FocusDPO: Dynamic Preference Optimization for Multi-Subject Personalized Image Generation via Adaptive Focus

Qiaoqiao Jin[1*], Siming Fu[1*†], Dong She[1*], Weinan Jia[2], Hualiang Wang[3], Mu Liu[1], Jidong Jiang[1‡]

[1]ByteDance, China
[2]University of Science and Technology of China
[3]Hong Kong University of Science and Technology
jinqiaoqiao@bytedance.com

Abstract

Multi-subject personalized image generation aims to synthesize customized images containing multiple specified subjects without requiring test-time optimization. However, achieving fine-grained independent control over multiple subjects remains challenging due to difficulties in preserving subject fidelity and preventing cross-subject attribute leakage. We present FocusDPO, a framework that adaptively identifies focus regions based on dynamic semantic correspondence and supervision image complexity. During training, our method progressively adjusts these focal areas across noise timesteps, implementing a weighted strategy that rewards information-rich patches while penalizing regions with low prediction confidence. The framework dynamically adjusts focus allocation during the DPO process according to the semantic complexity of reference images and establishes robust correspondence mappings between generated and reference subjects. Extensive experiments demonstrate that our method substantially enhances the performance of existing pre-trained personalized generation models, achieving state-of-the-art results on both single-subject and multi-subject personalized image synthesis benchmarks. Our method effectively mitigates attribute leakage while preserving superior subject fidelity across diverse generation scenarios, advancing the frontier of controllable multi-subject image synthesis.

Introduction

The rapid advancement of diffusion models (Ho, Jain, and Abbeel 2020) has revolutionized personalized image generation (Ruiz et al. 2023; Gal et al. 2022; Hu et al. 2022; Ye et al. 2023; Li, Li, and Hoi 2023; She et al. 2025; Chen et al. 2025a; Xie et al. 2023; Chen et al. 2025b), enabling the synthesis of high-quality images featuring specific subjects of interest. Among various personalization paradigms, multi-subject personalized image generation has emerged as a particularly compelling research direction, aiming to synthesize customized images containing multiple specified subjects without requiring computationally expensive test-time optimization. This capability holds significant practical value for applications ranging from creative content generation to personalized advertising and digital art creation.

The multi-subject personalization methods (Wu et al. 2025b; Chen et al. 2025a; Mou et al. 2025; Xie et al. 2023; Xiao et al. 2025; Kumari et al. 2023; Huang et al. 2024b; Mao et al. 2024; Liu et al. 2025) fundamental difficulty lies in achieving fine-grained independent control over multiple subjects while simultaneously preserving visual fidelity of each individual subject. Existing approaches (Wu et al. 2025b; Xie et al. 2023; Xiao et al. 2025) often struggle with cross-subject attribute confusion, where characteristics from one subject inadvertently influence the appearance of another, leading to inconsistent or corrupted generations. Moreover, maintaining precise details and distinctive features of each reference subject becomes increasingly complex as the number of subjects grows, particularly when subjects share similar semantic categories or visual attributes.

Recent efforts in multi-subject generation have explored various strategies, such as PatchDPO (Huang et al. 2025a), it estimates the quality of image patches within each generated image and accordingly trains the model. However, these methods (Huang et al. 2025a; Wallace et al. 2024) typically employ fixed treatment on different image regions across different training timesteps, failing to account for the varying complexity and semantic importance of different areas within the generated image. This limitation becomes particularly pronounced when dealing with subjects of different scales, positions, or semantic complexity levels. As shown in Fig. 2, when the noise strength is changed, the regions that require focused attention during model training should adapt accordingly. At higher noise levels, the model needs to concentrate on global structure and semantic features, while at lower noise levels, attention should shift toward fine-grained details and local texture preservation. ***This observation motivates our dynamic focus modulation mechanism, which adjusts the spatial focus regions based on the current denoising step and the complexity of the supervision signal.***

To address these challenges, we present FocusDPO, a novel framework that leverages dynamic Direct Preference Optimization to achieve superior multi-subject personalized

[*]Equal Contribution
[†]Project Lead
[‡]Corresponding Author

Figure 1: Our proposed FocusDPO demonstrates capabilities in single-subject and multi-subject driven generation tasks.

image generation. Our ***key insight*** is that effective multi-subject control requires adaptive attention allocation based on both the dynamic correspondence between generated and reference subjects and the semantic complexity of supervision images. Rather than applying uniform optimization pressure across all image regions, FocusDPO intelligently identifies regions of focus and adjusts the training dynamics accordingly. The core contribution of our approach lies in its weighted training strategy that rewards high-quality image patches while penalizing regions with low prediction confidence. This selective optimization enables the model to concentrate computational resources on challenging areas while maintaining efficiency in well-handled regions. Furthermore, our framework establishes robust correspondence mappings between generated and reference subjects, ensuring consistent identity preservation across diverse generation scenarios. Extensive experiments demonstrate that FocusDPO substantially enhances the performance of existing pre-trained personalized generation models, achieving state-of-the-art results on both single-subject and multi-subject personalized image synthesis benchmarks. The proposed framework effectively addresses attribute leakage while maintaining superior subject preservation, marking a significant step forward in controllable image generation. Our primary contributions are threefold:

- We introduce FocusDPO, which intelligently identifies "focus regions" characterized by high semantic complexity and detailed-preserving generation difficulty. By adaptively intensifying optimization on these key areas, it efficiently enhances overall image quality and training stability.
- By dynamic semantic and detail-preserving preference optimization, our method effectively mitigates identity confusion and attribute leakage in multi-subject scenarios, ensuring faithful subject preservation.
- Extensive experiments demonstrate that FocusDPO substantially boosts the performance of existing models, achieving state-of-the-art results on both single- and multi-subject personalized image synthesis benchmarks.

Related Work

Subject-driven Generation

Multi-subject personalized generation extends beyond single-subject methods like DreamBooth (Ruiz et al. 2023) and Textual Inversion (Gal et al. 2022), presenting significant challenges in maintaining subject integrity while preventing inter-subject interference. Recent approaches have explored various architectural strategies: OmniControl (Xie et al. 2023) and IC LoRA (Huang et al. 2024a) averages DiTs as image encoders for subject references, while MS-Diffusion (Wang et al. 2025) and MIP-Adapter (Huang et al. 2025b) introduce specialized adapters for multiple subjects. UNO (Wu et al. 2025b) employs progressive training with Universal Rotary Position Embedding to mitigate attribute confusion, and XVerse (Chen et al. 2025a) utilizes text-stream modulation for reference image processing. Despite these advances, existing methods struggle with inter-subject entanglement and attribute leakage, particularly when subjects share similar visual characteristics. This work addresses these limitations through dynamic semantic guidance, achieving robust multi-subject generation with enhanced semantic consistency.

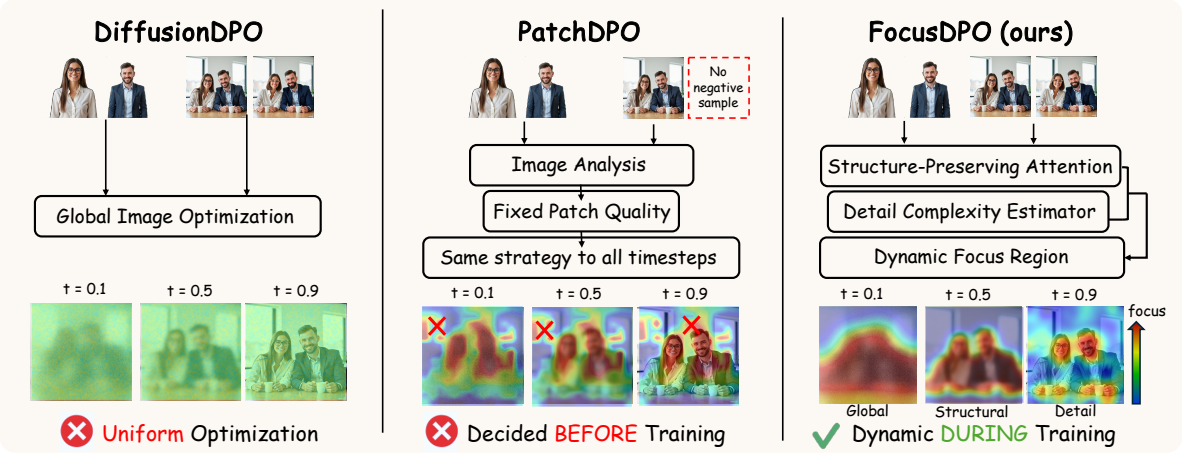

Figure 2: Comparison of training optimization strategy: DiffusionDPO's uniform optimization (left) vs. PatchDPO's fixed patch optimization (middle) FocusDPO's adaptive focus strategy (right).

Diffusion-based Preference Optimization

Direct Preference Optimization Direct Preference Optimization (DPO) (Rafailov et al. 2023; Zeng et al. 2024; Zhou et al. 2023) and Reinforcement Learning from Human Feedback (RLHF) (Bai et al. 2022; Casper et al. 2023; Fan et al. 2023; Lee et al. 2023), originally formulated for language model alignment, have been successfully adapted to diffusion-based image synthesis through methods including DPOK (Fan et al. 2023), DDPO (Black et al. 2023), DRaFT (Clark et al. 2023) and AlignProp (Prabhudesai et al. 2023) for enhancing generation quality. However, standard DPO frameworks fall short in consistency-sensitive tasks due to semantic confounds in global image comparisons. PatchDPO (Huang et al. 2025a) mitigates this by optimizing preferences at the patch level, promoting finer-grained consistency learning. Unlike PatchDPO, which relies on static supervision without explicit positive-negative pairs, our method builds semantically aligned pairs and employs dynamic objectives to better preserve consistency.

Preliminary: Diffusion-DPO

Our work builds upon Diffusion Direct Preference Optimization (Diffusion-DPO) (Wallace et al. 2024), a powerful framework for aligning text-to-image diffusion models with human preferences without requiring an explicit reward model. The training data for this process consists of preference pairs $\mathcal{D} = \{(x_0^w, x_0^l)\}$, where x_0^w is the preferred (winning) image and x_0^l is the dispreferred (losing) image for a given prompt. The core idea of Diffusion-DPO is to reinterpret the DPO loss in the context of diffusion model training:

$$\mathcal{L}(\theta) = -\mathbb{E}_{(x_0^w,x_0^l)\sim\mathcal{D},t\sim\mathcal{U}(0,T),x_t^w\sim q(x_t^w|x_0^w),x_t^l\sim q(x_t^l|x_0^l)} log\sigma(-\beta T\omega(\lambda_t)(||\epsilon^w - \epsilon_\theta(x_t^w,t)||_2^2 - ||\epsilon^w - \epsilon_{\text{ref}}(x_t^w,t)||_2^2 - (||\epsilon^l - \epsilon_\theta(x_t^l,t)||_2^2 - ||\epsilon^l - \epsilon_{\text{ref}}(x_t^l,t)||_2^2)), \quad (1)$$

where $x_t^* = \alpha_t x_0^* + \sigma_t \epsilon^*, \epsilon^* \sim \mathcal{N}(0, I)$. And $\lambda_t = \frac{\alpha_t^2}{\sigma_t^2}$ is the signal-to-noise ratio. However, this uniform weighting strategy in Diffusion-DPO proves suboptimal for the nuanced demands of multi-subject generation. By treating all spatial regions with equal importance, the approach is inherently susceptible to interference from irrelevant background details. Consequently, its efficacy in preserving the consistency and identity of multiple subjects within a single composition is significantly constrained. To overcome this limitation, a more targeted weighting mechanism is necessary—one that can distinguish between foreground subjects and the background, focusing optimization where most needed.

Method

We propose **D**isrupted-**I**nstance **P**air Dataset (**DIP**) and Focus Direct Preference Optimization (FocusDPO), a spatially-aware preference optimization framework for addressing subject-level inconsistencies in personalized image generation. Our method constructs high-quality subject-consistent pairs $\{(x_0^w, x_0^l)\}$ with controlled subject variation, utilizing a binary prior guidance $\mathbf{M}_{\text{prior}} \in \{0,1\}^{H\times W}$ to identify regions containing subject differences. We then introduce a spatial weighting mechanism $\mathbf{M} \in \mathbb{R}^{H\times W}$ that dynamically modulates the optimization process across different regions, where H and W denote the height and width of the latent feature maps.

Disrupted-Instance Pair Dataset

We construct the DIP Dataset as a semantically aligned collection of positive-negative image pairs designed to isolate subject-level inconsistencies. Each image pair maintains semantic content parity while introducing localized perturbations to subject regions. We begin by synthesizing reference and target pairs $\{(x_r, x_0^w)\}$ using the FLUX generator (Labs 2024), conditioned on identical subject prompts c but with varied auxiliary attributes such as background and lighting. From the resulting corpus, we manually curate 5,000 high-quality pairs for single-subject and multi-subject scenario, where x_0^w shows strong visual alignment with x_r in subject identity, serving as **positive samples**.

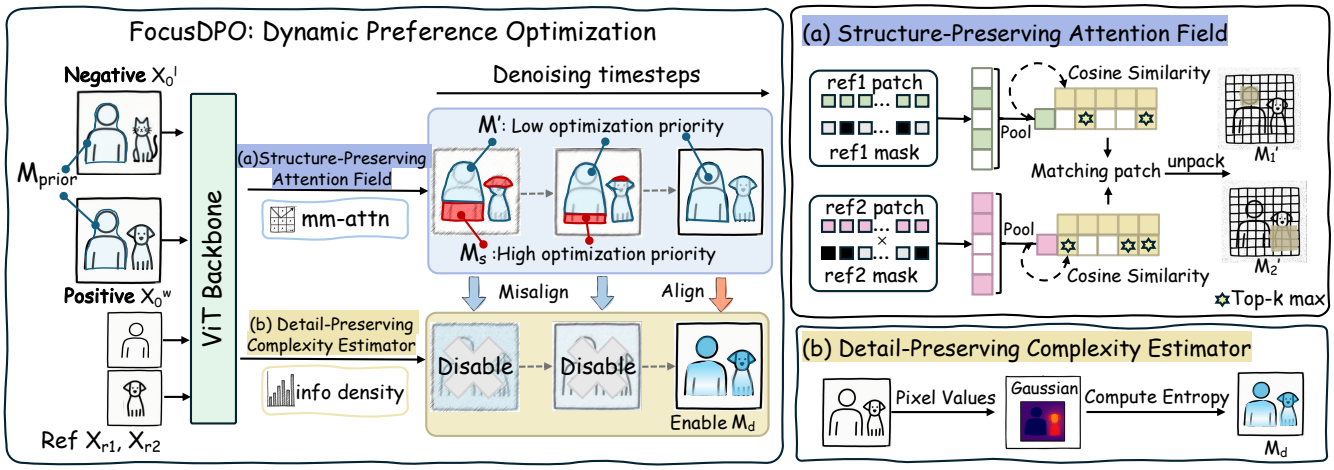

Figure 3: **The overall framework of FocusDPO.** FocusDPO introduce a *spatially-aware optimization framework (left)* that adaptively focuses on critical regions through dynamic semantic guidance, leveraging (a) *Structure-Preserving Attention Field* and (b) *Detail-Preserving Complexity Estimator*.

To construct **negative samples** x_0^l, we introduce controlled semantic disruptions to x_0^w while preserving surrounding context. We apply GroundingSAM2 (Liu et al. 2023; Ravi et al. 2024) to extract accurate binary segmentation maps $\mathbf{M}_{\text{prior}}$ that isolate subject regions. To maintain semantic consistency in the altered regions, we utilize GPT-4o to generate fine-grained captions for the segmented areas, providing detailed semantic descriptions. These captions are fed into the inpainting model (Labs 2024), which modifies subject-relevant pixels while preserving surrounding context. The resulting image x_0^l exhibits degraded visual consistency with respect to x_r while maintaining overall semantic coherence.

The final DIP dataset comprises quadruplets (c, x_r, x_0^w, x_0^l), with shared conditioning prompt c and reference image x_r specifying target subject identity, a high-consistency image x_0^w maintaining strong alignment with reference x_r, and a low-consistency counterpart x_0^l generated through localized semantic perturbation within $\mathbf{M}_{\text{prior}}$. This approach ensures subject-level consistency in preference learning while removing confounders, forming the basis for spatially-aware optimization in FocusDPO.

Focus Direct Preference Optimization

To overcome the limitations of uniform weighting in existing methods, we introduce a spatially-aware optimization framework, as shown in Fig. 3. The cornerstone of our approach is a dynamic semantic guidance strategy that adaptively focuses supervisory signals on critical regions by decomposing the task into two complementary sub-problems: preserving global consistency via a Structure-Preserving Attention Field and maintaining high-fidelity features with a Detail-Preserving Complexity Estimator. We leverage these components to inform a spatial weighting mechanism that dynamically modulates the optimization process across semantically critical and non-critical regions.

Structure-Preserving Attention Field: Contemporary personalized generation methods suffer from subject confusion, where target attributes propagate to semantically unrelated regions, especially in multi-subject generation scenarios. Our primary objective is to address these structural deficiencies in generated images. Multi-modal attention layers inherently encode semantic relationships between the noised latent representation x_t and the reference image x_r. For each attention layer i, we denote the target and reference token embeddings as:

$$\begin{cases} H_{x_t}^i \in \mathbb{R}^{p_{x_t} \times d}, \\ H_{x_r}^i \in \mathbb{R}^{p_{x_r} \times d}, \end{cases} \quad i \in \{0, 1, \ldots, N-1\}, \tag{2}$$

where p_{x_t} and p_{x_r} denote the number of patch tokens, d is the embedding dimension, and N is the total number of multi-modal attention layers.

To establish semantic correspondences between subject regions and target patches, we employ a cross-layer correspondence strategy. For each reference subject patch j, we identify the target location with maximal semantic alignment by computing:

$$\mathcal{S} = \sum_{i=1}^{N} \frac{1}{N} \cdot \langle \frac{CLS_{x_r}^i}{|CLS_{x_r}^i|^2}, \frac{H_{x_t}^i}{|H_{x_t}^i|^2} \rangle, \tag{3}$$
$$CLS_{x_r}^i = \text{pool}(H_{x_r}^i), \quad CLS_{x_r}^i \in \mathbb{R}^{1 \times d}.$$

Let S be the score vector of size $1 \times p_{x_t}$, with elements S_i for $i = 1, \ldots, p_{x_t}$. We identify the set of indices $\mathcal{J}$ corresponding to the K largest scores in S:

$$\mathcal{J} = \underset{i=1,\ldots,p_{x_t}}{\arg\max}(S_i), \tag{4}$$

where K is the number of subject tokens of positive sample x_0^w. Using this set, we construct a binary attention map M' where element $\mathbf{M}'_j$ is set to 1 if its index j is in the set of top-K indices $\mathcal{J}$, and 0 otherwise:

$$\mathbf{M}'_j = \mathbb{I}(j \in \mathcal{J}). \tag{5}$$

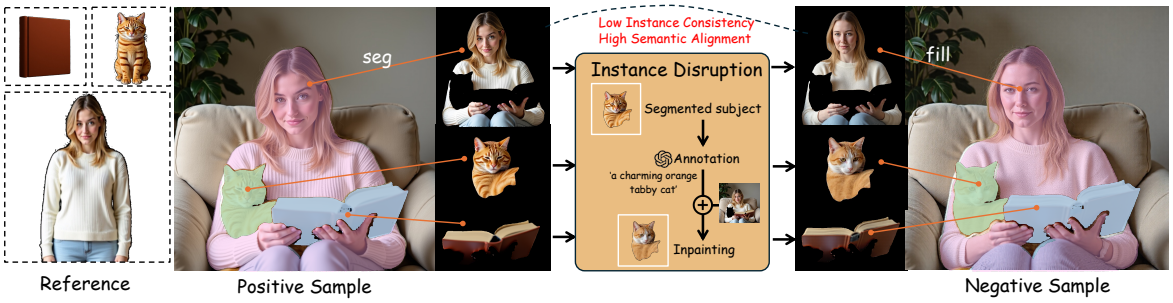

Figure 4: **Disrupted-Instance Pair Dataset (DIP) dataset construction workflow.** Preference pairs are generated by creating subject-consistent images, then introducing controlled perturbations to segmented areas to produce degraded counterparts.

Algorithm 1: FocusDPO Training Algorithm

procedure COMPUTEMASKS(x_0^w, x_t^w, x_r)
  Compute the Structure-Preserving Attention Field $\mathbf{M}_s$ based on Eq. 3-Eq. 6.
  Compute the Detail-Preserving Complexity Estimator $\mathbf{M}_d$ based on Eq. 7 and Eq. 8.
  return $\mathbf{M}_s, \mathbf{M}_d$
end procedure

for each training step **do**
  Sample preference pair $(x_0^w, x_0^l, \mathbf{M}_{\text{prior}}) \sim \mathcal{D}$, timestep $t \sim \mathcal{U}(1, T)$, noise $\epsilon \sim \mathcal{N}(0, I)$.
  Create noised latents x_t^w, x_t^l.

Step 1: Generate dynamic semantic masks
  $\mathbf{M}_s, \mathbf{M}_d \leftarrow$ COMPUTEMASKS(x_0^w, x_t^w, x_r)

Step 2: Calculate the focus coverage ratio
  $A_{\text{focus}} \leftarrow \|\mathbf{M}_s\|_1 / \|\mathbf{M}_{\text{prior}}\|_1$

Step 3: Determine fusion mask $\mathbf{M}$
  Determine fusion mask $\mathbf{M}$ using Eq. 9.

Step 4: Compute loss and update model
  Compute the final loss $\mathcal{L}_{\text{FocusDPO}}$ using the spatially weighted objective (Eq. 10).
  Update model parameters: $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}_{\text{FocusDPO}}$.
end for

Finally, we define Structure-Preserving Attention Field as:

$$\mathbf{M}_s = \mathbf{M}_{\text{prior}} \setminus \mathbf{M}'. \tag{6}$$

Detail-Preserving Complexity Estimator: Conventional preference optimization methods treat preference distinctions as globally distributed properties. However, for multi-subject generation in complex scenes, preference signals are often spatially localized. We observe that preference-critical regions correlate with areas of high visual complexity, which also pose challenges for diffusion model reconstruction. Motivated by this observation, we introduce a Detail-Preserving Complexity Estimator that prioritizes these complex regions during preference optimization through an information density weighting mechanism.

Our method generates this weight mask through two core steps. First, we compute a visual complexity score for each local region. For a patch centered at pixel p, denoted as patch_p, we obtain its complexity score $\mathbf{C}_p$ by calculating the Shannon entropy of its grayscale pixel intensity distribution. Higher entropy values signify richer textures and details corresponding to higher visual complexity, while smooth and uniform regions yield lower entropy:

$$\mathbf{C}_p = \text{CalculateComplexity}(\text{patch}_p). \tag{7}$$

Second, to ensure these scores are comparable and form a usable weighting scheme, we perform global normalization. We scale the complexity score $\mathbf{C}_p$ of each location to the range $[0, 1]$ to obtain the final information complexity $\mathbf{M}_d$:

$$\mathbf{M}_d = \frac{\mathbf{C}_p - C_{\min}}{C_{\max} - C_{\min}}, \tag{8}$$

where $C_{\min}$ and $C_{\max}$ represent the minimum and maximum complexity scores computed across all local regions of the entire image. The resulting matrix of weights $\mathbf{M}_d$ constitutes our Detail-Preserving Complexity Estimator.

By incorporating this complexity estimator into the model's optimization loss function, we guide the model to focus attention on regions with higher weights during training. This enables the model to preferentially learn from and correct errors in critical areas such as fine textures and facial details, thereby significantly enhancing the local quality and overall fidelity of generated images.

Final Loss: As shown in Alg. 1, we determine the dynamic fusion field $\mathbf{M}$ through an adaptive strategy that leverages both structural and detail-preserving components. The fusion mechanism employs a focus threshold $\tau = 0.1$ and tradeoff parameter $\gamma = 0.3$ to balance global consistency and local fidelity (*with ablation study in supplementary*):

$$\mathbf{M} = \begin{cases} \mathbf{M}_s, & A_{\text{focus}} > \tau \\ \gamma \mathbf{M}_s + (1-\gamma)\mathbf{M}_d \odot \mathbf{M}_{\text{prior}}, & A_{\text{focus}} \le \tau \end{cases} \tag{9}$$

The complete FocusDPO objective incorporates the spatial fusion field into the preference optimization framework:

$$\mathcal{L}_{\text{FocusDPO}}(\theta) = -\mathbb{E} \log \sigma(-\beta T \omega(\lambda_t)) \\ (||(\epsilon^w - \epsilon_\theta(x_t^w, t)) \odot \mathbf{M}||_2^2 - ||(\epsilon^w - \epsilon_{\text{ref}}(x_t^w, t)) \odot \mathbf{M}||_2^2 \\ - (||(\epsilon^l - \epsilon_\theta(x_t^l, t)) \odot \mathbf{M}||_2^2 - ||(\epsilon^l - \epsilon_{\text{ref}}(x_t^l, t)) \odot \mathbf{M}||_2^2)). \tag{10}$$

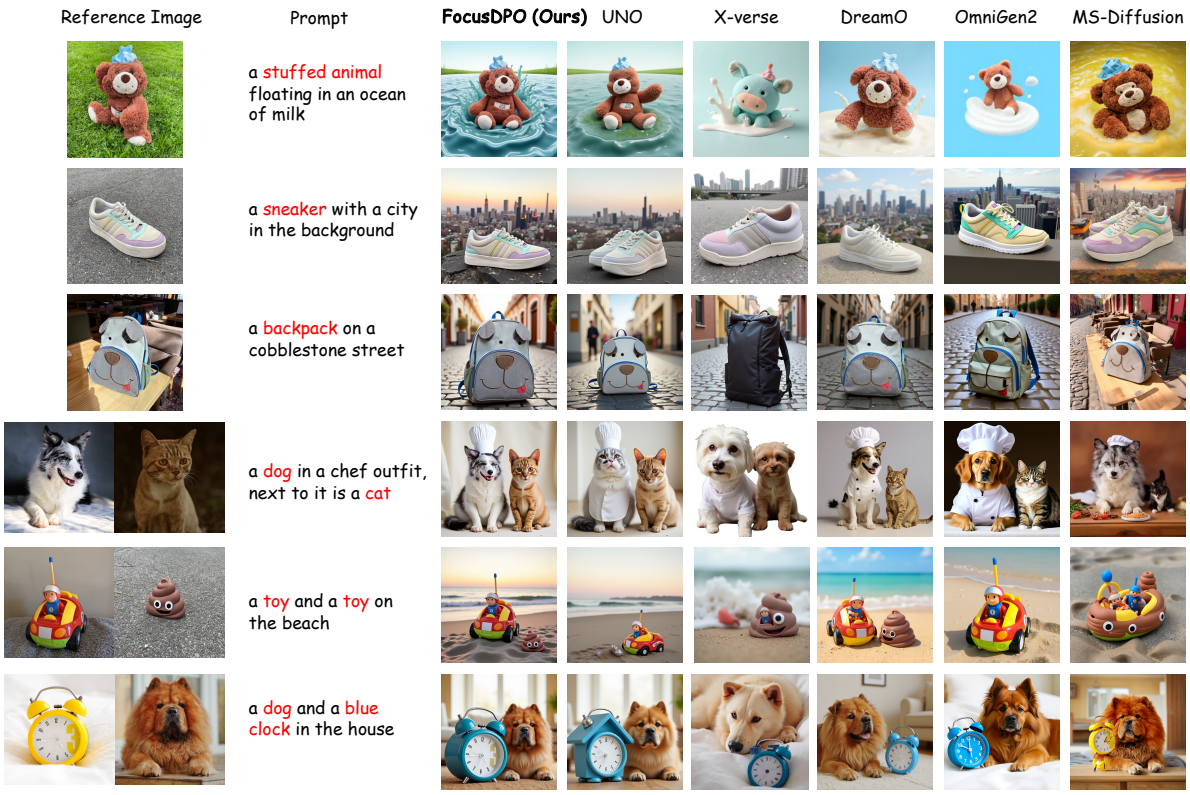

Figure 5: Qualitative comparison of single-subject and multi-subject generation with different methods on DreamBench.

This spatially-aware framework enables the model to preferentially learn from and correct errors in critical regions, significantly enhancing local quality and overall fidelity of generated images while maintaining subject consistency across complex multi-subject scenarios.

Experiments

Implementation Details We implement our FocusDPO methodology on two representative diffusion architectures: U-Net (Rombach et al. 2022) and DiT (Peebles and Xie 2023), ensuring broad applicability and generalizability. For SDXL, we use the pre-trained IP-Adapter-Plus (Ye et al. 2023) with the SDXL (Podell et al. 2024) text-to-image model, consistent with PatchDPO (Huang et al. 2025a). For DiT, we integrate FocusDPO into the UNO (Wu et al. 2025b) pipeline for multi-subject synthesis, utilizing FLUX.1 dev (Labs 2024) as the foundation model.

Our training uses rank-32 LoRA (Hu et al. 2022) modules in both backbones, trained for 2000 steps on 8 GPUs with AdamW (lr=1e-8, batch size=1 per GPU). The training dataset DIP contains 10,000 preference pairs (5000 each for single- and multi-subject generation). The test dataset is DreamBench (Ruiz et al. 2023), which includes both single-subject and multi-subject cases, with the multi-subject evaluation protocol following settings in (Ma et al. 2024; Huang et al. 2025b; Wu et al. 2025b).

Evaluation Metrics We evaluate subject fidelity and text-image alignment using established metrics. Subject fidelity is measured through CLIP-I (Radford et al. 2021) and DINO (Oquab et al. 2024) scores, computing cosine similarity between generated, and reference images and text-image alignment is assessed using CLIP-T score.

Qualitative Analysis

We conduct a comprehensive comparative evaluation of our proposed methodology against existing state-of-the-art approaches across both single-subject and multi-subject generation paradigms. Our comparative analysis encompasses five recent and competitive methods: Xverse (Chen et al. 2025a), DreamO (Mou et al. 2025), OmniGen2 (Wu et al. 2025a), MS-Diffusion (Wang et al. 2025), and our backbone architecture UNO (Wu et al. 2025b). As demonstrated in Fig. 5, FocusDPO exhibits superior performance in maintaining subject consistency and identity preservation while simultaneously ensuring global semantic coherence. The performance gains arise from our training paradigm, which dynamically targets complex, detail-rich regions to better capture subject-specific features and preserve fine-grained identity—especially in multi-subject generation.

Quantitative comparisons

Tables 1 and 2 present comprehensive quantitative evaluations of our method against competing approaches on the

Method	DINO	CLIP-I	CLIP-T
DreamBooth (Ruiz et al. 2023)	0.668	0.803	0.305
SSR-Encoder (Zhang et al. 2024)	0.612	0.821	0.308
RealCustom++ (Mao et al. 2024)	0.702	0.794	**0.318**
OmniGen (Xiao et al. 2025)	0.693	0.801	0.315
OminiControl (Tan et al. 2024)	0.684	0.799	0.312
DreamO (Mou et al. 2025)	0.712	0.809	0.314
IPAP (Ye et al. 2023)	0.692	0.826	0.281
IPAP + SFT	0.691	0.828	0.279
IPAP + DPO	0.695	0.831	0.276
IPAP + PatchDPO	0.727	0.838	0.292
IPAP + FocusDPO	0.751	<u>0.840</u>	0.303
UNO (Wu et al. 2025b)	0.760	0.835	0.304
UNO + SFT	0.761	0.832	0.299
UNO + DPO	<u>0.764</u>	0.835	0.301
UNO + FocusDPO	**0.802**	**0.842**	<u>0.316</u>

Table 1: Performance comparison for single-object personalized generation on DreamBench (Ruiz et al. 2023). IPAP stands for IP-Adapter-Plus. The best result is marked in **bold**, and the second-best is <u>underlined</u>.

Method	DINO	CLIP-I	CLIP-T
DreamBooth (Ruiz et al. 2023)	0.430	0.695	0.308
MS-Diffusion (Wang et al. 2025)	0.525	0.726	0.319
OmniGen (Xiao et al. 2025)	0.511	0.722	**0.331**
DreamO (Mou et al. 2025)	0.539	0.727	0.313
UNO (DiT Backbone)	0.542	<u>0.733</u>	0.322
UNO + SFT	0.542	0.732	0.321
UNO + DPO	<u>0.545</u>	0.731	0.322
UNO + FocusDPO	**0.570**	**0.739**	<u>0.328</u>

Table 2: Multi-subject sythesis results on DreamBench.

DreamBench dataset for single-object and multi-subject personalized generation, respectively.We conduct evaluations of FocusDPO across two distinct backbone architectures, demonstrating consistent performance gains across all metrics. When integrated with the UNO backbone, FocusDPO establishes new state-of-the-art results, achieving a 5.5% improvement on DINO (0.802 vs. 0.760) for single-object generation, and state-of-the-art performance (DINO: 0.570, CLIP-I: 0.739) for multi-subject synthesis. These empirical results substantiate the efficacy of our proposed framework.

Ablation Study

Ablation on Dynamic Fusion Components In Fig. 6, We conduct an ablation study to analyze the contribution of two critical weighting components within our dynamic optimization framework: the Structure-Preserving Attention Field $\mathbf{M}_s$ and the Detail-Preserving Complexity Estimator $\mathbf{M}_d$. The supplementary experimental results presented in Fig. 7 demonstrate that the absence of $\mathbf{M}_s$ results in substantial subject disambiguation failures, with this degradation being particularly pronounced in multi-subject synthesis scenarios. Conversely, the omission of $\mathbf{M}_d$ manifests as inconsistencies in fine-grained structural coherence across the generated outputs.These results confirm that $\mathbf{M}_s$ and $\mathbf{M}_d$ work in tandem to preserve semantic alignment and enhance structural fidelity in complex regions.

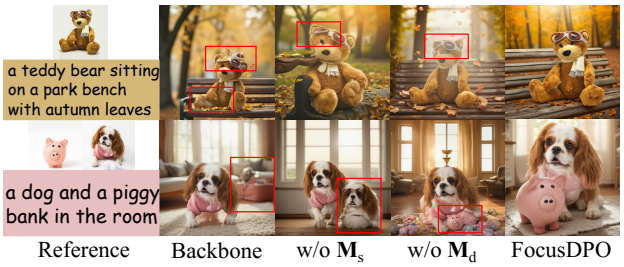

Figure 6: Ablation study on dynamic weighting components.

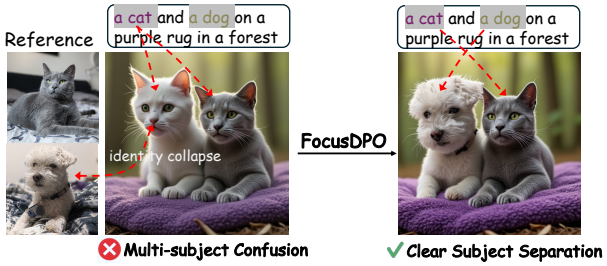

Figure 7: Analysis of resolving semantic confusion.

Ablation on Optimization Strategy We evaluate our dynamic optimization approach by comparing three training paradigms across two backbone architectures using the generated DIP dataset: supervised fine-tuning (SFT), standard DPO, and our FocusDPO methodology. As shown in Tab. 1, FocusDPO consistently outperforms both SFT and conventional DPO across all metrics. For single-subject scenarios, FocusDPO achieves 0.751 under IPAP (8.7% improvement over SFT, 8.1% over DPO) and 0.802 under UNO (5.4% improvement over SFT, 5.0% over DPO). For multi-subject scenarios under UNO, FocusDPO achieves 0.570 (DINO), 0.739 (CLIP-I), and 0.326 (CLIP-T), representing improvements of 5.2%/4.6% (DINO), 1.0%/1.1% (CLIP-I), and 1.6%/1.2% (CLIP-T) over SFT/DPO respectively. These results demonstrate that improvements stem from our optimization strategy rather than dataset biases.

Conclusion

We present FocusDPO, a framework for multi-subject personalized image generation that tackles the challenge of preserving subject identity while avoiding attribute confusion. Our method introduces dynamic focus modulation mechanism that dynamically prioritize regions based on supervision complexity and semantic alignment. Extensive experiments show that FocusDPO consistently improves pretrained models, achieving state-of-the-art results on both single- and multi-subject benchmarks, setting a new standard for controllable multi-subject generation.

References

Bai, Y.; Jones, A.; Ndousse, K.; Askell, A.; Chen, A.; DasSarma, N.; Drain, D.; Fort, S.; Ganguli, D.; Henighan, T.; et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Black, K.; Janner, M.; Du, Y.; Kostrikov, I.; and Levine, S. 2023. Training diffusion models with reinforcement learning. *arXiv preprint arXiv:2305.13301*.

Casper, S.; Davies, X.; Shi, C.; Gilbert, T. K.; Scheurer, J.; Rando, J.; Freedman, R.; Korbak, T.; Lindner, D.; Freire, P.; et al. 2023. Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv:2307.15217*.

Chen, B.; Zhao, M.; Sun, H.; Chen, L.; Wang, X.; Du, K.; and Wu, X. 2025a. XVerse: Consistent Multi-Subject Control of Identity and Semantic Attributes via DiT Modulation. *arXiv preprint arXiv:2506.21416*.

Chen, N.; Huang, M.; Chen, Z.; Zheng, Y.; Zhang, L.; and Mao, Z. 2025b. Customcontrast: A multilevel contrastive perspective for subject-driven text-to-image customization.

Clark, K.; Vicol, P.; Swersky, K.; and Fleet, D. J. 2023. Directly fine-tuning diffusion models on differentiable rewards. *arXiv preprint arXiv:2309.17400*.

Fan, Y.; Watkins, O.; Du, Y.; Liu, H.; Ryu, M.; Boutilier, C.; Abbeel, P.; Ghavamzadeh, M.; Lee, K.; and Lee, K. 2023. Dpok: Reinforcement learning for fine-tuning text-to-image diffusion models. *Advances in Neural Information Processing Systems*, 36: 79858–79885.

Gal, R.; Alaluf, Y.; Atzmon, Y.; Patashnik, O.; Bermano, A. H.; Chechik, G.; and Cohen-Or, D. 2022. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*.

Huang, L.; Wang, W.; Wu, Z.; Shi, Y.; Dou, H.; Liang, C.; Feng, Y.; Liu, Y.; and Zhou, J. 2024a. In-Context LoRA for Diffusion Transformers. *CoRR*, abs/2410.23775.

Huang, M.; Mao, Z.; Liu, M.; He, Q.; and Zhang, Y. 2024b. RealCustom: Narrowing Real Text Word for Real-Time Open-Domain Text-to-Image Customization. *CoRR*, abs/2403.00483.

Huang, Q.; Chan, L.; Liu, J.; He, W.; Jiang, H.; Song, M.; and Song, J. 2025a. PatchDPO: Patch-level DPO for Finetuning-free Personalized Image Generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, 18369–18378. Computer Vision Foundation / IEEE.

Huang, Q.; Fu, S.; Liu, J.; Jiang, H.; Yu, Y.; and Song, J. 2025b. Resolving Multi-Condition Confusion for Finetuning-Free Personalized Image Generation. In *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, 3707–3714.

Kumari, N.; Zhang, B.; Zhang, R.; Shechtman, E.; and Zhu, J.-Y. 2023. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1931–1941.

Labs, B. F. 2024. FLUX. https://github.com/black-forest-labs/flux.

Lee, K.; Liu, H.; Ryu, M.; Watkins, O.; Du, Y.; Boutilier, C.; Abbeel, P.; Ghavamzadeh, M.; and Gu, S. S. 2023. Aligning text-to-image models using human feedback. *arXiv preprint arXiv:2302.12192*.

Li, D.; Li, J.; and Hoi, S. C. H. 2023. BLIP-Diffusion: Pre-trained Subject Representation for Controllable Text-to-Image Generation and Editing. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Liu, M.; She, D.; Pang, J.; Huang, Q.; Ying, J.; He, W.; Hou, Y.; and Fu, S. 2025. TFCustom: Customized Image Generation with Time-Aware Frequency Feature Guidance. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, 2714–2723. Computer Vision Foundation / IEEE.

Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Li, C.; Yang, J.; Su, H.; Zhu, J.; et al. 2023. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*.

Ma, J.; Liang, J.; Chen, C.; and Lu, H. 2024. Subject-Diffusion: Open Domain Personalized Text-to-Image Generation without Test-time Fine-tuning. In *ACM SIGGRAPH 2024 Conference Papers, SIGGRAPH 2024, Denver, CO, USA, 27 July 2024- 1 August 2024*, 25. ACM.

Mao, Z.; Huang, M.; Ding, F.; Liu, M.; He, Q.; and Zhang, Y. 2024. Realcustom++: Representing images as real-word for real-time customization. *arXiv preprint arXiv:2408.09744*.

Mou, C.; Wu, Y.; Wu, W.; Guo, Z.; Zhang, P.; Cheng, Y.; Luo, Y.; Ding, F.; Zhang, S.; Li, X.; Li, M.; Zhao, S.; Zhang, J.; He, Q.; and Wu, X. 2025. DreamO: A Unified Framework for Image Customization. *CoRR*, abs/2504.16915.

Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H. V.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; Assran, M.; Ballas, N.; Galuba, W.; Howes, R.; Huang, P.; Li, S.; Misra, I.; Rabbat, M.; Sharma, V.; Synnaeve, G.; Xu, H.; Jégou, H.; Mairal, J.; Labatut, P.; Joulin, A.; and Bojanowski, P. 2024. DINOv2: Learning Robust Visual Features without Supervision. *Trans. Mach. Learn. Res.*, 2024.

Peebles, W.; and Xie, S. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4195–4205.

Podell, D.; English, Z.; Lacey, K.; Blattmann, A.; Dockhorn, T.; Müller, J.; Penna, J.; and Rombach, R. 2024. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. In *The Twelfth International Conference on*

Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024.

Prabhudesai, M.; Goyal, A.; Pathak, D.; and Fragkiadaki, K. 2023. Aligning text-to-image diffusion models with reward backpropagation.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139, 8748–8763. PMLR.

Rafailov, R.; Sharma, A.; Mitchell, E.; Manning, C. D.; Ermon, S.; and Finn, C. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36: 53728–53741.

Ravi, N.; Gabeur, V.; Hu, Y.-T.; Hu, R.; Ryali, C.; Ma, T.; Khedr, H.; Rädle, R.; Rolland, C.; Gustafson, L.; Mintun, E.; Pan, J.; Alwala, K. V.; Carion, N.; Wu, C.-Y.; Girshick, R.; Dollár, P.; and Feichtenhofer, C. 2024. SAM 2: Segment Anything in Images and Videos. arXiv:2408.00714.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; and Aberman, K. 2023. 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*, 22500–22510.

She, D.; Fu, S.; Liu, M.; Jin, Q.; Wang, H.; Liu, M.; and Jiang, J. 2025. MOSAIC: Multi-Subject Personalized Generation via Correspondence-Aware Alignment and Disentanglement. *CoRR*, abs/2509.01977.

Tan, Z.; Liu, S.; Yang, X.; Xue, Q.; and Wang, X. 2024. OminiControl: Minimal and Universal Control for Diffusion Transformer. *CoRR*, abs/2411.15098.

Wallace, B.; Dang, M.; Rafailov, R.; Zhou, L.; Lou, A.; Purushwalkam, S.; Ermon, S.; Xiong, C.; Joty, S.; and Naik, N. 2024. Diffusion Model Alignment Using Direct Preference Optimization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, 8228–8238.

Wang, X.; Fu, S.; Huang, Q.; He, W.; and Jiang, H. 2025. MS-Diffusion: Multi-subject Zero-shot Image Personalization with Layout Guidance. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*.

Wu, C.; Zheng, P.; Yan, R.; Xiao, S.; Luo, X.; Wang, Y.; Li, W.; Jiang, X.; Liu, Y.; Zhou, J.; Liu, Z.; Xia, Z.; Li, C.; Deng, H.; Wang, J.; Luo, K.; Zhang, B.; Lian, D.; Wang, X.; Wang, Z.; Huang, T.; and Liu, Z. 2025a. OmniGen2: Exploration to Advanced Multimodal Generation. *CoRR*, abs/2506.18871.

Wu, S.; Huang, M.; Wu, W.; Cheng, Y.; Ding, F.; and He, Q. 2025b. Less-to-more generalization: Unlocking more controllability by in-context generation. *arXiv preprint arXiv:2504.02160*.

Xiao, S.; Wang, Y.; Zhou, J.; Yuan, H.; Xing, X.; Yan, R.; Li, C.; Wang, S.; Huang, T.; and Liu, Z. 2025. OmniGen: Unified Image Generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, 13294–13304. Computer Vision Foundation / IEEE.

Xie, Y.; Jampani, V.; Zhong, L.; Sun, D.; and Jiang, H. 2023. Omnicontrol: Control any joint at any time for human motion generation. *arXiv preprint arXiv:2310.08580*.

Ye, H.; Zhang, J.; Liu, S.; Han, X.; and Yang, W. 2023. IP-Adapter: Text Compatible Image Prompt Adapter for Text-to-Image Diffusion Models. *CoRR*, abs/2308.06721.

Zeng, Y.; Liu, G.; Ma, W.; Yang, N.; Zhang, H.; and Wang, J. 2024. Token-level direct preference optimization. *arXiv preprint arXiv:2404.11999*.

Zhang, Y.; Song, Y.; Liu, J.; Wang, R.; Yu, J.; Tang, H.; Li, H.; Tang, X.; Hu, Y.; Pan, H.; and Jing, Z. 2024. SSR-Encoder: Encoding Selective Subject Representation for Subject-Driven Generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, 8069–8078. IEEE.

Zhou, Z.; Liu, J.; Yang, C.; Shao, J.; Liu, Y.; Yue, X.; Ouyang, W.; and Qiao, Y. 2023. Beyond one-preference-for-all: Multi-objective direct preference optimization.