaki. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. *Advances in Neural Information Processing Systems*, 34:18408–18419, 2021. 3

[91] Mingjun Zhao, Haijiang Wu, Di Niu, and Xiaoli Wang. Reinforced curriculum learning on pre-trained neural machine translation models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9652–9659, 2020. 3

[92] Yao Zhao, Rishabh Joshi, Tianqi Liu, Misha Khalman, Mohammad Saleh, and Peter J Liu. Slic-hf: Sequence likelihood calibration with human feedback. *arXiv preprint arXiv:2305.10425*, 2023. 3

[93] Yikai Zhou, Baosong Yang, Derek F Wong, Yu Wan, and Lidia S Chao. Uncertainty-aware curriculum learning for neural machine translation. In *Proceedings of the 58th Annual Meeting of the association for computational linguistics*, pages 6934–6944, 2020. 3