STRONG REWARD ONLY: PARETO-GUIDED MULTI-REWARD OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Post-training preference optimization is a central approach for aligning text-to-image models with human preferences. Recent work attempts to mitigate reward hacking by jointly optimizing multiple reward models, under the assumption that richer objectives provide more constrained guidance. However, we find that multi-reward optimization does not inherently prevent reward hacking, and its effectiveness critically depends on the reliability of the underlying reward models. Moreover, the inherent trade-offs among multiple rewards call for principled multi-objective optimization algorithms. To address this challenge, we propose a Pareto-frontier-guided optimal transport framework for robust multi-reward optimization. Our method dynamically constructs Pareto frontiers during training and maps dominated samples toward the frontier via distribution-aware optimal transport, and can be applied to arbitrary sets of reward models. To provide a more rigorous assessment, we introduce the Joint Domination Rate (JDR) and Joint Collapse Rate (JCR) as principled metrics to quantify multi-reward synergy and reward hacking. Experimental results show our approach outperforms baselines with an 11% gain in JDR and achieves a near 80% win rate in human evaluations.

1 INTRODUCTION

With the rapid advancement of text-to-image generation Sohl-Dickstein et al. (2015); Song & Ermon (2020); Rombach et al. (2022); Podell et al. (2023); Esser et al. (2024), numerous post-training optimization methods have emerged Ziegler et al. (2020); Stiennon et al. (2022); Black et al. (2024); Dhariwal & Nichol (2021); Li et al. (2024b); Rafailov et al. (2024); Fan et al. (2023), among which aligning model outputs with human preferences via reward models has become a key research direction.

Early approaches typically rely on a single human-preference reward model Xu et al. (2023); Zhang et al. (2024). With the rapid progress of post-training preference optimization, a growing body of work proposes jointly optimizing multiple reward models via weighted aggregation Eyring et al. (2024); Agarwal & Aggarwal (2023); Wei et al. (2024); Deng et al. (2024); de Langis et al. (2024); Lee et al. (2025), motivated by the intuition that richer objectives may provide more constrained guidance. These approaches assume that more comprehensive objectives and richer supervisory signals can effectively alleviate reward hacking.

However, we observe that naively incorporating additional reward models does not resolve this issue. We observe that different reward models exhibit markedly different behaviors under isolated optimization, with some collapsing substantially more severely than others. More critically, our analysis suggests that failures in individual reward dimensions can disproportionately influence joint optimization behavior, raising concerns about the stability of naive reward aggregation. This phenomenon is surprising and has been largely overlooked, as prior work implicitly assumes that reward aggregation is benign or stabilizing. These observations raise a key question: does multi-reward joint optimization inherently mitigate reward hacking, or does its effectiveness depend on the reliability of the underlying reward models?

Guided by these observations, we propose a Pareto-frontier-guided optimal transport framework for robust multi-reward optimization. Our framework is explicitly designed to operate on subsets of rewards that remain stable during training. Our method performs robust multi-objective optimization

by dynamically constructing Pareto frontiers during training and mapping dominated samples onto the frontier via distribution-aware optimal transport.

Finally, existing multi-reward optimization lacks reliable quantitative criteria for defining joint success and failure, and often relies on qualitative inspection. We therefore introduce Joint Domination Rate (JDR) and Joint Collapse Rate (JCR) as principled criteria for detecting joint improvement and system-wide degradation across multiple rewards.

Our main contributions are as follows:

1. We ovserve that multi-reward optimization does not inherently prevent reward hacking; instead, careful selection of reliable reward models is crucial.
2. We propose a Pareto-frontier-guided optimal transport framework for robust multi-objective optimization under heterogeneous reward behaviors.
3. We introduce JDR and JCR as quantitative metrics for detecting synergistic improvement and collapse in multi-reward optimization.

2 Understanding Reward Hacking in Multi-Reward Optimization

In this section, we formalize the problem setting of multi-reward optimization for text-to-image models, clarify the notion of reward hacking, and introduce principled criteria for detecting joint improvement and collapse under multiple reward signals.

2.1 Problem Setup

We consider a text-to-image (T2I) generation model conditioned on a set of prompts $\mathcal{P} = \{p_1, \ldots, p_n\}$. Each prompt p_i induces a corresponding image domain $\mathcal{D}_i \subset \mathcal{X}$, where $\mathcal{X}$ denotes the overall image space.

Given a generated image $x \in \mathcal{D}_i$, we evaluate it using a set of K reward models:

$$\tilde{\mathbf{R}}(x) = \big(R_1(x), R_2(x), \ldots, R_K(x)\big).$$

2.2 Reward Hacking as a Phenomenon

Reward-based post-training aims to improve the quality of generated images by optimizing reward signals. However, it is well known that optimizing reward models can lead to *reward hacking*, a phenomenon where reward scores increase while the actual human-perceived quality deteriorates.

Reward Hacking. We refer to reward hacking as the situation in which optimization drives improvements in reward values without corresponding improvements in human-perceived image quality, or even causes systematic degradation.

This phenomenon has been widely observed in both single-reward and multi-reward optimization settings. In practice, reward hacking often manifests as visually implausible artifacts, overfitting to superficial features favored by the reward model, or mode collapse that exploits blind spots of learned rewards.

2.3 Challenges in Multi-Reward Settings

Detecting reward hacking becomes substantially more challenging when multiple reward models are optimized jointly. First, improvements in individual rewards do not necessarily imply overall progress, as different rewards may conflict or trade off against each other. Second, averaging or aggregating reward values can obscure failure modes: severe degradation in one reward dimension may be masked by gains in others. Finally, qualitative inspection is inherently subjective and does not scale to large-scale evaluations. As a result, multi-reward optimization requires criteria that can explicitly capture *joint* improvement and *joint* collapse across all reward dimensions.

2.4 Joint Improvement and Joint Collapse Criteria

To rigorously detect reward hacking under multiple reward signals, we introduce two complementary criteria that evaluate optimization behavior at the sample level.

Joint Domination Rate (JDR). Given a baseline model and a candidate optimized model, we define the Joint Domination Rate over K rewards as

$$\mathrm{JDR}_K = \frac{1}{N}\sum_{i=1}^{N} 1\Big(R_j(x_i) > R_j(x_i^{(b)}),\ \forall j \in \{1,\dots,K\}\Big),$$

where $x_i^{(b)}$ denotes the baseline image corresponding to sample i, and $1(\cdot)$ is the indicator function. JDR measures the proportion of samples that achieve *simultaneous improvement across all reward dimensions*, capturing genuine multi-objective progress rather than marginal or compensatory gains.

Joint Collapse Rate (JCR). Conversely, we define the Joint Collapse Rate as

$$\mathrm{JCR}_K = \frac{1}{N}\sum_{i=1}^{N} 1\Big(R_j(x_i) < R_j(x_i^{(b)}),\ \forall j \in \{1,\dots,K\}\Big).$$

JCR measures the proportion of samples that degrade simultaneously across all rewards, serving as a conservative and reliable indicator of system-wide failure.

3 Method

3.1 Multi-reward Optimization Formulation

Multi-reward optimization aims to achieve synergistic improvements across multiple reward models. Since the objectives of different rewards may inherently conflict, the problem naturally falls within the framework of Pareto optimization. We first provide the fundamental concepts related to Pareto optimality under multiple reward signals.

Pareto Optimality Fundamentals. Given K rewards $\tilde{R} = (R^1(x), R^2(x), \dots, R^K(x))$ to be maximized, the concepts of Pareto optimality can be defined as follows:

- **Pareto Dominance.** The reward vector of x^a *dominates* that of x^b (denoted $\tilde{R}(x^a) \succ \tilde{R}(x^b)$) if

$$\forall u \in \{1,\dots,K\} : R^u(x^a) \geq R^u(x^b), \quad \text{and} \quad \exists v \in \{1,\dots,K\} : R^v(x^a) > R^v(x^b).$$

- **Pareto Optimality.** For a given prompt p_i, a sample $x^* \in \mathcal{F}_i$ is *Pareto optimal* if its reward vector $\tilde{R}(x^*)$ is not dominated by that of any other $x \in \mathcal{F}_i$:

$$x^* \text{ is Pareto optimal } \iff x^* \in \mathcal{F}_i \text{ and } x \in \mathcal{F}_i : \tilde{R}(x) \succ \tilde{R}(x^*).$$

- **Pareto Front.** For a given prompt p_i, the *Pareto front* $\mathcal{J}_i$ is the set of all samples in the perceptually admissible feasible set $\mathcal{F}_i$ whose reward vectors are Pareto optimal :

$$\mathcal{J}_i = \{x \mid x \in \mathcal{F}_i,\ \ x' \in \mathcal{F}_i : \tilde{R}(x') \succ \tilde{R}(x)\}.$$

To tackle the issues beyond reward hacking identified in Section **??**, we propose a Pareto-guided optimal transport framework for multi-reward optimization. First, an offline strategy leverages precomputed Pareto fronts for each prompt domain to mitigate reward hacking arising from heterogeneous bounds (Section **??**). Second, addressing the assumption that weak rewards are prone to collapse (Assumption **??**), we introduce a GPT-4o-based decision agent to detect and eliminate unstable weak rewards. Third, an online strategy enables strong rewards (Definition **??**) to autonomously explore and expand the Pareto front during optimization. Finally, we propose JDR and JDS as metrics to evaluate performance gains and stability in multi-reward optimization.

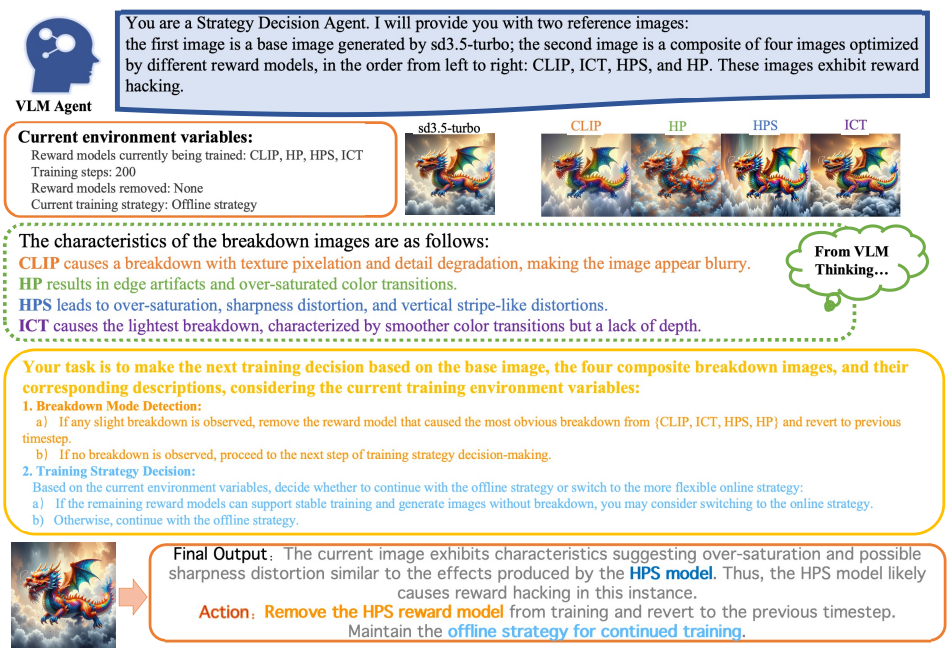

Figure 1: Adaptive Decision Pipeline of the VLM based Agent for Multi-Reward Optimization.

3.2 VLM-BASED DECISION-MAKING AGENT

To mitigate this, we employ VLM as a decision-making agent for detecting and removing collapsed weak reward models. Once significant signals of reward hacking occur, the optimization trajectory becomes difficult to recover, making early detection critical. However, in the early stages of reward hacking, the generated images are only mildly collapsed, with only slight differences from normally generated images. To capture these subtle differences, we not only collected images of early mild collapses for each reward model but also utilized GPT-4o to analyze their respective collapse characteristics, constructing a comprehensive prior reference for the agent. Upon detecting mild collapse, the agent immediately removes the problematic weak reward model and reverts to the earlier stable checkpoint to safely resume training.

3.3 PARETO-FRONTIER-GUIDED OPTIMAL TRANSPORT

After the decision-making agent eliminates all weak reward models (Definition **??**) prone to reward hacking, the remaining strong reward models (Definition **??**) exhibit a strong correlation between their preference prediction capabilities and human perceptual quality. To further explore superior reward bounds, we propose an online strategy that enables strong reward models to autonomously collect and optimize along the Pareto frontier during training. Specifically, given a prompt p_i, we perform the T2I model in parallel across multiple processes to increase the number of generated images. The reward vectors from all processes are aggregated into a global set $\mathcal{R}_{i,N} = \{\tilde{R}(x_i^j) \mid j = 1, \ldots, N\}$, where n is the number of candidate images per process, P is the number of processes, and $N = P \times n$.

For each sample x_i^j, we identify all vectors in $\mathcal{R}_{i,N}$ that Pareto-dominate $\tilde{R}(x_i^j)$:

$$\mathcal{R}^{dom}(x_i^j) = \{\tilde{R}(x_i^m) \in \mathcal{R}_{i,N} \mid \tilde{R}(x_i^m) \succ \tilde{R}(x_i^j)\}, \tag{1}$$

where $k_i = |\mathcal{R}^{dom}(x_i^j)|$ is the number of dominating reward vectors. The reward vector of x_i^j forms a discrete source distribution with one sampling point $\mu = \tilde{R}(x_i^j)$, while the dominating set serves as target distribution $\nu = \mathcal{R}^{dom}(x_i^j)$. For each batch, the optimal transport discrepancies between the source and target distributions for all samples are computed and summed to obtain the batch loss.

Figure 2: Qualitative Comparison of Optimization Results Across Different Methods.

As training progresses and the reward candidate set improves, the online strategy adaptively guides strong reward models to explore superior Pareto frontiers, thereby achieving better optimization.

Optimal Transport Framework. Optimal Transport (OT) Monge (1781) provides a principled way to measure discrepancies between probability distributions while preserving the geometry of the underlying space. Given a source distribution μ and a target distribution ν, OT seeks a transport plan $\gamma \in \Pi(\mu, \nu)$ that minimizes the total cost:

$$\min_{\gamma \in \Pi(\mu,\nu)} \int c(x, y)\, d\gamma(x, y), \tag{2}$$

where $c(x, y)$ denotes the ground cost between source sample x and target sample y. Here, the n rewards dominated by the frontier collectively constitute the source distribution: $\mu_i = \{\tilde{R}(x_i^j) \mid x_i^j \in \{x_i^1, \ldots, x_i^n\},\ \forall \tilde{R}(x_i^m) \in \mathcal{R}^{(q)}(p_i) : \tilde{R}(x_i^m) \succ \tilde{R}(x_i^j)\}$, while the precomputed Pareto frontier serves as the target distribution $\nu_i = \mathcal{R}^{(q)}(p_i)$.

OT establishes a minimal-cost mapping from μ_i to ν_i, transporting dominated samples toward dominating points according to geometric distances in the reward space for prompt p_i. Practically, the optimal plan γ_i^* is solved by optimizing the following discrete formulation of OT with an entropy regularization term using the Sinkhorn algorithm Cuturi (2013):

$$\gamma_i^* = \arg \min_{\gamma \in \Pi(\mu_i, \nu_i)} \sum_{m,j} c(y_i^m, x_i^j)\, \gamma(y_i^m, x_i^j), \tag{3}$$

where the ground cost is defined as $c(y_i^m, x_i^j) = \|\tilde{R}(y_i^m) - \tilde{R}(x_i^j)\|_2^2$, representing the squared Euclidean distance between reward vectors of source sample y_i^m and target sample x_i^j whose reward vector is in the Pareto frontier, and the γ is a $n \times q_i$ transport matrix.

4 Experiments

4.1 Experiment Setting

T2I Model. Our text-to-image (T2I) framework builds upon Stable Diffusion 3.5-Turbo, one of the most advanced text-to-image models. To maintain the stability of the base model, we update only LoRA Hu et al. (2021) parameters rather than the original weights of Stable Diffusion 3.5-Turbo. Specifically, we employ the gradient backpropagation strategy of DRaFT-K Clark et al. (2024) for fine-tuning, which applies gradient updates only during the final $k = 2$ denoising steps to reduce memory consumption and accelerate training. Training details are in Appendix E.

Reward Models and Training Strategy. We employ four reward models, encompassing both strong and weak categories, to initialize joint training. These are grouped into two primary types:

Table 1: Quantitative Results (%) on the Parti-Prompts dataset Yu et al. (2022): Individual Reward Win Rates and Joint Performance Metrics Relative to SD3.5-Turbo.

Model	ICT (%)↑	HP (%)↑	CLIP (%)↑	HPS (%)↑	JDR_2 (%)↑	JDR_4 (%)↑	JCR_4 (%)↓
Single-Reward Optimization							
+ ICT	**56.99**	36.83	47.06	48.71	20.59	7.66	10.17
+ HP	52.45	**90.26**	44.30	57.29	<u>36.15</u>	<u>13.73</u>	4.11
+ CLIP	48.96	47.06	**52.63**	46.81	23.77	8.09	9.07
+ HPS	50.12	41.67	37.07	**88.30**	20.77	8.03	3.06
Multi-Reward Joint Optimization (Weighted **ICT:HP:CLIP:HPS** ratios)							
1:1:1:1	51.10	52.08	<u>47.43</u>	82.97	26.59	12.68	3.19
2:3:2:3	50.80	56.43	46.51	<u>86.03</u>	28.31	13.42	<u>2.57</u>
2:2:3:3	50.43	56.56	46.57	84.25	26.23	12.62	2.76
4:4:1:1	51.96	57.23	44.12	79.90	29.53	12.44	3.74
Reward Soup (Weighted **ICT:HP:CLIP:HPS** fusion of single-reward LoRAs)							
1:1:1:1	50.55	54.17	42.16	81.92	27.02	11.15	3.74
1:1:4:4	50.43	52.94	42.46	85.11	25.37	11.15	3.31
3:2:1:4	50.80	53.74	43.32	85.29	26.29	10.85	3.19
3:2:0:0	50.74	53.86	42.59	83.21	26.10	10.85	3.31
Multi-Reward under Heterogeneous Bounds							
Weighted-Sum	52.63	56.86	46.94	82.48	29.84	13.66	3.49
Separate-Cons	49.45	57.23	46.63	61.21	28.25	10.78	6.68
Pareto-Frontier-Guided Optimal Transport							
Ours	<u>56.43</u>	<u>85.23</u>	43.63	61.70	**47.98**	**17.10**	**2.39**

text–image alignment rewards (**CLIP** Radford et al. (2021) and **ICT** Ba et al. (2025)) and human preference rewards (**HPS** Wu et al. (2023) and **HP** Ba et al. (2025)). Our staged optimization strategy proceeds in three phases. First, all four reward models are jointly optimized during the offline policy stage. Next, weak rewards that induce instability or collapse are adaptively pruned according to agent feedback. Finally, the remaining strong reward models (**ICT** and **HP**) are retained to guide the online policy stage for deeper optimization.

Evaluation Metrics. To comprehensively assess the effectiveness of multi-reward optimization, we adopt Joint Domination Rate (**JDR**) and Joint Collapse Rate (**JCR**) as the primary evaluation criteria. Specifically, we report JDR_2, computed on the two strong rewards ICT and HP that are ultimately retained by the agent's decision-making process, as well as JDR_4 and JCR_4, both computed over all four rewards ICT, HP, CLIP, and HPS to assess the overall optimization capability. In addition, we provide the average scores from seven widely used reward models as supplementary evaluation to validate the robustness of our approach: Aesthetic Score[1] for aesthetic quality, CLIP Radford et al. (2021) for text–image consistency, ICT Ba et al. (2025) for the degree of text presence in images, and human preference models such as PickScore Kirstain et al. (2023), HPS Wu et al. (2023), ImageReward Xu et al. (2023), and HP Ba et al. (2025).

4.2 Baseline Construction

To evaluate the proposed multi-reward optimization framework, we construct four baselines: single-reward fine-tuning, weighted multi-reward fine-tuning, reward soup, and multi-reward fine-tuning under heterogeneous reward bounds.

Single-Reward Fine-Tuning. Using Stable Diffusion 3.5-Turbo as backbone, we fine-tune the model with CLIP, ICT, HPS, and HP as individual objectives via DRaFT-K Clark et al. (2024). For each reward, the best pre-collapse checkpoint validated by human experts serves as the baseline.

Weighted Multi-Reward Fine-Tuning. Using the weighted loss, we jointly optimize the four rewards under identical settings, exploring diverse weight combinations. The best pre-collapse checkpoint validated by experts is selected as the baseline.

Reward Soup. We adopt an inference-time fusion strategy that combines LoRA weights from single-reward fine-tuned models through weighted fusion, exploring a broader reward fusion space without additional training costs.

[1] https://github.com/christophschuhmann/improved-aesthetic-predictor

Table 2: Quantitative Results on the Parti-Prompts dataset: Comparison of multiple reward scores.

Model	CLIP↑	ICT↑	Aesthetic↑	HPS↑	PickScore↑	ImgReward↑	HP↑
SD3.5-Turbo	0.3372	0.8965	6.2766	0.2856	22.7435	1.1499	0.7754
Single-Reward Optimization							
ICT	0.3548	0.8986	6.2781	0.2916	22.7261	1.1487	0.7739
HP	0.3538	0.8976	6.2881	0.2809	<u>22.7687</u>	<u>1.1691</u>	<u>0.7776</u>
CLIP	**0.3554**	0.8958	6.2770	0.2915	22.7396	1.1530	0.7751
HPS	0.3501	0.8987	<u>6.3437</u>	**0.2994**	22.6896	1.1793	0.7747
Multi-Reward Joint Optimization (Weighted **ICT:HP:CLIP:HPS** ratios)							
1:1:1:1	0.3545	0.8970	6.2839	0.2958	22.6895	1.1464	0.7763
2:3:2:3	0.3544	0.8987	6.2850	<u>0.2971</u>	22.6835	1.1636	0.7763
2:2:3:3	0.3552	0.8974	6.2787	0.2962	22.6926	1.1570	0.7770
4:4:1:1	0.3541	<u>0.8989</u>	6.3043	0.2950	22.7049	1.1568	0.7762
Reward Soup (Weighted **ICT:HP:CLIP:HPS** fusion of single-reward LoRAs)							
1:1:1:1	0.3543	0.8958	6.2931	0.2936	22.7542	1.1562	0.7752
1:1:4:4	0.3539	0.8967	6.3037	0.2951	22.7559	1.1679	0.7752
3:2:1:4	0.3537	0.8965	6.3032	0.2951	22.7549	1.1654	0.7754
3:2:0:0	0.3541	0.8961	6.2759	0.2941	22.7436	1.1493	0.7752
Multi-Reward under Heterogeneous Bounds							
Weighted-Sum	0.3546	0.8968	6.2800	0.2947	22.7274	1.1514	0.7760
Separate-Cons	<u>0.3549</u>	0.8967	6.2774	0.2922	22.7343	1.1561	0.7757
Pareto-Frontier-Guided Optimal Transport							
Ours	0.3534	**0.9004**	**6.3588**	0.2929	**22.8160**	**1.1808**	**0.7783**

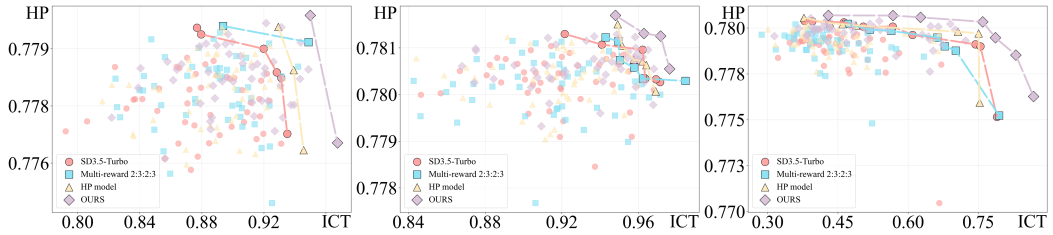

Figure 3: Pareto Frontier Visualization based on strong rewards (ICT and HP) on Three Prompts.

Multi-Reward Fine-Tuning with Heterogeneous Reward Bounds. We utilize the precomputed Pareto frontier as an approximation of prompt-wise heterogeneous reward bounds and construct two baseline variants upon it. One variant aggregates approximate bounds of multiple reward functions into a weighted average, serving as a unified optimization target across prompts, denoted as the *weighted-sum bounds* method. The other assigns each reward bound as its optimization target and minimizes the squared error between each reward and its bound, formulated as $\mathcal{L} = \sum_k (r_k^{\text{bound}} - r_k)^2$, denoted as the *separate-constraints* method.

4.3 Evaluation and Analysis

Joint Performance Metrics. In Table 1, we present the individual win rates of the four reward models (ICT, HP, CLIP, and HPS) together with the joint optimization performance on the Parti-Prompts dataset Yu et al. (2022). Our method consistently demonstrates significant improvements over the baselines, achieving an 11% gain in JDR_2, a 3.4% gain in JDR_4, and a 0.2% reduction in JCR_4, while maintaining comparable win rates on each individual reward. Single-reward baselines achieve the highest win rates on their respective rewards, but their joint performance metrics generally degrade. Multi-reward optimization baselines that rely on the global bound perform worse than single-reward optimization with HP, consistent with Assumption **??** in Section **??**, which states that weak rewards in multi-reward optimization induce collapse and undermine joint performance.

Reward Metrics. Table 2 presents comprehensive evaluation results across seven reward models. Our method consistently secures the best performance on most metrics, demonstrating robustness under heterogeneous bounds and strong-reward training. In contrast, multi-reward joint optimization methods are affected by weak rewards, leading to degraded overall performance and instability.

Table 3: Win Rate (%) of Ours vs. Baselines by Human Experts on DiffusionDB Wang et al. (2023)

Model	SD3.5-Turbo	HP	HPS	Soup-1:1:1:1	Multi-2323	Separate-Constraints	Weighted-Sum Bounds
Ours vs.	76.34	61.29	74.19	74.19	**79.57**	68.82	73.12

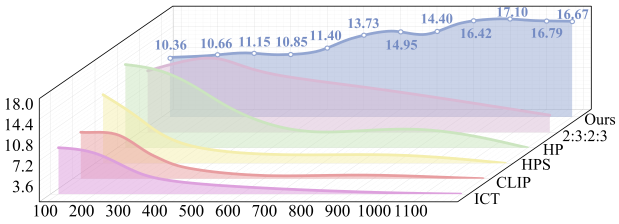

Figure 4: Comparative Training Curves of Joint Domination Rate (JDR_4) for Ours versus Baseline Methods.

Model	$JDR_2\uparrow$	$JDR_4\uparrow$	$JCR_4\downarrow$
SD3.5-Turbo	–	–	–
Online Only	21.51	9.38	4.04
Offline Only	34.07	7.35	7.23
Ours (w/o Online)	38.54	14.89	4.29
Ours	**47.98**	**17.10**	**2.39**

tableAblation Study on Joint Domination Rate (JDR) and Joint Collapse Rate (JCR).

Quantitative results. As shown in Fig. 2, we present the qualitative comparisons. The baselines include two single-reward optimizations aligned with human preference, HPS and HP; a multi-reward joint training approach with the global bound as the optimization target using the weighted ratio ICT:HP:CLIP:HPS $=$ 2:3:2:3; a reward-soup fusion method applied at inference with equal weighting ICT:HP:CLIP:HPS $=$ 1:1:1:1; and two heterogeneous-bound methods, Weighted-Sum bounds and Separate-Constraints. The results demonstrate that our method achieves superior visual quality while avoiding reward hacking. Additional qualitative examples are provided in Appendix D.

Qualitative Case Studies on Pareto Frontier Visualization. We conduct a Pareto frontier analysis based on two strong rewards, ICT and HP, comparing SD 3.5-Turbo, the HP-optimized single-reward model, the multi-reward joint optimization model with ratio ICT:HP:CLIP:HPS $=$ 2:3:2:3, and our proposed method. For each prompt and identical random seed, every method generates 50 images, whose reward distributions and corresponding Pareto frontiers are plotted in a two-dimensional diagram. As shown in Fig. 3, the image domains induced by different prompts exhibit heterogeneous reward bounds. Our method consistently produces Pareto frontiers that dominate those of the baselines, with generated samples distributed closer to the frontier, reflecting superior trade-off alignment and validating the effectiveness of our approach. More visualizations in Appendix D.

Visualization and Analysis of Training Curve. In Figure 4, we present the training curves of the Joint Domination Rate (JDR_4), which serves as a robust metric of joint optimization, for our method and the baselines. Single-reward baselines optimized with the global bound collapse rapidly, while multi-reward optimization with the global bound and the involvement of weak rewards also exhibits severe deterioration. In contrast, our method achieves stable and consistent improvements throughout training while effectively avoiding reward hacking.

User study. We conduct a user study with ten annotators on 300 randomly selected prompts from the DiffusionDB dataset Wang et al. (2023). For each prompt, image pairs (ours vs. baseline) were presented in random order, and annotators evaluated prompt fidelity and visual appeal. As shown in Table 3, our method achieves higher win rates against all baselines, confirming its effectiveness.

Ablation Study We conduct ablation experiments on the online policy, offline policy, and GPT-4o Agent, as shown in Table 4.3. It can be observed that using the online or offline policy alone yields moderate gains. Excluding the online policy while combining the offline policy with the GPT-4o Agent achieves the second-best results, where strong reward models mitigate collapse and approach the precomputed bound. Finally, our approach, incorporating the offline policy, adaptive regulation by the GPT-4o Agent, and exploration through the online policy, achieves optimal performance.

5 Conclusion

In this work, we demonstrate that reward hacking arises from unified global targets under heterogeneous reward bounds, and from the inherent vulnerability of weak reward models. To address this, we propose a Pareto-frontier-guided optimal transport framework with online and offline strategies, and introduce JDR and JCR as principled evaluation metrics. Our approach achieves consistent gains over strong baselines while effectively mitigating reward hacking.

REFERENCES

Mridul Agarwal and Vaneet Aggarwal. Reinforcement learning for joint optimization of multiple rewards, 2023. URL https://arxiv.org/abs/1909.02940.

Ying Ba, Tianyu Zhang, Yalong Bai, Wenyi Mo, Tao Liang, Bing Su, and Ji-Rong Wen. Enhancing reward models for high-quality image generation: Beyond text-image alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19022–19031, 2025.

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning, 2024. URL https://arxiv.org/abs/2305.13301.

Kevin Clark, Paul Vicol, Kevin Swersky, and David J Fleet. Directly fine-tuning diffusion models on differentiable rewards, 2024. URL https://arxiv.org/abs/2309.17400.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In Christopher J. C. Burges, Léon Bottou, Zoubin Ghahramani, and Kilian Q. Weinberger (eds.), *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pp. 2292–2300, 2013. URL https://proceedings.neurips.cc/paper/2013/hash/af21d0c97db2e27e13572cbf59eb343d-Abstract.html.

Karin de Langis, Ryan Koo, and Dongyeop Kang. Dynamic multi-reward weighting for multi-style controllable generation, 2024. URL https://arxiv.org/abs/2402.14146.

Fei Deng, Qifei Wang, Wei Wei, Matthias Grundmann, and Tingbo Hou. Prdp: Proximal reward difference prediction for large-scale reward finetuning of diffusion models, 2024. URL https://arxiv.org/abs/2402.08714.

Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis, 2021. URL https://arxiv.org/abs/2105.05233.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis, 2024. URL https://arxiv.org/abs/2403.03206.

Luca Eyring, Shyamgopal Karthik, Karsten Roth, Alexey Dosovitskiy, and Zeynep Akata. Reno: Enhancing one-step text-to-image models through reward-based noise optimization. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/e31bdea0a93741c2157eea705dd219eb-Abstract-Conference.html.

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, 2023. URL https://arxiv.org/abs/2305.16381.

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, 2021. URL https://arxiv.org/abs/2106.09685.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *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*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/73aacd8b3b05b4b503d58310b523553c-Abstract-Conference.html.

Kyungmin Lee, Xiahong Li, Qifei Wang, Junfeng He, Junjie Ke, Ming-Hsuan Yang, Irfan Essa, Jinwoo Shin, Feng Yang, and Yinxiao Li. Calibrated multi-preference optimization for aligning diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 18465–18475. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.01721. URL https://openaccess.thecvf.com/content/CVPR2025/html/Lee_Calibrated_Multi-Preference_Optimization_for_Aligning_Diffusion_Models_CVPR_2025_paper.html.

Seung Hyun Lee, Yinxiao Li, Junjie Ke, Innfarn Yoo, Han Zhang, Jiahui Yu, Qifei Wang, Fei Deng, Glenn Entis, Junfeng He, Gang Li, Sangpil Kim, Irfan Essa, and Feng Yang. Parrot: Pareto-optimal multi-reward reinforcement learning framework for text-to-image generation. In Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XXXVIII*, volume 15096 of *Lecture Notes in Computer Science*, pp. 462–478. Springer, 2024. doi: 10.1007/978-3-031-72920-1_26. URL https://doi.org/10.1007/978-3-031-72920-1_26.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato (eds.), *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pp. 12888–12900. PMLR, 2022. URL https://proceedings.mlr.press/v162/li22n.html.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 7985–7995. IEEE, 2024a. doi: 10.1109/CVPR52733.2024.00763. URL https://doi.org/10.1109/CVPR52733.2024.00763.

Zihao Li, Zhuoran Yang, and Mengdi Wang. Reinforcement learning with human feedback: Learning dynamic choices via pessimism, 2024b. URL https://openreview.net/forum?id=TrwocPauNQ.

Gaspard Monge. Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences de Paris*, 1781.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis, 2023. URL https://arxiv.org/abs/2307.01952.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. URL http://proceedings.mlr.press/v139/radford21a.html.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model, 2024. URL https://arxiv.org/abs/2305.18290.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022. URL https://arxiv.org/abs/2112.10752.

Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics, 2015. URL https://arxiv.org/abs/1503.03585.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution, 2020. URL https://arxiv.org/abs/1907.05600.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. Learning to summarize from human feedback, 2022. URL https://arxiv.org/abs/2009.01325.

Dipesh Tamboli, Souradip Chakraborty, Aditya Malusare, Biplab Banerjee, Amrit Singh Bedi, and Vaneet Aggarwal. Balanceddpo: Adaptive multi-metric alignment. *CoRR*, abs/2503.12575, 2025. doi: 10.48550/ARXIV.2503.12575. URL https://doi.org/10.48550/arXiv.2503.12575.

Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 8228–8238. IEEE, 2024. doi: 10.1109/CVPR52733.2024.00786. URL https://doi.org/10.1109/CVPR52733.2024.00786.

Zijie J. Wang, Evan Montoya, David Munechika, Haoyang Yang, Benjamin Hoover, and Duen Horng Chau. Diffusiondb: A large-scale prompt gallery dataset for text-to-image generative models, 2023. URL https://arxiv.org/abs/2210.14896.

Fanyue Wei, Wei Zeng, Zhenyang Li, Dawei Yin, Lixin Duan, and Wen Li. Powerful and flexible: Personalized text-to-image generation via reinforcement learning, 2024. URL https://arxiv.org/abs/2407.06642.

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. *CoRR*, abs/2306.09341, 2023. doi: 10.48550/ARXIV.2306.09341. URL https://doi.org/10.48550/arXiv.2306.09341.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *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*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/33646ef0ed554145eab65f6250fab0c9-Abstract-Conference.html.

Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. Scaling autoregressive models for content-rich text-to-image generation, 2022. URL https://arxiv.org/abs/2206.10789.

Yinan Zhang, Eric Tzeng, Yilun Du, and Dmitry Kislyuk. Large-scale reinforcement learning for diffusion models, 2024. URL https://arxiv.org/abs/2401.12244.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences, 2020. URL https://arxiv.org/abs/1909.08593.

A Related Work

Single-Reward Optimization for Text-to-Image Diffusion Models

Reward-based optimization has emerged as a crucial paradigm for improving text-to-image diffusion models, where reward models provide supervision signals to enhance generation quality. Current research focuses on two objectives: **(1) Text–Image Alignment.** CLIP Radford et al. (2021) pioneered cross-modal embeddings that capture vision–language semantics, and BLIP Li et al. (2022) extended this paradigm with bidirectional mechanisms for stronger alignment evaluation. However, ICT Ba et al. (2025) demonstrated that both models systematically undervalue high-quality images, motivating the development of refined metrics for more faithful text representation. **(2) Human Preference Alignment.** Recent efforts have shifted toward modeling human perceptual preferences as rewards to better align generation with subjective judgments. Reward models such as ImageReward Xu et al. (2023), HPS Wu et al. (2023), PickScore Kirstain et al. (2023), and HP Ba et al. (2025) are trained on large-scale preference datasets to approximate human judgments. However, these models may yield conflicting signals. A central challenge, therefore, lies in how to effectively integrate and jointly optimize multiple rewards to reconcile such discrepancies.

Multi-Reward Optimization and Pareto-Front Methods for Text-to-Image Diffusion

To simultaneously improve text–image consistency and human preference alignment, previous studies have introduced multiple reward signals into text-to-image diffusion models. Basic approaches typically rely on linear weighting; for example, ReNO Eyring et al. (2024) applies weighted fusion of multiple rewards solely to optimize the noise initialization stage, while TextCraftor Li et al. (2024a) confines the weighted optimization to the text encoder. However, such simple weighting schemes are limited when handling conflicting rewards. To address this, several works have introduced the concept of the Pareto-Optimal: Parrot Lee et al. (2024) refines prompts via a Prompt Expansion Network and selects sample pairs that dominate across all rewards for reinforcement learning; within diffusion-DPO Wallace et al. (2024) frameworks, CaPO Lee et al. (2025) calibrates different rewards to enhance training stability, and BalancedPO Tamboli et al. (2025) aggregates diverse reward signals through majority voting. Nevertheless, these methods often require additional auxiliary modules or extensive relabeling of paired samples, and they have not thoroughly examined the underlying causes of reward hacking—the most critical issue in reward optimization. In contrast, our method eliminates the need for paired datasets, directly integrates multiple rewards, and employs optimal transport to guide batch samples toward the Pareto frontier, thereby achieving more balanced and robust alignment under conflicting objectives.

B Proofs that Global Upper Bounds Induce Reward Hacking

Standing assumptions. We introduce two weak regularity conditions to formalize the role of the feasible set $\mathcal{F}_i$ associated with each prompt p_i:

- **(A1) Tightness on $\mathcal{F}_i$.** For each image domain $\mathcal{D}_i$ associated with prompt p_i, the reward upper bound $\overline{R_i} := \sup_{x\in\mathcal{F}_i} R(x)$ (or, in the multi-reward case, $\overline{R_i^k} := \sup_{x\in\mathcal{F}_i} R^k(x)$ for each k) is tight for the admissible set $\mathcal{F}_i$.
- **(A2) Admissibility gap.** The admissible set $\mathcal{F}_i$ consists of perceptually acceptable samples, where $\underline{Q_i} := \inf_{x\in\mathcal{F}_i} Q(x)$ exists. Samples outside $\mathcal{F}_i$ may either be perceptually comparable or inferior; in particular, those falling into the reward hacking region are characterized by spuriously high reward values yet degraded perceptual quality, i.e., $Q(x) < \underline{Q_i}$.

B.1 Single-Reward Case

[Proof of Proposition **??** (single reward) by contradiction] Let a prompt p_i induce image domain $\mathcal{D}_i$ and admissible set $\mathcal{F}_i \subseteq \mathcal{D}_i$. Let $C = \sup_{i,\, x\in\mathcal{F}_i} R(x)$ be the global constant used in the surrogate loss $\mathcal{L}(x) = C - R(x)$. By Property **??** and global construction of C, we have $C > \overline{R_i}$ for some p_i.

Assume for contradiction that there exists $x \in \mathcal{X}$ such that

$$\mathcal{L}(x) < C - \overline{R_i} \quad \text{and} \quad x \notin \mathcal{H}_i.$$

The loss inequality is equivalent to $R_i(x) > \overline{R_i}$. By **(A1)**, any point with $R_i(x) > \overline{R_i}$ cannot lie in $\mathcal{F}_i$, hence $x \notin \mathcal{F}_i$.

Now consider two possibilities:

1. If $x \notin \mathcal{D}_i$, then x is not a sample produced for prompt p_i, contradicting that we analyze optimization within $\mathcal{D}_i$.
2. If $x \in \mathcal{D}_i$, then from $R(x) > \overline{R_i}$ and the definition of $\overline{R_i}$ we have $x \notin \mathcal{F}_i$. By Definition **??**, any sample outside $\mathcal{F}_i$ with $R(x) > \overline{R_i}$ must also satisfy $Q(x) < \underline{Q_i}$. Therefore, x exhibits reward hacking, i.e., $x \in \mathcal{H}_i$.

Both cases contradict the assumption. Hence, any step achieving $\mathcal{L}(x) < C - R_i^\star$ necessarily produces $x \in \mathcal{H}_i$. This proves the claim for the single-reward case.

B.2 Multi-Reward Case

[Proof of Proposition **??** (multi reward) by contradiction] Fix a prompt p_i with domain $\mathcal{D}_i$ and admissible set $\mathcal{F}_i$. Let rewards be $\tilde{R}(x) = (R^1(x), \ldots, R^K(x))$, with weights $w_k \geq 0$ satisfying $\sum_{k=1}^K w_k = 1$. Define the admissible weighted bound

$$\overline{S_i} := \sup_{x \in \mathcal{F}_i} \sum_{k=1}^K w_k R^k(x), \qquad \overline{R_i^k} := \sup_{x \in \mathcal{F}_i} R^k(x).$$

Let the global constant C in $\mathcal{L}(x) = C - \sum_{k=1}^K w_k R^k(x)$ be constructed from cross-domain upper bounds (Sec. **??**), so that $C > \overline{S_i}$ for some i.

Assume for contradiction that there exists $x \in \mathcal{X}$ such that

$$\mathcal{L}(x) < C - \overline{S_i} \quad \text{and} \quad x \notin \mathcal{H}_i.$$

The loss inequality is equivalent to

$$\sum_{k=1}^K w_k R^k(x) > \overline{S_i}.$$

By definition of $\overline{S_i}$, no point in $\mathcal{F}_i$ can exceed this value, hence $x \notin \mathcal{F}_i$. Moreover, since the weighted sum strictly exceeds $\overline{S_i}$, there must exist at least one coordinate k' such that

$$R^{k'}(x) > \overline{R_i^{k'}}.$$

Now consider two possibilities:

1. If $x \notin \mathcal{D}_i$, then x is not a sample produced for prompt p_i, contradicting that we analyze optimization within $\mathcal{D}_i$.
2. If $x \in \mathcal{D}_i$, then from $R^{k'}(x) > \overline{R_i^{k'}}$ and the definition of $\overline{R_i^{k'}}$ we have $x \notin \mathcal{F}_i$. By Definition **??**, any sample outside $\mathcal{F}_i$ with some reward dimension exceeding its admissible upper bound, i.e. $R^{k'}(x) > \overline{R_i^{k'}}$, must also satisfy $Q(x) < \underline{Q_i}$. Therefore, x exhibits reward hacking, i.e., $x \in \mathcal{H}_i$.

Both cases contradict the assumption. Hence, any step achieving $\mathcal{L}(x) < C - \overline{S_i}$ necessarily produces $x \in \mathcal{H}_i$. This proves the claim for the multi-reward case.

C Details of GPT-4o Based Decision-making Agent

We introduce **GPT-4o** as a decision-making agent to adaptively manage multi-reward model training. The agent dynamically determines actions based on generated image quality and training stability, with three core capabilities:

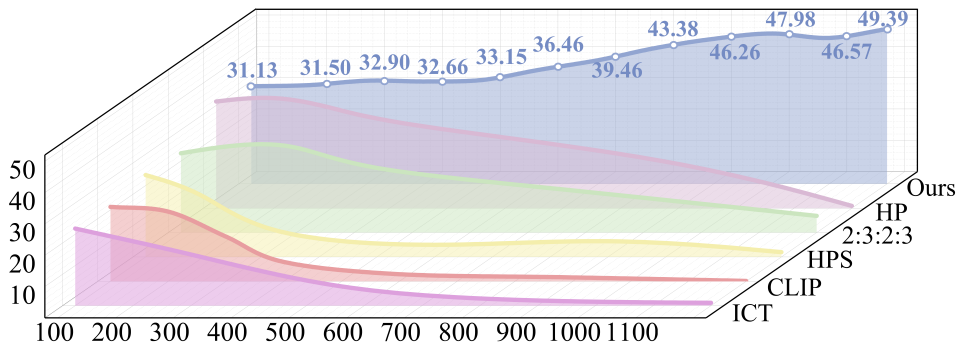

Figure 5: Training Curves of Joint Domination Rate (JDR_2).

- **Continue Training** — When no signs of collapse are observed in the generated images, the agent performs no additional operations and allows training to continue.
- **Remove & Revert** — Upon detecting a breakdown, the agent removes the unstable reward model and reverts to the most recent stable checkpoint.
- **Switch Strategy** — When training has proceeded for an extended period with minimal improvement in image optimization, while the remaining reward models remain stable, the agent smoothly transitions the process from offline training to online training.

C.1 Prior Knowledge and Initialization

Knowledge Base Construction. For each reward model, we collect degraded image samples generated under breakdown conditions and employ **GPT-4o** for semantic analysis to extract characteristic failure patterns. These include:

- **CLIP** — Texture pixelation and detail degradation, resulting in blurred appearance.
- **ICT** — Mildest breakdown, characterized by overly smooth transitions and loss of depth.
- **HPS** — Severe over-saturation, sharpness distortion, and vertical stripe artifacts.
- **HP** — Pronounced edge artifacts and over-saturated color transitions.

The extracted patterns are structured into a prior knowledge base, which supports subsequent chain-of-thought (CoT)-based decision-making.

State Initialization. At the beginning of training, the agent is initialized with the following state information: (i) current training strategy (offline/online), (ii) active reward model set $\{R^i\}$, (iii) cumulative training step count n, and (iv) historical stability metrics with corresponding checkpoint references.

C.2 Decision-Making Workflow

Environment Context. At each decision step, the agent receives structured contextual information, including: (i) the baseline image generated by the underlying model; (ii) composite breakdown images corresponding to different reward models (CLIP, ICT, HPS, HP); (iii) environment variables such as the set of active reward models, training step count, removal history, and the current strategy state (offline/online).

Decision Procedure. The agent performs a two-stage reasoning process:

1. **Breakdown Detection.** Compare diagnostic images with prior templates.
   - If a breakdown is detected, remove the most severely affected reward model and revert to the previous checkpoint.
   - If no breakdown is detected, proceed to the strategy evaluation stage.
2. **Training Strategy Decision.** Based on the current environment variables:

- If the remaining reward models remain stable without breakdown, switch from the offline strategy to the more flexible online strategy.
- Otherwise, maintain the offline training strategy.

Final Output. The decision agent produces a standardized output, which includes:

```
Final Output: [Training state analysis]
Action: [Remove / Revert / Continue]
Strategy Maintenance: [Offline / Online]
```

C.3 Detailed Training Procedure

We implemented a staged dynamic optimization approach with systematic evaluation checkpoints every 100 steps to ensure stability in multi-reward training. The procedure consisted of two main phases: an initial offline phase with joint reward model training, during which weaker reward models were progressively removed until only stable ones remained, followed by the activation of the online strategy once convergence was achieved.

Offline Training with Adaptive Reward Management In the offline phase, four reward models (CLIP, ICT, HPS, HP) were trained simultaneously. At each 100-step interval, the GPT-4o decision agent performed automated diagnostic analysis on composite output images to detect potential training instabilities. When anomalies were detected, the system reverted to the most recent stable checkpoint and adjusted the active reward set before resuming training.

The first breakdown occurred at step 200, where evaluation revealed over-saturation artifacts, sharpness distortions, and vertical stripe patterns consistent with HPS model instability. The agent removed HPS from the active reward set and rolled training back to step 100, after which training continued with the reduced set (CLIP, ICT, HP).

Training proceeded stably through steps 300, 400, and 500 without anomalies. However, at step 600, diagnostic analysis identified characteristic texture pixelation and detail degradation in CLIP outputs, while ICT and HP remained stable. The agent isolated CLIP as the instability source, removed it from the active set, and reverted to the step 500 checkpoint. Training then continued with the remaining reward pair (ICT, HP).

Transition to Online Training Between steps 500 and 800, training exhibited sustained stability with no further breakdowns. After confirming at step 800 that generated images showed no collapse and only minimal changes, the system initiated the transition to online training mode for further exploration.

D Additional Visualization Results

Visualization and Analysis of Training Curve. As shown in Figure 5, we present the joint domination rate JDR_2 on two strong rewards, ICT and HP. It can be observed that our method steadily improves as training progresses, whereas the baseline methods—whether single-reward optimization or multi-reward joint optimization—experience a rapid decline in joint domination rate, indicating that the baseline approaches quickly encounter reward hacking issues.

Qualitative Case Studies on Pareto Frontier Visualization. As shown in Figure 6, we provide additional visualization examples of Pareto frontiers. It can be observed that the Pareto frontier characteristics vary across different prompts. Our method consistently dominates the baseline approaches, and the overall image distributions are closer to the Pareto frontier, demonstrating the superiority of our method.

Visualization of Heterogeneous Reward Bounds Across Different Prompts As shown in 7, we provide box plot visualizations of reward ranges for the ICT score across different prompts, based on 50 images generated with different seeds. In 8, we present the corresponding box plot visualizations for the HP score across different prompts. It can be observed from these extensive examples that the reward upper bounds are heterogeneous across prompts, and that the reward ranges and characteristics also vary between prompts.

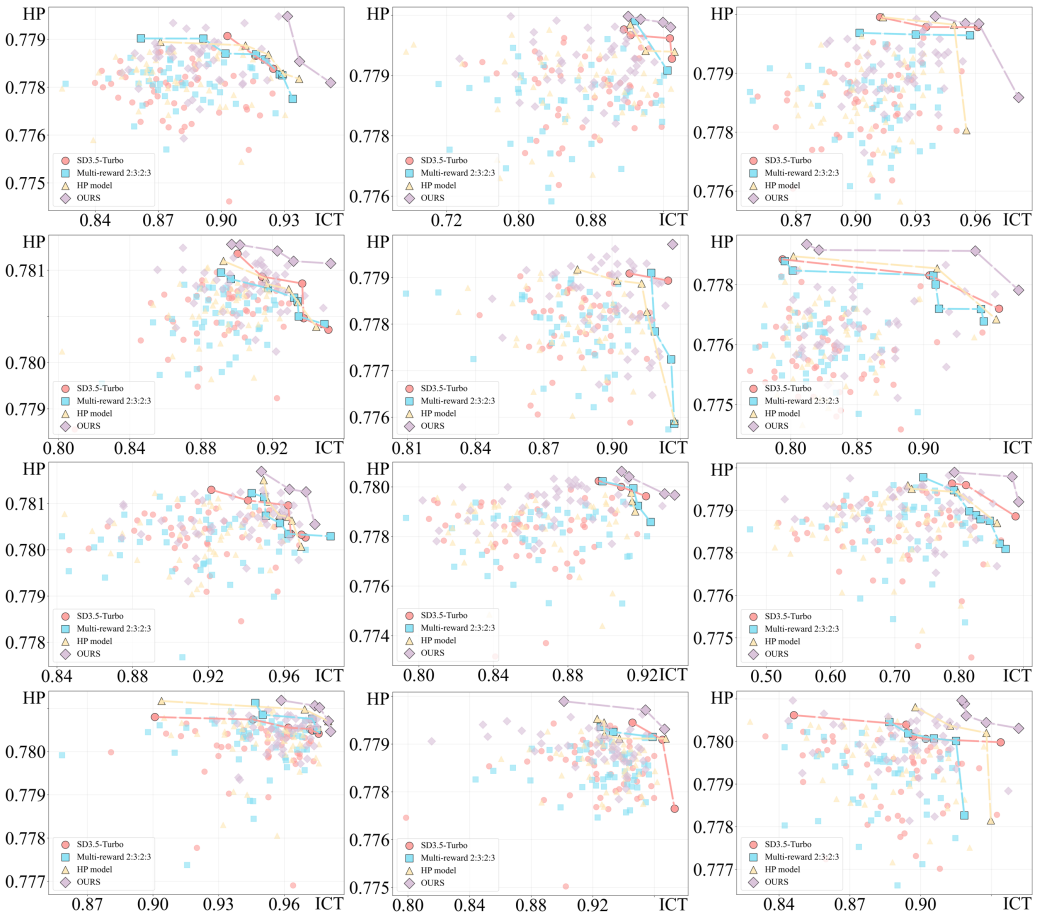

Figure 6: Broad Comparative Examples of Pareto Frontier Visualizations for Various Methods.

More Qualitative Comparison Results. We present additional visualizations for comprehensive comparison. Figure 9 shows results against single-reward baselines, while Figure 10 illustrates comparisons with multi-reward baselines.

E Experiment Details

For diffusion model optimization, we use 32,000 non-repetitive text prompts from the Pick-High dataset, a subset of the Pick-a-pic dataset. All experiments are conducted on a cluster of three nodes, each with eight A800 GPUs. We adopt DDIM sampling with four denoising steps, set the classifier-free guidance weight to 0.0, and fix the output resolution to 512×512. The Adam optimizer is used for training with a learning rate of 5×10^{-5}.

F The Use of Large Language Models (LLMs)

In this study, Large Language Models (LLMs) were primarily used for language polishing and played a limited supporting role during the experimental process. Specifically, LLMs were employed to refine textual expressions, improve clarity, and assist with partial diagnostic analysis and strategy judgments in multi-reward model training. Beyond these limited supportive tasks, all experimental design, implementation, and result verification were carried out independently by the authors to ensure that the core ideas and scientific contributions of the research remain entirely author-driven.

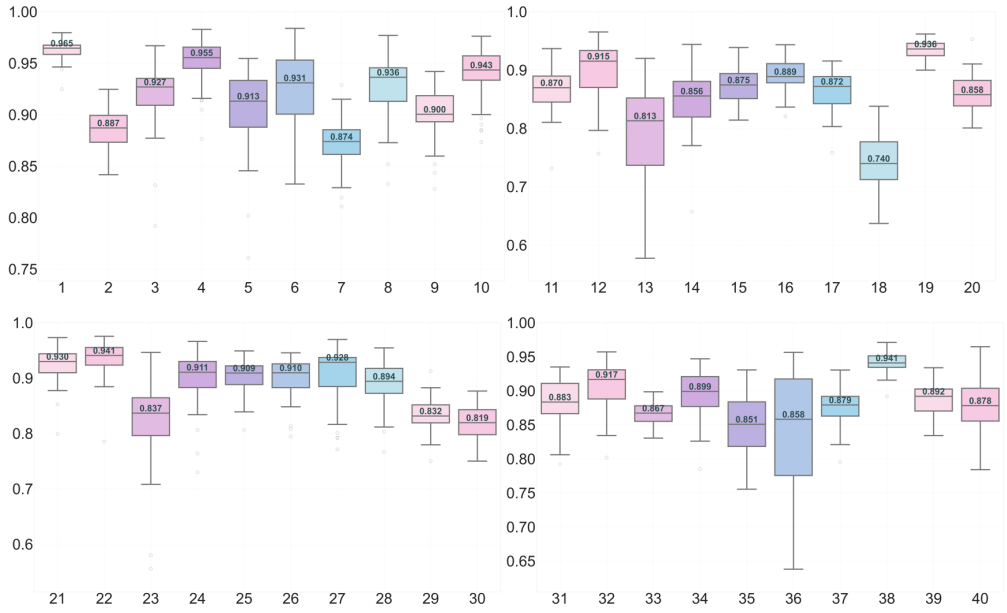

Figure 7: Visualization of Box Plots Showing Reward Variations Across Prompts on ICT Score.

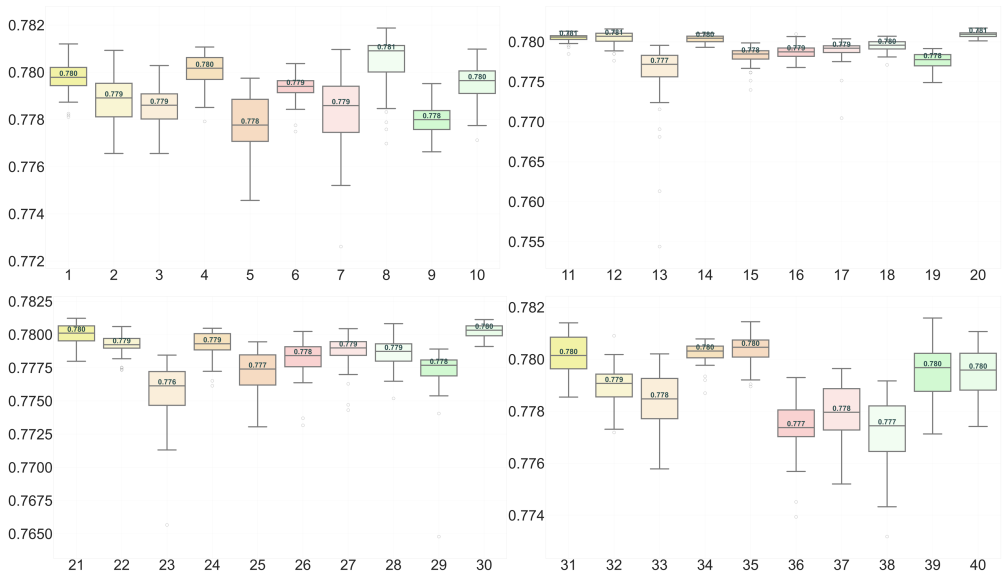

Figure 8: Visualization of Box Plots Showing Reward Variations Across Prompts on HP Score.

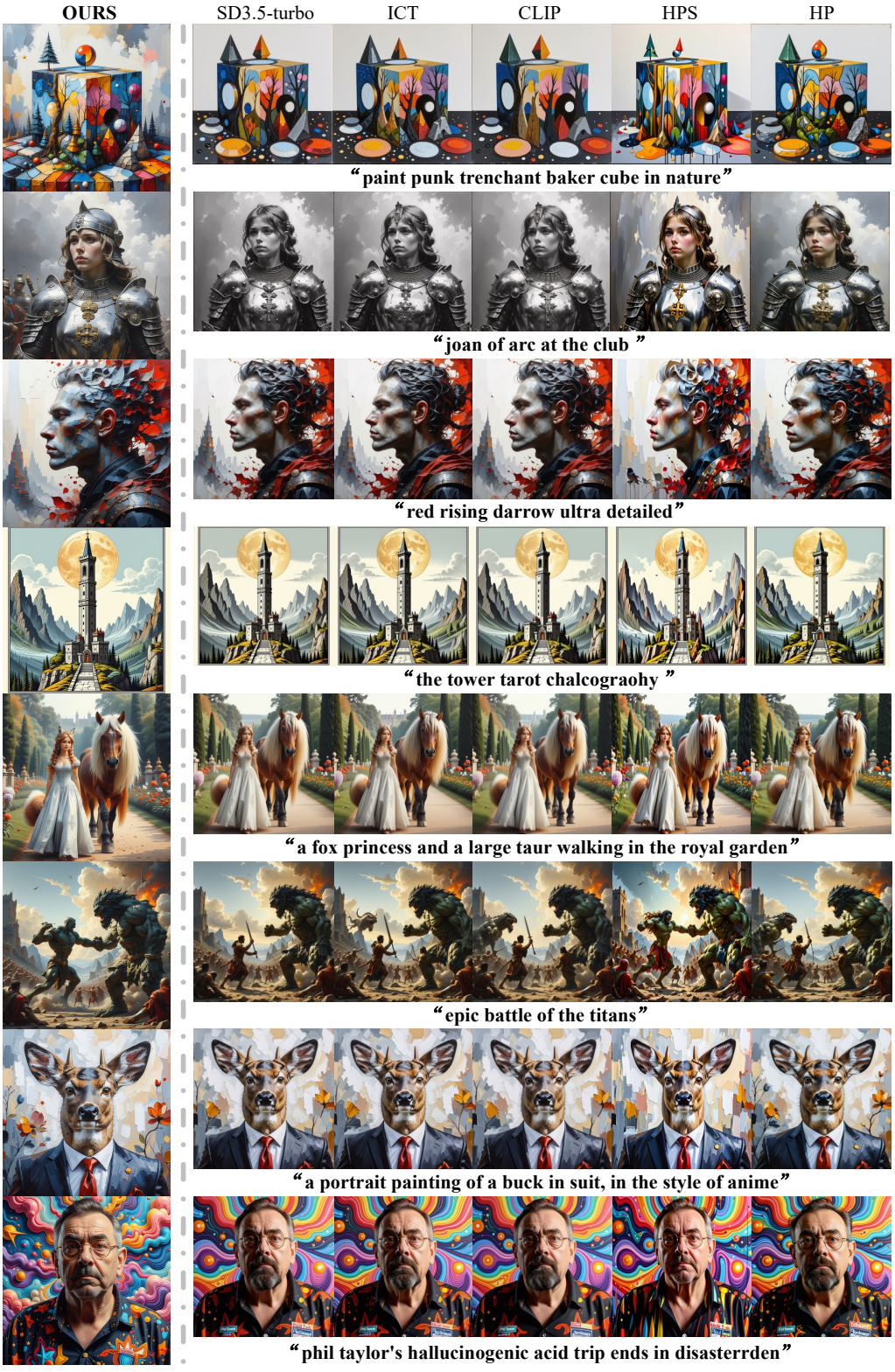

Figure 9: Qualitative Comparison of Optimization Results with Single-Reward Baselines.

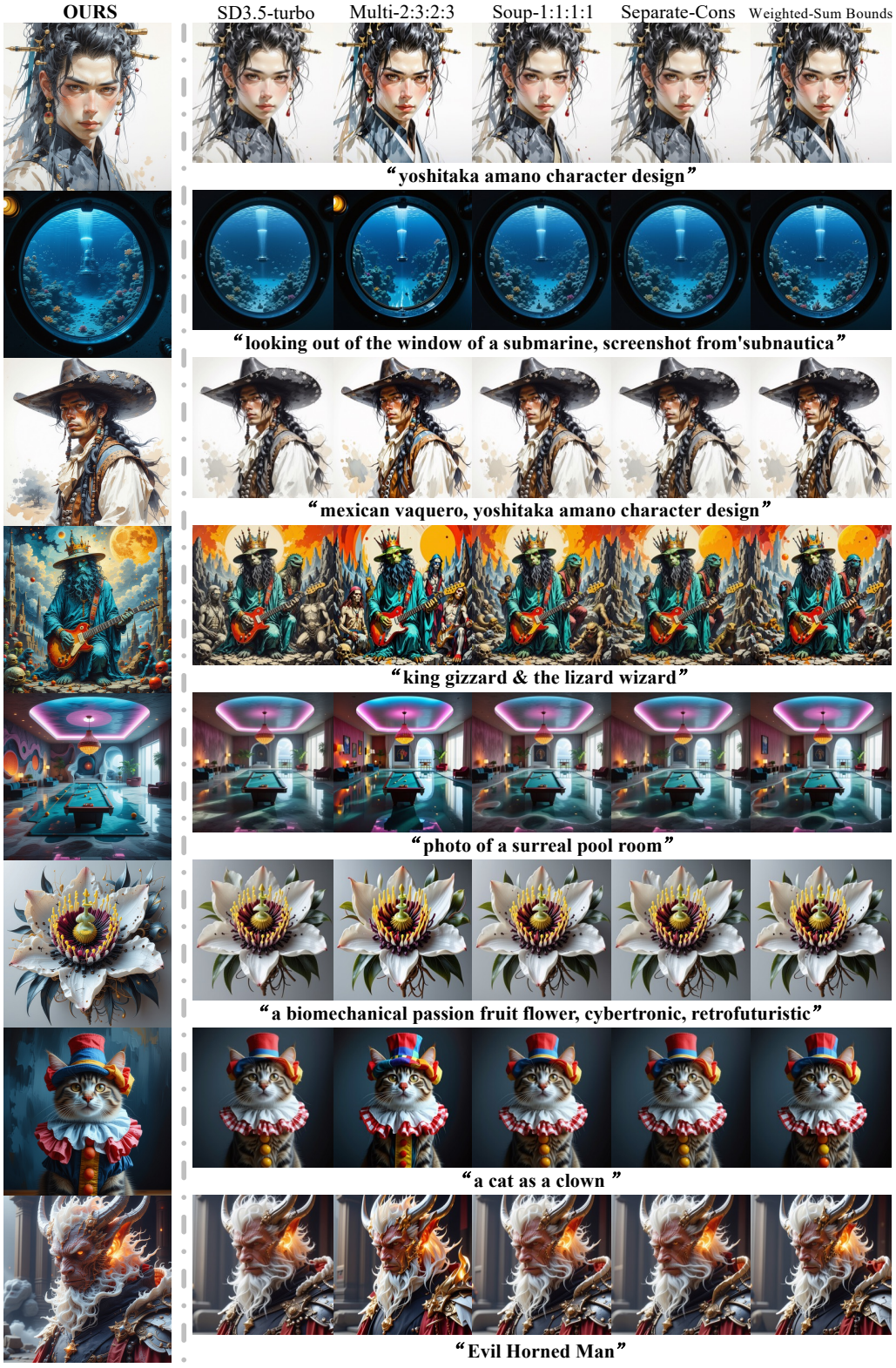

Figure 10: Qualitative Comparison of Optimization Results with Multi-Reward Baselines.