Multi-dimensional Preference Alignment by Conditioning Reward Itself

Jiho Jang[1], Jinyoung Kim[1], Kyungjune Baek[2]*, and Nojun Kwak[1]*

[1] Seoul National University
[2] Sejong University
{geographic,nojunk}@snu.ac.kr, kyungjune.baek@sejong.ac.kr

Abstract. Reinforcement Learning from Human Feedback has emerged as a standard for aligning diffusion models. However, we identify a fundamental limitation of the standard DPO formulation when optimizing multi-dimensional preferences: by relying on the Bradley-Terry model, it aggregates heterogeneous evaluation axes (e.g., aesthetic quality and semantic alignment) into a single scalar reward. This aggregation creates a reward conflict where the model is forced to unlearn desirable features of a specific dimension if they appear in a globally non-preferred sample. To address this issue, we propose Multi Reward Conditional DPO (MCDPO) which resolves reward conflicts by introducing a disentangled Bradley-Terry objective. MCDPO explicitly injects a preference outcome vector as a condition during training, which allows the model to learn the correct optimization direction for each reward axis independently within a single network. Extensive experiments on Stable Diffusion 1.5 and SDXL demonstrate that MCDPO achieves superior performance on benchmarks even with 18% and 3% training data respectively. Notably, our conditional framework enables dynamic and multiple-axis control at inference time using Classifier Free Guidance to amplify specific reward dimensions without additional training or external reward models.

Keywords: DPO · Diffusion · Multi-preference

1 Introduction

Reinforcement Learning from Human Feedback (RLHF) [15] has emerged as a key component in large-scale language model (LLM) training [11, 14, 27], and has recently been extended to generative image models [1, 8, 26]. Among these, Diffusion-DPO [23] applies Direct Preference Optimization (DPO) [18] to diffusion models using only binary win-lose labels via the Bradley-Terry (BT) model.

However, while a single image sample can possess various evaluation axes such as aesthetic quality, prompt fidelity, and safety, the BT model relies solely on a single win-lose label for the entire sample. This formulation makes it difficult to control or reflect specific, multi-dimensional preference axes during training. This fundamental conflict arises from the standard BT model's formulation, which models global preference as a linear combination of multiple reward axes.

* Co-corresponding Author

Consequently, when presented with a sample pair that is a global *win* but a local *lose* on a specific dimension, the model is forced to learn in the opposite direction of the intended preference for that dimension. Prior works address this through parameter merging [24], gradient regularization [13], SFT-based in-context reward modeling [28], or filtering to conflict-free pairs [3, 9], but none resolve the fundamental conflict within the DPO loss.

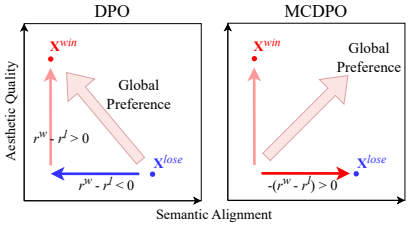

Fig. 1: DPO suffers from reward conflicts ($r^w - r^l < 0$) when global preference contradicts a specific dimension. MCDPO resolves this by disentangling axes to learn the correct direction for each dimension independently ($-(r^w - r^l) > 0$).

To resolve the issue, we propose Multi-reward Conditional Direct Preference Optimization (MCDPO), which injects a per-dimension preference outcome vector as a condition during the DPO loss calculation, transforming problematic reward-conflict pairs into a powerful supervision signal for disentanglement within a single network (See Figure 1). The key idea is a disentangled Bradley-Terry objective that models preferences independently for each dimension by explicitly encoding the win, lose, or tie status for each axis, ensuring every dimension is optimized in its correct direction.

By resolving the reward conflict, we experimentally demonstrate that MCDPO outperforms baselines and forms robust, efficient feature representations for each alignment axis. Notably, MCDPO trained solely on conflicting pairs surpasses standard DPO trained on randomly sampled data, where standard DPO actively degrades. This confirms that our method transforms reward conflicts from obstacles into a powerful learning signal. Furthermore, our conditional framework enables dynamic, multi-axis control at inference time by leveraging CFG [5] to amplify the score function towards higher-reward outcomes without additional training or external reward models.

In summary, the main contributions are as follows:

- We propose the disentangled BT objective and its practical implementation, Multi-reward Conditional DPO (MCDPO), to address the known reward conflict problem in BT-based DPO. Our method uses a preference outcome vector as a condition to disentangle reward axes.
- We experimentally show that MCDPO achieves superior performance and sample efficiency, learning robust representations for each alignment axis.
- We showcase MCDPO's conditional structure, which enables dynamic, multi-axis control during inference using Classifier-Free Guidance.

2 Background

2.1 Diffusion Models

Diffusion models are a class of generative models that learn to generate data by reversing a gradual noising process. [4] The forward process progressively adds

Gaussian noise to data $\mathbf{x}_0$ over T timesteps:

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}), \tag{1}$$

where β_t is a variance schedule. The reverse process learns to denoise by training a network ϵ_θ to predict the noise at each timestep. The training objective is:

$$\mathcal{L} = \mathbb{E}_{t,\mathbf{x}_0,\epsilon}\left[\|\epsilon - \epsilon_\theta(\mathbf{x}_t, t, c)\|^2\right], \tag{2}$$

where c represents conditioning information (e.g., text prompt), and $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ is the added noise to $\mathbf{x}_0$.

At inference, sampling starts from pure noise $\mathbf{x}_T \sim \mathcal{N}(0, \mathbf{I})$ and iteratively denoises to generate $\mathbf{x}_0$. The learned score function $\nabla_\mathbf{x} \log p_\theta(\mathbf{x}_t|c)$ guides this reverse process [22]. Classifier-Free Guidance (CFG) [5] strengthens conditioning by modifying the score function:

$$\nabla_\mathbf{x} \log p_\theta(\mathbf{x}_t) + \lambda_{\text{cfg}} \left(\nabla_\mathbf{x} \log p_\theta(\mathbf{x}_t|c) - \nabla_\mathbf{x} \log p_\theta(\mathbf{x}_t)\right), \tag{3}$$

where λ_{cfg} controls guidance strength. By amplifying the score difference, CFG enhances adherence to the conditioning signal c.

2.2 DPO in Diffusion Models

Direct Preference Optimization (DPO) [18] offers a simplified approach to align generative models with human preferences by directly optimizing policy models without requiring explicit reward models. DPO leverages the Bradley-Terry (BT) model [2] to model preference probabilities from pairwise comparisons. Given a pair of samples $\mathbf{x}^w$ (preferred) and $\mathbf{x}^l$ (less preferred) conditioned on prompt c, the BT model defines:

$$p_{\text{BT}}(\mathbf{x}^w > \mathbf{x}^l|c) = \sigma(r(\mathbf{x}^w, c) - r(\mathbf{x}^l, c)), \tag{4}$$

where $\sigma(\cdot)$ is the sigmoid function and $r(\mathbf{x}, c)$ is a reward function.

A key insight of DPO is that the reward can be expressed analytically in terms of the learned policy p_θ and a reference policy p_{ref}:

$$r(\mathbf{x}, c) = \beta \log\left(\frac{p_\theta(\mathbf{x}|c)}{p_{\text{ref}}(\mathbf{x}|c)}\right) + \beta \log Z(c), \tag{5}$$

where β is a KL regularization parameter and $Z(c)$ is the partition function. This formulation allows DPO to optimize preferences directly through the policy without training a separate reward model.

For diffusion models [23], the DPO objective operates at the denoising transition level. Given a preference (win-lose) pair $(\mathbf{x}^w, \mathbf{x}^l)$, the loss function is:

$$\begin{aligned}\mathcal{L}_{\text{DPO}} =& -\mathbb{E}\log\sigma\left(\beta\log\frac{p_\theta(\mathbf{x}_t^w|\mathbf{x}_{t+1}^w, c)}{p_{\text{ref}}(\mathbf{x}_t^w|\mathbf{x}_{t+1}^w, c)} - \beta\log\frac{p_\theta(\mathbf{x}_t^l|\mathbf{x}_{t+1}^l, c)}{p_{\text{ref}}(\mathbf{x}_t^l|\mathbf{x}_{t+1}^l, c)}\right) \\ =& -\mathbb{E}\log\sigma(-\beta T\omega_t\ (||\epsilon^w - \epsilon_\theta(\mathbf{x}_t^w, t)||^2 - ||\epsilon^w - \epsilon_{\text{ref}}(\mathbf{x}_t^w, t)||^2 \\ & - (||\epsilon^l - \epsilon_\theta(\mathbf{x}_t^l, t)||^2 - ||\epsilon^l - \epsilon_{\text{ref}}(\mathbf{x}_t^l, t)||^2)),\end{aligned} \tag{6}$$

where the expectation is over timesteps t and noisy samples from the diffusion forward process. This formulation trains the diffusion model to increase the likelihood of preferred samples while decreasing that of less preferred ones, relative to the reference policy.

In practice, generated samples are evaluated along multiple distinct axes, such as aesthetic quality, semantic alignment, and safety. However, the standard BT model aggregates these into a single scalar reward by a linear combination of different dimensions:

$$r(\mathbf{x}, c) = \sum_{i=1}^{D} w_i r_i(\mathbf{x}, c), \tag{7}$$

where r_i represents the reward for dimension i and w_i its weight. This leads to a fundamental reward conflict: when x^w is preferred globally but performs worse than x^l on a specific dimension j, the DPO loss may inadvertently degrade that dimension as shown in Figure 1 (Left).

Existing studies attempt to address or bypass this issue. In the LLM field, prominent approaches include: (1) dynamically combining expert model parameters based on user-specified reward weights [24], (2) training sequentially with regularization that encourages gradient alignment with previously learned preferences [13], and (3) avoiding DPO conflicts by using supervised fine-tuning (SFT) that incorporates reward scores directly in the prompt [28]. In the diffusion domain, a common strategy to avoid this conflict is to generate images until the samples that are dominant across all axes as preferred [9, 31]. Nevertheless, these approaches fail to resolve the fundamental conflict within the DPO loss formulation itself. By either training only on non-conflicting pairs (which is sample-inefficient) or using alternative objectives (like SFT or regularization), they rely on a limited learning signal rather than directly solving the ambiguity in DPO. In this work, we aim to analyze the problem and resolve the limitations.

3 Method

We first analyze the reward conflict and propose a disentangled BT objective (Section 3.1), then introduce MCDPO as its practical single-network implementation (Section 3.2), address gradient domination via reward dropout (Section 3.3), and show how our conditional framework enables test-time reward control (Section 3.4).

3.1 Multi-reward Disentanglement

As discussed in Section 2, the standard BT model collapses multi-dimensional preferences into a single scalar preference:

$$p_{BT}(\mathbf{x}^w > \mathbf{x}^l | c) = \sigma\left[\sum_i w_i \big(r_i(\mathbf{x}^w, c) - r_i(\mathbf{x}^l, c)\big)\right]. \tag{8}$$

This formulation leads to a reward conflict when a sample $\mathbf{x}^w$ is globally preferred over $\mathbf{x}^l$, but is worse than x^l on a particular dimension j, (i.e., $r_j(\mathbf{x}^w, c) < r_j(\mathbf{x}^l, c)$). Because the standard DPO loss optimizes only for the global win, it inadvertently compels the model to learn in the opposite direction for dimension j, actively degrading that specific quality. To resolve this, we first introduce a disentangled Bradley-Terry objective. The core idea is to model the preference for each dimension. We explicitly introduce a pairwise preference outcome vector $\gamma(\mathbf{x}, \mathbf{y}) \in \{-1, 0, 1\}^D$, where each entry indicates whether $\mathbf{x}$ wins, loses, or ties against $\mathbf{y}$ on reward axis i, *i.e.*, $\gamma_i(x, y) = \text{sign}(r_i(x) - r_i(y))$.

Using γ, we define the disentangled Bradley-Terry $p_{BT}^{\perp}$ target:

$$p_{BT}^{\perp}(\mathbf{x}^w > \mathbf{x}^l \mid c, \gamma(\mathbf{x}^w, \mathbf{x}^l)) = \sigma\left[\sum_i w_i \gamma_i(\mathbf{x}^w, \mathbf{x}^l)\big(r_i(\mathbf{x}^w, c) - r_i(\mathbf{x}^l, c)\big)\right]. \tag{9}$$

Eq. (9) specifies the desired preference probability once the per-dimension outcomes are known. This formulation ensures each dimension is optimized in the correct direction. In particular, if the j-th dimensional preference is inverted (e.g., $\mathbf{x}^l$ wins on aesthetics), γ_j becomes negative, which flips the sign for that dimension's contribution. By doing so, we can resolve the sign ambiguity in conflicting pairs and provide correct supervision across all dimensions.

3.2 Multi-reward Conditional DPO

Because standard DPO is structurally limited to modeling a single reward axis, modeling multiple axes would require training multiple separate implicit reward diffusion models.

$$r_i(\mathbf{x}, c) = \beta \log\left(\frac{p_i(\mathbf{x}|c)}{p_{\text{ref}}(\mathbf{x}|c)}\right) + \beta \log Z(c) \tag{10}$$

this approach would require optimizing a loss:

$$L_M = -\mathbb{E}\ \log\sigma\left[\beta \sum_i w_i \gamma_i \left(\log\frac{p_i(\mathbf{x}^w|c)}{p_{\text{ref}}(\mathbf{x}^w|c)} - \log\frac{p_i(\mathbf{x}^l|c)}{p_{\text{ref}}(\mathbf{x}^l|c)}\right)\right], \tag{11}$$

which has two major drawbacks. First, it does not fundamentally resolve conflicts, as each model is still trained on pairs where its dimension may conflict with others. Second, it requires training D separate models which is computationally prohibitive. The first issue can be addressed by incorporating the per-dimension preference indicator i from our disentangled BT objective (9), and the separate models can be further consolidated into a single network conditioned on i:

$$L'_M = -\mathbb{E}\ \log\sigma\left[\beta \sum_i w_i \gamma_i \left(\log\frac{p(\mathbf{x}^w|c, i)}{p_{\text{ref}}(\mathbf{x}^w|c)} - \log\frac{p(\mathbf{x}^l|c, i)}{p_{\text{ref}}(\mathbf{x}^l|c)}\right)\right]. \tag{12}$$

This formulation effectively resolves reward conflicts, as conditioning the model on i ensures that each specific quality dimension is optimized in its intended direction. However, it remains computationally prohibitive. The summation inside

the sigmoid requires evaluating the log-likelihood ratio independently for each dimension i, necessitating D separate forward-backward passes per training step.

To resolve this, we propose Multi-reward Conditional DPO (MCDPO), an efficient solution that models $p_{BT}^{\perp}$ (9) using a single conditional diffusion model. Following DPO, we interpret the policy/reference log-ratio as an implicit reward, but make it conditional on the pairwise outcome vector γ:

$$r_\theta(\mathbf{x}, c, \gamma) = \beta \log \frac{p_\theta(\mathbf{x} \mid c, \gamma)}{p_{\text{ref}}(\mathbf{x} \mid c)} + \beta \log Z(c, \gamma). \tag{13}$$

This induces a conditional Bradley-Terry probability

$$\begin{aligned} \hat{p}_\theta^{\perp}(\mathbf{x}^w > \mathbf{x}^l \mid c, \gamma) &= \sigma\big(r_\theta(\mathbf{x}^w, c, \gamma) - r_\theta(\mathbf{x}^l, c, \gamma)\big) \\ &= \sigma\left(\beta \log \frac{p_\theta(\mathbf{x}^w \mid c, \gamma)}{p_{\text{ref}}(\mathbf{x}^w \mid c)} - \beta \log \frac{p_\theta(\mathbf{x}^l \mid c, \gamma)}{p_{\text{ref}}(\mathbf{x}^l \mid c)}\right). \end{aligned} \tag{14}$$

Under sufficient model capacity and an expressive conditioning mechanism, the conditional policy $p_\theta(\mathbf{x}|c, \gamma)$ can, in principle, learn a distinct implicit reward function for each value of $\gamma \in \{-1, 0, 1\}^D$.

Let $r_{\text{eff}}(\mathbf{x}, c, \gamma) \triangleq \sum_i w_i \gamma_i \, r_i(\mathbf{x}, c)$ be the effective reward under condition γ. By construction, the effective reward difference satisfies:

$$r_{\text{eff}}(\mathbf{x}^w, c, \gamma) - r_{\text{eff}}(\mathbf{x}^l, c, \gamma) = \sum_i w_i |r_i(\mathbf{x}^w, c) - r_i(\mathbf{x}^l, c)| \geq 0, \tag{15}$$

which guarantees that the training label ($\mathbf{x}^w$ preferred) is consistent with the effective reward $r_{\text{eff}}(\mathbf{x}, c, \gamma)$ under the Bradley-Terry model for every γ. Since each fixed γ reduces L_{MC} (Eq. (17)) to a standard DPO objective with ground-truth reward r_{eff}, the DPO optimality result [18] directly implies that at the optimum θ^*, the induced reward satisfies:

$$r_{\theta^*}(\mathbf{x}, c, \gamma) = \sum_i w_i \gamma_i \, r_i(\mathbf{x}, c) + g(c, \gamma), \tag{16}$$

where $g(c, \gamma)$ is independent of $\mathbf{x}$ and thus cancels in the pairwise difference $r_{\theta^*}(\mathbf{x}^w, c, \gamma) - r_{\theta^*}(\mathbf{x}^l, c, \gamma)$. Substituting this into Eq. (14) recovers the disentangled target $p_{BT}^{\perp}$ in Eq. (9), establishing $\hat{p}_{\theta^*}^{\perp} = p_{BT}^{\perp}$.

Intuitively, the training dynamics naturally drive the model toward this decomposition. Consider a conflicting pair where $\gamma = [\text{aesthetic: } +1, \text{ semantic: } -1]$. The loss requires the model to increase the likelihood of $\mathbf{x}^w$ along the aesthetic axis while *decreasing* it along the semantic. The same pair also appears with the flipped condition $\gamma = [\text{aesthetic: } -1, \text{ semantic: } +1]$ via our final symmetric loss Eq. (17), which imposes the opposite constraint. To satisfy both simultaneously, the model must learn to modulate its implicit reward for each dimension independently as a function of γ, which is precisely the structure captured by Eq. (16): While Eq. (16) may not hold exactly in practice due to the finite capacity of the conditioning mechanism, Table 7 empirically validates that our architecture approximates this decomposition well. The single MCDPO model

achieves implicit reward accuracy comparable to specialist models trained on individual dimensions.

$$L_{MC} = -\mathbb{E}\left[\log\sigma\left(\beta\log\frac{p(\mathbf{x}^w|c,\gamma^{wl})}{p_{\text{ref}}(\mathbf{x}^w|c)} - \beta\log\frac{p(\mathbf{x}^l|c,\gamma^{wl})}{p_{\text{ref}}(\mathbf{x}^l|c)}\right) + \log\sigma\left(\beta\log\frac{p(\mathbf{x}^l|c,\gamma^{lw})}{p_{\text{ref}}(\mathbf{x}^l|c)} - \beta\log\frac{p(\mathbf{x}^w|c,\gamma^{lw})}{p_{\text{ref}}(\mathbf{x}^w|c)}\right)\right], \tag{17}$$

where γ^{wl} and γ^{lw} denote $\gamma(x^w, x^l)$ and $\gamma(x^l, x^w)$, respectively. To feed γ into the model, we devise a reward conditioning module described in Section 3.5.

3.3 Mitigating Gradient Domination

While our L_{MC} objective Equation (17) successfully disentangles reward axes, naively training with it can lead to gradient domination. This occurs when dimensions that are easier to learn suppress the learning signals for harder dimensions, leading to unbalanced optimization.

Under the reward decomposition at optimum (Eq. (16)), analyzing the gradient of L_{MC} reduces to an equivalent analysis of L'_M (Eq. (12)), since the conditional policy $p_\theta(\mathbf{x}|c,\gamma)$ decomposes into the same per-dimension structure. We can write the gradient of the loss as follows:

$$\begin{aligned}\nabla_\theta L'_M &= (\sigma(z) - 1)\nabla_\theta z,\\ \nabla_\theta z &= \sum_i \beta w_i(\nabla_\theta \log p_\theta(\mathbf{x}^w|c,i) - \nabla_\theta \log p_\theta(\mathbf{x}^l|c,i)),\end{aligned} \tag{18}$$

where z aggregates the per-dimension log-likelihood ratios. We can find two sources of the issue. First, if any single dimension becomes highly confident, $\sigma(z) \to 1$, the global gradient goes to zero; therefore, the entire training converges to the local optima. Second, within $\nabla_\theta z$, if the model becomes highly discriminative for dimension i, which causes a large gradient norm for that dimension; it can dominate the sum and overwhelm gradients from other dimensions.

To address this, we introduce dimensional reward dropout. During training, we randomly drop individual reward dimensions by setting their condition $\gamma_i = 0$. By doing so, we can zero out the contribution of the dropped dimension to both the $\sigma(z)$ term and the summed gradient $\nabla_\theta z$. This simple technique prevents any single dimension from consistently dominating the training signal, ensuring balanced optimization across all axes.

3.4 Test-time Reward Optimization

A significant advantage of our conditional framework is that it enables dynamic, multi-axis reward optimization at inference time. Our training objective implicitly guides the model to represent two distinct distributions, including the distribution of preferred samples $p(\mathbf{x}|c,\gamma^w)$ and the distribution of non-preferred samples $p(\mathbf{x}|c,\gamma^l)$:

$$p(\mathbf{x}|c,\gamma^w) = p_{\text{ref}}(\mathbf{x}|c)\exp(r^w(\mathbf{x},c)/\beta)/Z(c). \tag{19}$$

This structure is a natural fit for Classifier-Free Guidance (CFG). By computing the difference between the score functions for the "win" and "lose" conditions, we can derive the gradient of the implicit reward difference ($r^w - r^l$):

$$\nabla_{\mathbf{x}} \log\ p(\mathbf{x}|c, \gamma^w) - \nabla_{\mathbf{x}} \log\ p(\mathbf{x}|c, \gamma^l) = \nabla_{\mathbf{x}}(r^w - r^l)/\beta. \tag{20}$$

The gradient can be incorporated into the standard CFG sampling as follows:

$$\nabla_{\mathbf{x}} \log\ p(\mathbf{x}|\gamma^l) + \lambda_{cfg}(\nabla_{\mathbf{x}} \log\ p(\mathbf{x}|c, \gamma^w) - \nabla_{\mathbf{x}} \log\ p(\mathbf{x}|\gamma^l)). \tag{21}$$

By doing so, we can steer the generation process away from the non-preferred (γ^l) and toward the preferred (γ^w). This enables fine-grained, dynamic control, allowing a user to specify a weighted combination of preference gradients at test time without requiring any external reward models.

3.5 Context-aware Reward Conditioning

We encode the preference outcome vector γ as natural language tokens (e.g., "win tie", meaning win for the first (*e.g.*, aesthetic) score and tie for the second (*e.g.*, semantic alignment) via the text encoder to get an initial embedding e_γ. This embedding is then injected into the model.

First, the U-Net latent z_t is updated using parallel cross-attention (CA) blocks with the text prompt c_{prompt} and the final context-aware reward (CAR) embedding c_{reward}:

$$z'_t = \text{CA}(z_t, c_{\text{prompt}}) + \lambda \cdot \text{CA}'(z_t, c_{\text{reward}}), \tag{22}$$

where z'_t is the updated latent, CA$'$ is a copy of the pretrained CA module [29], and λ is a weighting scalar, which we set to 1 in all our experiments.

For dimensions requiring both text and image context (e.g., semantic alignment), the embedding c_{reward} is computed using a context-aware conditioning module. This module refines the initial γ embedding e_γ by attending to both the text prompt and the image latent sequentially:

$$h_{\text{SA}} = \text{SA}(e_\gamma, c_{\text{prompt}}),\ h_{\text{CA}} = \text{CA}''(h_{\text{SA}}, z_t),\ c_{\text{reward}} = \text{MLP}(h_{\text{CA}}), \tag{23}$$

where SA, CA$''$, and MLP are zero-initialized to preserve pretrained knowledge. This final c_{reward} embedding is then fed into Equation (22) as the input to CA$'$.

Training Initialization The modules introduced above (CA$'$ and the context-aware reward conditioning) are newly added parameters that are zero-initialized to preserve pretrained knowledge. As a result, the conditioning signal γ has no effect at initialization. Just as SFT before DPO in LLMs [18] helps the model learn basic instruction-following before preference optimization, we perform a supervised fine-tuning stage (MCSFT) that trains the conditioning modules to reconstruct preferred samples $\mathbf{x}^w$ given γ^w with a frozen U-Net, allowing the model to learn basic condition-following before the MCDPO alignment begins. We emphasize that MCSFT serves solely as initialization, as shown in Tabs. 2 and 3. MCSFT alone is insufficient and the subsequent MCDPO alignment stage is critical for final performance.

4 Experiments

4.1 Experimental Setup

Training Dataset. Following the previous works, we adopt Pick-a-Pic v2 [7], containing 1M human-preference image pairs as the training dataset. We augment each pair with four model-based reward dimensions: aesthetic quality (Aesthetic Predictor [20]), semantic alignment (CLIP [17]), overall quality (HPSv2 [25]), and prompt alignment (PickScore [7]). For our method, the 5-dimensional preference outcome vector γ is then constructed by encoding the win/lose/tie status for the original human label alongside these four proxy reward dimensions by discretized into 100 bins (same bin = ties) after normalization on training dataset. We use 18% of the data (∼180K pairs) for StableDiffusion1.5 (SD1.5) [19] and 3% (∼30K pairs) for SDXL [16] while baselines use the full dataset.
Configuration. Training proceeds in two phases: (1) MCSFT pre-training and (2) MCDPO alignment . The total computational time, including the prior MCSFT pretraining, was 16 and 32 A100 GPU hours for SD1.5 and SDXL, respectively. Other implementation details are provided in the Appendix.
Evaluation. To evaluate MCDPO's effectiveness, we compare against existing baselines trained on the Pick-a-Pic v2 dataset: Diffusion-DPO [23], Diffusion-KTO [10], MAPO [6], and DSPO [32]. Following standard convention, we test text-to-image generation on Pick-a-Pic v2 test set [7], PartiPrompts [30], and HPDv2 test set [25], measuring performance with PickScore, LAION Aesthetic Score, HPSv2, CLIP, ImageReward [26], and MPS [31].

4.2 Main Results

As shown in Table 1, MCDPO demonstrates dominant performance on the SD1.5, significantly outperforming all baselines (Diffusion-DPO, Diffusion-KTO, and DSPO) across every test set. We observe substantial gains in both PickScore and HPSv2. The most notable improvement is in the aesthetic score, which shows

Table 1: Main Results on SD1.5 (win rates vs. SD1.5 baseline). * indicates the model checkpoint released by the authors. **Bold** and underline indicate the best and second-best results, respectively. We will use the same notation throughout the paper.

Datasets	Models	PickScore	Aesthetic	HPSv2	CLIP	ImageReward	MPS	Average
PickV2	Diff. DPO*	75.52	65.08	70.28	57.80	64.44	67.33	66.74
	Diff. KTO*	73.40	71.24	82.24	56.64	76.04	69.30	71.48
	DSPO	76.36	75.88	81.92	**59.16**	77.36	68.80	73.24
	MCSFT	78.24	78.16	89.60	57.36	79.20	69.53	75.35
	MCDPO	**86.20**	**91.88**	**93.44**	57.64	**82.92**	**76.75**	**81.47**
PartiPrompts	Diff. DPO*	66.72	60.72	64.58	53.49	62.62	64.69	62.14
	Diff. KTO*	65.74	69.85	80.14	55.82	72.48	63.86	67.98
	DSPO	68.07	75.18	79.47	**56.67**	72.85	65.56	69.63
	MCSFT	60.42	68.75	69.18	48.49	65.63	60.88	62.22
	MCDPO	**80.94**	**92.52**	**86.46**	50.65	**76.35**	**72.52**	**76.57**
HPDv2	Diff. DPO*	76.80	66.60	70.90	**56.70**	62.50	66.66	66.69
	Diff. KTO*	75.55	74.40	86.35	53.20	78.20	69.17	72.81
	DSPO	78.10	78.40	86.55	55.45	79.80	70.25	74.75
	MCSFT	73.25	78.75	81.40	50.60	74.45	70.17	71.43
	MCDPO	**87.75**	**94.25**	**93.00**	51.75	**83.50**	**76.75**	**81.16**

Table 2: Main Results on SDXL (win rates vs. SDXL baseline).

Datasets	Models	PickScore	Aesthetic	HPSv2	CLIP	ImageReward	MPS	Average
PickV2	Diff. DPO*	71.60	49.20	72.92	**61.24**	68.64	60.36	63.99
	MAPO*	50.65	56.00	80.00	59.15	74.19	52.26	62.02
	DSPO	68.80	44.76	71.80	60.60	74.24	44.77	60.82
	MCSFT	50.72	37.24	72.36	59.12	66.20	45.96	55.26
	MCDPO	**74.92**	**63.40**	**85.84**	58.16	**77.72**	**64.31**	**70.72**
PartiPrompts	Diff. DPO*	64.09	54.66	67.34	**58.21**	68.44	62.70	62.57
	MAPO*	50.20	**68.07**	76.47	52.38	69.73	53.10	61.65
	DSPO	64.58	56.92	71.26	56.67	**78.49**	53.49	63.56
	MCSFT	46.69	46.32	71.51	55.27	64.28	51.96	56.81
	MCDPO	**72.06**	66.67	**82.41**	55.56	73.47	**64.01**	**69.03**
HPDv2	Diff. DPO*	67.80	54.85	72.10	**55.20**	67.85	57.53	62.55
	MAPO*	50.80	56.55	84.35	54.25	70.75	48.35	60.84
	DSPO	70.30	50.35	78.80	54.75	73.85	53.75	63.63
	MCSFT	51.85	40.10	78.30	54.75	66.95	46.98	56.48
	MCDPO	**76.40**	**69.25**	**90.60**	54.85	**75.20**	**69.26**	**72.59**

a massive performance gain over existing methods. Furthermore, MCDPO achieves consistent gains on ImageReward and MPS, metrics not used as a direct reward dimension, indicating strong generalization capabilities. This significant advantage is maintained on SDXL, as detailed in Table 2. MCDPO again outperforms all baselines on PickScore, HPSv2, and MPS, achieving the highest average score by a large margin.

Table 3: Fair comparison across training regimes. All models are trained on an identical 18% data budget. MCDPO significantly outperforms all baselines, including filtered and conflict-only regimes, demonstrating its robust conflict resolution.

	Pick	Aes	HPS	CLIP	IR	MPS	Avg
(1) DPO	75.1	62.6	71.7	60.4	62.6	68.1	66.7
(2) DPO merged	73.4	63.5	70.0	56.1	62.0	67.9	63.8
(3) CDPO merged	63.4	84.8	71.7	32.3	61.1	65.0	63.1
(4) Filtered DPO	83.6	84.1	85.7	**62.8**	79.2	76.6	78.7
(5) MCSFT (full)	78.1	78.3	81.7	55.4	77.0	68.7	73.2
(6) MCDPO	**86.2**	**91.8**	**93.4**	57.6	**82.9**	**76.7**	**81.1**
(7) Conflict DPO	62.0	57.0	60.4	56.4	56.6	61.9	59.0
(8) Conflict MCDPO	85.2	83.2	89.4	56.0	74.0	74.3	77.0

While MCDPO achieves substantial improvements across human-preference metrics, CLIP scores are slightly lower than standard DPO. This reflects a natural multi-objective trade-off, where the moderate CLIP enables dramatic gains in aesthetics (91.9%), HPSv2 (93.4%), ImageReward (82.9%), and MPS (76.7%). Importantly, Table 6 confirms that MCDPO has learned a disentangled CLIP representation that can be amplified at test-time. In Table 6, targeting CLIP alone yields 59.0%, and Aes-CLIP temporal scheduling, which assigns aesthetic 'tie' for the first 25 steps and 'win' for the remaining 25, further recovers the gap to 59.6% while preserving the highest overall average (81.5%).

To ensure a fair comparison across various training regimes, we evaluate our method using an identical 18% subset of the Pick-a-Pic v2 dataset. Table 3 compares eight distinct configurations to demonstrate how effectively MCDPO resolves reward conflicts. (1) Standard DPO on randomly sampled pairs, (2) merged checkpoints of DPO models trained independently on each reward axis — same supervision, (3) merged checkpoints of Conditional DPO (CDPO) models independently trained on each axis — same supervision and parameters (4) DPO-Filtered on conflict-free pairs only, (5) full MCSFT training (unfreezing the U-Net after the initial MCSFT phase) — same supervision and parameters (6) our proposed MCDPO on randomly sampled pairs, (7) DPO trained strictly on conflict-only pairs, and (8) MCDPO trained strictly on conflict-only pairs.

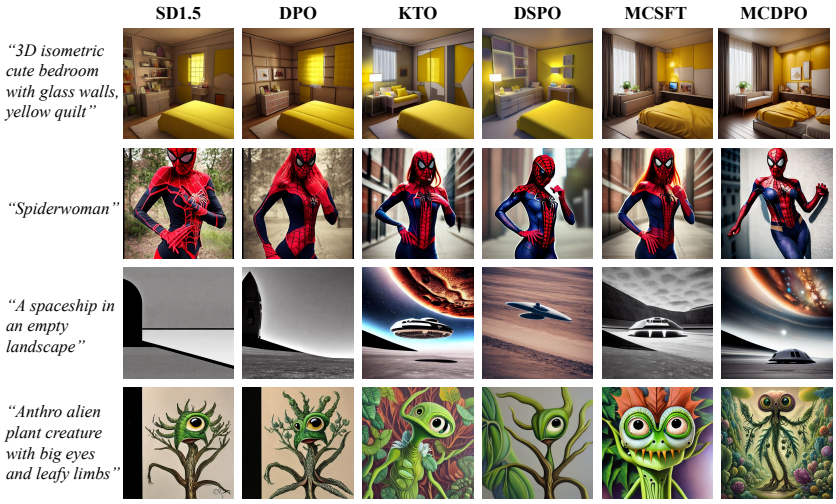

Fig. 2: Qualitative results of SD1.5.

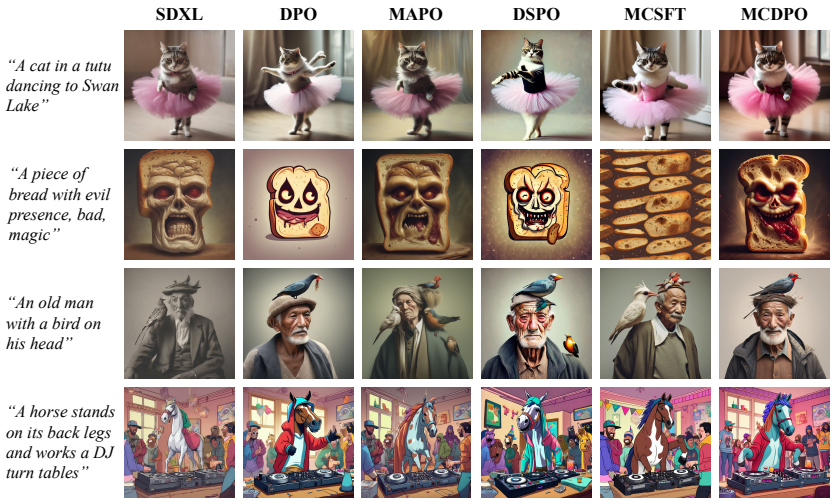

Fig. 3: Qualitative results of SDXL.

Notably, MCDPO surpasses DPO-Filtered, which explicitly avoids conflicting signals by discarding 82% of training data, demonstrating that MCDPO successfully transforms conflict pairs from obstacles into highly effective training signals. Furthermore, we observe that full MCSFT training alone actually degrades performance, confirming that the MCDPO alignment objective is indispensable for stable optimization. To explicitly demonstrate that our model fundamentally resolves reward conflicts, we compare the models trained exclusively on conflicting pairs ((7) and (8)). When restricted to conflict-only data,

Table 4: Standard DPO specialists.

	Pick	Aes	HPS	CLIP	IR	MPS	Avg
Human	75.1	62.6	71.7	**60.4**	62.6	68.1	66.7
Pick	72.6	63.4	72.0	59.2	57.6	66.6	65.2
Aes	69.8	66.3	66.4	56.4	59.5	66.0	64.1
HPS	69.8	63.9	70.1	54.5	59.5	65.9	64.0
CLIP	69.3	62.4	67.7	58.5	58.9	63.2	63.3
Merged	73.4	63.5	70.0	56.1	62.0	67.9	63.8
MCDPO	**86.2**	**91.8**	**93.4**	57.6	**82.9**	**76.7**	**81.1**

Table 5: Cond. DPO specialists.

	Pick	Aes	HPS	CLIP	IR	MPS	Avg
Human	79.0	79.8	86.5	46.8	75.9	73.2	73.5
Pick	80.7	82.8	88.5	49.0	76.8	74.7	75.4
Aes	79.6	88.1	87.2	49.1	76.2	72.5	75.4
HPS	77.8	81.8	88.0	50.2	77.6	73.3	74.7
CLIP	75.3	79.3	84.5	49.3	76.0	70.9	72.5
Merged	63.4	84.8	71.7	32.3	61.1	65.0	63.1
MCDPO	**86.2**	**91.8**	**93.4**	**57.6**	**82.9**	**76.7**	**81.1**

standard DPO suffers a substantial relative performance degradation of 11.5% compared to its randomly sampled baseline. In contrast, MCDPO experiences a minor drop of only 5.1%. In fact, Conflict MCDPO (Avg 77.0) trained solely on conflicting pairs surpasses standard DPO on randomly sampled data (Avg 66.7) by a large margin, confirming that MCDPO does not merely tolerate reward conflicts but actively leverages them as a stronger supervision signal.

Remarkably, the conflict-trained MCDPO maintains high performance even when conditioned on "all-win" across every dimension, a scenario never encountered during training. We provide further analysis in Figure 4 and agreement ratio over reward dimensions in appendix.

Furthermore, we visualize the qualitative result of MCDPO against baseline methods on both SD1.5 and SDXL, in Figure 2 and Figure 3, respectively. Across both models, our method demonstrates a clear superiority in generating images that are more aesthetically pleasing and detailed.

4.3 Analysis

Multi-dimensional Learning. Tables 4 and 5 compare MCDPO against specialist models trained on a single reward axis, using both the standard DPO objective (Table 4) and our Conditional DPO paradigm with identical architecture and hyperparameters (Table 5).

Remarkably, the MCDPO model not only achieves the highest average score but also outperforms every individual specialist model on its own specialized metric, with the sole exception of the CLIP score in Table 4. For example, the MCDPO model (Aesthetic win-rate 91.8%) significantly surpasses the model trained exclusively on 'Aesthetic' rewards (Aesthetic win-rate 88.1%) in Table 5.

We conjecture that MCDPO mitigates potential training reward overfitting by leveraging cross-dimensional references to learn robust features rather than train-specific shortcuts. For example, one dimension may exploit spurious cues that improve training reward but harm semantics. The CLIP dimension can provide a semantic reference that helps mitigate this.

Multi-dimensional Inference. As demonstrated in Table 6, our single model, trained to disentangle reward axes, can perform dynamic preference-enhancing sampling at test-time as described in Section 3.4. When steering the sampling to amplify only a specific or multiple dimensions, we observe distinct and controlled behaviors. Critically, when enhancing CLIP, the model achieves 59.0%, recovering from the all-dimension default (57.6%) and demonstrating that MCDPO has learned a disentangled CLIP representation that can be amplified when desired.

Table 6: Test-time reward optimization by guiding the MCDPO model towards specific target dimensions.

	Pick	Aes	HPS	CLIP	IR	MPS	Avg
Human	78.6	79.0	83.3	55.2	73.7	67.7	72.9
Pick	78.1	79.0	85.1	57.6	78.4	69.9	74.7
Aes	77.3	**95.8**	79.1	49.3	73.1	68.8	73.8
HPS	77.0	79.5	85.6	**59.0**	77.8	71.1	74.9
CLIP	76.4	76.5	85.1	**59.0**	75.0	67.8	73.3
Aes-CLIP Cont.	88.0	90.6	91.2	59.6	81.0	78.4	81.5
ALL	**86.2**	91.8	**93.4**	57.6	**82.9**	**76.7**	**81.1**

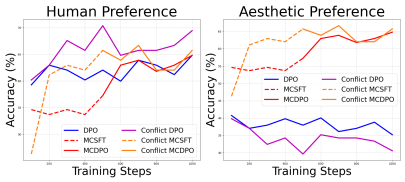

Fig. 4: Implicit reward accuracy on Human-Aesthetic conflict-only pairs.

This demonstrates that MCDPO has successfully learned a disentangled representation of CLIP rewards and can boost this dimension when desired. Similarly, enhancing Aes (Aesthetic) generates samples with exceptionally high aesthetic scores (95.8%), validating dimension-specific control.

Notably, when enhancing PickScore and HPSv2, their corresponding scores increase as expected. However, we also observe a concurrent rise in CLIP, ImageReward, and MPS scores. We attribute this to the nature of PickScore and HPSv2 as holistic metrics that evaluate overall image quality, which naturally correlates with other reward axes.

While coarse axis-level optimization is detailed in the supplementary material, we highlight MCDPO's capability for fine-grained temporal control. To mitigate the natural trade-off between aesthetics and semantics, we leverage the step-wise frequency properties of diffusion models. Semantic structures primarily form during early low-frequency steps, whereas aesthetic details emerge in later high-frequency steps. By assigning an aesthetic tie for the first 25 steps and a win condition for the remaining 25 steps, our Aes-CLIP Control (Table 6) significantly boosts the CLIP score while maintaining the average win rate. We further provide an incremental analysis in Appendix, where adding each reward axis as "win" progressively improves the corresponding metric, confirming that each dimension contributes independently to the final performance.

Implicit Reward Modeling Table 7 demonstrates that MCDPO accurately models implicit reward functions for each dimension. MCDPO achieves comparable accuracy to specialist models trained on individual dimensions, while being significantly more computationally efficient by learning all dimensions within a single network. Furthermore, our conditional DPO framework simultaneously models both r^w (preferred) and r^l (non-preferred) implicit reward models as described in Equation (19). Leveraging both models together yields performance gains over using either model alone. For example, in the Aesthetic dimension, while the win-only model achieves 50.8% and the lose-only model achieves 53.3%, combining both models substantially improves accuracy to 66.2%. Similarly, for Human and PickScore dimensions, despite the lose models showing relatively lower accuracies (42.2% and 41.0% respectively), combining them with the win models consistently yields performance improvements.

Reward Conflicts To demonstrate that MCDPO actively resolves reward conflicts beyond mere dimensional disentanglement, we plot the implicit rewards during training, on a filtered subset where Human and Aesthetic preferences

Table 7: Implicit reward accuracy comparison between MCDPO and specialist.

	Human	Pick	Aes	HPS	CLIP
Specialists	57.8	61.0	67.1	58.2	59.9
MCDPO	58.7	61.7	66.2	55.0	59.4
MCDPO win	58.5	58.7	50.8	52.9	59.4
MCDPO lose	42.2	41.0	53.3	46.8	56.8

Table 8: Ablation study of MCDPO components on PickV2.

	Pick	Aes	HPS	CLIP	IR	MPS	Avg
MCDPO	86.2	91.8	**93.4**	**57.6**	**83.0**	76.7	**81.1**
reward dropout 0.1	85.8	93.5	92.4	55.6	82.4	74.8	80.7
reward dropout 0.0	**87.3**	**96.0**	92.1	49.2	79.2	**77.1**	80.1
w/o MCSFT	53.8	79.2	55.1	48.8	55.7	70.7	60.5
w/o CAR module	80.1	90.0	80.7	45.7	69.8	54.4	70.1

conflict Figure 4. While standard DPO and Conflict DPO suffer active aesthetic feature degradation dropping to 35% and 30% respectively, MCDPO successfully rectifies the gradient direction via its axis-wise guidance γ. This isolates the aesthetic dimension from the global penalty and enables convergence towards the ideal disentangled distribution $p_{BT}^{\perp}$.

Component Ablation Study. We conduct an ablation study to validate the contribution of MCDPO's core components: dimensional reward dropout, the context-aware reward module, and the MCSFT pre-training stage in Table 8.

First, removing the dimensional reward dropout leads to a seemingly dominant aesthetic score (95.8%) but causes a catastrophic collapse in the CLIP score (49.3%). We believe this is a clear case of the gradient domination problem discussed in Section 3.3. The easier aesthetic dimension appears to saturate the learning signal, preventing the harder CLIP dimension from being optimized. Our reward dropout technique effectively mitigates this issue, ensuring balanced optimization and contributing significantly to the superior overall performance. Applying a reward dropout of just 0.1 significantly mitigates this domination.

Removing MCSFT pre-training causes a drastic performance decline, showing non-initialized conditioning is disruptive. However, MCSFT alone is insufficient. On SDXL, MCSFT performs poorly (Avg 57.1), while the full MCDPO model achieves 72.0 (Table 2), proving the MCDPO alignment stage itself is critical.

Finally, ablating the Context-Aware Reward conditioning module Eq. (23) causes a severe performance drop in all metrics except for aesthetics. These affected metrics (PickScore, HPS, CLIP, IR) measure alignment between the image and the text prompt. This result confirms our hypothesis: to condition on rewards like semantic alignment effectively, the model must be aware of both the image and the text context, validating the design of our module.

5 Conclusion

We introduced Multi Reward Conditional DPO (MCDPO), a framework that resolves the reward conflict in the Bradley-Terry formulation by conditioning on preference outcome vectors, transforming conflicting pairs into robust training signals for disentangled alignment. MCDPO achieves state-of-the-art performance with superior sample efficiency, and enables dynamic multi-dimensional control at inference without additional training or external reward models. One limitation is the reliance on proxy reward models to construct the preference outcome vector, which upper-bounds alignment quality by their accuracy, and introduces preprocessing overhead compared to standard DPO's binary labels.

6 Acknowledgement

This work was supported by the Korean Government through the grants from IITP (RS-2021-II211343, RS-2022-II220320, RS-2025-25442338) and NRF (RS-2026-25470591).

References

1. Black, K., Janner, M., Du, Y., Kostrikov, I., Levine, S.: Training diffusion models with reinforcement learning. arXiv preprint arXiv:2305.13301 (2023)
2. Bradley, R.A., Terry, M.E.: Rank analysis of incomplete block designs: I. the method of paired comparisons. Biometrika **39**(3/4), 324–345 (1952)
3. Gu, H., Wang, H., Mei, Y., Zhang, M., Jin, Y.: Paretohqd: Fast offline multiobjective alignment of large language models using pareto high-quality data. arXiv preprint arXiv:2504.16628 (2025)
4. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
5. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)
6. Hong, J., Paul, S., Lee, N., Rasul, K., Thorne, J., Jeong, J.: Margin-aware preference optimization for aligning diffusion models without reference (2024)
7. Kirstain, Y., Polyak, A., Singer, U., Matiana, S., Penna, J., Levy, O.: Pick-a-pic: An open dataset of user preferences for text-to-image generation. Advances in Neural Information Processing Systems **36**, 36652–36663 (2023)
8. Lee, K., Liu, H., Ryu, M., Watkins, O., Du, Y., Boutilier, C., Abbeel, P., Ghavamzadeh, M., Gu, S.S.: Aligning text-to-image models using human feedback. arXiv preprint arXiv:2302.12192 (2023)
9. Lee, K., Li, X., Wang, Q., He, J., Ke, J., Yang, M.H., Essa, I., Shin, J., Yang, F., Li, Y.: Calibrated multi-preference optimization for aligning diffusion models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 18465–18475 (2025)
10. Li, S., Kallidromitis, K., Gokul, A., Kato, Y., Kozuka, K.: Aligning diffusion models by optimizing human utility. In: The Thirty-eighth Annual Conference on Neural Information Processing Systems (2024)
11. Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., et al.: Deepseek-v3 technical report. arXiv preprint arXiv:2412.19437 (2024)
12. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
13. Lou, X., Zhang, J., Xie, J., Liu, L., Yan, D., Huang, K.: Spo: Multi-dimensional preference sequential alignment with implicit reward modeling. arXiv preprint arXiv:2405.12739 (2024)
14. OpenAI, :, Agarwal, S., Ahmad, L., Ai, J., Altman, S., Applebaum, A., Arbus, E., Arora, R.K., Bai, Y., Baker, B., Bao, H., Barak, B., Bennett, A., Bertao, T., Brett, N., Brevdo, E., Brockman, G., Bubeck, S., Chang, C., Chen, K., Chen, M., Cheung, E., Clark, A., Cook, D., Dukhan, M., Dvorak, C., Fives, K., Fomenko, V., Garipov, T., Georgiev, K., Glaese, M., Gogineni, T., Goucher, A., Gross, L., Guzman, K.G., Hallman, J., Hehir, J., Heidecke, J., Helyar, A., Hu, H., Huet, R., Huh, J., Jain, S., Johnson, Z., Koch, C., Kofman, I., Kundel, D., Kwon, J., Kyrylov, V., Le, E.Y.,

Leclerc, G., Lennon, J.P., Lessans, S., Lezcano-Casado, M., Li, Y., Li, Z., Lin, J., Liss, J., Lily, Liu, Liu, J., Lu, K., Lu, C., Martinovic, Z., McCallum, L., McGrath, J., McKinney, S., McLaughlin, A., Mei, S., Mostovoy, S., Mu, T., Myles, G., Neitz, A., Nichol, A., Pachocki, J., Paino, A., Palmie, D., Pantuliano, A., Parascandolo, G., Park, J., Pathak, L., Paz, C., Peran, L., Pimenov, D., Pokrass, M., Proehl, E., Qiu, H., Raila, G., Raso, F., Ren, H., Richardson, K., Robinson, D., Rotsted, B., Salman, H., Sanjeev, S., Schwarzer, M., Sculley, D., Sikchi, H., Simon, K., Singhal, K., Song, Y., Stuckey, D., Sun, Z., Tillet, P., Toizer, S., Tsimpourlas, F., Vyas, N., Wallace, E., Wang, X., Wang, M., Watkins, O., Weil, K., Wendling, A., Whinnery, K., Whitney, C., Wong, H., Yang, L., Yang, Y., Yasunaga, M., Ying, K., Zaremba, W., Zhan, W., Zhang, C., Zhang, B., Zhang, E., Zhao, S.: gpt-oss-120b & gpt-oss-20b model card (2025), https://arxiv.org/abs/2508.10925

15. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al.: Training language models to follow instructions with human feedback. Advances in neural information processing systems **35**, 27730–27744 (2022)
16. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023)
17. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
18. Rafailov, R., Sharma, A., Mitchell, E., Manning, C.D., Ermon, S., Finn, C.: Direct preference optimization: Your language model is secretly a reward model. Advances in neural information processing systems **36**, 53728–53741 (2023)
19. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10684–10695 (June 2022)
20. Schuhmann, C.: Laion-aesthetics. https://laion.ai/blog/laion-aesthetics/ (2022), accessed: 2023 - 11- 10
21. Shazeer, N., Stern, M.: Adafactor: Adaptive learning rates with sublinear memory cost. In: International conference on machine learning. pp. 4596–4604. PMLR (2018)
22. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. arXiv preprint arXiv:2011.13456 (2020)
23. Wallace, B., Dang, M., Rafailov, R., Zhou, L., Lou, A., Purushwalkam, S., Ermon, S., Xiong, C., Joty, S., Naik, N.: Diffusion model alignment using direct preference optimization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8228–8238 (2024)
24. Wang, K., Kidambi, R., Sullivan, R., Agarwal, A., Dann, C., Michi, A., Gelmi, M., Li, Y., Gupta, R., Dubey, K.A., et al.: Conditional language policy: A general framework for steerable multi-objective finetuning. In: Findings of the Association for Computational Linguistics: EMNLP 2024. pp. 2153–2186 (2024)
25. Wu, X., Hao, Y., Sun, K., Chen, Y., Zhu, F., Zhao, R., Li, H.: Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. CoRR (2023)

26. Xu, J., Liu, X., Wu, Y., Tong, Y., Li, Q., Ding, M., Tang, J., Dong, Y.: Imagereward: Learning and evaluating human preferences for text-to-image generation. Advances in Neural Information Processing Systems **36**, 15903–15935 (2023)
27. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., Zheng, C., Liu, D., Zhou, F., Huang, F., Hu, F., Ge, H., Wei, H., Lin, H., Tang, J., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Zhou, J., Lin, J., Dang, K., Bao, K., Yang, K., Yu, L., Deng, L., Li, M., Xue, M., Li, M., Zhang, P., Wang, P., Zhu, Q., Men, R., Gao, R., Liu, S., Luo, S., Li, T., Tang, T., Yin, W., Ren, X., Wang, X., Zhang, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Zhang, Y., Wan, Y., Liu, Y., Wang, Z., Cui, Z., Zhang, Z., Zhou, Z., Qiu, Z.: Qwen3 technical report (2025), https://arxiv.org/abs/2505.09388
28. Yang, R., Pan, X., Luo, F., Qiu, S., Zhong, H., Yu, D., Chen, J.: Rewards-in-context: Multi-objective alignment of foundation models with dynamic preference adjustment. arXiv preprint arXiv:2402.10207 (2024)
29. Ye, H., Zhang, J., Liu, S., Han, X., Yang, W.: Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. arXiv preprint arXiv:2308.06721 (2023)
30. Yu, J., Xu, Y., Koh, J.Y., Luong, T., Baid, G., Wang, Z., Vasudevan, V., Ku, A., Yang, Y., Ayan, B.K., et al.: Scaling autoregressive models for content-rich text-to-image generation. Trans. Mach. Learn. Res. (2022)
31. Zhang, S., Wang, B., Wu, J., Li, Y., Gao, T., Zhang, D., Wang, Z.: Learning multi-dimensional human preference for text-to-image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8018–8027 (2024)
32. Zhu, H., Xiao, T., Honavar, V.G.: DSPO: Direct score preference optimization for diffusion model alignment. In: The Thirteenth International Conference on Learning Representations (2025), https://openreview.net/forum?id=xyfb9HHvMe