RSTFA: Efficient Training-Free Human Preference Alignment via Rejection Sampling for Text-to-Image Diffusion Models

Hongzheng Yang, Jason Chun-Lok Li, Li Kun, Wenao Ma, Mingjie Xu, Yuzhi Zhao, Lai-Man Po, *Senior Member, IEEE,*

***Abstract*—Given a text-to-image diffusion model pretrained on large-scale text-image pairs, can we align the model with human preferences without further fine-tuning? In this paper, we analyze the effect of alignment tuning in diffusion models by comparing the diffusion denoising trajectory between base and aligned models. Our findings reveal that alignment tuning primarily affects superficial stylistic aspects during denoising, rather than fundamental content, suggesting superficial alignment behaviors. Based on this discovery, we introduce a novel, training-free alignment approach (RSTFA) that leverages rejection sampling at specific stylistic timesteps, ensuring human preference alignment without fine-tuning or heavy inference overhead. We provide a theoretical analysis and derive a bias bound for our rejection-sampling alignment scheme. Empirically, we show that RSTFA better preserves sample diversity than reinforcement-learning-based tuning methods. Extensive experiments on Pick-a-Pic, COCO, HPD V2, and PartiPrompts show that our method not only achieves superior alignment with human preferences compared to state-of-the-art methods, but also reduces computational demands, establishing efficient, human-centered diffusion model alignment.**

***Index Terms*—Diffusion Model, Human Preference Alignment, Computer Vision for Social Good.**

I. Introduction

TEXT-to-image diffusion models have rapidly advanced the field of generative modeling, enabling the creation of high-quality and diverse images from textual descriptions. These models have found widespread applications in art generation [1]–[3], design prototyping [4]–[6], and virtual reality [7]–[10], effectively bridging the gap between linguistic concepts and visual representations. However, these models may still lack an explicit understanding of nuanced human aesthetics and context [11]–[13]. Aligning these models with human preferences remains a significant challenge. This gap can manifest as the production of undesirable or low-quality images, limiting the utility and acceptance of these models in practical applications. Therefore, alignment is crucial to ensure that the content generated by diffusion models is not only high-quality but also aligned with human values [14]–[17].

Current approaches to aligning diffusion models with human preferences can be broadly categorized into training-based and training-free methods. Training-based methods, such as direct preference optimization (DiffDPO) [18], step-aware preference optimization (SPO) [19], and denoising diffusion policy optimization (DDPO) [20], involve fine-tuning models using pairwise human preference datasets [21], [22] or a reward model [11] to capture human preference. However, these methods often incur substantial training costs and are therefore less flexible. Moreover, Barceló et al. [23] pointed out that diffusion models fine-tuned with reinforcement learning are susceptible to training instability and mode collapse issues, resulting in reduced diversity of image generation. In contrast, training-free methods were proposed to remove the need for model retraining. For example, Direct Noise Optimization (DNO) [24] proposed to maximize a differentiable reward model via optimizing the intermediate noise at inference time. These methods generally incur a high inference cost per query (e.g., $\sim 5\times$ the base model's inference time), which limits practical deployment.

To overcome the above challenges, there is a pressing need to better understand the effect and mechanism of human preference alignment tuning in text-to-image diffusion models. Recent studies [25]–[30] in large language models (LLMs) have provided an important insight: alignment-tuned models show almost identical decoding across most token positions, with distribution shifts primarily occurring at stylistic tokens. This observation motivates us to investigate whether similar patterns exist in diffusion models, which could potentially lead to more efficient alignment methods. Through our investigation of diffusion denoising trajectory, we demonstrate that alignment adjustments in diffusion models primarily occur at specific stylistic timesteps, affecting only surface-level attributes of generated images without significantly altering core content. If alignment primarily operates at a superficial level and at specific timesteps, then perhaps we do not need extensive model fine-tuning or complex inference-time optimization procedures to achieve effective alignment.

Inspired by [31], we first explore a simple training-free approach called RandSeed, which generates multiple images using different random seeds and selects the one with the highest human preference score. Our experimental results show that RandSeed achieves alignment performance surpassing or comparable to both training-based and training-free methods. Although RandSeed is a brute-force sampling approach with high inference cost, its success demonstrates that base diffusion models could generate preference-aligned images via base models' inherent diversity and stochasticity. Based on the insight provided by our diffusion denoising trajectory analysis and RandSeed baseline, we propose RSTFA

Manuscript received July 17, 2025.

H. Yang, W. Ma are with the Chinese University of Hong Kong, Hong Kong, China. (e-mail: hzyang22@cse.cuhk.edu.hk; wama@link.cuhk.edu.hk)
J. Li is with the University of Hong Kong, Hong Kong, China. (e-mail: jasonlcl@connect.hku.hk)
M. Xu is with the Hong Kong University of Science and Technology (Guangzhou), Guangzhou, China. (e-mail: mingjiexu@hkust-gz.edu.cn)
K. Li, Y. Zhao, L.-M. Po are with the City University of Hong Kong, Hong Kong, China (e-mail: kunli25-c@my.cityu.edu.hk; yzzhao2-c@my.cityu.edu.hk; eelmpo@cityu.edu.hk).

(Rejection Sampling for Training-free Alignment), an efficient training-free alignment method with small inference cost. RSTFA selectively applies rejection sampling at critical stylistic timesteps during the diffusion process. By employing rejection sampling, we utilize the inherent randomness in base diffusion models to deliver preference-aligned images. By leveraging a timestep-aware reward model to estimate density ratios and compute acceptance probabilities, RSTFA efficiently guides the generation process toward outputs that better reflect human preferences. This method achieves alignment efficiently, requiring no model modifications or iterative optimization during inference. Theoretical analysis provides a bias bound for RSTFA, and RSTFA preserves base model diversity better than existing training-based and training-free methods in practice.

Our main contributions can be summarized as follows:

- We conduct an in-depth analysis of alignment effects in text-to-image diffusion models, demonstrating that alignment primarily impacts superficial stylistic attributes.
- We propose a novel training-free approach RSTFA, which efficiently achieves human preference alignment built on rejection sampling theory, with a bias bound provided.
- Extensive experiments on Pick-a-Pic, COCO, HPD V2 and PartiPrompts datasets with different backbones demonstrate that our method matches or surpasses state-of-the-art alignment techniques while significantly reducing computational costs and preserving image diversity. The user study suggests that our method achieves win rate of 65.4% compared to previous state-of-the-art methods.

II. Related Work

A. Text-to-Image Diffusion Model

Text-to-image diffusion models have emerged as a powerful class of generative models capable of producing high-quality images conditioned on textual descriptions. Built upon the foundations of diffusion probabilistic models [32], [33], these models iteratively refine random noise into coherent images through a reverse diffusion process guided by textual embeddings. Notable advancements include DALL-E [34]–[36], Imagen [37], [38], and Stable Diffusion series [39]–[41], all of which have demonstrated impressive capabilities in generating diverse, high-fidelity images from text prompts. However, [11], [18] found that these models may struggle to fully capture nuanced human aesthetics and context. As a result, the generated outputs may not always align with human preferences or expectations. Developing techniques to better align text-to-image models with human values and aesthetic sensibilities remains an open problem.

B. Aligning Text-to-Image Model with Human Preferences

Methods to align text-to-image diffusion models with human preferences generally fall into training-based and training-free categories. Training-based methods include reinforcement learning (DDPO) [20], direct preference optimization (DiffDPO) [18], step-aware preference optimization (SPO) [19], and MaPO [42]. In DDPO, diffusion models are optimized using a reward objective applied via policy gradient techniques. DiffDPO adapts the Direct Preference Optimization (DPO) [15] method from LLMs to diffusion models, achieving improved alignment with human preferences compared to supervised fine-tuning or reinforcement tuning methods. However, DiffDPO aligns all timesteps to the final timestep's preference order, a limitation identified by SPO, which introduces a step-aware preference optimization approach. MaPO applies the reference-free alignment method ORPO [43] from LLMs to diffusion models, eliminating the need for a reference model during preference optimization and reducing memory costs. These training-based methods rely on fine-tuning with preference-based datasets or training toward maximizing reward models to capture human preference judgments, but can be computationally intensive and may suffer from instability, causing mode collapse [23], and reducing generation diversity. Training-free approaches like direct noise optimization (DNO) [24] avoid fine-tuning by optimizing noise iteratively during inference, thereby bypassing fine-tuning overhead. However, this method requires additional optimization at inference time, increasing latency and limiting its practical applicability.

C. Diffusion Rejection Sampling

Diffusion Rejection Sampling (DRS) [44] improves unconditional or class-conditional generation quality by applying accept–reject at every diffusion step using statistics derived from diffusion likelihoods or classifier energies. In contrast, our setting is text-to-image preference alignment, where the target distribution is defined implicitly by human preferences rather than a tractable likelihood. Methodologically, we estimate per-step log density ratios by calibrating ordinal rewards via Bregman minimization. We use a multiple-try Metropolis–Hastings kernel with a small fixed proposal budget to bound latency. This yields a principled acceptance probability tied to human preferences and a different computational trade-off.

III. Preliminaries

A. Diffusion Background and Notations

We adopt the ϵ-parameterization of diffusion models with a variance-preserving noise schedule $\{\bar{\alpha}_t\}_{t=0}^{T}$. Let x_t denote the latent at timestep t and c the text condition. The forward process corrupts clean image x_0 via Gaussian noise, $x_t \sim \mathcal{N}(\sqrt{\bar{\alpha}_t}\, x_0, (1-\bar{\alpha}_t)I)$. The reverse sampler uses a denoiser $\epsilon_\theta(x_t, t, c)$ to estimate the noise and form the transition $x_t \to x_{t-1}$ under the chosen solver (SDE-DPMSolver for SDXL/SDv1.5 and EulerDiscrete for FLUX). For analysis only, we project a noisy state (x_t, t) to an estimated clean image via a one-step DDIM estimate:

$$\hat{x}_0(x_t, t; \epsilon_\theta) \;=\; \frac{x_t - \sqrt{1-\bar{\alpha}_t}\,\epsilon_\theta(x_t, t, c)}{\sqrt{\bar{\alpha}_t}} \tag{1}$$

We denote this quantity by $x_{\text{clean}}(t)$. This is a local projection used to compare states across timesteps.

B. Problem Definition

Given a pre-trained text-to-image diffusion model with denoising network $\epsilon_{\text{base}}(x_t, t, c)$, we aim to align the model's

output distribution with human preferences without fine-tuning. Formally, we seek to transform the base distribution $p_{\text{base}}(x_0|c)$ to an aligned distribution $p_{\text{aligned}}(x_0|c)$ that better reflects human aesthetic preferences as measured by a preference score function. Our approach focuses on training-free alignment, which modifies the sampling process during inference to produce samples from $p_{\text{aligned}}(x_0|c)$ while keeping the model parameters θ_{base} fixed. This alignment is guided by a reward function $r(x, c)$ that assigns preference scores to generated images x given prompt c, trained on human preference data following the Bradley-Terry model. Human preferences are modeled consistently through pairwise comparisons, where for any two images x^w and x^l generated from the same prompt c, the preference relation follows:

$$P(x^w \succ x^l|c) = \frac{\exp(r(x^w, c))}{\exp(r(x^w, c)) + \exp(r(x^l, c))} \tag{2}$$

To enable rejection sampling at intermediate diffusion steps, we extend this framework to a timestep-aware preference model $r(x_t, t, c)$ that evaluates preference scores for intermediate denoised images x_t at timestep t, enabling density ratio estimation throughout the diffusion process.

We refer to the alignment change at timestep t as superficial when it is confined to low-level stylistic attributes (e.g., tone, palette, texture granularity, lighting, local contrast) while semantics and global structure remain stable. This pattern manifests as a change in VGG [45] based style features together with consistently high SAM [46] derived structure similarities at the same t. We use the term *fundamental* to describe changes that alter semantics and structure, which co-occur with noticeable changes in CLIP and SAM features. This characterization is supported by our feature-wise decompositions, the similarity curves in Figure 2 and Figure 7, and the qualitative comparisons in Figure 1 and Figure 3 across backbones and prompt sources.

IV. Motivation

A. Superficial Alignment Behaviors in Training-based Methods

In this section, we investigate the alignment tuning behaviors in text-to-image diffusion models. We demonstrate that the alignment tuning primarily affects superficial stylistic attributes of the generated images, coinciding with the superficial alignment hypothesis [25], [26] observed in LLMs alignment literature. By uncovering these behaviors, we motivate the development of training-free methods that can achieve desired alignment effects efficiently without extensive fine-tuning.

First, we define two models sharing the same text-to-image architecture but different weights:

- **Base Model (ϵ_{base})**: A pretrained diffusion model without any alignment procedures applied.
- **Aligned Model ($\epsilon_{\text{aligned}}$)**: The same model after alignment training using a dataset of human-preference image pairs.

We want to analyze the diffusion denoising trajectory for ϵ_{base} and $\epsilon_{\text{aligned}}$. Recent studies in LLMs [25], [26] suggest that alignment tuning often concentrates on stylistic tokens while leaving the majority of decoding unchanged. We investigate whether a similar superficial alignment phenomenon exists in diffusion model denoising trajectory: alignment-tuned deviations are sparse in timesteps and localize to stylistic timesteps that adjust superficial attributes (e.g., tone, palette, texture granularity, lighting, local contrast) while preserving semantics and structure. To test this superficial alignment hypothesis, we design a localized, causal probe that isolates the effect of substituting the denoiser at exactly one timestep.

Given a prompt c, starting from the pure Gaussian noise $x_T \sim \mathcal{N}(0, I)$, the aligned model generates the reference denoising trajectory:

$$x_T, x_{T-1}, \ldots, x_0, \quad x_t \sim \mathcal{K}_t^{\text{aligned}}(x_{t+1}, c), \tag{3}$$

where $\mathcal{K}_t^{\text{aligned}}$ is the reverse transition at timestep t implemented by the aligned denoiser.

To isolate the base model's next step behavior, we construct a cross-model trajectory probed only on one timestep. We compute a single base reverse step ($t+1 \to t$) from the exact same aligned latent x_{t+1} for $t \in \{T-1, \ldots, 2, 1, 0\}$:

$$y_t \sim \mathcal{K}_t^{\text{base}}(x_{t+1}, c), \tag{4}$$

where $\mathcal{K}_t^{\text{base}}$ is the reverse transition at timestep t implemented by the base denoiser. We do not continue sampling from y_t. Each t is probed independently. This yields the sequence $y_{T-1}, y_{T-2}, \ldots, y_0$. This cross-model trajectory avoids compounding sampling noise, providing a clean attribution to the denoiser at only one timestep t.

For quantitative attribution, we project x_t and y_t to their one-step DDIM clean estimates (Eq. (1)):

$$x_{\text{clean}}(t) := \hat{x}_0(x_t, t; \epsilon_{\text{aligned}}), \quad y_{\text{clean}}(t) := \hat{x}_0(y_t, t; \epsilon_{\text{base}}). \tag{5}$$

We then compute a feature space similarity:

$$s(t) = \cos\big(\phi(x_{\text{clean}}(t)), \phi(y_{\text{clean}}(t))\big), \tag{6}$$

where $\phi(\cdot)$ concatenates CLIP image embeddings [47], VGG features [45], and SAM encoder features [46], with ℓ_2 normalized before concatenation. We use the CLIP image encoder to extract global semantic embedding, VGG to extract low-level visual patterns and SAM to extract content structure features. The clean projections defined in Eq. (5) yield image-like representations at identical signal-to-noise ratios, ensuring a fair comparison in the feature space. Furthermore, we conduct a step-by-step qualitative comparison between the sequence $\{y_t\}_{t=T-1}^{t=0}$ and $\{x_t\}_{t=T-1}^{t=0}$ to visualize how the base model would locally differ from the aligned model.

Findings. As shown in Figure 2, the similarity $s(t)$ remains remarkably high across most timesteps, indicating that the base model's outputs are very similar to those of the aligned model under the cross-model denoising trajectory at most timesteps. Interestingly, we observed that the similarity drops in certain stylistic timesteps periodically. We provide an intuitive explanation for this phenomenon in Appendix B. As shown in Figure 1, the differences between base and aligned model output at those timesteps are mainly superficial stylistic elements, rather than the core content of the generated images. Furthermore, at other timesteps, the differences between the base and aligned models are small and not noticeable upon visual inspection. This suggests that alignment tuning only affects a small portion of the diffusion process and supports the superficial alignment hypothesis.

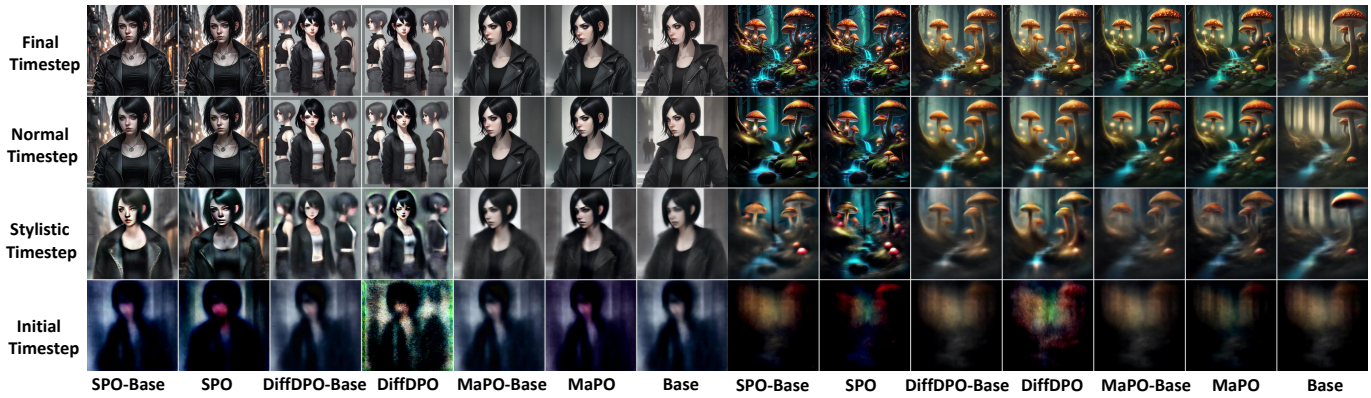

Fig. 1. Qualitative comparisons of denoising output at initial timestep, stylistic timestep, normal timestep and final timestep for base and aligned models. At the normal timestep, no discernible differences are visible between the denoised outputs. The stylistic timestep reveals minor but notable stylistic variations between models, where alignment tuning affects only superficial attributes such as color tone or texture.

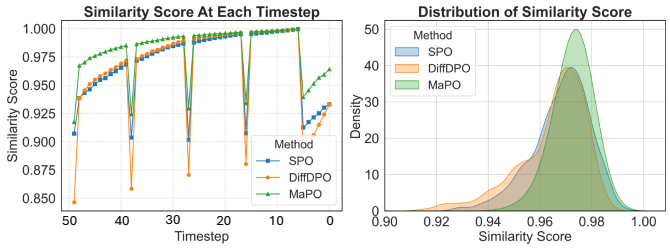

Fig. 2. Local one-step switching probe of alignment effects along the reverse denoising trajectory (SDXL, 50 steps). For a given timestep t, we switch only the single reverse transition $t{+}1 \to t$ by applying the base denoiser to the same aligned latent x_{t+1}, producing y_t (Eq. (4)). We compare the one-step clean projections $x_{\text{clean}}(t)$, $y_{\text{clean}}(t)$ and the cosine similarity $s(t)$ (Eq. (5), Eq. (6)). (a) $s(t)$ in a reverse-time direction from $t = T-1$ to $t = 0$; (b) aggregated distribution of $s(t)$ over prompts and timesteps.

B. A Compute-Matched Baseline: RandSeed

The superficial alignment behaviors observed in Section IV-A motivate us to rethink the necessity of computationally heavy alignment fine-tuning or inference-time optimization (e.g., DNO [24]). To evaluate whether the high inference overhead introduced by existing training-free methods is justified, we adopt a compute-matched baseline, **RandSeed**: generate k samples from the base model using different random seeds and select the top-1 by a preference scorer. We choose k to match the FLOPs of prior training-free optimizers. As shown in Table I, despite its simplicity, RandSeed achieves alignment performance comparable to or exceeding existing methods.

The success of RandSeed can be attributed to unlocking the potential of base diffusion models. By exploiting the inherent randomness and diversity of diffusion sampling, RandSeed effectively unlocks the capabilities of base models and improves alignment. This observation indicates that base diffusion models, when properly utilized, can deliver competitive alignment performance, motivating the development of a more efficient and targeted training-free method.

V. RSTFA: Training-free Alignment via Stylistic Rejection Sampling

The strong empirical results of RandSeed reveal that preference–aligned images are already present in the distribution of the base diffusion model p^{base}. RSTFA turns this observation into a training-free[1] sampling scheme that intervenes the generation process only at a handful of stylistic timesteps where Sec. IV-A showed a measurable divergence between the base and alignment fine-tuned models.

A. Density Ratio Estimation

Building on the insight that alignment effects are localized, RSTFA proposes to apply rejection sampling at these stylistic timesteps implemented via Metropolis-Hastings (MH) accept–reject correction. To guide this process, we first train a timestep-aware reward model $r_\phi(x_t, t, c)$ on a dataset of human-preference pairs. Every training instance consists of a prompt c, a timestep t and two latent images (x_t^w, x_t^ℓ) produced at the same timestep t and prompt c. The preferred (winner) latent is denoted as x_t^w and the less preferred (loser) is denoted as x_t^ℓ. The details can be found in Section VI-B. The model $r_\phi(x_t, t, c)$ outputs a scalar reward, which is trained to maximize the likelihood that winners outrank losers under the Bradley–Terry model:

$$\mathcal{L}_{\text{BT}}(\phi) = -\sum_{(w,\ell)} \log \frac{\exp(r_\phi(x_t^w, t, c))}{\exp(r_\phi(x_t^w, t, c)) + \exp(r_\phi(x_t^\ell, t, c))}. \tag{7}$$

After convergence, we freeze ϕ. In practice, we could adapt an off-the-shelf pretrained reward model into a timestep-aware reward model at low cost.

The reward is only monotonic with respect to the preference. To perform rejection sampling, we require the density ratio between the aligned and base distributions $\rho_t(x_t) = p_t^{\text{aligned}}(x_t)/p_t^{\text{base}}(x_t)$. Although we generally lack explicit density functions, we could estimate this ratio using the timestep-aware reward model $r_\phi(x_t, t, c)$. Following the density ratio estimation literature [48], [49], we calibrate this reward score into a log-density ratio using a shallow MLP g_{ψ_t}. The calibrator is trained by minimizing the Bregman divergence [48]:

$$\begin{aligned}\mathcal{L}_{\text{Bregman}}(\psi_t) = \mathbb{E}_{x_t \sim p_t^{\text{base}}}\left[e^{g_{\psi_t}(r_\phi(x_t,t,c))}\right] \\ - \mathbb{E}_{x_t \sim p_t^{\text{aligned}}}\left[g_{\psi_t}(r_\phi(x_t,t,c))\right].\end{aligned} \tag{8}$$

[1]Training-free refers to the denoiser ϵ_θ. We only train a reward model and a ratio calibrator once, amortized across backbones.

The optimal $g^*_{\psi_t}$ recovers the log-density ratio $\log(p_t^{\text{aligned}}/p_t^{\text{base}})$ up to an additive constant. The calibrated log-density ratio used at inference is $\log \rho_t(x_t) = g^\star_{\psi_t}\big(r_\phi(x_t, t, c)\big) + C_t$.

B. Stylistic Metropolis-Hastings Sampling

Let $\mathcal{T}_{\text{stylistic}}$ be the set of timesteps where the base and aligned models differ noticeably. For timesteps $t \notin \mathcal{T}_{\text{stylistic}}$, we sample the base model as usual. For stylistic timesteps $t \in \mathcal{T}_{\text{stylistic}}$, we apply a Metropolis-Hastings correction step to align the trajectory towards human-preferred outputs. The procedure operates through two operations: proposal generation and probabilistic acceptance.

First, for each $t \in \mathcal{T}_{\text{stylistic}}$, we sample a candidate denoised state x'_t from the base model transition distribution:

$$x'_t \sim p^{\text{base}}_{t|t+1}(\cdot\,|x_{t+1}, c), \tag{9}$$

This candidate represents a potential alternative denoising path at the current timestep.

Second, we compute the acceptance probability using ρ_t:

$$\alpha_t(x_t, x'_t) = \min\left(1,\ \frac{\rho_t(x'_t)}{\rho_t(x_t)}\right). \tag{10}$$

Here, $\rho_t(\cdot)$ quantifies the relative preference alignment of a given state. When $\rho_t(x'_t) > \rho_t(x_t)$, the candidate is always accepted since it represents strict alignment improvement. When $\rho_t(x'_t) \le \rho_t(x_t)$, acceptance occurs probabilistically proportional to the relative gain $\rho_t(x'_t)/\rho_t(x_t)$. This ensures the sampling process selects trajectories that maximize human preference alignment at critical stylistic timesteps.

To maintain predictable computational latency while preserving the quality of alignment, we employ a multiple-try [50] variant of MH sampling with $m = 5$ proposals per stylistic timestep. This ensures an average of fewer than m additional model evaluations while bounding the maximum computational overhead at a constant factor. This design preserves the stochastic benefits of rejection sampling while eliminating unpredictable computational spikes that would occur with a native repeat-until-accept rejection sampling implementation.

Conceptually, this MH correction acts as a lightweight steering mechanism. By selectively intervening only where alignment tuning induces measurable divergence between base and aligned models, this achieves significant alignment improvement with minimal computational cost, typically requiring just $1.6\times$ more time than base sampling compared to $6.7\times$ for alternative training-free methods.

VI. Experiments

A. Experimental Setup

Baselines and Datasets. Following the settings of [18], [19], [42], we employed the Stable Diffusion XL-1.0 (SDXL) [51], SDv1.5 and DiT [52] as the base model. We evaluated the effectiveness of our proposed method, RSTFA, through extensive experiments using the Pick-a-Pic V2 dataset [22], HPD V2 and COCO prompts. Pick-a-Pic V2 dataset contains human preference image pairs labeled by human raters and has three splits, *train*, *validation_unique* and *test_unique*. In each pair, one image is marked as more aligned with human preference, while the other is less preferred. Our approach was compared with the state-of-the-art methods for aligning diffusion models with human preferences. These baselines included training-based methods, DiffDPO [18], MaPO, and SPO [19], as well as the training-free method DNO [24]. We also include Diffusion Rejection Sampling (DRS) [44] as a baseline, which was proposed to improve the quality of class-conditional image generation. We adopted the train split of Pick-a-Pic V2 with 851,293 pairs as training data for these training-based methods. For DNO, we observed that the generated images did not change by visual inspection after optimizing for 10 steps. Therefore, we fixed the number of test-time optimization steps as 10. For RandSeed, we set the number of random seeds as $k = 10$. This ensures RandSeed inference computation FLOPs per instance do not exceed those of DNO. For DRS, we apply the DRS implementation of accept–reject at all timesteps using diffusion likelihood ratio as the acceptance statistic. We ensured fair comparisons by adhering to the optimal settings provided in the official implementations of each baseline method. We tune baseline hyperparameters on *validation_unique* split of Pick-a-Pic V2.

Evaluation Prompts. For evaluation, we generated images conditioned on 500 unique prompts from the Pick-a-Pic V2 *test_unique* split and evaluated them using various AI feedback models. We also adopt HPD V2 and COCO prompts for a more comprehensive evaluation as the reward model is trained on the Pick-a-Pic V2 dataset. This cross-dataset evaluation helps establish the broad applicability of our method across diverse text prompts and visual scenarios.

Evaluation Metrics. First, following [18], we measured the prompt-aware human preference alignment via four metrics, including Text-Image CLIP score (TI-CLIP) [47], [53], ImageReward [54], Human Preference Score v2 (HPS-v2) [21], and PickScore [22]. Second, we evaluated the prompt-independent human preference alignment using Aesthetic score calculated with the LAION Aesthetics Predictor [55]. The Aesthetic score [56], [57] quantifies the visual appeal of images independent of the textual prompt. Third, we measured the diversity of generated images using four metrics. Following [19], we used Image-Image CLIP (II-CLIP) score [58] to measure the similarity between images based on features extracted by the CLIP vision encoder. Following [59]–[63], standard metrics, including Vendi score, Mean-similarity score (MSS) and LPIPS were used to evaluate diversity. For text-to-image generation alignment, we do not have a generic and fixed reference set. Thus we do not adopt the metrics, such as Recall and Coverage, which are not reference-free. Finally, we measure the inference cost (GenTime) by benchmarking the average single image generation time on one A800 GPU with the same environment configurations for all methods.

B. Implementation Details

We adopted the SDE-DPMSolver [64] as the sampler with classifier-free guidance scale of 5.0 for SDXL and SDv1.5 models. For DiT, we maintained the default configurations recommended by the model authors. For the Rectified Flow experiments, we used the EulerDiscrete sampler [65]. For all

TABLE I
COMPARISON WITH STATE-OF-THE-ART ALIGNMENT METHODS ON PICK-A-PIC V2 *test_unique* SPLIT. THE TABLE PROVIDES METRICS FOR HUMAN PREFERENCE ALIGNMENT AND IMAGE DIVERSITY. HUMAN PREFERENCE METRICS INCLUDE AESTHETIC, TI-CLIP, IMAGEREWARD, PICKSCORE, AND HPSV2. IMAGE DIVERSITY IS ASSESSED USING II-CLIP, VENDI SCORE, MSS, AND LPIPS. THE BASE MODEL IS SDXL. THE NUMBERS IN PARENTHESES ARE P-VALUES COMPUTED PER PROMPT.

Workflow		Human Preference Alignment					Image Diversity				Computation Cost	
Category	Method	Aesthetic ↑	TI-CLIP ↑	ImageReward ↑	PickScore ↑	HPSV2 ↑	II-CLIP ↓	Vendi ↑	MSS ↓	LPIPS ↑	Train (GPU-H)	GenTime (ms)
N/A	Base Model	5.8959	0.3595	0.5284	0.2184	0.2781	0.4822	9.6310	0.3214	0.5141	0	1127
Training-based	SFT on Chosen	5.7804	0.3576	0.5072	0.2177	0.2785	0.5241	9.0572	0.3642	0.4957	1100	1127
Training-based	DiffDPO [18]	6.0142	0.3738	0.8038	0.2244	0.2845	0.4962	9.2415	0.3819	0.4901	1405	1127
Training-based	SPO [19]	6.1712	0.3514	1.0104	0.2260	0.2863	0.5355	9.2722	0.4217	0.5086	350	1127
Training-based	MaPO [42]	6.1504	0.3621	0.7048	0.2224	0.2831	0.5009	9.2915	0.3903	0.5049	1362	1127
Training-free	DNO [24]	6.0649	0.3578	0.7834	0.2251	0.2847	0.5471	9.2592	0.4051	0.5034	0	7562
Training-free	DRS [44]	5.8432	0.3580	0.7261	0.2214	0.2804	0.5329	9.2711	0.4089	0.5067	0	4702
Training-free	RandSeed	6.1019	0.3611	0.9826	0.2256	0.2852	0.5038	9.3512	0.4051	0.5004	0	6547
Training-free	RSTFA (ours)	6.0845 (0.08)	0.3722 (0.04)	1.0322 (0.02)	0.2265 (0.05)	0.2859 (0.07)	0.4851 (0.01)	9.5708 (0.03)	0.3512 (0.01)	0.5149 (0.02)	215	1748

methods, including ours, the number of sampling timesteps was set to 50 for primary experiments, with additional tests at 30 and 20 steps to evaluate robustness to different sampling configurations. We adapted the PickScore [22] model into a timestep-aware reward model by incorporating time conditions into the vision encoder of PickScore. We applied adaptive layer normalization after each attention and feed-forward layer, following the approach in [19]. Empirically, we intervene every 10 timesteps based on the observed periodic dips as shown in Figure 2. We acknowledge that current stylistic-step identification lacks a theoretical approach and is based primarily on empirical similarity calculations. Nevertheless, we validate the heuristic both in SDXL and DiT with prompts from different sources, and show that RSTFA is relatively robust to variations in stylistic timestep interval. To address variability introduced by rejection sampling during inference, which affects individual image generation speed, we maintained consistent inference time by setting the maximum number of rejection attempts per timestep as 5. When the limit is reached, our method chooses the candidate with the highest acceptance probability. For these hyperparameters, we conducted a comprehensive ablation study. Following [42], we used the prompt format as in Figure 4 to acquire the image preference via GPT-4o.

Reward Model Training. Following [19], the timestep-independent reward model was initialized from the PickScore [22] CLIP backbone. To provide the timestep-independent reward model with data reflecting different stages of the diffusion process, we started with pair-wise clean images from the Pick-a-Pic V2 [22] dataset and added noise corresponding to various timesteps of the diffusion process to obtain noisy images. Then, using the DDIM scheme, we predicted the original samples (denoised images) from these noisy images. This process simulates the intermediate denoising steps of the diffusion model. The reward model was then trained on these predicted (denoised) images without timestep condition to predict the preference order for different noise levels. The training was conducted for 1000 steps with a batch size of 2048. The learning rate was set as 1e-6 with a cosine decay scheduler.

For the timestep-aware reward model, timestep conditioning was incorporated using adaptive layer normalization applied after each attention and feed-forward layer, following the method described in [19]. The timestep-aware reward model was then fine-tuned using the same training procedure as the timestep-independent model, but with timestep conditions included. Unless otherwise stated, the timestep-aware reward model is trained once on Pick-a-Pic V2 pairs and reused across SDXL, SDv1.5, DiT and FLUX backbones, and across Pick-a-Pic V2 test_unique, HPD V2, COCO, and PartiPrompts. No retraining or per-prompt adaptation is performed.

Compute accounting. We report a Training Compute (GPU-hours) column alongside per-image GenTime (ms) for all methods. For RSTFA, this column contains the cost of adapting PickScore into a timestep-aware reward and the small calibration of g_{ψ_t}. For training-based baselines, we report the denoiser fine-tuning cost per backbone using the official training recipes on Pick-a-Pic V2. All measurements were run on the same hardware as our inference timing to ensure consistency. For training-free baselines, they also rely on a shared external reward scorer (e.g., PickScore/ImageReward). As no additional adaptation of the reward model is needed for these baselines, we report the additional training cost as 0.

C. Comparison with the State-of-the-art Methods

In this section, we compared our proposed method, RSTFA, with state-of-the-art training-based and training-free methods across multiple base models and evaluation datasets. We first present results using SDXL on the Pick-a-Pic V2 dataset,

Fig. 3. Qualitative comparison of baselines and our method. Prompts are from Pick-a-Pic V2 *test_unique* split. Our method produces images that capture fine-grained prompt details, emphasizing vivid colors and aligning closely with human aesthetic and contextual preferences.

TABLE II
COMPARISON WITH STATE-OF-THE-ART ALIGNMENT METHODS ON PARTIPROMPTS. THE TABLE PROVIDES METRICS FOR HUMAN PREFERENCE ALIGNMENT AND IMAGE DIVERSITY. HUMAN PREFERENCE METRICS INCLUDE AESTHETIC, TI-CLIP, IMAGEREWARD, PICKSCORE, AND HPSV2. IMAGE DIVERSITY IS ASSESSED USING II-CLIP, VENDI SCORE, MSS, AND LPIPS. THE BASE MODEL IS SDXL. THE NUMBERS IN PARENTHESES ARE P-VALUES COMPUTED PER PROMPT.

Workflow		Human Preference Alignment					Image Diversity				Computation Cost	
Category	Method	Aesthetic ↑	TI-CLIP ↑	ImageReward ↑	PickScore ↑	HPSV2 ↑	II-CLIP ↓	Vendi ↑	MSS ↓	LPIPS ↑	Train (GPU-H)	GenTime (ms)
N/A	Base Model	5.7127	0.3512	0.5314	0.2189	0.2784	0.4815	13.0345	0.3098	0.5280	0	1127
Training-based	SFT on Chosen	5.7921	0.3591	0.5093	0.2183	0.2788	0.5228	12.4691	0.3440	0.5070	1100	1127
Training-based	DiffDPO [18]	5.8285	0.3651	0.8115	0.2248	0.2849	0.4948	12.4532	0.3292	0.5131	1405	1127
Training-based	SPO [19]	5.8743	0.3628	1.0187	0.2234	0.2867	0.5342	12.2855	0.3170	0.5116	350	1127
Training-based	MaPO [42]	5.8621	0.3634	0.7125	0.2228	0.2835	0.4996	12.5041	0.3302	0.5190	1362	1127
Training-free	DNO [24]	5.8782	0.3592	1.0121	0.2255	0.2851	0.5458	12.2718	0.3257	0.5140	0	7562
Training-free	DRS [44]	5.7564	0.3595	0.7345	0.2218	0.2808	0.5316	12.2834	0.3376	0.5088	0	4702
Training-free	RandSeed	5.8552	0.3625	0.9913	0.2250	0.2856	0.5025	12.3645	0.3315	0.5090	0	6547
Training-free	RSTFA (ours)	5.9578 (0.04)	0.3735 (0.01)	1.1408 (0.01)	0.2269 (0.05)	0.2893 (0.04)	0.4838 (0.01)	12.7841 (0.05)	0.3101 (0.01)	0.5261 (0.03)	215	1748

followed by cross-architecture and cross-dataset evaluations to demonstrate the generality of our approach.

Quantitative Analysis. We demonstrated that RSTFA achieved superior or comparable results across various metrics of human preference alignment and image diversity. As shown in Table I, for prompt-aware human preference alignment, our approach obtained improvement of 5.92% in TI-CLIP compared with the SOTA training-based method SPO. For prompt-independent human preference alignment, our method achieved 3.20% aesthetic score improvement over base model and comparable aesthetic score compared with method SPO and DNO. We additionally evaluate on the PartiPrompts to reduce any perceived dependence on Pick-a-Pic. We also included comprehensive evaluations on HPD V2 and COCO using different base models. The full evaluation results can be found in Table II, Table VII and Table X.

Computation Cost. We report both per-image inference time (GenTime) and training GPU-hours for all methods. Compared with training-based methods, the training cost of the timestep-aware reward model is smaller than the training-

GPT-4o Judgement Prompt

Please select the Image (a) or (b) you prefer given the prompt. You should select the preferred image based on prompt, image composition, visual appeal and your own preference. Your answer should ONLY contain: Image (a) or Image (b).

Prompt:
{prompt}

Image (a):
The first image attached.

Image (b):
The second image attached.

Fig. 4. The GPT-4o prompt used to evaluate preference between two images based on a given instruction. The model was instructed to choose either Image (a) or Image (b).

based methods. Importantly, this reward model is trained once and then amortized across all backbones and all datasets. This one-time model is then used, without any further training or adaptation, for SDXL, SD v1.5, and DiT, and for different prompt sources: Pick-a-Pic V2 test_unique, HPD V2, COCO and PartiPrompts. In contrast, training-based alignment methods run model fine-tuning per backbone and often per setting. For inference cost, our method had 5814 ms less inference time per image compared with DNO. The quantitative results indicate that RSTFA effectively improves alignment with human preferences and produces diverse images without fine-tuning and heavy inference overhead. Metropolis-Hastings interventions are confined to stylistic timesteps identified by our trajectory analysis, the number of additional U-Net evaluations is both bounded and small, yielding an average runtime overhead of $\sim$1.6$\times$ relative to base sampling (versus $\sim$6.7$\times$ for DNO). This makes RSTFA Pareto-efficient in the quality–latency trade-off among training-free alignment methods.

Qualitative Analysis. We conducted a user study to qualitatively compare our method, RSTFA, with the state-of-the-art training-based method, SPO. The study involved human labelers evaluating images generated under criteria: General Preference (*Which image do you prefer given the prompt?*). The images were generated conditioned on 30 randomly sampled prompts from HPSV2 benchmark [21]. We requested 25 users to complete the survey. We also used GPT-4o as an evaluator to judge the pairwise images between SPO and RSTFA. As shown in Figure 6, RSTFA achieved a higher win rate than SPO judged by both humans and GPT-4o [66]. We present some image generation samples in Figure 3, with additional results provided in the Appendix D. As in Figure 3, our method generated more appealing images with vivid colors, lighting, better composition and details.

D. Additional Results

Generality to Different Base Model and Evaluation Prompts. We applied RSTFA to the DiT-based model (DiT) [52]. The observed improvements in Table III confirm that RSTFA generalizes well to non-U-Net architectures. We also tested using a weaker model SDv1.5, as the losing images in preference dataset Pick-a-Pic V2 are generated by similarly performing and mostly weaker models. The weaker backbone can also benefit from our method. This demonstrates that the superficial alignment is not merely an artifact of aligning with one specific preference dataset or base model. It is a general artifact of the training-based alignment methods.

The training of the reward model relies on Pick-a-Pic V2. To test the effectiveness of RSTFA facing new scenarios that differ from the training data distribution, we also adopted HPD V2 and COCO prompts [67] for evaluation. The results (Table III) confirm generalization. Our method achieves consistent improvements under different prompt source and backbone models. The full evaluation results on COCO and HPD V2 can be found in Table VII and Table X.

TABLE III
PERFORMANCE OF RSTFA ON HPD V2 AND COCO PROMPTS ACROSS DIFFERENT MODEL ARCHITECTURES. METRICS INCLUDE HUMAN PREFERENCE ALIGNMENT (TI-CLIP, HPSV2, AESTHETIC) AND IMAGE DIVERSITY (II-CLIP).

	Human Preference Alignment			**Image Diversity**
Method	TI-CLIP ↑	HPSv2 ↑	Aesthetic ↑	II-CLIP ↓
SDXL Base	0.3481	0.2821	5.8963	0.4910
SPO	0.3550	0.2840	5.9800	0.4870
RSTFA (ours)	**0.3615**	**0.2854**	**6.0425**	**0.4820**
SDv1.5 Base	0.3408	0.2607	5.6567	0.5039
SPO	0.3477	0.2777	5.7378	0.5576
RSTFA (ours)	**0.3532**	**0.2807**	**5.9167**	**0.5009**
DiT Base	0.3309	0.2811	6.0267	0.5121
SPO	0.3350	0.2850	6.1400	0.5040
RSTFA (ours)	**0.3389**	**0.2890**	**6.2341**	**0.4951**

Application on Rectified Flow Models with Euler Samplers. To further demonstrate the versatility of our proposed RSTFA method, we applied it to Rectified Flow models [68], a class of generative models that rely on flow matching techniques [69]. In applying RSTFA to Rectified Flow models, we focused on incorporating rejection sampling at the initial sampling stage. We adopted the FLUX.1 [dev] [70] as the backbone model. We adopted EulerDiscrete sampler [65] with classifier-free guidance of 3.5. As in Table IV, compared with base model, RSTFA significantly enhanced the alignment performance of Rectified Flow models, achieving 1.50%, 0.40% and 1.24% improvements in TI-CLIP, PickScore and Aesthetic score respectively. These gains were achieved while maintaining the deterministic nature of flow-based generation, highlighting RSTFA's ability to adapt across different generative models.

However, compared with DiT or SDXL, the performance gains on FLUX.1 [dev] are smaller and closer to RandSeed baseline. This behavior is expected as FLUX.1 [dev] is an open-weight, guidance-distilled student directly distilled from FLUX.1 [pro] (which is not public). Both classifier-free guidance and its distillation variants are known to suffer from

TABLE IV
PERFORMANCE METRICS OF RSTFA APPLIED TO RECTIFIED FLOW MODELS ON PICK-A-PIC V2 *test_unique* SPLIT.

	Human Preference Alignment					Image Diversity				Computation Cost	
Method	Aesthetic ↑	TI-CLIP ↑	ImageReward ↑	PickScore ↑	HPSV2 ↑	II-CLIP ↓	Vendi ↑	MSS ↓	LPIPS ↑	Train (GPU-H)	GenTime (ms)
Base Model	6.2267	0.3339	1.0426	0.2259	0.2881	0.4875	9.6833	0.3529	0.5095	0	1689
RandSeed	6.2837	0.3357	1.0673	0.2272	0.2895	0.4892	9.1891	0.3658	0.5096	0	8816
RSTFA (ours)	6.3041	0.3389	1.0862	0.2268	0.2920	0.4878	9.1949	0.3544	0.5050	215	2260
RSTFA w. jitter	6.3307	0.3507	1.2002	0.2315	0.2973	0.4905	9.1231	0.3522	0.5022	215	6745

TABLE V
ABLATION STUDIES OF REJECTION SAMPLING TIMESTEPS AND REWARD MODEL CONFIGURATION ON PICK-A-PIC V2 *validation_unique* SPLIT.

	Methods	PickScore ↑	GenTime (ms)
	Base Model	0.2217	1127
(a)	stylistic+timestep-aware	0.2265	1748
(b)	stylistic+timestep-independent	0.2231	1748
(c)	all+timestep-aware	0.2269	4585
(d)	all+timestep-independent	0.2228	4585

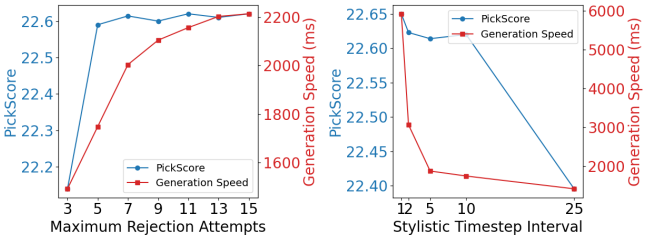

Fig. 5. Ablation studies on Pick-a-Pic V2 *validation_unique* split. (a) The effect of maximum rejection attempts per timestep; (b) the effect of stylistic timestep interval.

reduced sample diversity. This reduces the improvement room for MH corrections at fixed compute, thereby limiting the gap to RandSeed. Even in this constrained-diversity regime, RSTFA remains the more compute-efficient route to exploit available diversity, matching or slightly exceeding RandSeed's preference metrics at roughly 3–4× lower latency on the same hardware. To restore much of RSTFA's advantage without retraining, we propose stochastic guidance sampling. When generating candidate proposals, we sample the classifier-free guidance scale uniformly from $[\gamma-1, \gamma+1]$ around the default γ for each candidate. We then run multiple-try MH at stylistic timesteps. On FLUX.1 [dev], stochastic guidance sampling yields consistent improvements in PickScore, HPSv2, and TI-CLIP, demonstrating that modest inference-time diversity boosts amplify RSTFA's gains.

E. Ablation Study

SFT on Chosen Images. We performed Supervised Fine-Tuning (SFT) on the base diffusion model using only the chosen images from the human preference dataset. As shown in Table I, we found that naive fine-tuning on chosen images did not effectively capture the complex nuances of human preferences. The alignment metrics did not show improvement over the base model. This suggests that human preferences cannot be effectively captured through simple fine-tuning on preferred images alone. As a proof of concept, this highlights the necessity of approaches like RSTFA that consider the alignment process in a more nuanced manner.

Effect of Rejection Sampling Steps. We investigated the impact of performing rejection sampling at stylistic timesteps versus all timesteps. As shown in Table V, focusing on critical timesteps significantly reduced generation computation from 4585 ms to 1748 ms while maintaining alignment quality, achieving 2.17% improvement of PickScore compared with base model. Performing rejection sampling at all timesteps offered marginal 0.18% improvements but at a much higher latency of 4585 ms. This suggests that addressing the discrepancies at these stylistic timesteps is sufficient for achieving superior human preference alignment. RSTFA streamlines the overall computational burden and allows for more targeted adjustments to the generated content.

Effect of Timestep-aware Reward Model. We conducted an ablation study to evaluate the contribution of our timestep-aware reward model compared with a timestep-independent reward model. For timestep-independent reward model, we removed the timestep condition and additional adaptive layer normalization. In Table V, we observed that using the timestep-aware reward model outperformed using the timestep-independent reward model, as evidenced by 1.52% higher PickScore. This demonstrates the importance of timestep-aware module to accurately estimate the acceptance probability for our rejection sampling scheme.

Maximum Rejection Attempts. We conducted an ablation study to examine the effect of varying the maximum rejection attempts per timestep in our method. We tested values in the set {3, 5, 7, 9, 11, 13, 15} using the Pick-a-Pic V2 *validation_unique* split. As shown in Figure 5, we observed a trade-off between improvements in PickScore and decrease in generation speed. We found that above 5 rejection attempts, the improvements in PickScore were marginal. Therefore, we chose 5 rejection attempts for our main experiment.

Stylistic Timestep Interval. We performed an ablation study to evaluate the impact of varying stylistic timestep intervals. Specifically, we evaluated intervals of {1, 2, 5, 10, 25} timesteps. As shown in Figure 5, decreasing the interval improves alignment performance, as measured by PickScore, but significantly increases generation time. The improvement

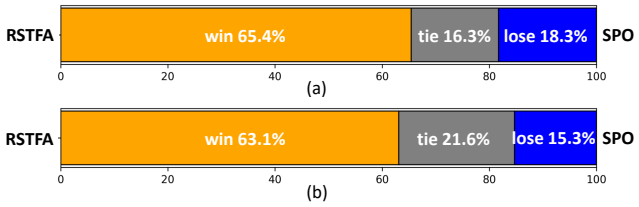

Fig. 6. User study results comparing SPO and our method RSTFA. We follow [18] and generate images conditioned on prompts from the HPSV2 benchmark. (a) Real Human Users as evaluator; (b) GPT-4o as evaluator.

in PickScore from an interval of 10 to 1 is marginal, while the associated increase in generation time is substantial. Based on this analysis, we selected an interval of 10 timesteps for our main experiments to balance performance and efficiency.

Robustness to Generation Steps. To evaluate the practical applicability of RSTFA in scenarios where computational efficiency is prioritized, we investigated its performance across different sampling step configurations. Table VI presents the PickScore results for the base SDXL model and our RSTFA method using 50, 30, and 20 sampling steps. The consistent performance gap between RSTFA and the base model across all step configurations demonstrates that our method maintains its alignment advantages even with reduced sampling steps. Notably, RSTFA with only 20 steps achieves a PickScore improvement of 0.0067 over the base model, indicating that the benefits of our approach persist in computationally constrained settings. This robustness to different step counts further supports the practical value of our method, as it can be adapted to various inference time requirements without sacrificing its core alignment benefits.

TABLE VI
PICKSCORE PERFORMANCE WITH DIFFERENT NUMBERS OF SAMPLING STEPS USING SDXL.

# Steps	Base Model	RSTFA (Ours)	Δ
50	0.2184	**0.2265**	+0.0081
30	0.2132	**0.2217**	+0.0085
20	0.2083	**0.2150**	+0.0067

VII. DISCUSSIONS

In this section, we examine the broader implications and generality of our findings on superficial alignment in text-to-image diffusion models. The superficial alignment hypothesis, that alignment tuning primarily affects stylistic aspects rather than fundamental content, represents an important insight into the behavior of diffusion models during human preference alignment. Our analysis confirms this hypothesis across different model architectures, prompt sources, and alignment techniques, suggesting it is a general characteristic of current alignment approaches rather than an artifact specific to our experimental setup. This section explores various dimensions of this hypothesis, providing additional evidence and examining its implications for alignment strategies.

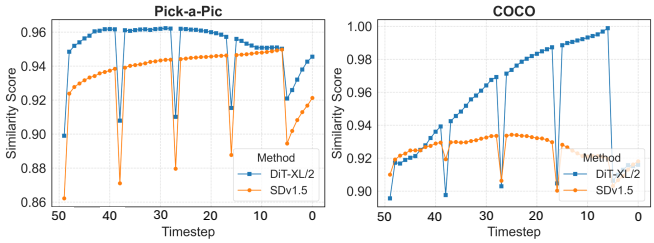

Fig. 7. Similarity score at each timestep between base and aligned models for different architectures (SDv1.5 and DiT-XL/2) and prompt sources (Pick-a-Pic and COCO).

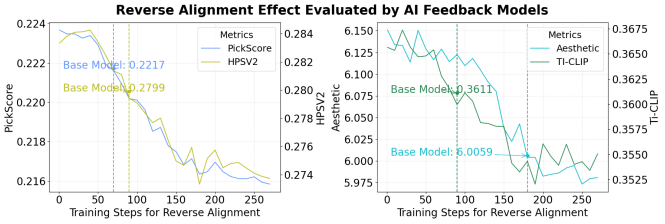

Fig. 8. The figure shows the progression of human preference alignment (PickScore, HPSV2, TI-CLIP) and image aesthetic (Aesthetic) during reverse alignment process.

Generality of Stylistic Alignment Findings. We repeated the denoising trajectory analysis from Section IV using SDv1.5 and DiT-XL/2 (DiT-based model) on Pick-a-Pic and COCO prompts (Figure 7). In both models, we observed the same pattern: high cross-model similarity across most timesteps, with periodic drops at stylistic points.

Feature-wise Sampling Impact. We decomposed the feature similarity between the base and aligned models into three components: global (CLIP), low-level (VGG), and structural (SAM). As shown in Table VIII, all components exhibit small similarity drops, but the differences are more pronounced in style-related features (VGG) compared to the structure-preserving SAM features. This supports our claim that alignment mainly alters stylistic attributes while preserving core image structure.

Alignment Loss (Effect of β). We vary $\beta \in \{50, 500, 5000\}$ in the DiffDPO loss using SDXL. We measured the style-only change rate, defined as the proportion of timesteps where the CLIP similarity between base and aligned features exceeds 0.85. In Table IX, we observe that even with a small β, the style-only rate remains high. These findings suggest that diffusion alignment remains largely superficial even with a lower KL penalty.

Reverse Alignment Mechanism. To further explore the superficial nature of alignment, we introduce a reverse alignment mechanism. We consider image pairs $(x_0^{\mathrm{w}}, x_0^{\ell})$ under a prompt c, where x_0^{w} is preferred over x_0^{ℓ}. For the reverse alignment, we swap preferences, treating x_0^{ℓ} as preferred. Starting from the DiffDPO [18] aligned model, which was fine-tuned using 850K image pairs over 2000 steps, we retrain it using the reversed image pairs with the DiffDPO objective. We evaluate the impact of reverse alignment by assessing image aesthetics, visual appeal, and image-text alignment using AI feedback models (detailed in Section VI-A).

TABLE VII
COMPARISON WITH STATE-OF-THE-ART ALIGNMENT METHODS ON COCO PROMPTS. THE TABLE PROVIDES METRICS FOR HUMAN PREFERENCE ALIGNMENT AND IMAGE DIVERSITY. HUMAN PREFERENCE METRICS INCLUDE AESTHETIC, TI-CLIP, IMAGEREWARD, PICKSCORE, AND HPSV2. IMAGE DIVERSITY IS ASSESSED USING II-CLIP, VENDI SCORE, MSS, AND LPIPS. THE BASE MODEL IS SDV1.5. THE NUMBERS IN PARENTHESES ARE P-VALUES COMPUTED PER PROMPT.

Workflow		Human Preference Alignment					Image Diversity				Computation Cost	
Category	Method	Aesthetic ↑	TI-CLIP ↑	ImageReward ↑	PickScore ↑	HPSV2 ↑	II-CLIP ↓	Vendi ↑	MSS ↓	LPIPS ↑	Train (GPU-H)	GenTime (ms)
N/A	Base Model	5.5500	0.3400	0.3421	0.2172	0.2605	0.5050	9.9821	0.4489	0.3194	0	1089
Training-based	SFT on Chosen	5.5234	0.3378	0.3215	0.2165	0.2809	0.5372	9.4218	0.4715	0.3101	915	1089
Training-based	DiffDPO [18]	5.6512	0.3542	0.6247	0.2231	0.2868	0.5198	9.3987	0.4691	0.3147	976	1089
Training-based	SPO [19]	5.6300	0.3470	0.6421	0.2243	0.2825	0.5587	9.3254	0.4680	0.3032	310	1089
Training-based	MaPO [42]	5.6187	0.3456	0.6234	0.2211	0.2819	0.5241	9.3562	0.4778	0.3095	1055	1089
Training-free	DNO [24]	5.6891	0.3425	0.6032	0.2238	0.2832	0.5694	9.4135	0.4624	0.3080	0	7321
Training-free	DRS [44]	5.6721	0.3412	0.5452	0.2201	0.2791	0.5561	9.3327	0.4862	0.3013	0	4558
Training-free	RandSeed	5.6524	0.3448	0.6502	0.2245	0.2837	0.5271	9.4178	0.4824	0.3144	0	6334
Training-free	RSTFA (ours)	5.8100 (0.02)	0.3525 (0.04)	0.6634 (0.02)	0.2252 (0.04)	0.2865 (0.06)	0.4920 (0.02)	9.7412 (0.04)	0.4587 (0.02)	0.3187 (0.08)	215	1692

TABLE VIII
FEATURE-LEVEL SIMILARITY SHIFT (MEAN DROP) BETWEEN BASE AND ALIGNED MODELS ACROSS DIFFERENT FEATURE REPRESENTATIONS.

Feature	Base vs Aligned-DPO	Base vs SPO
CLIP (Global)	0.085	0.088
VGG (Low-level)	0.062	0.065
SAM (Structure)	0.017	0.013

TABLE IX
EFFECT OF β IN DIFFDPO ALIGNMENT (SDXL, 50 STEPS)

β	PickScore ↑	CLIP Similarity ↑	Style-Only Rate ↑
50	0.2037	0.904	67.2%
500	0.2120	0.922	69.5%
5000	**0.2244**	**0.948**	**75.4%**

As shown in Figure 8, the reverse alignment leads to a significant degradation in image aesthetics, visual appeal, and image-text alignment. Notably, performing reverse alignment for only 200 training steps, a small fraction of the original 2000 steps training, already leads to a substantial drop in performance relative to both the aligned and base models. The ease with which alignment effects are completely reversed with only 200 training steps further highlights the superficiality of the alignment process.

VIII. CONCLUSION

In conclusion, our work analyzes the effect of alignment tuning in text-to-image diffusion models, uncovering superficial alignment behaviors that are consistent across multiple model architectures (SDXL, SDv1.5, DiT-XL/2, and Rectified Flow models) and prompt sources (Pick-a-Pic V2, HPD V2, and COCO). Based on this general finding, we present an efficient training-free approach to align text-to-image diffusion models via rejection sampling technique, called RSTFA. By leveraging rejection sampling at stylistic timesteps, we effectively balance alignment quality with computational efficiency, avoiding common pitfalls in previous methods like extensive fine-tuning, high inference cost, and mode collapse. Our extensive experimental results highlight that this method achieves state-of-the-art performance while reducing computational overhead.

APPENDIX

A. TV Distance Guarantee

This section establishes the theoretical foundations of RSTFA, providing guarantees on bias control while identifying critical conditions for effective alignment. We first formalize the bias bound that ensures proximity to the ideal aligned distribution. Then, we specify practical conditions that ensure the approach's efficacy across diverse diffusion architectures.

The core theoretical innovation of RSTFA lies in its training-free alignment via rejection sampling at strategically identified stylistic timesteps. This approach requires estimating the density ratio $\rho_t(x) = \frac{p_t^{\text{aligned}}(x)}{p_t^{\text{base}}(x)}$ between aligned and base distributions. By Bregman minimization, the solution $g^*_{\psi_t}$ satisfies $e^{g^*_{\psi_t}(x)} \propto \frac{p_t^{\text{aligned}}(x)}{p_t^{\text{base}}(x)}$ [49], [71], effectively recovering the log-density ratio up to an additive constant. Crucially, when the logarithmic error in the calibrated ratio estimator $\widehat{\rho}_t$ exhibits sub-Gaussian behavior, we obtain a formal guarantee of distributional proximity.

Theorem 1 (Bias Bound for RSTFA)**.** *Let $\widehat{\rho}_t$ be the calibrated density ratio estimator with logarithmic error $\varepsilon_t(x) = \log \widehat{\rho}_t(x) - \log \rho_t(x)$. If ε_t is sub-Gaussian with variance proxy σ_t^2 under p_t^{base} and q_t, then the total variation distance between*

TABLE X
COMPARISON WITH STATE-OF-THE-ART ALIGNMENT METHODS ON HPD V2. THE TABLE PROVIDES METRICS FOR HUMAN PREFERENCE ALIGNMENT AND IMAGE DIVERSITY. HUMAN PREFERENCE METRICS INCLUDE AESTHETIC, TI-CLIP, IMAGEREWARD, PICKSCORE, AND HPSV2. IMAGE DIVERSITY IS ASSESSED USING II-CLIP, VENDI SCORE, MSS, AND LPIPS. THE BASE MODEL IS SDV1.5. THE NUMBERS IN PARENTHESES ARE P-VALUES COMPUTED PER PROMPT.

Workflow		Human Preference Alignment					Image Diversity				Computation Cost	
Category	Method	Aesthetic ↑	TI-CLIP ↑	ImageReward ↑	PickScore ↑	HPSV2 ↑	II-CLIP ↓	Vendi ↑	MSS ↓	LPIPS ↑	Train (GPU-H)	GenTime (ms)
N/A	Base Model	5.7634	0.3415	0.2389	0.2076	0.2609	0.5028	7.9954	0.5486	0.2811	0	1089
Training-based	SFT on Chosen	5.5367	0.3392	0.2087	0.2169	0.2613	0.5349	7.4351	0.5698	0.2818	915	1089
Training-based	DiffDPO [18]	5.8645	0.3556	0.6195	0.2135	0.2772	0.5176	7.4121	0.5874	0.2764	976	1089
Training-based	SPO [19]	5.8456	0.3484	0.6358	0.2147	0.2729	0.5565	7.6387	0.5784	0.2749	310	1089
Training-based	MaPO [42]	5.8321	0.3471	0.6189	0.2115	0.2723	0.5218	7.4695	0.5961	0.2712	1055	1089
Training-free	DNO [24]	5.9024	0.3439	0.6241	0.2142	0.2736	0.5671	7.4268	0.5907	0.2797	0	7321
Training-free	DRS [44]	5.8854	0.3426	0.5408	0.2105	0.2695	0.5538	7.5461	0.5945	0.2730	0	4558
Training-free	RandSeed	5.8657	0.3462	0.6056	0.2149	0.2741	0.5248	7.6311	0.5807	0.2761	0	6334
Training-free	RSTFA (ours)	6.0234 (0.04)	0.3539 (0.04)	0.6340 (0.01)	0.2156 (0.05)	0.2749 (0.04)	0.5097 (0.01)	7.7545 (0.07)	0.5570 (0.04)	0.2804 (0.06)	215	1692

the RSTFA-generated distribution p_0^{accept} and the ideal aligned distribution $p_0^{aligned}$ satisfies:

$$\mathrm{TV}\left(p_0^{accept}, p_0^{aligned}\right) \le \sum_{t\in\mathcal{T}_{stylistic}} \frac{1}{2}\sigma_t \tag{11}$$

Proof. At each $t \in \mathcal{T}_{\text{stylistic}}$, define:

$$q_t(x) = \frac{\rho_t(x)p_t^{\text{base}}(x)}{Z_2}, \quad Z_2 = \int \rho_t dp_t^{\text{base}}$$
$$\widehat{q}_t(x) = \frac{\widehat{\rho}_t(x)p_t^{\text{base}}(x)}{Z_1}, \quad Z_1 = \int \widehat{\rho}_t dp_t^{\text{base}}$$

The KL divergence is bounded using the sub-Gaussian property:

$$\mathrm{KL}(\widehat{q}_t \parallel q_t) = \mathrm{E}_{\widehat{q}_t}\left[\log\frac{\widehat{q}_t}{q_t}\right] = \mathrm{E}_{\widehat{q}_t}\left[\varepsilon_t - \log\frac{Z_1}{Z_2}\right] \tag{12}$$

By the Donsker-Varadhan inequality and sub-Gaussianity:

$$\mathrm{KL}(\widehat{q}_t \parallel q_t) \le \log \mathrm{E}_{q_t}\left[e^{\varepsilon_t}\right] \le \log\left(e^{\frac{1}{2}\sigma_t^2}\right) = \frac{1}{2}\sigma_t^2 \tag{13}$$

Applying Pinsker's inequality:

$$\mathrm{TV}(\widehat{q}_t, q_t) \le \sqrt{\frac{1}{2}\mathrm{KL}(\widehat{q}_t \parallel q_t)} \tag{14}$$
$$\le \sqrt{\frac{1}{2}\cdot\frac{1}{2}\sigma_t^2} \tag{15}$$
$$= \frac{\sigma_t}{2} \tag{16}$$

Let $K_{t-1}, \ldots, K_0$ be the ideal diffusion kernels for subsequent steps. By the data-processing inequality for Markov kernels [72]:

$$\mathrm{TV}\left(\widehat{q}_t \cdot K_{t-1}\cdots K_0,\ q_t \cdot K_{t-1}\cdots K_0\right) \le \mathrm{TV}(\widehat{q}_t, q_t) \tag{17}$$
$$\le \frac{\sigma_t}{2} \tag{18}$$

This bounds the TV at x_0 introduced at step t.

By the triangle inequality for TV distance:

$$\mathrm{TV}\left(p_0^{\text{accept}}, p_0^{\text{aligned}}\right) \tag{19}$$
$$\le \sum_{t\in\mathcal{T}_{\text{stylistic}}} \mathrm{TV}\left(\widehat{q}_t \cdot K_{t-1}\cdots K_0,\ q_t \cdot K_{t-1}\cdots K_0\right) \tag{20}$$
$$\le \sum_{t\in\mathcal{T}_{\text{stylistic}}} \frac{\sigma_t}{2} \tag{21}$$

□

B. Extended Discussion on Periodic Stylistic Timesteps

In our main experiments, we observed that stylistic timesteps tend to appear periodically rather than concentrating exclusively at the earliest or latest stages of the diffusion process. Here, we provide an intuitive explanation for this phenomenon.

Let $\epsilon_\theta(\mathbf{x}_t, t)$ be a denoising network parameterized by θ. The model is aligned using the DiffDPO framework with the following loss function:

$$\mathcal{L}_{\text{DiffDPO}}(\theta) = -\mathbb{E}_{\mathbf{c},\mathbf{x}_0^w,\mathbf{x}_0^\ell,t}\left[\log\sigma\left(-\beta T\left(\Delta_t^w - \Delta_t^\ell\right)\right)\right], \tag{22}$$

where the differences in prediction errors are given by:

$$\Delta_t^w = \left\|\epsilon^w - \epsilon_\theta(\mathbf{x}_t^w, t)\right\|^2 - \left\|\epsilon^w - \epsilon_{\text{ref}}(\mathbf{x}_t^w, t)\right\|^2, \tag{23}$$
$$\Delta_t^\ell = \left\|\epsilon^\ell - \epsilon_\theta(\mathbf{x}_t^\ell, t)\right\|^2 - \left\|\epsilon^\ell - \epsilon_{\text{ref}}(\mathbf{x}_t^\ell, t)\right\|^2. \tag{24}$$

Here, $\epsilon^w, \epsilon^\ell \sim \mathcal{N}(0, I)$ are the noise samples corresponding to the winning and losing data points, respectively.

The denoising network ϵ_θ employs sinusoidal timestep embeddings $\tau(t)$, which are periodic functions of t:

$$\tau(t) = \left[\sin(\omega_1 t), \cos(\omega_1 t), \ldots, \sin(\omega_N t), \cos(\omega_N t)\right], \tag{25}$$

where ω_n are predefined frequencies.

We hypothesize that due to the interaction between these periodic timestep embeddings and the network architecture

(specifically, time-conditioned U-Nets [32], [73] or Diffusion Transformers [74]), the network develops an uneven sensitivity to parameter changes across the diffusion trajectory. Rather than a uniform distribution of gradient updates, the architecture may exhibit increased sensitivity at specific, periodically occurring timesteps $\{t_k\}$.

Under this hypothesis, the gradient magnitude of the alignment loss with respect to the model parameters would peak at these intervals:

$$\|\nabla_\theta L_{t_k}(\theta)\| \gg \|\nabla_\theta L_t(\theta)\|, \quad \forall t \notin \{t_k\}. \tag{26}$$

Consequently, this uneven gradient landscape leads to significant parameter updates primarily at the timesteps $\{t_k\}$, resulting in the noticeable, periodic differences we observe in the denoised outputs during inference.

C. Key Conditions

For our rejection sampling approach to be effective, the following conditions must hold:

Condition 1 (Stylistic Timesteps Existence). There exist specific timesteps $\mathcal{T}_{\text{stylistic}} \subset \{1, 2, \ldots, T\}$ where alignment adjustments primarily occur:

$$\text{sim}(\,\phi\big(x_{\text{clean}}(t)\big),\ \phi\big(y_{\text{clean}}(t)\big)) < \tau, \quad \forall t \in \mathcal{T}_{\text{stylistic}} \tag{27}$$

$$\text{sim}(\,\phi\big(x_{\text{clean}}(t)\big),\ \phi\big(y_{\text{clean}}(t)\big)) \geq \tau, \quad \forall t \notin \mathcal{T}_{\text{stylistic}} \tag{28}$$

where $\text{sim}(\cdot,\cdot)$ denotes similarity measure and τ is a threshold. Here τ is used to formalize that a subset of timesteps exhibits noticeably stronger divergence between the base and aligned models than the remainder of the trajectory. In our implementation, RSTFA does not depend on finely tuning τ. Instead, we select intervention timesteps using a compute-budget rule with a fixed intervention interval. Empirically, the similarity curves exhibit pronounced periodic dips, so moderate changes of τ within a reasonable range identify a near-identical set of timesteps.

Condition 2 (Base Model Diversity). The base model $p_{\text{base}}(x_0|c)$ generates sufficiently diverse outputs such that preference-aligned samples exist within its natural distribution.

Condition 3 (Reward Model Accuracy). The timestep-aware reward model $r(x_t, t, c)$ provides reliable estimates of the density ratio:

$$\frac{p_t^{\text{aligned}}(x_t|c)}{p_t^{\text{base}}(x_t|c)} \approx f(r(x_t, t, c)) \tag{29}$$

where $f(\cdot)$ is a transformation derived from the Bregman Divergence minimization.

These conditions ensure that rejection sampling at stylistic timesteps can effectively guide the generation process toward human-preferred outputs while maintaining computational efficiency. Our extensive experiments across different backbone models (SDXL, SDv1.5, DiT), different prompt sources (Pick-a-Pic V2, HPD V2, COCO), and various hyperparameters demonstrate that these conditions generally hold in practical text-to-image diffusion models, providing empirical support for the theoretical foundation of our approach.

In practice, the realized gain at a fixed proposal budget depends on the diversity of the base proposals at the intervention timesteps. In particular, by Condition 2 (Base Model Diversity), the proposal distribution places non-trivial mass on regions favored by the aligned density. Properties of backbones, like distillation under strong guidance, can yield more concentrated proposals, which may lower acceptance gains. We therefore view distilled backbones as a challenging regime where proposal diversification becomes important. Our stochastic guidance sampling results on FLUX.1 [dev] (Table IV) show that modest inference-time diversification can further increase RSTFA's effectiveness without any denoiser fine-tuning.

D. Additional Results

We provide additional examples generated by our method and the baseline method SPO [19]. The prompt for each image was sampled from the Pick-a-Pic V2 *test_unique* split. We also requested 20 users to acquire human preferences for these generations. As shown in Figure 9, images generated by our method could achieve a high win rate, above 50% for almost all prompts.

E. Baselines

All baseline implementations followed the official settings to ensure fair comparison. DiffDPO and MaPO methods involved fine-tuning the SDXL model using the full Pick-a-Pic V2 training set, which consists of over 850K win-lose image pairs. Adafactor [75] was adopted as the optimizer. The training was conducted with a total batch size of 2048. The learning rate was set as $\frac{2000}{\beta} \cdot 2.048 \cdot 10^{-8}$ with 25% linear warmup for 2000 training steps. β was the strength of KL regularization and set as 10. In contrast, SPO was fine-tuned using a randomly sampled subset of 4K prompts from Pick-a-Pic V2, and win-lose pairs were generated online for these prompts. LoRA [76] fine-tuning was conducted for 10 epochs. The LoRA rank was set to 64.

REFERENCES

[1] J. Chen, C. Ge, E. Xie, Y. Wu, L. Yao, X. Ren, Z. Wang, P. Luo, H. Lu, and Z. Li, "Pixart-\sigma: Weak-to-strong training of diffusion transformer for 4k text-to-image generation," *arXiv preprint arXiv:2403.04692*, 2024.

[2] Z. Evans, C. Carr, J. Taylor, S. H. Hawley, and J. Pons, "Fast timing-conditioned latent audio diffusion," *arXiv preprint arXiv:2402.04825*, 2024.

[3] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," *Advances in neural information processing systems*, vol. 34, pp. 8780–8794, 2021.

[4] Y. Hao, Z. Chi, L. Dong, and F. Wei, "Optimizing prompts for text-to-image generation," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[5] A. Hertz, A. Voynov, S. Fruchter, and D. Cohen-Or, "Style aligned image generation via shared attention," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 4775–4785.

[6] A. Schneuing, C. Harris, Y. Du, K. Didi, A. Jamasb, I. Igashov, W. Du, C. Gomes, T. L. Blundell, P. Lio *et al.*, "Structure-based drug design with equivariant diffusion models," *Nature Computational Science*, vol. 4, no. 12, pp. 899–909, 2024.

[7] R. Po, W. Yifan, V. Golyanik, K. Aberman, J. T. Barron, A. Bermano, E. Chan, T. Dekel, A. Holynski, A. Kanazawa *et al.*, "State of the art on diffusion models for visual computing," in *Computer Graphics Forum*, vol. 43, no. 2. Wiley Online Library, 2024, p. e15063.

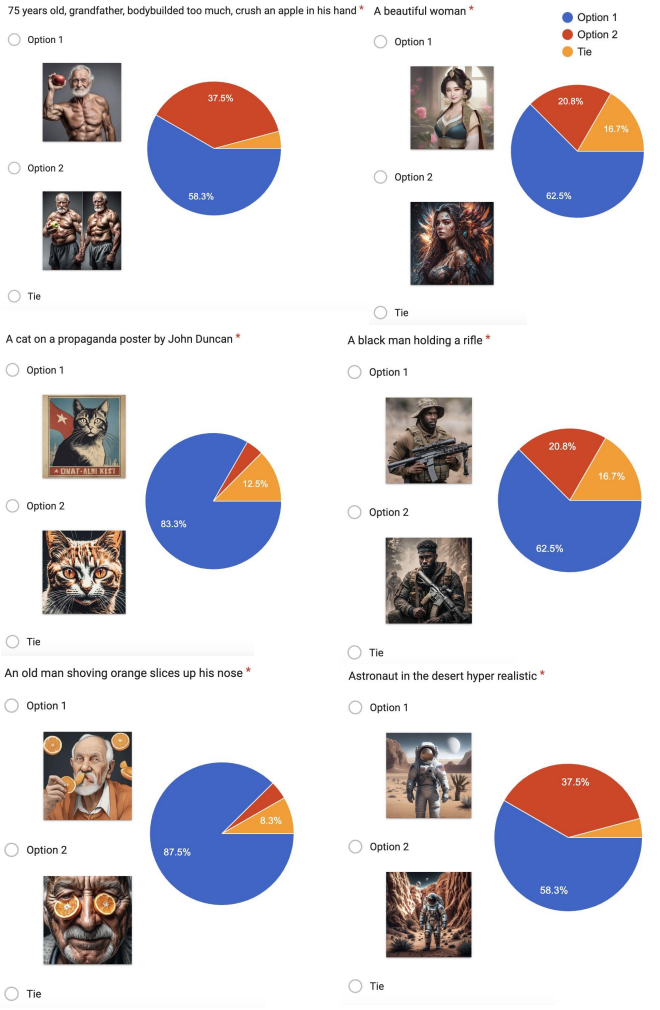

Fig. 9. Qualitative comparison of RSTFA and SPO. Option 1 was generated by RSTFA. Option 2 was generated by SPO.

[8] Y. Shen, M. Xu, and W. Liang, "Context-aware head-and-eye motion generation with diffusion model," in *2024 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*. IEEE, 2024, pp. 157–167.

[9] J. Li, S. Liu, Z. Liu, Y. Wang, K. Zheng, J. Xu, J. Li, and J. Zhu, "Instructpix2nerf: instructed 3d portrait editing from a single image," *arXiv preprint arXiv:2311.02826*, 2023.

[10] A. Hertz, R. Mokady, J. Tenenbaum, K. Aberman, Y. Pritch, and D. Cohen-Or, "Prompt-to-prompt image editing with cross attention control," *arXiv preprint arXiv:2208.01626*, 2022.

[11] K. Clark, P. Vicol, K. Swersky, and D. J. Fleet, "Directly fine-tuning diffusion models on differentiable rewards," in *The Twelfth International Conference on Learning Representations*.

[12] B. Wallace, A. Gokul, S. Ermon, and N. Naik, "End-to-end diffusion latent optimization improves classifier guidance," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 7280–7290.

[13] M. Prabhudesai, A. Goyal, D. Pathak, and K. Fragkiadaki, "Aligning text-to-image diffusion models with reward backpropagation," *arXiv preprint arXiv:2310.03739*, 2023.

[14] P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei, "Deep reinforcement learning from human preferences," *Advances in neural information processing systems*, vol. 30, 2017.

[15] R. Rafailov, A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, and C. Finn, "Direct preference optimization: Your language model is secretly a reward model," *Advances in Neural Information Processing Systems*, vol. 36, pp. 53 728–53 741, 2023.

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

[17] Y. Bai, S. Kadavath, S. Kundu, A. Askell, J. Kernion, A. Jones, A. Chen, A. Goldie, A. Mirhoseini, C. McKinnon, C. Chen, C. Olsson, C. Olah, D. Hernandez, D. Drain, D. Ganguli, D. Li, E. Tran-Johnson, E. Perez, J. Kerr, J. Mueller, J. Ladish, J. Landau, K. Ndousse, K. Lukosuite, L. Lovitt, M. Sellitto, N. Elhage, N. Schiefer, N. Mercado, N. DasSarma, R. Lasenby, R. Larson, S. Ringer, S. Johnston, S. Kravec, S. E. Showk, S. Fort, T. Lanham, T. Telleen-Lawton, T. Conerly, T. Henighan, T. Hume, S. R. Bowman, Z. Hatfield-Dodds, B. Mann, D. Amodei, N. Joseph, S. McCandlish, T. Brown, and J. Kaplan, "Constitutional ai: Harmlessness from ai feedback," 2022. [Online]. Available: https://arxiv.org/abs/2212.08073

[18] B. Wallace, M. Dang, R. Rafailov, L. Zhou, A. Lou, S. Purushwalkam, S. Ermon, C. Xiong, S. Joty, and N. Naik, "Diffusion model alignment using direct preference optimization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 8228–8238.

[19] Z. Liang, Y. Yuan, S. Gu, B. Chen, T. Hang, J. Li, and L. Zheng, "Step-aware preference optimization: Aligning preference with denoising performance at each step," *arXiv preprint arXiv:2406.04314*, 2024.

[20] K. Black, M. Janner, Y. Du, I. Kostrikov, and S. Levine, "Training diffusion models with reinforcement learning," in *The Twelfth International Conference on Learning Representations*.

[21] X. Wu, Y. Hao, K. Sun, Y. Chen, F. Zhu, R. Zhao, and H. Li, "Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis," *arXiv preprint arXiv:2306.09341*, 2023.

[22] Y. Kirstain, A. Polyak, U. Singer, S. Matiana, J. Penna, and O. Levy, "Pick-a-pic: An open dataset of user preferences for text-to-image generation," *Advances in Neural Information Processing Systems*, vol. 36, pp. 36 652–36 663, 2023.

[23] R. Barceló, C. Alcázar, and F. Tobar, "Avoiding mode collapse in diffusion models fine-tuned with reinforcement learning," *arXiv preprint arXiv:2410.08315*, 2024.

[24] Z. Tang, J. Peng, J. Tang, M. Hong, F. Wang, and T.-H. Chang, "Tuning-free alignment of diffusion models with direct noise optimization," *arXiv preprint arXiv:2405.18881*, 2024.

[25] B. Y. Lin, A. Ravichander, X. Lu, N. Dziri, M. Sclar, K. Chandu, C. Bhagavatula, and Y. Choi, "The unlocking spell on base llms: Rethinking alignment via in-context learning," in *The Twelfth International Conference on Learning Representations*, 2023.

[26] C. Zhou, P. Liu, P. Xu, S. Iyer, J. Sun, Y. Mao, X. Ma, A. Efrat, P. Yu, L. Yu *et al.*, "Lima: Less is more for alignment," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[27] X. Chen, J. Ye, C. Zu, N. Xu, R. Zheng, M. Peng, J. Zhou, T. Gui, Q. Zhang, and X. Huang, "How robust is gpt-3.5 to predecessors? a comprehensive study on language understanding tasks," *arXiv preprint arXiv:2303.00293*, 2023.

[28] M. Raghavendra, V. Nath, and S. Hendryx, "Revisiting the superficial alignment hypothesis," *arXiv preprint arXiv:2410.03717*, 2024.

[29] S. Xu, W. Fu, J. Gao, W. Ye, W. Liu, Z. Mei, G. Wang, C. Yu, and Y. Wu, "Is dpo superior to ppo for llm alignment? a comprehensive study," 2024. [Online]. Available: https://arxiv.org/abs/2404.10719

[30] J. Chen, R. Cong, Y. Zhao, H. Yang, G. Hu, H. H. S. Ip, and S. Kwong, "Sefe: Superficial and essential forgetting eliminator for multimodal continual instruction tuning," *arXiv preprint arXiv:2505.02486*, 2025.

[31] K. Xu, L. Zhang, and J. Shi, "Good seed makes a good crop: Discovering secret seeds in text-to-image diffusion models," *arXiv preprint arXiv:2405.14828*, 2024.

[32] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[33] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," in *International Conference on Learning Representations*.

[34] G. Goh, J. Betker, L. Jing, A. Ramesh, T. Brooks, J. Wang, L. Li, L. Ouyang, J. Zhuang, J. Lee, P. Dhariwal, C. Chu, J. Jiao, J. W. Kim, A. Nichol, Y. Song, L. Wang, and T. Xu, "Improving image generation with better captions," 2023.

[35] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," *arXiv preprint arXiv:2204.06125*, vol. 1, no. 2, p. 3, 2022.

[36] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation," in *International conference on machine learning*. Pmlr, 2021, pp. 8821–8831.

[37] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans

et al., "Photorealistic text-to-image diffusion models with deep language understanding," *Advances in neural information processing systems*, vol. 35, pp. 36 479–36 494, 2022.

[38] J. Ho, T. Salimans, A. Gritsenko, W. Chan, M. Norouzi, and D. J. Fleet, "Video diffusion models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 8633–8646, 2022.

[39] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," 2021.

[40] P. Esser, S. Kulal, A. Blattmann, R. Entezari, J. Müller, H. Saini, Y. Levi, D. Lorenz, A. Sauer, F. Boesel *et al.*, "Scaling rectified flow transformers for high-resolution image synthesis," in *Forty-first International Conference on Machine Learning*, 2024.

[41] A. Blattmann, R. Rombach, H. Ling, T. Dockhorn, S. W. Kim, S. Fidler, and K. Kreis, "Align your latents: High-resolution video synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 22 563–22 575.

[42] J. Hong, S. Paul, N. Lee, K. Rasul, J. Thorne, and J. Jeong, "Margin-aware preference optimization for aligning diffusion models without reference," *arXiv preprint arXiv:2406.06424*, 2024.

[43] J. Hong, N. Lee, and J. Thorne, "Orpo: Monolithic preference optimization without reference model," *arXiv preprint arXiv:2403.07691*, vol. 2, no. 4, p. 5, 2024.

[44] B. Na, Y. Kim, M. Park, D. Shin, W. Kang, and I.-C. Moon, "Diffusion rejection sampling," in *International Conference on Machine Learning*. PMLR, 2024, pp. 37 097–37 121.

[45] K. Simonyan, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[46] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4015–4026.

[47] G. Ilharco, M. Wortsman, R. Wightman, C. Gordon, N. Carlini, R. Taori, A. Dave, V. Shankar, H. Namkoong, J. Miller, H. Hajishirzi, A. Farhadi, and L. Schmidt, "Openclip," Jul. 2021. [Online]. Available: https://doi.org/10.5281/zenodo.5143773

[48] A. Menon and C. S. Ong, "Linking losses for density ratio and class-probability estimation," in *International Conference on Machine Learning*. PMLR, 2016, pp. 304–313.

[49] S. Nowozin, B. Cseke, and R. Tomioka, "f-gan: Training generative neural samplers using variational divergence minimization," *Advances in neural information processing systems*, vol. 29, 2016.

[50] J. S. Liu, F. Liang, and W. H. Wong, "The multiple-try method and local optimization in metropolis sampling," *Journal of the American Statistical Association*, vol. 95, no. 449, pp. 121–134, 2000.

[51] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach, "Sdxl: Improving latent diffusion models for high-resolution image synthesis," in *The Twelfth International Conference on Learning Representations*.

[52] H. Li, S. Lal, Z. Li, Y. Xie, Y. Wang, Y. Zou, O. Majumder, R. Manmatha, Z. Tu, S. Ermon *et al.*, "Efficient scaling of diffusion transformers for text-to-image generation," *arXiv preprint arXiv:2412.12391*, 2024.

[53] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[54] J. Xu, X. Liu, Y. Wu, Y. Tong, Q. Li, M. Ding, J. Tang, and Y. Dong, "Imagereward: Learning and evaluating human preferences for text-to-image generation," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[55] C. Schuhmann, "Laion-aesthetics," https://laion.ai/blog/laion-aesthetics/, 2022, accessed: 2023 - 11- 10.

[56] S. He, Y. Zhang, R. Xie, D. Jiang, and A. Ming, "Rethinking image aesthetics assessment: Models, datasets and benchmarks." in *IJCAI*, 2022, pp. 942–948.

[57] X. Tian, Z. Dong, K. Yang, and T. Mei, "Query-dependent aesthetic model with deep learning for photo quality assessment," *IEEE Transactions on Multimedia*, vol. 17, no. 11, pp. 2035–2048, 2015.

[58] J. Hessel, A. Holtzman, M. Forbes, R. L. Bras, and Y. Choi, "Clipscore: A reference-free evaluation metric for image captioning," *arXiv preprint arXiv:2104.08718*, 2021.

[59] S. Sadat, J. Buhmann, D. Bradley, O. Hilliges, and R. M. Weber, "Cads: Unleashing the diversity of diffusion models through condition-annealed sampling," in *The Twelfth International Conference on Learning Representations*.

[60] G. Corso, Y. Xu, V. De Bortoli, R. Barzilay, and T. S. Jaakkola, "Particle guidance: non-iid diverse sampling with diffusion models," in *The Twelfth International Conference on Learning Representations*.

[61] J. Lu, R. Teehan, and M. Ren, "Procreate, don't reproduce! propulsive energy diffusion for creative generation," in *European Conference on Computer Vision*. Springer, 2024, pp. 397–414.

[62] M. Kirchhof, J. Thornton, L. Béthune, P. Ablin, E. Ndiaye *et al.*, "Shielded diffusion: Generating novel and diverse images using sparse repellency," in *Forty-second International Conference on Machine Learning*.

[63] M. M. Morshed and V. Boddeti, "Diverseflow: Sample-efficient diverse mode coverage in flows," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 23 303–23 312.

[64] C. Lu, Y. Zhou, F. Bao, J. Chen, C. Li, and J. Zhu, "Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models," *arXiv preprint arXiv:2211.01095*, 2022.

[65] T. Karras, M. Aittala, T. Aila, and S. Laine, "Elucidating the design space of diffusion-based generative models," *Advances in neural information processing systems*, vol. 35, pp. 26 565–26 577, 2022.

[66] A. Hurst, A. Lerer, A. P. Goucher, A. Perelman, A. Ramesh, A. Clark, A. Ostrow, A. Welihinda, A. Hayes, A. Radford *et al.*, "Gpt-4o system card," *arXiv preprint arXiv:2410.21276*, 2024.

[67] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick, "Microsoft coco captions: Data collection and evaluation server," *arXiv preprint arXiv:1504.00325*, 2015.

[68] X. Liu, C. Gong *et al.*, "Flow straight and fast: Learning to generate and transfer data with rectified flow," in *The Eleventh International Conference on Learning Representations*.

[69] Y. Lipman, R. T. Chen, H. Ben-Hamu, M. Nickel, and M. Le, "Flow matching for generative modeling," in *The Eleventh International Conference on Learning Representations*.

[70] B. F. Labs, "Flux," https://blackforestlabs.ai/, 2024.

[71] M. Sugiyama, T. Suzuki, S. Nakajima, H. Kashima, P. Von Bünau, and M. Kawanabe, "Direct importance estimation for covariate shift adaptation," *Annals of the Institute of Statistical Mathematics*, vol. 60, pp. 699–746, 2008.

[72] W. R. Gilks, S. Richardson, and D. Spiegelhalter, *Markov chain Monte Carlo in practice*. CRC press, 1995.

[73] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*. Springer, 2015, pp. 234–241.

[74] W. Peebles and S. Xie, "Scalable diffusion models with transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4195–4205.

[75] N. Shazeer and M. Stern, "Adafactor: Adaptive learning rates with sublinear memory cost," in *International Conference on Machine Learning*. PMLR, 2018, pp. 4596–4604.

[76] E. J. Hu, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen *et al.*, "Lora: Low-rank adaptation of large language models," in *International Conference on Learning Representations*.

Hongzheng Yang received the B.Eng. degree in Artificial Intelligence from Beihang University in 2022. He is currently pursuing the Ph.D. degree at The Chinese University of Hong Kong. His research interests include AI alignment, and AI for social good (such as healthcare applications). He has authored 2 first-authored works and has contributed to over 10 papers in prestigious venues, including ICML, NeurIPS, TMI and Nature Communications.

Jason Chun Lok Li received the B.Eng. degree in Electrical and Electronic Engineering from University of Hong Kong in 2020, and the Ph.D. degree in Electronic Engineering from University of Hong Kong in 2024. His research interests include deep learning, CNN, efficient neural network design, compression, tensor decomposition, and mobile neural networks.

Kun Li received the B.Eng. degree in IoT engineering from the University of Electronic Science and Technology of China, Chengdu, China, in 2022. She is currently pursuing the Ph.D. degree with the Department of Electrical Engineering, City University of Hong Kong. Her research interests include deep learning and computer vision.

Wenao Ma received the B.Eng. degree in Electronic and Information Engineering from Xiamen University in 2017, and the Ph.D. degree in Computer Science and Engineering from The Chinese University of Hong Kong in 2024. His research interests include Deep Learning, Computer Vision and Medical Image Analysis. He has authored 5 first-authored works and has contributed to over 20 papers in prestigious venues, including MICCAI, TMI, ISBI and MIA.

Mingjie Xu received the B.Sc. degree in Computer Science and Mathematics from Victoria University of Wellington, New Zealand, and completed an exchange program at the University of California, San Diego. He is currently a research assistant at The Hong Kong University of Science and Technology (Guangzhou), and his research interests include multimodal large language models, vision–language understanding, scene graph generation, data-centric AI, and image/content quality assessment.

Yuzhi Zhao (S'19) received the B.Eng. degree in Electronic and Information Engineering from Huazhong University of Science and Technology (HUST) in 2018, and the Ph.D. degree in Electronic Engineering from City University of Hong Kong (CityU) in 2023.

His research interests include low-level vision, computational photography, and generative models (such as diffusion models and multimodal large language models). He has authored 10 first-authored works and has contributed to over 35 papers in prestigious venues. He also serves as a peer reviewer for international conferences and journals, including CVPR, ICCV, ECCV, SIGGRAPH, AAAI, PR, IJCV, CVIU, IEEE TMM, IEEE TCSVT, IEEE TNNLS, IEEE TIP, ACM TOG, etc.

Lai-Man Po (M'92–SM'09) received the B.S. and Ph.D. degrees in electronic engineering from the City University of Hong Kong, Hong Kong, in 1988 and 1991, respectively. He has been with the Department of Electronic Engineering, City University of Hong Kong, since 1991, where he is currently an Associate Professor of Department of Electrical Engineering. He has authored over 150 technical journal and conference papers. His research interests include image and video coding with an emphasis on deep learning based computer vision algorithms.

Dr. Po is a member of the Technical Committee on Multimedia Systems and Applications and the IEEE Circuits and Systems Society. He was the Chairperson of the IEEE Signal Processing Hong Kong Chapter in 2012 and 2013. He was an Associate Editor of HKIE Transactions in 2011 to 2013. He also served on the Organizing Committee, of the IEEE International Conference on Acoustics, Speech and Signal Processing in 2003, and the IEEE International Conference on Image Processing in 2010.