Generalizing Alignment Paradigm of Text-to-Image Generation with Preferences through f-divergence Minimization

Haoyuan Sun, Bo Xia, Yongzhe Chang*, Xueqian Wang*
Tsinghua Shenzhen International Graduate School, Tsinghua University
{sun-hy23, xiab21}@mails.tsinghua.edu.cn; {changyongzhe, wang.xq}@sz.tsinghua.edu.cn

Abstract

Direct Preference Optimization (DPO) has recently expanded its successful application from aligning large language models (LLMs) to aligning text-to-image models with human preferences, which has generated considerable interest within the community. However, we have observed that these approaches rely solely on minimizing the reverse Kullback-Leibler divergence during alignment process between the fine-tuned model and the reference model, neglecting the incorporation of other divergence constraints. In this study, we focus on extending reverse Kullback-Leibler divergence in the alignment paradigm of text-to-image models to f-divergence, which aims to garner better alignment performance as well as good generation diversity. We provide the generalized formula of the alignment paradigm under the f-divergence condition and thoroughly analyze the impact of different divergence constraints on alignment process from the perspective of gradient fields. We conduct comprehensive evaluation on image-text alignment performance, human value alignment performance and generation diversity performance under different divergence constraints, and the results indicate that alignment based on *Jensen-Shannon divergence* achieves the best trade-off among them. The option of divergence employed for aligning text-to-image models significantly impacts the trade-off between alignment performance (especially human value alignment) and generation diversity, which highlights the necessity of selecting an appropriate divergence for practical applications.

1 Introduction

Text-to-image generative models have witnessed significant advancements in recent years [1–4]. When presented with appropriate textual prompts, they are capable of generating high-fidelity images that are semantically coherent with the provided descriptions, which spans a diverse range of topics, piquing significant public interest in their potential applications and societal implications. Existing self-supervised pre-trained generators, although advanced, still exhibit imperfections, with a significant challenge being their alignment with human preferences [5].

Reinforcement Learning from Human Feedback (RLHF) has established itself as a pivotal research endeavor, demonstrating notable efficacy in aligning text-to-image models with human preferences [6–8]. Faced with the intricate challenge of defining an objective that authentically encapsulates human preferences in the realm of Reinforcement Learning from Human Feedback (RLHF), researchers conventionally assemble a dataset to mirror such preferences through comparative assessments of model-generated outputs [6, 9]. Then, a reward model is trained based on Bradley-Terry model

*Corresponding Authors

38th Workshop on Fine-Tuning in Machine Learning (NeurIPS 2024).

[10], inferring human preferences from the collected dataset. And the text-to-image model is fine-tuned with a reinforcement learning (RL) pipeline. It is noteworthy that such process is conducted while ensuring the model remains closely with its original form, which is achieved by employing a *reverse Kullback-Leibler* divergence penalty. Significant complexity has been introduced to the RLHF pipeline due to the requirement to train a separate reward model, even though it is somewhat effective. Moreover, Reinforcement learning pipelines also present notable challenges in terms of stability and memory demands towards alignment process of text-to-image models.

Recent research has demonstrated significant success in fine-tuning large language models (LLMs) using methods based on implicit rewards, specially the Direct Preference Optimization (DPO) [11]. Application of similar fine-tuning techniques to text-to-image models has also produced promising results, such as Diffusion-DPO [12], D3PO [13]. Such results have raisen significant interest within the community regarding the alignment of text-to-image models with human value through the methodology of utilizing implicit rewards. Furthermore, researchers have devoted significant efforts to applying such paradigm of aligning human value to text-to-image models, including SPO [14], NCPPO [15], DNO[16], and so on. However, it is the situation that existing research of text-to-image generation alignment predominantly targets solutions subject to the constraint of the *reverse Kullback-Leibler* divergence, with notable underexploitation of strategies that integrate other types of divergences.

It has been pointed out that models would overfit due to repeated fine-tuning on a few images, thus leading to reduced output diversity [17]. In the alignment of large language models, similar challenges exist; and some studies [18, 19] have highlighted that the mode-seeking property of reverse KL divergence tends to reduce diversity in generated outputs, which can constrain the model's potential. Studies on aligning large language models [20, 21] indicate that the problem of diversity reduction caused by fine-tuning can be alleviated by incorporating diverse divergence constraints. Therefore, in this study, we also explore the effects of employing diverse divergence constraints on the generation diversity.

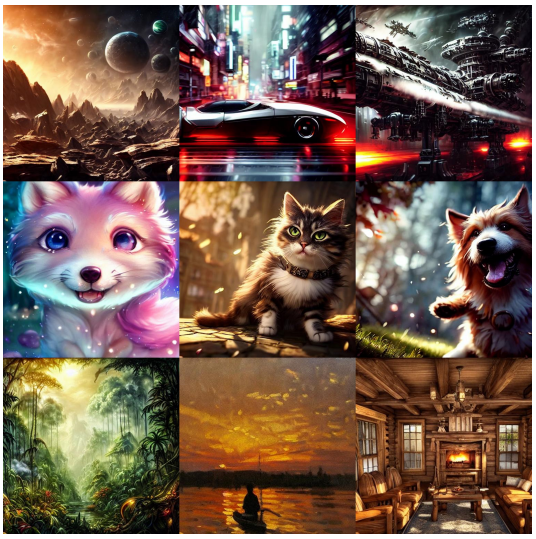


Figure 1: Examples of image generated by the model aligned using the Jensen-Shannon divergence constraint.

In this study, we generalize the alignment of text-to-image models based on *reverse Kullback-Leibler* divergence to a framework based on f-*divergence* constraints, which encompasses a wider range of divergences, including Jensen-Shannon divergence, forward Kullback-Leibler divergence, α-divergence, and so on. We comprehensively analyze the impact of diverse divergence constraints on the alignment process from the perspective of gradient fields. Furthermore, we set Step-aware Preference Optimization (SPO) [14] as our benchmark method, utilize Stable Diffusion V1.5 [22] as our benchmark model, and assess on the test split of HPS-V2 [9] with different divergence constraints. Evaluations are carried out to examine the performance of image-text alignment, human value alignment, and generation diversity, which also aim to discern the certain divergence most effectively balances these three aspects. Our results indicate that *Jensen-Shannon divergence* successfully strikes the ideal equilibrium among the three criteria examined, while also achieving the highest standard in human value alignment performance. Therefore, in text-to-image alignment, judicious selection of the divergence constraint, tailored to the specific alignment requirements, is paramount. In Figure 1, we present several images generated by the model that have been aligned under the Jensen-Shannon divergence.

To the best of our knowledge, this is the first work to apply different divergence constraints to text-to-image alignment paradigm. Our contributions are summarized as follows: (1) *Generalized alignment formula*: we propose a generalized formula for text-to-image generation alignment, aiming to provide more choices on divergence constraints in alignment execution. (2) *Thorough alignment*

process analysis: we comprehensively analyze the impact of different divergence constraints on alignment process from the perspective of gradient fields. (3) *Extensive alignment evaluations:* we conducted extensive evaluations on text-to-image generation alignment, meticulously assessing both alignment performance (image-text alignment and human value alignment) and generation diversity.

2 Related Work

2.1 Aligning Text-to-Image Model with Preferences

Recently, inspired by the alignment approaches based on human preferences, notably exemplified by methods such as direct preference optimization (DPO) [11], eliminating the need for explicit reward models and showing their significant success on Large Language Models (LLMs), and then garnering substantial attention within the community on the development of offline alignment for text-to-image diffusion models. Diffusion-DPO [12] enables text-to-image diffusion models to directly learn from human feedback in an open-vocabulary setting, and fine-tunes them on the contains Pick-a-Pic [6] dataset with image preference pairs. Direct Preference for Denoising Diffusion Policy Optimization (D3PO) [13] proposes a method on generating pairs of images from the same prompt and identifying the preferred and dispreferred images with the help of human evaluators. Step-aware Preference Optimization (SPO) [14] propose an approach that preferences at each step should be assessed and it utilizes a step-aware preference model and a step-wise resampler to ensure accurate step-aware preference alignment. DenseReward method [23] proposes enhancing the DPO scheme by incorporating a temporal discounting approach, which prioritizes the initial denoising steps. Noise-Conditioned Perceptual Preference Optimization (NCPPO) [15] proposes that the optimization process should aligns with human perceptual features, instead of the less informative pixel space. Direct Noise Optimization (DNO) [16] optimizes noise during the sampling process of text-to-image diffusion models. PopAlign [24] is an approach for population-level preference optimization, mitigating the biases of pretrained text-to-image diffusion models. Diffusion-KTO [25] generalizes the human utility maximization framework to the alignment of text-to-image diffusion models. While these studies have demonstrated impressive results in addressing the text-to-image alignment challenge, we also notice that they all rely on *reverse Kullback-Leibler divergence* to minimize the discrepancy between the fine-tuned model and the reference model.

2.2 f-divergence utilized in Generation Models

In previous studies, researchers have extensively examined the application of f-divergences in generative models. In the classical work done by [26], the concept of Generative Adversarial Networks (GANs) and their relationship to the Jensen-Shannon divergence are introduced. f-GAN [27] proposes that the variational expression of the f-divergence can be regarded as the loss function for Generative Adversarial Networks (GANs). Wasserstein-GAN [28] offers theoretical insights into the connection between the choice of divergences and the convergence of probability distributions. Moreover, in the work [29], it is proposed that utilizing various divergences and metrics can result in divergent trade-offs, and distinct evaluations tend to favor specific models. The application of f-divergence has also been observed in large language model alignment tasks. f-DPG [20] shows that Jensen-Shannon divergence strikes a good balance between different competing objectives, and often significantly outperforming the reverse Kullback-Leibler divergence. f-DPO [21] generalizes the framework of DPO by incorporating diverse divergence constraints; and it shows that by adjusting the divergence regularization, we can achieve a better balance between the alignment performance and the generation diversity.

3 Preliminary

3.1 f-divergence

For any convex function $f(x) : \mathbb{R}^+ \to \mathbb{R}$ with $f(1) = 0$, and p_1, p_2 are two distributions over a discrete set $\mathcal{X}$, the f-divergence between p_1 and p_2 can be defined as [30]:

$$D_f(p_1||p_2) = \mathbb{E}_{x\sim p_2}\left[f\left(\frac{p_1(x)}{p_2(x)}\right) + f'(\infty)p_1(p_2 = 0)\right],$$

where $f'(\infty) = \lim_{t\to 0} tf(\frac{1}{t})$ [31], $p_1(p_2 = 0) = 0$ is the p_1-mass of the set $\{x \in \mathcal{X} : p_2(x) = 0\}$. Under normal circumstances, we can make the assumption that the support set of p_1 is dominated by the support set of p_2, i.e. $Supp(p_1) \subset Supp(p_2)$, and then we can have $p_1(p_2 = 0) = 0$. Hence, the aforementioned definition can be simplified as:

$$D_f(p_1||p_2) = \mathbb{E}_{x\sim p_2}\left[f\left(\frac{p_1(x)}{p_2(x)}\right)\right]$$

For different functions $f(x)$, the f-divergence class encompasses a wide range of commonly employed divergence measures, such as reverse Kullback-Leibler (KL) divergence, forward Kullback-Leibler (KL) divergence, α-divergence ($\alpha \in (0, 1)$), Jensen-Shannon (JS) divergence, and so on.

$f-$divergence	$f(x)$	$f'(x)$	$f''(x)$
Reverse KL	$x \log x$	$\log x + 1$	$\frac{1}{x}$
Forward KL	$-\log x$	$-\frac{1}{x}$	$\frac{1}{x^2}$
α-divergence	$\frac{x^{1-\alpha}-(1-\alpha)x-\alpha}{\alpha(\alpha-1)}$	$\frac{1-x^{-\alpha}}{\alpha}$	$\frac{1}{x^{\alpha+1}}$
JS divergence	$x \log \frac{2x}{x+1} + \log \frac{2}{x+1}$	$\log \frac{2x}{1+x}$	$\frac{1}{x(1+x)}$

Table 1: Several commonly used f-divergence with their derivatives and second derivatives.

In previous studies, reverse KL divergence can be regarded as a specific instance of α-divergence with $\alpha = 0$; and forward KL divergence as a specific instance of α-divergence with $\alpha = 1$. We summarize several commonly used f-divergence, the derivatives and the second derivatives in Table 1.

4 Method

Much like in the alignment tasks of large language models, there are many concepts that are analogous in the alignment tasks of text-to-image models, and we start by elucidating these parallels. Firstly, the question input of LLMs is akin to the text (condition) input of T2I models, i.e. $x \to c$; and the output answer of LLMs is akin to the generated image of T2I models, i.e. $y \to x_0$. Moreover, the policy of LLMs parallels the sampling probability of T2I models (especially diffusion models), i.e. $\pi(y|x) \to p(x_{0:T}|c)$. Finally, the preference data for output answers of LLMs is analogous to the preference data for generated images of T2I models, i.e. $(x, y_w, y_l) \to (c, x_0^w, x_0^l)$. In the following subsections, we first derive the generalized formula of alignment objective function. Then, we analyze the gradient field of different divergences on the alignment process with respect to the objective function and comprehensively analyze the impact of diverse divergence constraints on alignment performance.

4.1 Generalized Formula

In previous studies of Reinforcement Learning from Human Feedback (RLHF), researchers typically aim to maximize the reward function ($r(c, x_{0:T})$) while penalizing the reverse KL divergence between the fine-tuned model and the original model to prevent it from collapsing during training. In our study, we generalize such penalty constraint from the reverse KL divergence ($D_{\text{KL}}\left(p_\theta(x_{0:T}|c), p_{\text{ref}}(x_{0:T}|c)\right)$) to the f-divergence ($D_f\left(p_\theta(x_{0:T}|c), p_{\text{ref}}(x_{0:T}|c)\right)$).

We reframe the reinforcement learning objective function as an optimal problem, presenting its formulation as follows:

$$\arg\max_{p_\theta} \mathbb{E}_{c\sim p_c, x_{0:T}\sim p_\theta(x_{0:T}|c)}\left[r(c, x_{0:T})\right] - \beta D_f\Big(p_\theta(x_{0:T}|c), p_{\text{ref}}(x_{0:T}|c)\Big)$$
$$\text{s.t.} \quad \sum_{x_{0:T}} p_\theta(x_{0:T}|c) = 1; \forall x_{0:T} \quad p_\theta(x_{0:T}|c) \geq 0$$

Such optimization problem can be addressed through the Karush-Kuhn-Tucker (KKT) conditions. Firstly, according to the definition of f-divergence, we construct the following Lagrangian function:

$$\mathcal{L}(p_\theta(x_{0:T}|c), \lambda, \zeta(x_{0:T})) = \mathbb{E}_{c\sim p_c, x_{0:T}\sim p_\theta}\left[r(c, x_{0:T})\right] - \beta\mathbb{E}_{p_{\text{ref}}} f\left(\frac{p_\theta(x_{0:T}|c)}{p_{\text{ref}}(x_{0:T}|c)}\right) - \lambda\left(\sum_{x_{0:T}} p_\theta(x_{0:T}) - 1\right) + \sum_{x_{0:T}} \zeta(x_{0:T}) p_\theta(x_{0:T}|c)$$

Furthermore, we can derive the Theorem 1 from the *Stationarity Condition* and *Complementary Slackness* of the Karush-Kuhn-Tucker (KKT) conditions, i.e.

$$\begin{cases} \nabla_{p_\theta(x_{0:T}|c)} \mathcal{L}(p_\theta(x_{0:T}|c), \lambda, \zeta(x_{0:T})) = 0; \\ \forall x_{0:T}, \zeta(x_{0:T}) p_\theta(x_{0:T}|c) = 0. \end{cases}$$

Theorem 1. *If $p_{\text{ref}}(x_{0:T}|c) > 0$ holds for all condition c, $f'(x)$ is an invertible function and 0 is not in the definition domain of function $f'(x)$, the reward class consistent with Bradley-Terrry model can be reparameterized with the sampling probability $p_\theta(x_{0:T}|c)$ and the reference sampling probability $p_{\text{ref}}(x_{0:T}|c)$ as:*

$$r(c, x_{0:T}) = \beta f'\left(\frac{p_\theta(x_{0:T}|c)}{p_{ref}(x_{0:T}|c)}\right) + \text{const} \tag{1}$$

As shown in Theorem 1, the reward function can be represented by a sampling probability $p_\theta(x_{0:T}|c)$, a reference sampling probability $p_{\text{ref}}(x_{0:T}|c)$, and a constant λ that is independent of $x_{0:T}$. Finally, substituting Equation (1) into the Bradley-Terrry model [10] enables us to derive the generalized formula of text-to-image generation with preferences in Theorem 2.

Theorem 2. *In the substitution process of Bradley-Terry model, the constant λ is independent of $x_{0:T}$ and thus can be canceled out, resulting in the following form:*

$$\mathcal{L}(\theta) = \mathbb{E}_{\substack{(c,x_0^w,x_0^l)\sim\mathcal{D}, \\ x_{1:T}^w\sim p_\theta(x_{1:T}^w|x_0^w,c), \\ x_{1:T}^l\sim p_\theta(x_{1:T}^l|x_0^l,c).}} -\log\sigma\left[\beta f'\left(\frac{p_\theta(x_{0:T}^w|c)}{p_{ref}(x_{0:T}^w|c)}\right) - \beta f'\left(\frac{p_\theta(x_{0:T}^l|c)}{p_{ref}(x_{0:T}^l|c)}\right)\right] \tag{2}$$

where $\sigma(\cdot)$ is the Sigmoid function; $f'(\cdot)$ represents the derivatives of $f(\cdot)$, as listed in Table 1; β is the penalty coefficient.

So far, we have derived the generalized formula for text-to-image generation alignment with preferences. With different divergence constraint choices, we can obtain diverse alignment objectives, thereby offering more options for the alignment process.

4.2 Analysis on Gradient Fields of Alignment Process

In this section, we delve into the gradient fields of alignment objective functions derived from various f-divergence, which aims to further elucidate the intricate mechanisms underlying the alignment process.

Let's abstract from the specific details of $f'(\cdot)$, and concentrate instead on a more general formulation of the loss function:

$$\mathcal{L}_f\left(\mathrm{X}_1, \mathrm{X}_2\right) = -\mathbb{E}\Big[\log\sigma\big(\beta f'(\mathrm{X}_1) - \beta f'(\mathrm{X}_2)\big)\Big] \tag{3}$$

where X_1 is the training win ratio, and is equivalent to $\frac{p_\theta(x_{0:T}^w|c)}{p_{\text{ref}}(x_{0:T}^w|c)}$; similarly, X_2 is the training loss ratio, and is identical to $\frac{p_\theta(x_{0:T}^l|c)}{p_{\text{ref}}(x_{0:T}^l|c)}$. We present the gradients of Equation (3) with respect to X_1 and X_2 in the ensuing Theorem 3:

Theorem 3. *The partial derivatives (gradients) of X_1 and X_2 resulting from Equation* (3) *can be expressed as follows:*

$$\begin{cases} \dfrac{\partial\mathcal{L}_f(\mathrm{X}_1, \mathrm{X}_2)}{\partial \mathrm{X}_1} = -\beta\left(1 - \sigma\left(\beta f'(\mathrm{X}_1) - \beta f'(\mathrm{X}_2)\right)\right) f''(\mathrm{X}_1) \\ \dfrac{\partial\mathcal{L}_f(\mathrm{X}_1, \mathrm{X}_2)}{\partial \mathrm{X}_2} = \beta\left(1 - \sigma\left(\beta f'(\mathrm{X}_1) - \beta f'(\mathrm{X}_2)\right)\right) f''(\mathrm{X}_2) \end{cases}$$

Thus, the gradient ratio of $\mathcal{L}_f(\mathrm{X}_1, \mathrm{X}_2)$ between enhancement in probability for human-preferred responses (X_1) and reduction in probability for human-dispreferred responses (X_2) has the expression:

$$\left| \frac{\partial \mathcal{L}_f(\mathrm{X}_1, \mathrm{X}_2)}{\partial \mathrm{X}_1} \Big/ \frac{\partial \mathcal{L}_f(\mathrm{X}_1, \mathrm{X}_2)}{\partial \mathrm{X}_2} \right| = \frac{f''(\mathrm{X}_1)}{f''(\mathrm{X}_2)} \tag{4}$$

Referencing Table 1, different divergences yield distinct gradient ratios. If selected divergence is reverse Kullback-Leibler divergence, the gradient ratio is $\frac{\mathrm{X}_2}{\mathrm{X}_1}$; if selected divergence is Jensen-Shannon divergence, the gradient ratio is $\frac{\mathrm{X}_2 \cdot (\mathrm{X}_2+1)}{\mathrm{X}_1 \cdot (\mathrm{X}_1+1)}$; if selected divergence is α-divergence, the gradient ratio is $\frac{\mathrm{X}_2^{1+\alpha}}{\mathrm{X}_1^{1+\alpha}}$; if selected divergence is forward Kullback-Leibler divergence, the gradient ratio is $\frac{\mathrm{X}_2^2}{\mathrm{X}_1^2}$. Previous studies [32, 33] present the results of original DPO framework, focusing particularly on its application in the context of reverse Kullback-Leibler divergence; while our outcomes show the generalization under diverse divergences.

Furthermore, as the alignment advances, the value of X_1 tends to increase to more than 1, whereas X_2 tends to decrease to less than 1. Hence, for any pairwise preference data, $\mathrm{X}_2/\mathrm{X}_1 < 1$ holds during the alignment process. Then, Theorem 4 can be easily derived.

Theorem 4. *As alignment progresses, we have $\mathrm{X}_2/\mathrm{X}_1 < 1$. Hence,*

$$0 < \frac{\mathrm{X}_2^2}{\mathrm{X}_1^2} < \frac{\mathrm{X}_2 \cdot (\mathrm{X}_2+1)}{\mathrm{X}_1 \cdot (\mathrm{X}_1+1)} < \frac{\mathrm{X}_2}{\mathrm{X}_1} < 1 \quad \textit{and} \quad 0 < \frac{\mathrm{X}_2^2}{\mathrm{X}_1^2} < \frac{\mathrm{X}_2^{1.8}}{\mathrm{X}_1^{1.8}} < \frac{\mathrm{X}_2^{1.6}}{\mathrm{X}_1^{1.6}} < \frac{\mathrm{X}_2^{1.4}}{\mathrm{X}_1^{1.4}} < \frac{\mathrm{X}_2^{1.2}}{\mathrm{X}_1^{1.2}} < \frac{\mathrm{X}_2}{\mathrm{X}_1} < 1$$

Theorem 4 presents the inequality of gradient ratio of different divergences. A lower gradient ratio results in a swifter alteration in the probability of a dispreferred image compared to that of a preferred one, indicating a more pronounced decrease in the probability of dispreferred images. Hence, the decline varies in intensity, with *forward KL divergence (α=1)* exhibiting the highest decrease, *reverse KL divergence (α=0)* the lowest, and both *Jensen-Shannon divergence* and *α-divergence ($\alpha \in$(0,1))* falling in between.

In order to obtain a more intuitive understanding of the impact of different divergence choices during the alignment process, we visualize the landscape of alignment objective functions with different divergences in Equation (3) in Appendix F.

5 Experiments

In this section, we present extensive experimental evaluations to answer the following questions:

Q1: When choosing different divergence constraints, would it have a significant impact on the final *image-text* alignment performance?

Q2: When choosing different divergence constraints, would it have a significant impact on the alignment of *human value*? If so, which divergence constraint achieves the best performance?

Q3: When choosing different divergence constraints, would it have a significant impact on the *generation diversity*? Which divergence can achieve the best trade-off between alignment performance and generation diversity?

5.1 Experimental Settings

5.1.1 Benchmark.

Step-aware Preference Optimization (SPO) [14] employs a step-aware preference model and a step-wise resampler to guarantee precise step-aware preference alignment. Consequently, to support a more tangible experimental assessment, we select SPO as our benchmark approach. To establish a fair basis for comparison with prior methods, we select Stable Diffusion v1-5 model [22] as our benchmark model. In order to conduct a more comprehensive evaluation, we utilize the test set of HPS-V2 [9] as our evaluation benchmark dataset, which comprises 400 prompts. We report the mean and standard deviation of metrics of the generated image for these prompts.

5.1.2 Evaluation Metrics.

We evaluate the generated images from three aspects (for the aforementioned three questions).

In terms of model's image-text alignment performance (for Q1), we adopt the widely used evaluation metrics in Text-to-Image models, i.e., Text-Image CLIP score [34] and VQAScore [35]. CLIP Score is fundamentally based on the CLIP model, transforming input text and images into distinct text and image vectors, and then followed by calculation of the dot product of these vectors. VQAScore, an efficacy metric, emerges from the training of generative vision-language models designed for visual-question-answering endeavors, where an image and a query converge to yield a response. Hence, higher Text-Image CLIP score and VQAScore indicate better alignment between the text and the image.

In terms of model's human value alignment performance (for Q2), we adopt four metrics for comprehensive evaluation. Aesthetic score is obtained using the LAION Aesthetics Predictor [36], which quantifies the average human appreciation for the visual appeal of generated images. ImageReward [7], leveraging a structure that combines ViT-L for image encoding and a 12-layer Transformer for text encoding, which effectively models the human value and preference. PickScore [6] is an advanced scoring function built upon a meticulously curated comprehensive dataset named "Pick-a-Pic". Human Preference Score v2 (HPS-v2) [9] has been developed through the refinement of the CLIP model on HPD-v2, which enhances the precision of assessing human preferences for generated images. Furthermore, higher Aesthetics Score, ImageReward, Pickscore and HPS-V2 suggest better alignment with human value.

In terms of diversity of generated images from the aligned model (for Q3), we adopt eight metrics for a further comprehensive evaluation. Image-Image CLIP score [34] serves as a reliable metric for assessing similarity between images. RMSE, PSNR, and SSIM are conventional metrics used to evaluate image similarity, we also utilize them to assess the diversity of generated images. Feature Similarity Index Measure (FSIM) [37] quantifies the similarity between images by assessing the alignment of edges, shapes, visual patterns, and surface attributes. Learned Perceptual Image Patch Similarity (LPIPS) [38] utilizes the feature representations learned by a deep neural network, which is capable of capturing details of human visual perception such as texture, color, and structure; then the computation of perceptual similarity between two images can be conducted. Furthermore, it's worth noting that these six metrics all initially describe the similarity between images; and when they are used to describe generation diversity, their properties are the opposite of their properties when describing similarity. Moreover, we opt for Image Entropy, encompassing both Entropy 1D and Entropy 2D, to evaluate the information content diversity within images themselves; they quantifies the average information per pixel, with higher entropy values indicative of a greater diversity and richness in the image's information content.

5.2 Image-Text Alignment (For Q1)

For text-to-image models, the alignment performance between text prompt and generated images is a crucial evaluation metric. Therefore, we test the Text-Image CLIP score and VQAScore of all models fine-tuned under different divergence constraints to assess the alignment performance in Table 2. The results indicate that the reverse Kullback-Leibler divergence achieves the best text-image alignment performance; while it is also worth noting that different divergences do not significantly affect the final text-image alignment performance.

5.3 Human Value Alignment (For Q2)

Evaluating how well the aligned models are with human values and preferences is crucial. To comprehensively assess various divergences in aligning with human values, we compare their performances systematically on four metrics: Aesthetic score, ImageReward, PickScore, and HPS-V2 in Table 2. The comparison between the results of the fine-tuned models and the original model indicates that the alignment process effectively enhances the model in terms of its performance in human values. Furthermore, in comparing the influence of diverse divergence constraints on human value alignment, the results reveal that different divergence would significantly affect human value alignment; remarkably, the *Jensen-Shannon (JS) divergence* exhibits the best performance across all four human value alignment metrics, suggesting that it serves as a more potent constraint specifically for the scenario of human value alignment. Actually, it also aligns with our previous analysis of the

Model		CLIPScore ↑	VQAScore ↑	Aesthetics Score ↑	ImageReward ↑	PickScore ↑	HPS-V2 ↑
Original Model		0.352±0.049	0.625±0.239	5.648±0.526	0.173±1.011	20.908±1.228	26.933±1.454
Reverse KL Divergence		**0.363±0.049**	**0.676±0.236**	5.812±0.514	0.619±0.921	21.621±1.151	27.801±1.352
α-Divergence	α=0.2	0.360±0.048	0.673±0.228	5.827±0.546	0.561±0.957	21.528±1.177	27.848±1.391
	α=0.4	0.361 ±0.047	0.675±0.233	5.755±0.518	0.622±0.911	21.569±1.204	27.762±1.385
	α=0.6	0.358 ±0.047	0.657±0.232	5.769±0.481	0.491±0.943	21.357±1.180	27.712±1.350
	α=0.8	0.361±0.050	0.662±0.236	5.821±0.511	0.561±0.965	21.483±1.175	27.675±1.379
Forward KL Divergence		0.362±0.050	0.670±0.231	5.844±0.528	0.551±0.946	21.552±1.170	27.822±1.355
Jensen-Shannon Divergence		0.361±0.049	0.665±0.231	**5.884±0.514**	**0.631±0.939**	**21.635±1.149**	**27.850±1.388**

Table 2: Evaluations of the alignment performance, where the Text-Image CLIP score and VQAScore evaluate image-text alignment performance, and the remaining four metrics evaluate human value alignment performance.

Model		Image-Image CLIP score ↓	Entropy 1D ↑	Entropy 2D ↑	LPIPS ↑
Original Model		0.8052 ± 0.0824	3.8235 ± 0.2960	7.5474 ± 0.6516	0.2972 ± 0.0419
Reverse KL Divergence		0.8448 ± 0.0774	3.9613 ± 0.1467	7.8347 ± 0.3694	0.2907 ± 0.0363
α-Divergence	α =0.2	0.8436 ± 0.0854	3.9411 ± 0.1885	7.7836 ± 0.4400	0.3047 ± 0.0377
	α =0.4	0.8377 ± 0.0824	**3.9784 ± 0.1464**	7.8206 ± 0.3729	0.2959 ± 0.0349
	α =0.6	**0.8372 ± 0.0825**	3.9275 ± 0.1991	7.7937 ± 0.4625	**0.3109 ± 0.0373**
	α =0.8	0.8423 ± 0.0795	3.9594 ± 0.1666	7.7563 ± 0.4179	0.3001 ± 0.0377
Forward KL Divergence		0.8454 ± 0.0821	3.9477 ± 0.1555	7.7750 ± 0.3619	0.2962 ± 0.0347
Jensen-Shannon Divergence		0.8448 ± 0.0798	3.9632 ± 0.1487	**7.8767 ± 0.3801**	0.2989 ± 0.0361
Model		RMSE ↑	PSNR ↓	SSIM ↓	FSIM ↓
Original Model		0.0132 ± 0.0028	37.745 ± 1.843	0.8839 ± 0.0382	0.3791 ± 0.0230
Reverse KL Divergence		0.0132 ± 0.0028	36.398 ± 1.573	0.8512 ± 0.0372	0.3813 ± 0.0182
α-Divergence	α =0.2	0.0163 ± 0.0027	35.856 ± 1.467	0.8404 ± 0.0368	**0.3759 ± 0.0212**
	α =0.4	0.0154 ± 0.0025	36.363 ± 1.427	0.8530 ± 0.0348	0.3821 ± 0.0180
	α =0.6	**0.0166 ± 0.0026**	**35.705 ± 1.373**	**0.8357 ± 0.0349**	0.3778 ± 0.0209
	α =0.8	0.0155 ± 0.0028	36.320 ± 1.566	0.8517 ± 0.0374	0.3806 ± 0.0215
Forward KL Divergence		0.0157 ± 0.0026	36.171 ± 1.457	0.8468 ± 0.0351	0.3780 ± 0.0185
Jensen-Shannon Divergence		0.0158 ± 0.0026	36.104 ± 1.431	0.8449 ± 0.0354	0.3817 ± 0.0195

Table 3: Evaluations of the generation diversity. The metrics originally utilized for evaluating image similarity exhibit an opposite property when evaluating generation diversity.

gradient fields, where the Jenson-Shannon (JS) divergence shows the smoothest loss function surface and suboptimal gradient ratio, resulting in a more stable alignment process.

5.4 Generation Diversity (For Q3)

We evaluate the generation diversity of aligned models using different divergence constraints from multiple perspectives (embedding diversity, pixel-level diversity, structural diversity, perceptual diversity, information complexity, and so on), and the corresponding results are shown in Table 3. From the results, we can observe that different divergence constraints exhibit advantages in different aspects when evaluated with different generation diversity metrics. Firstly, we would like to compare the aligned models under different divergence constraints to the original model: it is demonstrated that the aligned models show a decrease in embedding diversity; however, they exhibit improvements in other aspects such as pixel-level diversity, structural diversity, information complexity. Such observation reveals a transformation in the alignment process where the variety of the primary subject diminishes, yet the intricacy and breadth of details and structures of the generated images expand, echoing findings from DreamBooth [17].

Furthermore, it has also indicated that increased generative diversity is associated with a decline in alignment performance (both image-text alignment and human value alignment). Therefore, careful consideration of the trade-off between alignment performance and generation diversity is essential when choosing the divergence constraint. Through a deeper comparison and analysis, we can observe that Jensen-Shannon divergence outperforms or matches reverse Kullback-Leibler divergence across most diversity metrics. Combining such observation with the previous evaluation of alignment performance where it achieves the best human value alignment, we believe Jensen-Shannon divergence is a better trade-off between alignment performance and generation diversity.

6 Conclusion

In this paper, we extend the alignment framework for text-to-image models, transitioning from a criterion based on the reverse Kullback-Leibler (KL) divergence to a more inclusive framework grounded in f-divergence constraints. Through the analysis of gradient fields (gradient ratio and loss function surface) under diverse divergence constraints, we further illustrate the advantages of different divergence constraints in the alignment process. Regarding image-text alignment, minimal differentiation is observed among the diverse divergence constraints; conversely, for human value alignment, Jensen-Shannon (JS) divergence excels, showcasing its superior performance across all four evaluation metrics. In generative diversity evaluation, we observe that diverse divergence constraints demonstrate strengths in various aspects of diversity. Furthermore, it has been observed that increased generation diversity consistently correlates with a decrease in alignment performance. After thorough comparison, we advocate for the selection of Jensen-Shannon (JS) divergence as the foremost option in practice, which is a better trade-off between alignment performance and generation diversity.

Acknowledgment

This work is partly supported by the National Natural Science Foundation of China (No.62103225), Natural Science Foundation of Shenzhen (No.JCYJ20230807111604008), Natural Science Foundation of Guangdong Province (No.2024A1515010003) and National Key Research and Development Program (No.2022YFB4701402).

References

[1] 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. In *The Twelfth International Conference on Learning Representations*, 2024.

[2] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024.

[3] Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. Autoregressive image generation without vector quantization. *arXiv preprint arXiv:2406.11838*, 2024.

[4] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8, 2023.

[5] Xiao Liu, Fanjin Zhang, Zhenyu Hou, Li Mian, Zhaoyu Wang, Jing Zhang, and Jie Tang. Self-supervised learning: Generative or contrastive. *IEEE transactions on knowledge and data engineering*, 35(1):857–876, 2021.

[6] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:36652–36663, 2023.

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

[8] Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. In *The Twelfth International Conference on Learning Representations*, 2024.

[9] Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023.

[10] Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952. ISSN 00063444.

[11] 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. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 53728–53741, 2023.

[12] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8228–8238, 2024.

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

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

[15] Alexander Gambashidze, Anton Kulikov, Yuriy Sosnin, and Ilya Makarov. Aligning diffusion models with noise-conditioned perception. *arXiv preprint arXiv:2406.17636*, 2024.

[16] Zhiwei Tang, Jiangweizhi Peng, Jiasheng Tang, Mingyi Hong, Fan Wang, and Tsung-Hui Chang. Tuning-free alignment of diffusion models with direct noise optimization. *arXiv preprint arXiv:2405.18881*, 2024.

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

[18] Gian Wiher, Clara Meister, and Ryan Cotterell. On decoding strategies for neural text generators. *Transactions of the Association for Computational Linguistics*, 10:997–1012, 2022.

[19] Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. Red teaming language models with language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3419–3448, 2022.

[20] Dongyoung Go, Tomasz Korbak, Germàn Kruszewski, Jos Rozen, Nahyeon Ryu, and Marc Dymetman. Aligning language models with preferences through f-divergence minimization. In *International Conference on Machine Learning*, pages 11546–11583. PMLR, 2023.

[21] Chaoqi Wang, Yibo Jiang, Chenghao Yang, Han Liu, and Yuxin Chen. Beyond reverse kl: Generalizing direct preference optimization with diverse divergence constraints. In *The Twelfth International Conference on Learning Representations*, 2024.

[22] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, June 2022.

[23] Shentao Yang, Tianqi Chen, and Mingyuan Zhou. A dense reward view on aligning text-to-image diffusion with preference. In *Forty-first International Conference on Machine Learning*, 2024.

[24] Shufan Li, Harkanwar Singh, and Aditya Grover. Popalign: Population-level alignment for fair text-to-image generation. *arXiv preprint arXiv:2406.19668*, 2024.

[25] Shufan Li, Konstantinos Kallidromitis, Akash Gokul, Yusuke Kato, and Kazuki Kozuka. Aligning diffusion models by optimizing human utility. *arXiv preprint arXiv:2404.04465*, 2024.

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

[27] Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-gan: Training generative neural samplers using variational divergence minimization. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

[28] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017.

[29] L Theis, A van den Oord, and M Bethge. A note on the evaluation of generative models. In *International Conference on Learning Representations (ICLR 2016)*, pages 1–10, 2016.

[30] Friedrich Liese and Igor Vajda. On divergences and informations in statistics and information theory. *IEEE Transactions on Information Theory*, 52(10):4394–4412, 2006.

[31] Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. *Convex analysis and minimization algorithms I: Fundamentals*, volume 305. Springer science & business media, 1996.

[32] Duanyu Feng, Bowen Qin, Chen Huang, Zheng Zhang, and Wenqiang Lei. Towards analyzing and understanding the limitations of dpo: A theoretical perspective. *arXiv preprint arXiv:2404.04626*, 2024.

[33] Yuzi Yan, Yibo Miao, Jialian Li, Yipin Zhang, Jian Xie, Zhijie Deng, and Dong Yan. 3d-properties: Identifying challenges in dpo and charting a path forward. *arXiv preprint arXiv:2406.07327*, 2024.

[34] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021.

[35] Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. In *European Conference on Computer Vision*, pages 366–384. Springer, 2025.

[36] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.

[37] Lin Zhang, Lei Zhang, Xuanqin Mou, and David Zhang. Fsim: A feature similarity index for image quality assessment. *IEEE transactions on Image Processing*, 20(8):2378–2386, 2011.

[38] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

[39] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

[40] 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. *arXiv preprint arXiv:1909.08593*, 2019.

[41] Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, et al. Open problems and fundamental limitations of reinforcement learning from human feedback. *Transactions on Machine Learning Research*, 2023.

[42] Donald Joseph Hejna III and Dorsa Sadigh. Few-shot preference learning for human-in-the-loop rl. In *Conference on Robot Learning*, pages 2014–2025. PMLR, 2023.

[43] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

[44] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[45] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[46] Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024.

[47] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017.

[48] Rajkumar Ramamurthy, Prithviraj Ammanabrolu, Kianté Brantley, Jack Hessel, Rafet Sifa, Christian Bauckhage, Hannaneh Hajishirzi, and Yejin Choi. Is reinforcement learning (not) for natural language processing: Benchmarks, baselines, and building blocks for natural language policy optimization. In *The Eleventh International Conference on Learning Representations*, 2023.

[49] Joar Skalse, Nikolaus Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward gaming. *Advances in Neural Information Processing Systems*, 35:9460–9471, 2022.

[50] Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning*, pages 10835–10866. PMLR, 2023.

[51] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015.

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

[53] Tianyi Zhang, Zheng Wang, Jing Huang, Mohiuddin Muhammad Tasnim, and Wei Shi. A survey of diffusion based image generation models: Issues and their solutions. *arXiv preprint arXiv:2308.13142*, 2023.

[54] Huan Liao, Haonan Han, Kai Yang, Tianjiao Du, Rui Yang, Zunnan Xu, Qinmei Xu, Jingquan Liu, Jiasheng Lu, and Xiu Li. Baton: Aligning text-to-audio model with human preference feedback. *arXiv preprint arXiv:2402.00744*, 2024.

[55] Zhen Xing, Qijun Feng, Haoran Chen, Qi Dai, Han Hu, Hang Xu, Zuxuan Wu, and Yu-Gang Jiang. A survey on video diffusion models. *arXiv preprint arXiv:2310.10647*, 2023.

[56] Eric R Chan, Koki Nagano, Matthew A Chan, Alexander W Bergman, Jeong Joon Park, Axel Levy, Miika Aittala, Shalini De Mello, Tero Karras, and Gordon Wetzstein. Generative novel view synthesis with 3d-aware diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4217–4229, 2023.

[57] Ivan Kapelyukh, Vitalis Vosylius, and Edward Johns. Dall-e-bot: Introducing web-scale diffusion models to robotics. *IEEE Robotics and Automation Letters*, 8(7):3956–3963, 2023.

[58] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

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

[60] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr, 2021.

[61] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023.

[62] Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. Compositional visual generation with composable diffusion models. In *European Conference on Computer Vision*, pages 423–439. Springer, 2022.

[63] Weixi Feng, Xuehai He, Tsu-Jui Fu, Varun Jampani, Arjun Reddy Akula, Pradyumna Narayana, Sugato Basu, Xin Eric Wang, and William Yang Wang. Training-free structured diffusion guidance for compositional text-to-image synthesis. In *The Eleventh International Conference on Learning Representations*, 2023.

[64] Kimin Lee, Hao Liu, Moonkyung Ryu, Olivia Watkins, Yuqing Du, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, and Shixiang Shane Gu. Aligning text-to-image models using human feedback. *arXiv preprint arXiv:2302.12192*, 2023.

[65] Bram Wallace, Akash Gokul, Stefano Ermon, and Nikhil Naik. End-to-end diffusion latent optimization improves classifier guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7280–7290, 2023.

[66] Kevin Clark, Paul Vicol, Kevin Swersky, and David J Fleet. Directly fine-tuning diffusion models on differentiable rewards. In *The Twelfth International Conference on Learning Representations*, 2024.

[67] Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. Reinforcement learning for fine-tuning text-to-image diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024.

[68] Amelia Carolina Sparavigna. Entropy in image analysis, 2019.

[69] Cort J. Willmott and Kenji Matsuura. Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance. *Climate Research*, 30: 79–82, 2005. URL `https://api.semanticscholar.org/CorpusID:120556606`.

[70] Alain Horé and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In *2010 20th International Conference on Pattern Recognition*, pages 2366–2369, Aug 2010. doi: 10.1109/ICPR.2010.579.

[71] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. URL https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

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

[73] Baiqi Li, Zhiqiu Lin, Deepak Pathak, Jiayao Li, Yixin Fei, Kewen Wu, Tiffany Ling, Xide Xia, Pengchuan Zhang, Graham Neubig, et al. Genai-bench: Evaluating and improving compositional text-to-visual generation. *arXiv preprint arXiv:2406.13743*, 2024.

[74] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *Transactions on Machine Learning Research*, 2022.

A Additional Related Work Statement

A.1 Reinforcement from Human Feedback (RLHF).

Reinforcement learning from human feedback (RLHF) [39, 40] is a crucial method for aligning artificial intelligence systems with human values, ensuring that AI systems operate and make decisions in accordance with human goals. It often integrates three core components [41]: feedback collection, reward modeling, and policy optimization. It facilitates humans in communicating goals without the need for manually specifying a reward function. And it leverages human judgments, which are often easier to obtain than demonstrations. Furthermore, RLHF can mitigate reward hacking compared to manually specified proxies, making reward shaping more natural and implicit. Hence, RLHF has been proven to be a valuable tool for assisting policies in learning intricate solutions in control environments [42] and for fine-tuning large scale models [43, 8]. Despite its widespread adoption, it still faces several limitations and open problems. In the work [41], they are summarized as four aspects: challenges with obtaining human feedback; challenges with the reward model training; challenges from policy optimization and challenges with jointly training process. Moreover, it is pointed out that several of such weaknesses can be mitigated through the enhancement of the RLHF approach; and alternatively, some of these weaknesses can be offset by implementing additional safety measures; while others requires avoiding or compensating for with non-RLHF approaches.

A.2 Fine-tuning Large Language Models with Reinforcement Learning .

Before RLHF, LLMs are typically aligned with human preferences through supervised fine-tuning (SFT) on demonstration data. The integration of RLHF into the training process of large language models (LLMs) has marked a significant milestone to the field of foundation model development. It has enabled LLMs to achieve human-level performance on various tasks, including text summarization, machine translation, question answering, and so on. In RLHF based fine-tuning pipeline, LLMs are trained by using human feedback as reward signals, guiding the models towards generating more accurate, relevant, and informative responses. Such iterative process allows LLMs to continuously learn and improve their performance, and this paradigm has led to the emergence of numerous remarkable models, such as OpenAI's GPT-4 [44], Meta's Llama 3 [45], Google's Gemma [46], and so on. Prior works has used policy-gradient methods [47] to this end. While they are indeed quite successful, they often come with high cost training, require extensive hyperparameter tuning process [48], and can be vulnerable to reward hacking, as demonstrated in various studies [49, 50]. Recent methods have emerged that fine-tune policy models by directly training them with a ranking loss on preference data, such as direct preference optimization (DPO) [11], which have been shown to achieve performance on par with RLHF.

A.3 Denoising Diffusion Probabilistic Models.

Denoising diffusion probabilistic models (DDPMs) have become a leading force in generative modeling due to their remarkable ability to generate diverse data formats. Diffusion model class utilizes an iterative denoising process to transform Gaussian noise into samples that adhere to a learned data distribution. Initially introduced in [51], further develop and promote in [52], they have been proved to be highly effective in a range of domains, including image generation [53], audio generation [54], video generation [55], 3D synthesis [56], robotics [57], and so on. Diffusion models, integrating with large-scale language encoders, have demonstrated remarkable performance in text-to-image generation [58, 59]. Advancements in text-to-image generation diffusion models have revolutionized the creation of lifelike visual representations based on written descriptions [60] and such breakthrough has opened exciting opportunities in digital art and design. In order to achieve more precise control over the outputs generated by diffusion models, researchers are exploring innovative methods to guide the diffusion process. While existing text-to-image models have achieved impressive results, they still exhibit several limitations, including challenges with compositionality, attribute binding, and so on. Researchers have also conducted extensive work to improve these aspects. Adapters [61] have been developed to impose additional input constraints, ensuring that the generated content aligns more precisely with specific standards. For the sake of enhancing image quality and generation control, compositional approaches [62, 63] have been developed to integrate multiple models effectively.

Considering the data distribution $x_0 \sim q_0(x_0), x_0 \in \mathbb{R}^n$. DDPM algorithm approximates the data distribution q_0 with a parameterized model with the form of $p_\theta(x_0) = \int p_\theta(x_{0:T}|c)dx_{1:T}$, where $p_\theta(x_{0:T}|c)$=$p_T(x_T)\prod_{t=1}^{T} p_\theta(x_{t-1}|x_t, c)$, and c is the conditioning information, i.e., image category and image caption. Then, we can describe the reverse process to be an Markov chain with dynamics as follows:

$$p(x_T) = \mathcal{N}(0, I), p_\theta(x_{t-1}|x_t, c) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, c), \Sigma_t)$$

Furthermore, DDPMs exploits an approximate posterior $q(x_{1:T}|x_0, c)$, namely the forward process, adding Gaussian noise to the data acccording to the variance coefficients $\beta_1, ..., \beta_T$:

$$q(x_{1:T}|x_0, c) = \prod_{t=1}^{T} q(x_t|x_{t-1}),$$

$$q(x_t|x_{t-1}, c) = \mathcal{N}(\sqrt{1-\beta_t}x_{t-1}, \beta_t I),$$

$$\alpha_t = 1 - \beta_t, \widetilde{\alpha_t} = \prod_{i=1}^{t} \alpha_i, \widetilde{\beta_t} = \frac{1-\widetilde{\alpha}_{t-1}}{1-\widetilde{\alpha_t}}.$$

Based on these, in the work [52], parameterization is applied as follows:

$$\mu_\theta(x_t, c) = \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{\beta_t}{\sqrt{1-\widetilde{\alpha_t}}}\epsilon_\theta(x_t, c))$$

A.4 Fine-tuning Text-to-Image Diffusion Models with Reinforcement Learning.

Although reinforcement learning from human feedback has been widely used to align large language models, its application to diffusion models remains largely unexplored. Reward-weighted likelihood maximization [64] proposes a three-stage fine-tuning method that leverages RLHF to enhance the alignment of text-to-image models. Rather than utilizing the reward model in dataset construction process, it is leveraged for the coefficients of loss function. DOODL [65] optimizes the initial diffusion noise vectors with respect to the loss on images generated from the full-chain diffusion, meaning that it improves a single generation iteratively at inference time. DRAFT [66] proposes a simple and effective method for fine-tuning generative models to maximize differentiable reward functions. ReFL [7] utilizes a two-stage approach for diffusion model fine-tuning. In the first stage, leveraging human preference data, a reward model named ImageReward is trained, which is used for guiding the subsequent fine-tuning process. During fine-tuning, ReFL randomly selects timesteps to predict the final image with the purpose of stabilizing the training process and preventing it from focusing solely on the last step. DDPO [8] proposes a reinforcement learning (RL) framework for fine-tuning diffusion models. By defining the denoising process of diffusion models as a MDP problem, it update the pre-trained model with policy gradients to maximize the feedback-trained reward. DPOK [67] is also a RL-based approach to similarly maximize the scored reward; furthermore, DPOK integrates policy optimization with reverse KL regularization for both RL fine-tuning and supervised fine-tuning.

B Typical DPO-based Text-to-Image Diffusion Alignment

Advancing from the significant accomplishments of Direct Preference Optimization (DPO) in alignment, previous researches have explored its application in the application of text-to-image diffusion models, particularly Diffusion-DPO and D3PO, whose efforts established robust paradigms.

B.1 Diffusion-DPO.

Diffusion-DPO [12] offers an enhanced solution to the text-to-image alignment problem by leveraging the DPO algorithm, which is initially proposed for LLM alignment. We can begin by implementing the following symbol conversions:

- The input question to the text input: $x \rightarrow c$;
- The output answer to the generated image: $y \rightarrow x_0$;

- The policy of large language models to the sampling probability of diffusion models: $\pi(y|x) \rightarrow p(x_0|c)$;
- The preference data for output answers to the preference data for generated images: $(x, y_w, y_l) \rightarrow (c, x_0^w, x_0^l)$.

Given the settings, our goal is to optimize $p(x_0|c)$. However, when it comes to applying DPO, challenges are presented particularly for the sake that calculating sampling probability $p(x_0|c)$ requires integration over the whole sampling path $(x_1, x_2, ..., x_T)$ and $p(x_0|c)$ therefore is not computable. Consequently, it modifies the objective into optimizing the distribution of the sampling paths. Based on this, a reinforcement learning (RL)-based objective function is formulated as follows:

$$\arg\max_{p_\theta} \mathbb{E}_{c\sim\mathcal{D}_c, x_{0:T}\sim p_\theta(x_{0:T}|c)} \left[r(c, x_0)\right] - \beta\mathbb{D}_{\text{KL}} \left[p_\theta(x_{0:T}|c)||p_{ref}(x_{0:T}|c)\right] \tag{5}$$

Furthermore, the loss function for Diffusion-DPO can be derived as follows:

$$\mathcal{L}_{\text{Diffusion-DPO}}(\theta) = -\mathbb{E}_{\substack{(c,x_0^w,x_0^l)\sim\mathcal{D},\\ x_{1:T}^w\sim p_\theta(x_{1:T}^w|x_0^w,c),\\ x_{1:T}^l\sim p_\theta(x_{1:T}^l|x_0^l,c)}} \log\sigma\left(\beta\log\frac{p_\theta(x_{0:T}^w|c)}{p_{\text{ref}}(x_{0:T}^w|c)} - \beta\log\frac{p_\theta(x_{0:T}^l|c)}{p_{\text{ref}}(x_{0:T}^l|c)}\right) \tag{6}$$

To enhance the training efficiency, Diffusion-DPO utilizes Jensen's inequality and the convexity of the function $-\log\sigma$ to optimize an upper bound of the original objective function as follows:

$$-\mathbb{E}_{\substack{(c,x_0^w,x_0^l)\sim\mathcal{D},t\sim\mathcal{U}(0,T),\\ x_{t-1,t}^w\sim p_\theta(x_{t-1,t}^w|x_0^w,c),\\ x_{t-1,t}^l\sim p_\theta(x_{t-1,t}^l|x_0^l,c).}} \log\sigma\left(\beta T\log\frac{p_\theta(x_{t-1}^w|x_t^w,c)}{p_{\text{ref}}(x_{t-1}^w|x_t^w,c)} - \beta T\log\frac{p_\theta(x_{t-1}^l|x_t^l,c)}{p_{\text{ref}}(x_{t-1}^l|x_t^l,c)}\right) \tag{7}$$

B.2 Direct Preference for Denoising Diffusion Policy Optimization (D3PO).

D3PO [13] approaches the denoising process as a multi-step Markov Decision Process (MDP) and using the following mapping relationship:

$$\begin{gathered} s_t = (c, t, x_{T-t}); a_t = x_{T-1-t}; \\ P(s_{t+1}|s_t, a_t) = (\delta_c, \delta_{t+1}, \delta_{x_{T-1-t}}); \rho_0(s_0) = (p(c), \delta_0, \mathcal{N}(0, I)); \\ \pi(a_t|s_t) = p_\theta(x_{T-1-t}|x_{T-t}, c); r(s_t, a_t) = r((c, t, x_{T-t}), x_{T-t-1}) \end{gathered} \tag{8}$$

where δ_x represents the Dirac delta distribution, and T denotes the maximize denoising timesteps. It sets up a kind of sparse reward: $\forall s_t, a_t$, $r(s_t, a_t) = 1$ for preferred, while $r(s_t, a_t) = -1$ for dispreferred.

Furthermore, D3PO posits that preference for one segment implies that all state-action pairs within the segment are considered superior to those in the other segment. Under such assumption, T sub-segments can be conducted for the alignment process efficiently:

$$\sigma_i = \{s_i, a_i, s_{i+1}, a_{i+1}, ..., s_{T-1}, a_{T-1}\} \quad 0 \leq i \leq T-1$$

And the overall loss of D3PO algorithm can be calculated with these sub-segments as follows:

$$\mathcal{L}_i(\theta) = -\mathbb{E}_{(s_i,\sigma_w^i,\sigma_l^i)} \log\rho\Big(\beta\log\frac{\pi_\theta(a_i^w|s_i^w)}{\pi_{\text{ref}}(a_i^w|s_i^w)} - \beta\log\frac{\pi_\theta(a_i^l|s_i^l)}{\pi_{\text{ref}}(a_i^l|s_i^l)}\Big) \tag{9}$$

where $i \in [0, T-1]$; $\sigma_w^i = \{s_i^w, a_i^w, ..., s_{T-1}^w, a_{T-1}^w\}$ denotes the segment preferred over the other segment $\sigma_l^i = \{s_i^l, a_i^l, ..., s_{T-1}^l, a_{T-1}^l\}$.

B.3 Step-aware Preference Optimization (SPO).

Contrary to the prevailing assumption that a uniform preference ordering across all stages of the diffusion process aligns with the final output images, Step-aware Preference Optimization (SPO) posits that this assumption fails to account for the nuanced effectiveness of denoising at each individual stage. SPO addresses such limitation by employing a step-aware preference model and a step-wise resampler. At the t-th denoising timestep, a small set $\{x_{t-1}^1, x_{t-1}^2, ..., x_{t-1}^k\}$ is sampled, from which a preference pair (x_{t-1}^w, x_{t-1}^l) is established by selecting the most preferred item x_{t-1}^w

and the most dispreferred one x_{t-1}^l. A set of preference pairs can be obtained at the t-th timestep by sampling from various prompts. And the DPO loss at the t-th timestep can be expressed as follows:

$$\mathcal{L}_t(\theta) = -\mathbb{E}_{(x_{t-1}^w, x_{t-1}^l)\sim p_\theta(x_{t-1}|c,t,x_t)} \log \sigma \left(\beta \log \frac{p_\theta(x_{t-1}^w|c,t,x_t)}{p_{\text{ref}}(x_{t-1}^w|c,t,x_t)} - \beta \log \frac{p_\theta(x_{t-1}^l|c,t,x_t)}{p_{\text{ref}}(x_{t-1}^l|c,t,x_t)} \right) \tag{10}$$

where c refers to the prompt and $p(c)$ is the distribution of the prompts.
Furthermore, the final SPO objective for all T timesteps can be derived as:

$$\begin{aligned} \mathcal{L}(\theta) = -\mathbb{E}_{t\sim\mathcal{U}[1,T],c\sim p(c),x_T\sim\mathcal{N}(0,I)(x_{t-1}^w,x_{t-1}^l)\sim p_\theta(x_{t-1}|c,t,x_t)} \\ \log \sigma \left(\beta \log \frac{p_\theta(x_{t-1}^w|c,t,x_t)}{p_{\text{ref}}(x_{t-1}^w|c,t,x_t)} - \beta \log \frac{p_\theta(x_{t-1}^l|c,t,x_t)}{p_{\text{ref}}(x_{t-1}^l|c,t,x_t)} \right) \end{aligned} \tag{11}$$

C Detailed Mathematical Derivation

In this section, we will provide detailed proofs of Theorems.

Theorem 1. *If $p_{\text{ref}}(x_{0:T}|c) > 0$ holds for all condition c, $f'(x)$ is an invertible function and 0 is not in definition domain of function $f'(x)$, the reward class consistent with Bradley-Terrry model can be reparameterized with the policy preference $p_\theta(x_{0:T})$ and the reference preference $p_{\text{ref}}(x_{0:T}|c)$ as:*

$$r(c, x_0) = \beta f'\Big(\frac{p_\theta(x_{0:T})}{p_{\text{ref}}(x_{0:T}|c)}\Big) + \text{const} \tag{12}$$

Proof. Consider the following optimal problem:

$$\begin{aligned} \max_{p_\theta} \mathbb{E}_{c\sim p_c, x_{0:T}\sim p_\theta(x_{0:T}|c)}[r(c, x_{0:T})] - \beta D_f\Big(p_\theta(x_{0:T}|c), p_{\text{ref}}(x_{0:T}|c)\Big) \\ \text{s.t.} \sum_{x_{0:T}} p_\theta(x_{0:T}|c) = 1;\ \forall x_{0:T} \quad p_\theta(x_{0:T}|c) \geq 0 \end{aligned} \tag{13}$$

Defining the Lagrange function as:

$$\begin{aligned} \mathcal{L}(p_\theta(x_{0:T}|c), \lambda, \zeta(x_{0:T})) = \mathbb{E}_{c\sim p_c, x_{0:T}\sim p_\theta(x_{0:T}|c)}[r(c, x_{0:T})] \\ -\beta\mathbb{E}_{p_{\text{ref}}(x_{0:T}|c)}\left[f\left(\frac{p_\theta(x_{0:T}|c)}{p_{\text{ref}}(x_{0:T}|c)}\right)\right] - \lambda(\sum_{x_{0:T}} p_\theta(x_{0:T}) - 1) + \sum_{x_{0:T}} \zeta(x_{0:T}) p_\theta(x_{0:T}|c) \end{aligned} \tag{14}$$

We conduct the analysis using *Karush-Kuhn-Tucker (KKT) condition* as follows.

Firstly, the stationarity condition necessitates that the gradient of the Lagrangian function with respect to the primal variables be equal to zero:

$$\nabla_{p_\theta(x_{0:T}|c)} \mathcal{L}(p_\theta(x_{0:T}|c), \lambda, \zeta(x_{0:T})) = 0;$$

After performing the calculations, it can be determined that:

$$r(c, x_0) - \beta f'\left(\frac{p_\theta(x_{0:T}|c)}{p_{\text{ref}}(x_{0:T}|c)}\right) - \lambda + \zeta(x_{0:T}) = 0 \tag{15}$$

Hence, we can get the formula of reward class preliminarily:

$$r(c, x_0) = \beta f'\left(\frac{p_\theta(x_{0:T}|c)}{p_{\text{ref}}(x_{0:T}|c)}\right) + \lambda - \zeta(x_{0:T})$$

Furthermore, we would like to consider the dual feasibility, which means the Lagrange multiplier corresponding to inequality constraint must be non-negative:

$$\forall x_{0:T}, \zeta(x_{0:T}) \geq 0$$

And the primal feasibility holds:

$$\sum_{x_{0:T}} p_\theta(x_{0:T}|c) = 1;\ \forall x_{0:T}\ p_\theta(x_{0:T}|c) \geq 0$$

Finally, we would like to consider the complementary slackness, which shows the fact that the inequality constraint must either meet with equality or have Lagrange multipliers that are zero:

$$\forall x_{0:T}, \quad \zeta(x_{0:T}) \cdot p_\theta(x_{0:T}|c) = 0 \tag{16}$$

Since we have made the assumption that 0 is not in the definition domain of function $f'(x)$, which shows the fact that $\frac{p_\theta(x_{0:T}|c)}{p_{\text{ref}}(x_{0:T}|c)} > 0$ always holds true. Moreover, we have assumed that $p_{\text{ref}}(x_{0:T}|c) > 0$ holds for all condition c; hence, we can draw the conclusion that $p_\theta(x_{0:T}|c) > 0$ always holds true. Therefore, we must have:

$$\forall x_{0:T}; \zeta(x_{0:T}) = 0$$

The formula of reward class can be written as:

$$r(c, x_0) = \beta f'\left(\frac{p_\theta(x_{0:T}|c)}{p_{\text{ref}}(x_{0:T}|c)}\right) + \lambda$$

The constant λ in the formula is independent of $x_{0:T}$, which could be canceled out when applying into the Bradley-Terry model. So far, we have completed the proof. □

Theorem 2. *In the substitution process of Bradley-Terry model, the constant λ is independent of $x_{0:T}$ and thus can be canceled out, resulting in the following form:*

$$\mathcal{L}(\theta) = \mathbb{E}_{\substack{(c,x_0^w,x_0^l)\sim\mathcal{D},\\ x_{1:T}^w\sim p_\theta(x_{1:T}^w|x_0^w,c),\\ x_{1:T}^l\sim p_\theta(x_{1:T}^l|x_0^l,c).}} - \log\sigma\left[\beta f'\left(\frac{p_\theta(x_{0:T}^w|c)}{p_{ref}(x_{0:T}^w|c)}\right) - \beta f'\left(\frac{p_\theta(x_{0:T}^l|c)}{p_{ref}(x_{0:T}^l|c)}\right)\right] \tag{17}$$

where $\sigma(\cdot)$ is the Sigmoid function; $f'(\cdot)$ represents the derivatives of $f(\cdot)$; β is the penalty coefficient.

Proof. We know that the Bradley-Terry (BT) model provides a framework for representing human preferences as a function of pairwise comparisons:

$$p_{\text{BT}}(x_{0:T}^w \succ x_0^l) = \sigma(r_\phi(c, x_{0:T}^w) - r_\phi(c, x_{0:T}^l))$$

where $r_\phi(c, \cdot)$ represents reward function reparameterized by network ϕ. Furthermore, the loss function can be written as maximum likelihood formula for binary classification:

$$\mathcal{L}_{\text{BT}} = -\mathbb{E}_{c,x_{0:T}^w,x_{0:T}^l}[\log\sigma(r_\phi(c, x_{0:T}^w) - r_\phi(c, x_{0:T}^l))].$$

Plugging Equation (12) into aforementioned loss function, canceling out the constant λ, and we can get the generalized formula:

$$\mathcal{L}(\theta) = \mathbb{E}_{\substack{(c,x_0^w,x_0^l)\sim\mathcal{D},\\ x_{1:T}^w\sim p_\theta(x_{1:T}^w|x_0^w,c),\\ x_{1:T}^l\sim p_\theta(x_{1:T}^l|x_0^l,c)}} - \log\sigma\left[\beta f'\left(\frac{p_\theta(x_{0:T}^w|c)}{p_{\text{ref}}(x_{0:T}^w|c)}\right) - \beta f'\left(\frac{p_\theta(x_{0:T}^l|c)}{p_{\text{ref}}(x_{0:T}^l|c)}\right)\right]$$

□

Concentrating instead on a more general formulation of the loss function as follows:

$$\mathcal{L}_f\left(\mathrm{X}_1, \mathrm{X}_2\right) = -\mathbb{E}\Big[\log\sigma\big(\beta f'(\mathrm{X}_1) - \beta f'(\mathrm{X}_2)\big)\Big], \tag{18}$$

where X_1 is the training win ratio, and is equivalent to $\frac{p_\theta(x_{0:T}^w|c)}{p_{\text{ref}}(x_{0:T}^w|c)}$; similarly, X_2 is the training loss ratio, and is identical to $\frac{p_\theta(x_{0:T}^l|c)}{p_{\text{ref}}(x_{0:T}^l|c)}$.

Theorem 3. *The partial derivatives (gradients) of* X_1 *and* X_2 *resulting from Equation* (3) *can be expressed as follows:*

$$\begin{cases} \dfrac{\partial \mathcal{L}_f(\mathrm{X}_1, \mathrm{X}_2)}{\partial \mathrm{X}_1} = -\beta \left(1 - \sigma\left(\beta f'(\mathrm{X}_1) - \beta f'(\mathrm{X}_2)\right)\right) f''(\mathrm{X}_1) \\ \dfrac{\partial \mathcal{L}_f(\mathrm{X}_1, \mathrm{X}_2)}{\partial \mathrm{X}_2} = \beta \left(1 - \sigma\left(\beta f'(\mathrm{X}_1) - \beta f'(\mathrm{X}_2)\right)\right) f''(\mathrm{X}_2) \end{cases}$$

Thus, the gradient ratio of $\mathcal{L}_f(\mathrm{X}_1, \mathrm{X}_2)$ *between enhancement in probability for human-preferred responses (*X_1*) and reduction in probability for human-dispreferred responses (*X_2*) has the expression:*

$$\left| \frac{\partial \mathcal{L}_f(\mathrm{X}_1, \mathrm{X}_2)}{\partial \mathrm{X}_1} \Big/ \frac{\partial \mathcal{L}_f(\mathrm{X}_1, \mathrm{X}_2)}{\partial \mathrm{X}_2} \right| = \frac{f''(\mathrm{X}_1)}{f''(\mathrm{X}_2)} \tag{19}$$

Proof. It is known that derivative of sigmoid function is given by the following equation:

$$\sigma(x)' = \sigma(x) \cdot (1 - \sigma(x))$$

Hence,

$$\begin{aligned} &\frac{\partial \mathcal{L}_f(\mathrm{X}_1, \mathrm{X}_2)}{\partial \mathrm{X}_1} \\ =& -\frac{1}{\sigma\left(\beta f'(\mathrm{X}_1) - \beta f'(\mathrm{X}_2)\right)} \cdot \sigma\left(\beta f'(\mathrm{X}_1) - \beta f'(\mathrm{X}_2)\right) \cdot \left(1 - \sigma\left(\beta f'(\mathrm{X}_1) - \beta f'(\mathrm{X}_2)\right)\right) \cdot \beta f''(\mathrm{X}_1) \\ =& -\left(1 - \sigma\left(\beta f'(\mathrm{X}_1) - \beta f'(\mathrm{X}_2)\right)\right) \cdot \beta f''(\mathrm{X}_1) \\ =& -\beta(1 - \sigma\left(\beta f'(\mathrm{X}_1) - \beta f'(\mathrm{X}_2)\right)) f''(\mathrm{X}_1) \end{aligned}$$

$$\begin{aligned} &\frac{\partial \mathcal{L}_f(\mathrm{X}_1, \mathrm{X}_2)}{\partial \mathrm{X}_2} \\ =& \frac{1}{\sigma\left(\beta f'(\mathrm{X}_1) - \beta f'(\mathrm{X}_2)\right)} \cdot \sigma\left(\beta f'(\mathrm{X}_1) - \beta f'(\mathrm{X}_2)\right) \cdot \left(1 - \sigma\left(\beta f'(\mathrm{X}_1) - \beta f'(\mathrm{X}_2)\right)\right) \cdot \beta f''(\mathrm{X}_2) \\ =& \left(1 - \sigma\left(\beta f'(\mathrm{X}_1) - \beta f'(\mathrm{X}_2)\right)\right) \cdot \beta f''(\mathrm{X}_2) \\ =& \beta(1 - \sigma\left(\beta f'(\mathrm{X}_1) - \beta f'(\mathrm{X}_2)\right)) f''(\mathrm{X}_2) \end{aligned}$$

Thus,

$$\left| \frac{\partial \mathcal{L}_f(\mathrm{X}_1, \mathrm{X}_2)}{\partial \mathrm{X}_1} \Big/ \frac{\partial \mathcal{L}_f(\mathrm{X}_1, \mathrm{X}_2)}{\partial \mathrm{X}_2} \right| = \left| \frac{-\beta \left(1 - \sigma\left(\beta f'(\mathrm{X}_1) - \beta f'(\mathrm{X}_2)\right)\right) f''(\mathrm{X}_1)}{\beta \left(1 - \sigma\left(\beta f'(\mathrm{X}_1) - \beta f'(\mathrm{X}_2)\right)\right) f''(\mathrm{X}_2)} \right| = \frac{f''(\mathrm{X}_1)}{f''(\mathrm{X}_2)}$$

which completes the proof. □

Remark 3.1. *If the divergence is* ***Reverse KL divergence****, the aforementioned equation* (4) *transforms into:*

$$\left| \frac{\partial \mathcal{L}_f(\mathrm{X}_1, \mathrm{X}_2)}{\partial \mathrm{X}_1} \Big/ \frac{\partial \mathcal{L}_f(\mathrm{X}_1, \mathrm{X}_2)}{\partial \mathrm{X}_2} \right| = \frac{\mathrm{X_2}}{\mathrm{X_1}}$$

If the divergence is ***Jensen-Shannon divergence****, the aforementioned equation* (4) *transforms into:*

$$\left| \frac{\partial \mathcal{L}_f(\mathrm{X}_1, \mathrm{X}_2)}{\partial \mathrm{X}_1} \Big/ \frac{\partial \mathcal{L}_f(\mathrm{X}_1, \mathrm{X}_2)}{\partial \mathrm{X}_2} \right| = \frac{\mathrm{X_2} \cdot (\mathrm{X_2} + \mathbf{1})}{\mathrm{X_1} \cdot (\mathrm{X_1} + \mathbf{1})}$$

If the divergence is α***-divergence****, the aforementioned equation* (4) *transforms into:*

$$\left| \frac{\partial \mathcal{L}_f(\mathrm{X}_1, \mathrm{X}_2)}{\partial \mathrm{X}_1} \Big/ \frac{\partial \mathcal{L}_f(\mathrm{X}_1, \mathrm{X}_2)}{\partial \mathrm{X}_2} \right| = \frac{\mathrm{X_2^{1+\alpha}}}{\mathrm{X_1^{1+\alpha}}}$$

If the divergence is ***forward KL divergence****, the aforementioned equation* (4) *transforms into:*

$$\left| \frac{\partial \mathcal{L}_f(\mathrm{X}_1, \mathrm{X}_2)}{\partial \mathrm{X}_1} \Big/ \frac{\partial \mathcal{L}_f(\mathrm{X}_1, \mathrm{X}_2)}{\partial \mathrm{X}_2} \right| = \frac{\mathrm{X_2^2}}{\mathrm{X_1^2}}$$

Fact 1. *For any pairwise preference data,* $\mathrm{X}_2/\mathrm{X}_1 < 1$ *always holds. As optimization advances, the value of* X_1 *tends to increase to more than 1, whereas* X_2 *tends to decrease to less than 1.*

Theorem 4. *As optimization progresses, we have* $\mathrm{X}_2/\mathrm{X}_1 < 1$*. Hence,*

$$0 < \frac{\mathrm{X}_2^2}{\mathrm{X}_1^2} < \frac{\mathrm{X}_2 \cdot (\mathrm{X}_2+1)}{\mathrm{X}_1 \cdot (\mathrm{X}_1+1)} < \frac{\mathrm{X}_2}{\mathrm{X}_1} < 1 \quad \textit{and} \quad 0 < \frac{\mathrm{X}_2^2}{\mathrm{X}_1^2} < \frac{\mathrm{X}_2^{1.8}}{\mathrm{X}_1^{1.8}} < \frac{\mathrm{X}_2^{1.6}}{\mathrm{X}_1^{1.6}} < \frac{\mathrm{X}_2^{1.4}}{\mathrm{X}_1^{1.4}} < \frac{\mathrm{X}_2^{1.2}}{\mathrm{X}_1^{1.2}} < \frac{\mathrm{X}_2}{\mathrm{X}_1} < 1$$

Proof. Setting $g(x) = a^x$, where $0 < a < 1$, and we have $g'(x) = lna \cdot a^x \ < 0$ always holds.

Thus, $g(x)$ is a monotone decreasing function, and we can easily derive that

$$0 < \left(\frac{\mathrm{X}_2}{\mathrm{X}_1}\right)^2 < \left(\frac{\mathrm{X}_2}{\mathrm{X}_1}\right)^{1.8} < \left(\frac{\mathrm{X}_2}{\mathrm{X}_1}\right)^{1.6} < \left(\frac{\mathrm{X}_2}{\mathrm{X}_1}\right)^{1.4} < \left(\frac{\mathrm{X}_2}{\mathrm{X}_1}\right)^{1.2} < \frac{\mathrm{X}_2}{\mathrm{X}_1} < 1$$

Furthermore, we know that $\mathrm{X}_2/\mathrm{X}_1 < 1$, i.e. $\mathrm{X}_2 < \mathrm{X}_1$, and then $(\mathrm{X}_2+1)/(\mathrm{X}_1+1) < 1$; therefore,

$$0 < \frac{\mathrm{X}_2^2}{\mathrm{X}_1^2} < \frac{\mathrm{X}_2 \cdot (\mathrm{X}_2+1)}{\mathrm{X}_1 \cdot (\mathrm{X}_1+1)} < \frac{\mathrm{X}_2}{\mathrm{X}_1} < 1$$

□

D Alternate Derivation: From the Rinforcement Learning Perspective

In this section, we aim to further derive the generalized formula under f-divergence from a reinforcement learning perspective. Here, we adopt the premise and setup of D3PO [13], which regarding the process as a multi-step Markov Decision Process (MDP) and using the following mapping relationship:

$$\begin{gathered} s_t = (c, t, x_{T-t}); a_t = x_{T-1-t}; \\ P(s_{t+1}|s_t, a_t) = (\delta_c, \delta_{t+1}, \delta_{x_{T-1-t}}); \rho_0(s_0) = (p(c), \delta_0, \mathcal{N}(0, I)); \\ \pi(a_t|s_t) = p_\theta(x_{T-1-t}|x_{T-t}, c); r(s_t, a_t) = r((c, t, x_{T-t}), x_{T-t-1}) \end{gathered}$$

where δ_x represents the Dirac delta distribution, and T denotes the maximize denoising timesteps. It sets up a kind of sparse reward: $\forall s_t, a_t,\ r(s_t, a_t) = 1$ for preferred, while $r(s_t, a_t) = -1$ for dispreferred.

Theorem 5. *If* $\pi_{ref}(a|s)$ *holds for all* s*,* $f'(x)$ *is an invertible function and* 0 *is not in the definition domain of function* $f'(x)$*, the optimal policy* $\pi^*(a|s)$ *has the expression of:*

$$\pi^*(a|s) = \pi_{ref}(a|s) \cdot (f')^{-1}\left(\frac{Q^*(s,a) - \lambda}{\beta}\right)$$

where $(f')^{-1}$ *is the inverse function of the derivative of function* $f(x)$*;* λ *is a fixed, constant term that is independent of* a*.*

Proof. Consider the following optimal problem:

$$\max_\pi \mathbb{E}_{s\sim d^\pi, a\sim\pi(\cdot|s)}\left[Q^*(s,a)\right] - \beta D_f\left[\pi(a|s), \pi_{\text{ref}}(a|s)\right]$$

$$s.t. \sum_a \pi(a|s) = 1; \forall a\, \pi(a|s) \geq 0$$

where $Q^*(s,a)$ is the optimal action-value function; $d^\pi = (1-\gamma)\sum_{t=0}^\infty \gamma^t P_t^\pi(s)$ represents the state visitation distribution; s, a, π, P_t^π adhere to the definitions outlined in equation (8).
Defining the following Lagrange function:

$$\begin{aligned} \mathcal{L}(\pi(a|s), \lambda, \xi(a)) = {} & \mathbb{E}_{s\sim d^\pi, a\sim\pi(\cdot|s)}\left[Q^*(s,a)\right] \\ & -\beta\mathbb{E}_{\pi_{\text{ref}}(a|s)}\left[f\left(\frac{\pi(a|s)}{\pi_{\text{ref}}(a|s)}\right)\right] - \lambda\left(\sum_a \pi(a|s) - 1\right) + \xi(a)\pi(a|s) \end{aligned} \tag{20}$$

Employing the *Karush-Kuhn-Tucker (KKT) conditions* for analysis:
Firstly, the stationarity condition necessitates that the gradient of the Lagrangian function with respect to the primal variables should be zero:

$$\nabla_{\pi(a|s)}\mathcal{L}(\pi(a|s), \lambda, \xi(a)) = 0$$

Performing the calculation, we can get:

$$Q^*(s,a) - \beta f'\left(\frac{\pi(a|s)}{\pi_{\text{ref}}(a|s)}\right) = \lambda - \xi(a)$$

Furthermore, considering the dual feasibility, which stipulates that the Lagrange multiplier associated with an inequality constraint must adhere to a non-negative condition:

$$\forall a,\ \xi(a) \geq 0$$

And the primal feasibility shows that:

$$\sum_a \pi(a|s) = 1; \forall a\, \pi(a|s) \geq 0$$

Moreover, thinking about the complementary slackness, which dictates that for an inequality constraint, either the constraint must be satisfied with equality, or its corresponding Lagrange multiplier must be zero:

$$\forall a;\ \pi(a|s) \cdot \xi(a) = 0$$

Given that 0 is not in the definition domain of $f'(x)$, it follows the fact that $\frac{\pi(a|s)}{\pi_{\text{ref}(a|s)}} > 0$ always holds. Moreover, we have the assumption that $\pi_{\text{ref}}(a|s) > 0$ is satisfied. Hence, there must be $\pi(a|s) > 0$. From the analysis that has been conducted, the subsequent conclusion is attainable:

$$\forall a;\ \xi(a) = 0$$

Substituting the above conclusion into the stationarity condition yields:

$$f'\left(\frac{\pi(a|s)}{\pi_{\text{ref}}(a|s)}\right) = \frac{Q^*(s,a) - \lambda}{\beta}$$

Through certain algebraic computation, we can derive:

$$\pi^*(a|s) = \pi_{\text{ref}}(a|s) \cdot (f')^{-1}\left(\frac{Q^*(s,a) - \lambda}{\beta}\right)$$

So far, we have completed the proof. □

Remark 5.1. *Rearranging the equation in Theorem 5, we can obtain the following formula:*

$$Q^*(s,a) = \beta f'\left(\frac{\pi(a|s)}{\pi_{\textit{ref}}(a|s)}\right) + \lambda$$

Substituting the result we obtained in Remark 5.1 into the Bradley-Terry model [10], we can similarly eliminate the constant λ and gain the final generalized formula.

E Further Analysis on the Gradient Fields

In the work [33], it is shown that the original DPO (alignment of LLMs) increasingly loses its ability to steer the direction of response optimization in LLM alignment as the alignment process advances; in other words, it risks degenerating into a mechanism that merely learns the rejected responses, rather than actively shaping the chosen responses' trajectory towards alignment. We would like to further investigate whether such phenomena continued exist in the context of text-to-image generation alignment with diverse divergence constraints.

Remark 5.2. *If the divergence is* ***Reverse KL divergence****, equations in Theorem 3 can be simplified as:*

$$\begin{cases} \dfrac{\partial \mathcal{L}_f\left(\mathrm{X}_1, \mathrm{X}_2\right)}{\partial \mathrm{X}_1} = -\beta \dfrac{\mathrm{X}_2^{\beta}}{\mathrm{X}_1 \cdot \left({\mathrm{X}_1}^{\beta} + {\mathrm{X}_2}^{\beta}\right)} \\ \dfrac{\partial \mathcal{L}_f\left(\mathrm{X}_1, \mathrm{X}_2\right)}{\partial \mathrm{X}_2} = \beta \dfrac{\mathrm{X}_2^{\beta-1}}{\left({\mathrm{X}_1}^{\beta} + {\mathrm{X}_2}^{\beta}\right)} \end{cases}$$

The main cause of the aforementioned described phenomenon, known as *model learning degradation*, occurs as $\mathrm{X}_2 \to 0$ and $\beta < 1$, resulting in $\frac{\partial \mathcal{L}_f(\mathrm{X}_1, \mathrm{X}_2)}{\partial \mathrm{X}_1}$ tends 0 for the sake of $\mathrm{X}_2^{\beta} \to 0$, while $\frac{\partial \mathcal{L}_f(\mathrm{X}_1, \mathrm{X}_2)}{\partial \mathrm{X}_2}$ tends infinity as a consequence of $\mathrm{X}_2^{\beta-1} \to \infty$. In fact, in our scenario, the aforementioned phenomenon is mitigated by our typical practice of assigning a relative large value to β (β =10 in the experiments of our work), thus prevent it focus on unlearning rejected items only.

Remark 5.3. *If the divergence is* ***Jensen-Shannon divergence****, equations in Theorem 3 can be simplified as:*

$$\begin{cases} \dfrac{\partial \mathcal{L}_f\left(\mathrm{X}_1, \mathrm{X}_2\right)}{\partial \mathrm{X}_1} = -\beta \cdot \dfrac{1}{\mathrm{X}_1} \cdot \dfrac{\mathrm{X}_2^{\beta}\left(1+\mathrm{X}_1\right)^{\beta-1}}{\mathrm{X}_2^{\beta}(1+\mathrm{X}_1)^{\beta} + \mathrm{X}_1^{\beta}(1+\mathrm{X}_2)^{\beta}} \\ \dfrac{\partial \mathcal{L}_f\left(\mathrm{X}_1, \mathrm{X}_2\right)}{\partial \mathrm{X}_2} = -\beta \cdot \dfrac{1}{\mathrm{X}_2+1} \cdot \dfrac{\mathrm{X}_2^{\beta-1}\left(1+\mathrm{X}_1\right)^{\beta}}{\mathrm{X}_2^{\beta}(1+\mathrm{X}_1)^{\beta} + \mathrm{X}_1^{\beta}(1+\mathrm{X}_2)^{\beta}} \end{cases}$$

When $\mathrm{X}_2 \to 0$*, if* $\beta > 1$*, both* $\frac{\partial \mathcal{L}_f(\mathrm{X}_1, \mathrm{X}_2)}{\partial \mathrm{X}_1} \to 0$ *and* $\frac{\partial \mathcal{L}_f(\mathrm{X}_1, \mathrm{X}_2)}{\partial \mathrm{X}_2} \to 0$ *simultaneously; conversely, if* $\beta < 1$*, the result would be* $\frac{\partial \mathcal{L}_f(\mathrm{X}_1, \mathrm{X}_2)}{\partial \mathrm{X}_1} \to 0$ *but* $\frac{\partial \mathcal{L}_f(\mathrm{X}_1, \mathrm{X}_2)}{\partial \mathrm{X}_2} \to \infty$.

Remark 5.3 indicates that when the *Jensen-Shannon divergence* is selected as an regularization, choosing a value of β *greater than 1* is advantageous.

Remark 5.4. *If the divergence is* α***-divergence****, equations in Theorem 3 can be simplified as:*

$$\begin{cases} \dfrac{\partial \mathcal{L}_f\left(\mathrm{X}_1, \mathrm{X}_2\right)}{\partial \mathrm{X}_1} = -\beta \cdot \dfrac{1}{\mathrm{X}_1^{1+\alpha}} \cdot \dfrac{e^{\beta/\alpha \cdot \frac{1}{\mathrm{X}_1^{\alpha}}}}{e^{\beta/\alpha \cdot \frac{1}{\mathrm{X}_1^{\alpha}}} + e^{\beta/\alpha \cdot \frac{1}{\mathrm{X}_2^{\alpha}}}} \\ \dfrac{\partial \mathcal{L}_f\left(\mathrm{X}_1, \mathrm{X}_2\right)}{\partial \mathrm{X}_2} = \beta \cdot \dfrac{1}{\mathrm{X}_2^{1+\alpha}} \cdot \dfrac{e^{\beta/\alpha \cdot \frac{1}{\mathrm{X}_1^{\alpha}}}}{e^{\beta/\alpha \cdot \frac{1}{\mathrm{X}_1^{\alpha}}} + e^{\beta/\alpha \cdot \frac{1}{\mathrm{X}_2^{\alpha}}}} \end{cases}$$

Given that both α *and* β *are positive values, it follows that* $\frac{\partial \mathcal{L}_f(\mathrm{X}_1, \mathrm{X}_2)}{\partial \mathrm{X}_1} \to 0$ *and* $\frac{\partial \mathcal{L}_f(\mathrm{X}_1, \mathrm{X}_2)}{\partial \mathrm{X}_2} \to 0$ *always hold true when* $\mathrm{X}_2 \to 0$*.*

Remark 5.5. *If the divergence is* ***Forward KL divergence****, equations in Theorem 3 can be simplified as:*

$$\begin{cases} \dfrac{\partial \mathcal{L}_f\left(\mathrm{X}_1, \mathrm{X}_2\right)}{\partial \mathrm{X}_1} = -\beta \cdot \dfrac{1}{\mathrm{X}_1^2} \cdot \dfrac{e^{\frac{\beta}{\mathrm{X}_1}}}{e^{\frac{\beta}{\mathrm{X}_1}} + e^{e^{\frac{\beta}{\mathrm{X}_2}}}} \\ \dfrac{\partial \mathcal{L}_f\left(\mathrm{X}_1, \mathrm{X}_2\right)}{\partial \mathrm{X}_2} = \beta \cdot \dfrac{1}{\mathrm{X}_2^2} \cdot \dfrac{e^{\frac{\beta}{\mathrm{X}_1}}}{e^{\frac{\beta}{\mathrm{X}_1}} + e^{\frac{\beta}{\mathrm{X}_2}}} \end{cases}$$

$\forall \beta$*,* $\frac{\partial \mathcal{L}_f(\mathrm{X}_1, \mathrm{X}_2)}{\partial \mathrm{X}_1} \to 0$ *and* $\frac{\partial \mathcal{L}_f(\mathrm{X}_1, \mathrm{X}_2)}{\partial \mathrm{X}_2} \to 0$ *always hold true when* $\mathrm{X}_2 \to 0$*.*

According to Remark 5.4 and Remark 5.5, if the regularization is in terms of α-divergence or Forward KL divergence, then no matter the value of β, it will not result in model training degradation.

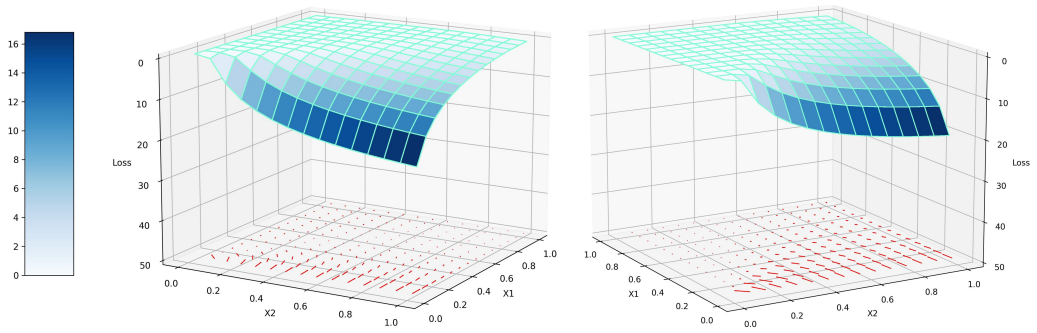
Moreover, we would like to further discuss the relationship between generation diversity and gradient field. Within the work [13], the property of dispersion effect on unseen generations of original DPO has been proposed. It elucidates that as X_2 rapidly decreases to 0, the gradient on X_1 will gradually diminish, consequently leading to a stochastic decline in the likelihood of the selected response.

Such dispersion effect contributes to the genration diversity. Hence, for the sake that forward KL divergence has the minimal gradient ratio and reverse KL divergence has the maximal gradient ratio, should theoretically resulting in optimal alignment diversity for forward KL divergence and poorest alignment diversity for reverse KL divergence. In fact, [21, 20] does reach such a analogous conclusion from a practical point of view in the LLM alignment task.

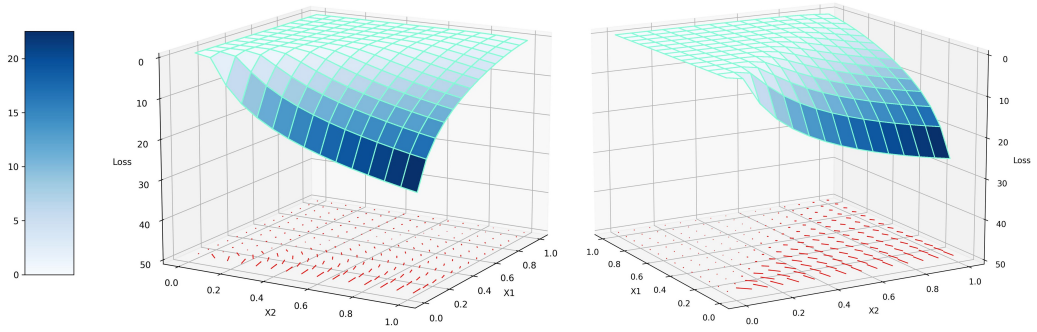
However, we should also be mindful of two aspects. Firstly, in the task of LLM alignment, typically only one epoch is conducted, thus conforming well to the aforementioned theory. Nevertheless, in the task of Text-to-Image generation alignment, multiple epochs are often performed (e.g., 10 epochs in Diffusion-DPO, SPO and experiments of our work; 1000 epochs in D3PO), which renders the diversity variations caused by the gradient ratio negligible after training for multiple epochs. Secondly, the contextual dimension of images is higher than that of text, and the evaluation indicators for image diversity often focus on different aspects of images. In our experiments, we observe that after sufficient training, $\alpha-$divergence (α=0.6) generally achieves the best generation diversity. However, it is worth noting that while it achieves the optimal generation diversity, it performs worst in terms of human value alignment performance. And we can intuitively find that generation diversity and alignment performance are a pair of conflicting entities. To achieve the best trade-off between the two in alignment, we should first pursue better alignment performance, and then, on the basis of assured alignment performance, pursue better generation diversity. Based on a comprehensive theoretical examination and empirical evidence from experimental results, we advocate for regarding **Jensen-Shannon divergence** as the first choice in practice.

F Plot of Gradient Fields and Visualization of Landscapes

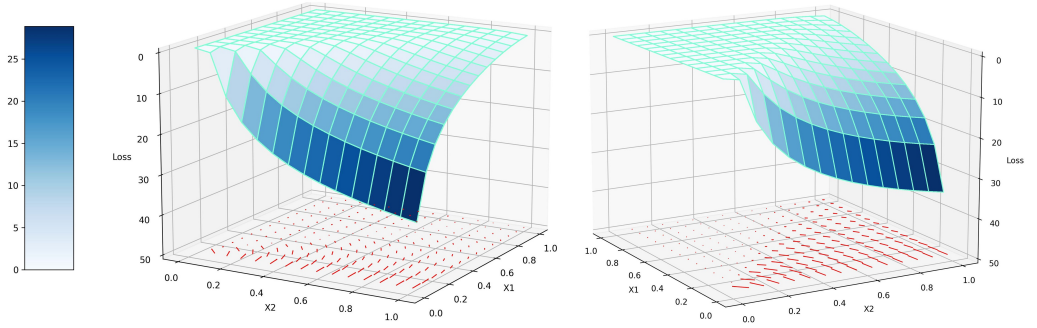
In order to obtain a more intuitive understanding of the impact of different divergence choices during the alignment process, we visualize the landscape of alignment objective functions with different divergences from two viewing angles, as shown in Figure 2 (the penalty coefficient β is selected as 10). Furthermore, to enhance intuition, we plot the gradient field of corresponding loss function on the plane where Z equals 50. When it comes to consider the smoothness within loss function landscape, surface of *Jensen-Shannon divergence* exhibits the best smoothness, which suggests a more stable alignment process. Moreover, this indicates a more robust alignment mechanism, which helps prevent the process from merely unlearning undesired outputs rather than actively steering chosen outputs towards optimization; and this also mitigates the phenomenon that the gradient on X_1 gradually diminishes as X_2 rapidly decreases to 0, which consequently leads to a stochastic decline in the likelihood of the selected response [33].



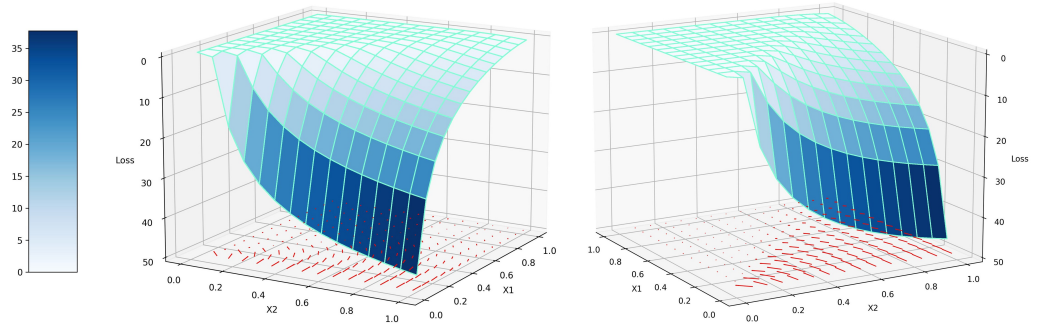
Jensen–Shannon Divergence



Reverse Kullback-Leibler Divergence(α=0)



α Divergence(α=0.2)



α Divergence(α=0.4)

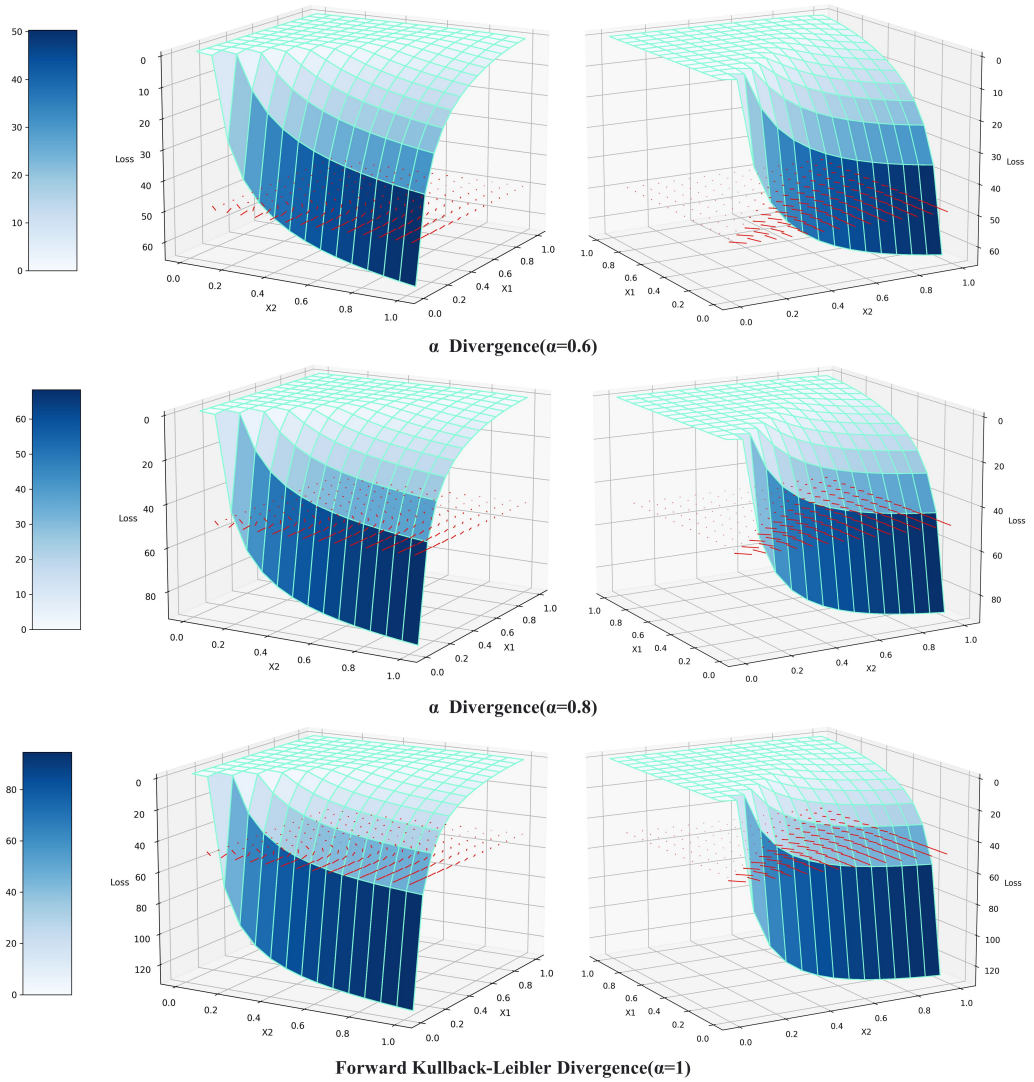


Figure 2: Landscapes' visualization of alignment objective functions with different divergences from two viewing angles and gradient fields' visualization of the corresponding loss function on the plane $Z = 50$.

G Detailed Metric Description

G.1 Alignment Performance Metric.

In this paper, we utilize six metrics for evaluating the alignment performance. We employ the text-image CLIP score [34] and VQAScore [35] to evaluate the performance of text-image alignment and the Aesthetics score [36], ImageReward [7], PickScore [6] and HPS-v2 [9] to evaluate the performance of human value alignment.

Text-Image CLIP score. The Text-Image CLIP score serves as a quantitative measure for evaluating the likeness between text-image pairs. CLIPScore is fundamentally based on the CLIP model, which transforms input text and images into distinct text and image vectors, followed by calculation of the dot product of these vectors. Foundational aim of the CLIP model is to cultivate versatile multimodal representations, free from specialized domain expertise, through the integration of linguistic indicators and visual data. Training approach of CLIP model mainly hinges on contrastive learning, where the system partitions the incoming text-image pairs into two categories: one cluster includes similar pairs

to the input, whereas the other assembles dissimilar pairs. The model then learns representations of these inputs, with the objective to increment similarity within matching pairs while reducing it between non-matching pairs. Benefiting from its pre-training strategy, it enables the extraction of significant image and text features from vast unsupervised datasets. The CLIP model and CLIPScore have demonstrated commendable performance across a wide range of tasks, encompassing image classification, semantic segmentation, image generation, object localization, video interpretation, and so on.

VQAScore. VQAScore meticulously transforms textual cues into precise inquiries, deploying the generative vision-language models with visual-question-answering (VQA) tasks to evaluate the congruence between the image and the descriptive text. Such innovative approach streamlines the assessment process while markedly enhancing the precision and dependability of evaluations. Furthermore, by utilizing the CLIP-FlanT5 model, VQAScore fosters a reciprocal influence between the visual content and the textual query, aligning more closely with human comprehension of the interplay between the image and the text.

Aesthetics score. The LAION Aesthetics Predictor is utilized to estimate an image's aesthetic score, quantifying the mean human appreciation for its visual appeal. It leverages a neural network architecture (MLP) that takes CLIP embeddings as inputs to ascertain the average preference level for the image. Each image is assigned a score on the scale of 0 to 10, with 0 signifying the least visually attractive and 10 denoting the highest level of visual appeal.

ImageReward. ImageReward, leveraging a structure that combines ViT-L for image encoding and a 12-layer Transformer for text encoding, tackles the challenges of text-to-image generation to some extent, especially regarding the quality of pre-training data, which are plagued by noise and a skewed distribution that doesn't match the data users input in prompts. Notably, as a zero-shot evaluation tool, ImageReward often aligns with human judgments, demonstrating the capability to make nuanced quality comparisons between individual samples.

PickScore. A comprehensive, natural dataset, dubbed "Pick-a-Pic," is compiled and utilizing the dataset, an advanced scoring function, namely "PickScore", is built. "PickScore" excels in assessing generated images against prompts, surpassing not only machine learning models but also expert human evaluations. Its utility spans multiple domains such as model evaluation, image generation enhancement, text-to-image dataset refinement, and the optimization of text-to-image models through methodologies such as Reinforcement Learning Human Feedback (RLHF). PickScore follows the architecture of CLIP; provided a prompt x and an image y, PickScore s calculates a real number through the representation of x with a text encoder and y with an image encoder as two d-dimensional vectors, and subsequently returns their inner product:

$$score(x, y) = \mathrm{E}_{\text{txt}}(x) \cdot \mathrm{E}_{\text{img}}(y) \cdot T$$

where T is the learned temperature parameter of CLIP.

HPS-v2. Human Preference Dataset v2 (HPD-v2) encapsulates human preferences for images sourced from a multitude of platforms. It consists of 798,090 individual human preference choices for 433,760 paired image comparisons. The dataset has been taken care to deliberately collect the text prompts and images to minimize potential biases, a common pitfall in previous datasets. Whereafter, through fine-tuning the CLIP model on HPD-v2, the Human Preference Score v2 (HPS-v2) is derived, a scoring model that can more accurately gauge human preferences for generated images. HPS-v2 has been shown to generalize more effectively than earlier metrics across a variety of image datasets, and it is responsive to algorithmic improvements of text-to-image generative models.

G.2 Generation Diversity Metric

In this work, we utilize eight metrics for comprehensively evaluating generation diversity from diverse aspects: Image-Image CLIP score [34], Image Entropy (Entropy 1D and Entropy 2D) [68], LPIPS [38], RMSE [69], PSNR [70], SSIM [70], FSIM [37].

Image-Image CLIP score. Text-Image CLIP score and Image-Image CLIP score are both grounded in the evaluation of high-dimensional embeddings produced by the CLIP model. Similarly, the Image-Image CLIP score functions as an efficacious metric for evaluating the structural congruity between images, thereby enabling assessments of images' similarity. Hence, we select Image-Image CLIP score as an indicator for the diversity of images produced by the trained diffusion model: a

diminutive CLIP score between two generated images signifies a pronounced disparity, implying that the model demonstrates a heightened capacity for generating diverse content.

Image Entropy (Entropy 1D and Entropy 2D). Image Entropy is a statistical metric employed to evaluate the information content and complexity within an image. It quantifies the average information per pixel, with higher entropy values signaling a greater diversity and richness in the image's information content. One-dimensional image entropy (Entropy 1D) quantifies the information encapsulated within the distribution's clustering properties of gray levels:

$$H_{1d} = \sum_{i=0}^{255} P_i \log P_i$$

where P_i presents the proportion of pixels in the image with gray level value i.

The one-dimensional image entropy (Entropy 1D) successfully captures the aggregation properties of gray level distribution, yet neglects spatial attributes. To rectify this discrepancy, supplementary feature metrics are incorporated, which, in conjunction with the one-dimensional entropy, serve as the cornerstone for the evolution of two-dimensional image entropy (Entropy 2D). Such augmentation facilitates a more holistic evaluation that merges both spatial and distributional data within an image. The neighborhood gray level mean, when chosen as a spatial feature quantity in conjunction with the pixel gray levels, constitutes a feature tuple denoted as (i, j):

$$P_{i,j} = \frac{f(i,j)}{N^2}$$

where i is the gray value of the pixel, and j is the mean gray value of its neighborhood; $f(i, j)$ is the occurrence frequency of characteristic binary (i, j) and N is the dimension of the image. Then the two-dimensional image entropy (Entropy 2D) can be defined as:

$$H_{2d} = \sum_{i=0}^{255} \sum_{j=0}^{255} P_{ij} \log P_{ij}$$

LPIPS. Learned Perceptual Image Patch Similarity (LPIPS) is a deep learning-based metric designed for assessing image similarity, which is calculated based on features output by a deep convolutional neural network (AlexNet in our work). Utilizing the feature representations learned by a deep neural network, which is capable of capturing details of human visual perception such as texture, color, and structure, the computation of perceptual similarity between two images can be conducted. Firstly, pre-trained deep neural networks, notably those like AlexNet [71] or VGG [72], are employed to extract features from input images. The outputs from the network's intermediate layers are then typically utilized, as they encapsulate a spectrum of abstract features, spanning from rudimentary edge and texture details to more sophisticated representations of objects and scenes, denoted as $\hat{y_1^i}$, $\hat{y_2^i} \in \mathbb{R}^{H_l \times W_l \times C_l}$. Subsequently, the distances between extracted features in the feature space can be calculated:

$$d(y_1, y_2) = \sum_l \frac{1}{H_l W_l} \sum_{h,w} \| w_l \odot (\hat{y_1^i} - \hat{y_2^i}) \|_2^2$$

and converted into a comprehensible similarity score by standardizing them to the range of 0 to 1. For our evaluation, LPIPS is selected as a metric; a lower score indicates a higher similarity between images, while a higher score suggests greater diversity or disparity between them.

RMSE. Root Mean Square Error (RMSE) is a statistical metric that primarily employed in statistical analysis and machine learning disciplines, serving as a benchmark for gauging the accuracy of predictions. In our scenario, we utilize the RMSE as a criterion to measure the pixel-level differences between pairs of generated images. Such pixel-level variance can be considered as an indicator of the images' diversity, thereby offering a quantifiable assessment of the generated images' diversity:

$$RMSE = \sqrt{\frac{1}{L \times W} \sum_{i=1}^{L \times W} (p_i - q_i)^2}$$

where L is the length of the image, W is the width of the image; p_i and q_i are i-th pixels of two generated images.

PNSR. Peak Signal-to-Noise Ratio (PSNR), a prevalent metric in image processing and compression, quantitatively assesses the discrepancy between an original image and its modified counterpart. This metric, typically reported in decibels (dB), is characterized by a higher value indicative of a diminished difference between the two images, effectively serving as a metric for evaluating the diversity of generated images:

$$PSNR = 10 \log_{10} \left(\frac{Max^2}{MSE} \right)$$

where Max denotes the maximum pixel value that an image attain, MSE denotes the mean squared error between two images.

SSIM. Structural Similarity Index Measure (SSIM) is a metric designed to assess the likeness between images by emulating the human visual system's perception of image quality. Traditional metrics (e.g. RMSE, PNSR, Image Entropy) usually focus on disparities in pixel values, whereas SSIM incorporates the structural aspects of images for evaluation. It is often executed from three aspects: luminance similarity, contrast similarity, and structural similarity - on a scale from 0 to 1:

$$\begin{aligned} l(x,y) &= \frac{2\mu_x\mu_y + c_1}{\mu_x^2 + \mu_y^2 + c_1}; \\ c(x,y) &= \frac{2\sigma_x\sigma_y + c_2}{\sigma_x^2 + \sigma_y^2 + c_2}; \\ s(x,y) &= \frac{\sigma_{xy} + c_3}{\sigma_x\sigma_y + c_3} \end{aligned}$$

where μ_x and μ_y are means of x and y; σ_x and σ_y are variances of x and y, respectively; and σ_{xy} is the covariance of x and y. SSIM calculates similarity between two images across these three dimensions, providing an overall similarity index ranging from 0 to 1. The closer the value is to 1, the more similar the two images are:

$$SSIM(x,y) = \left[l(x,y)^\alpha \cdot c(x,y)^\beta \cdot s(x,y)^\gamma \right]$$

In typical situations, α, β and γ are all set to 1.

FSIM. Feature Similarity Index Measure (FSIM) utilizes feature similarity for assessment. The Human Visual System (HVS) bases its perception on essential visual attributes, and the phase congruency (PC) feature excels in depicting local structures. Remarkably, PC's resilience to changes in the image context guarantees the stability of feature extraction. Nonetheless, it's recognized that modifications in the image can influence visual perception. Therefore, to augment the comprehensive analysis, gradient features, particularly gradient magnitude (GM), are incorporated. Consequently, in FSIM, both PC and GM features collaborate to serve complementary roles, synergistically capturing a holistic evaluation. For two images, the calculations for $PC1$, $GM1$, $PC2$, and $GM2$ are firstly performed. Subsequently, compute the similarity for PC and for GM as follows:

$$\begin{aligned} S_{PC}(\mathrm{x}) &= \frac{2PC_1(\mathrm{x}) \cdot PC_2(\mathrm{x}) + T_1}{PC_1(\mathrm{x})^2 + PC_2(\mathrm{x})^2 + T_1}; \\ S_{GM}(\mathrm{x}) &= \frac{2GM_1(\mathrm{x}) \cdot GM_2(x) + T_2}{GM_1(x)^2 + GM_2(x)^2 + T_2}. \end{aligned}$$

Furthermore, the similarity expressed by fusion of PC and GM can be given as:

$$S_L(\mathrm{x}) = [S_{PC}(\mathrm{x})]^\alpha \cdot [S_{GM}(\mathrm{x})]^\beta$$

Finally, the calculation of FSIM is described as follows:

$$FSIM = \frac{\sum_{\mathrm{x}\in\Omega} S_L(\mathrm{x}) \cdot PC_m(\mathrm{x})}{\sum_{\mathrm{x}\in\Omega} PC_m(\mathrm{x})}$$

H Further Larger-Scale Evaluation

To further demonstrate the persuasiveness of our conclusions, we conduct additional evaluation on the aligned models. GenAI-Bench [73], serves as a comprehensive benchmark for compositional text-to-visual generation, and we report the evaluation results on GenAI-Bench in Table 4 (alignment performance) and Table 5 (diversity performance). Furthermore, we employ the parti-prompts [74] training dataset and the entire HPS-v2 [9] training set for evaluation, by combining the generated images from these sets together, we report the evaluation results *on all 4832 prompts* in Table 6 (alignment performance) and Table 7 (generation diversity). The results obtained are similar to that in the main paper. **Jensen-Shannon divergence** exhibits the best alignment performance and suboptimal generation diversity, achieving the best trade-off.

Model		CLIPScore ↑	VQAScore ↑	Aesthetics Score ↑	ImageReward ↑	Pickscore ↑	HPS-V2 ↑
Original Model		0.334±0.046	0.638±0.268	5.433±0.427	0.195±0.996	21.446±1.143	27.148±1.463
Reverse KL Divergence		**0.344±0.045**	**0.669±0.265**	5.582±0.408	0.556±0.947	21.889±1.153	27.827±1.452
α-Divergence	α=0.2	0.344±0.046	0.665±0.270	5.607±0.428	0.535±0.949	21.850±1.168	***27.900±1.475***
	α=0.4	0.343±0.045	0.666±0.269	5.547±0.405	0.563±0.918	21.874±1.162	27.803±1.407
	α=0.6	0.340±0.046	0.650±0.275	5.585±0.399	0.486±0.960	21.764±1.158	27.785±1.446
	α=0.8	0.343±0.045	0.661±0.268	5.582±0.436	0.491±0.943	21.821±1.169	27.709±1.448
Forward KL Divergence		0.344±0.046	0.664±0.266	5.589±0.416	0.517±0.942	21.852±1.138	27.854±1.446
Jensen-Shannon Divergence		0.342±0.045	0.661±0.268	**5.649±0.409**	**0.573±0.940**	**21.904±1.158**	**27.880±1.436**

Table 4: Evaluations of the alignment performance with **Gen-AI Benchmark** experiments, where the CLIPScore and VQAScore evaluates image-text alignment performance, and the remaining four metrics evaluate human value alignment performance.

Model		Image-Image CLIPScore ↓	Entropy 1D ↑	Entropy 2D ↑	LPIPS ↑
Original Model		0.8358 ± 0.0916	3.8889 ± 0.1875	7.6543 ± 0.4906	0.3031 ± 0.0388
Reverse KL Divergence		0.8668 ± 0.0843	3.9865 ± 0.1107	7.8165 ± 0.3640	0.3020 ± 0.0355
α-Divergence	α =0.2	0.8683 ± 0.0834	3.9724 ± 0.1298	7.8136 ± 0.4026	0.3121 ± 0.0371
	α =0.4	0.8646 ± 0.0857	**4.0095 ± 0.1018**	7.8478 ± 0.3396	0.3058 ± 0.0336
	α =0.6	**0.8608 ± 0.0864**	3.9634 ± 0.1252	7.8141 ± 0.3904	**0.3185 ± 0.0371**
	α =0.8	0.8653 ± 0.0870	3.9733 ± 0.1398	7.7432 ± 0.4297	0.3088 ± 0.0380
Forward KL Divergence		0.8682 ± 0.0847	3.9804 ± 0.1143	7.7786 ± 0.3660	0.3072 ± 0.0350
Jensen-Shannon Divergence		0.8685 ± 0.0818	3.9886 ± 0.1145	**7.8590 ± 0.3808**	0.3103 ± 0.0359
Model		RMSE ↑	PSNR ↓	SSIM ↓	FSIM ↓
Original Model		0.0133 ± 0.0026	37.686 ± 1.797	0.8838 ± 0.0323	0.3795 ± 0.0213
Reverse KL Divergence		0.0153 ± 0.0025	36.415 ± 1.501	0.8550 ± 0.0326	0.3803 ± 0.0188
α-Divergence	α =0.2	0.0161 ± 0.0025	35.943 ± 1.453	0.8445 ± 0.0339	**0.3755 ± 0.0208**
	α =0.4	0.0155 ± 0.0023	36.288 ± 1.363	0.8540 ± 0.0308	0.3806 ± 0.0180
	α =0.6	**0.0165 ± 0.0025**	**35.757 ± 1.384**	**0.8394 ± 0.0330**	0.3771 ± 0.0209
	α =0.8	0.0153 ± 0.0026	36.417 ± 1.632	0.8553 ± 0.0349	0.3783 ± 0.0210
Forward KL Divergence		0.0157 ± 0.0026	36.170 ± 1.410	0.8501 ± 0.0315	0.3762 ± 0.0189
Jensen-Shannon Divergence		0.0158 ± 0.0026	36.092 ± 1.372	0.8473 ± 0.0313	0.3793 ± 0.0201

Table 5: Evaluations of the generation diversity with **Gen-AI Benchmark**. The metrics originally utilized for evaluating image similarity exhibit an opposite property when evaluating generation diversity.

Model		CLIPScore ↑	VQAScore ↑	Aesthetics Score ↑	ImageReward ↑	Pickscore ↑	HPS-V2 ↑
Original Model		0.343±0.054	0.658±0.251	5.575±0.556	0.231±1.047	21.059±1.216	27.082±1.541
Reverse KL Divergence		**0.353±0.053**	**0.710±0.234**	5.698±0.534	0.661±0.940	21.682±1.194	27.907±1.519
α-Divergence	α=0.2	0.352±0.054	0.705±0.236	5.712±0.527	0.627±0.962	21.581±1.200	27.953±1.543
	α=0.4	0.351±0.053	0.700±0.239	5.659±0.519	0.626±0.957	21.611±1.205	27.840±1.504
	α=0.6	0.349±0.053	0.691±0.241	5.666±0.509	0.569±0.971	21.456±1.205	27.831±1.513
	α=0.8	0.351±0.054	0.697±0.239	5.701±0.537	0.598±0.969	21.555±1.195	27.786±1.510
Forward KL Divergence		0.353±0.054	0.706±0.236	5.735±0.524	0.626±0.952	21.640±1.184	27.941±1.510
Jensen-Shannon Divergence		0.352±0.053	0.707±0.235	**5.765±0.513**	**0.672±0.942**	**21.708±1.194**	**27.954±1.502**

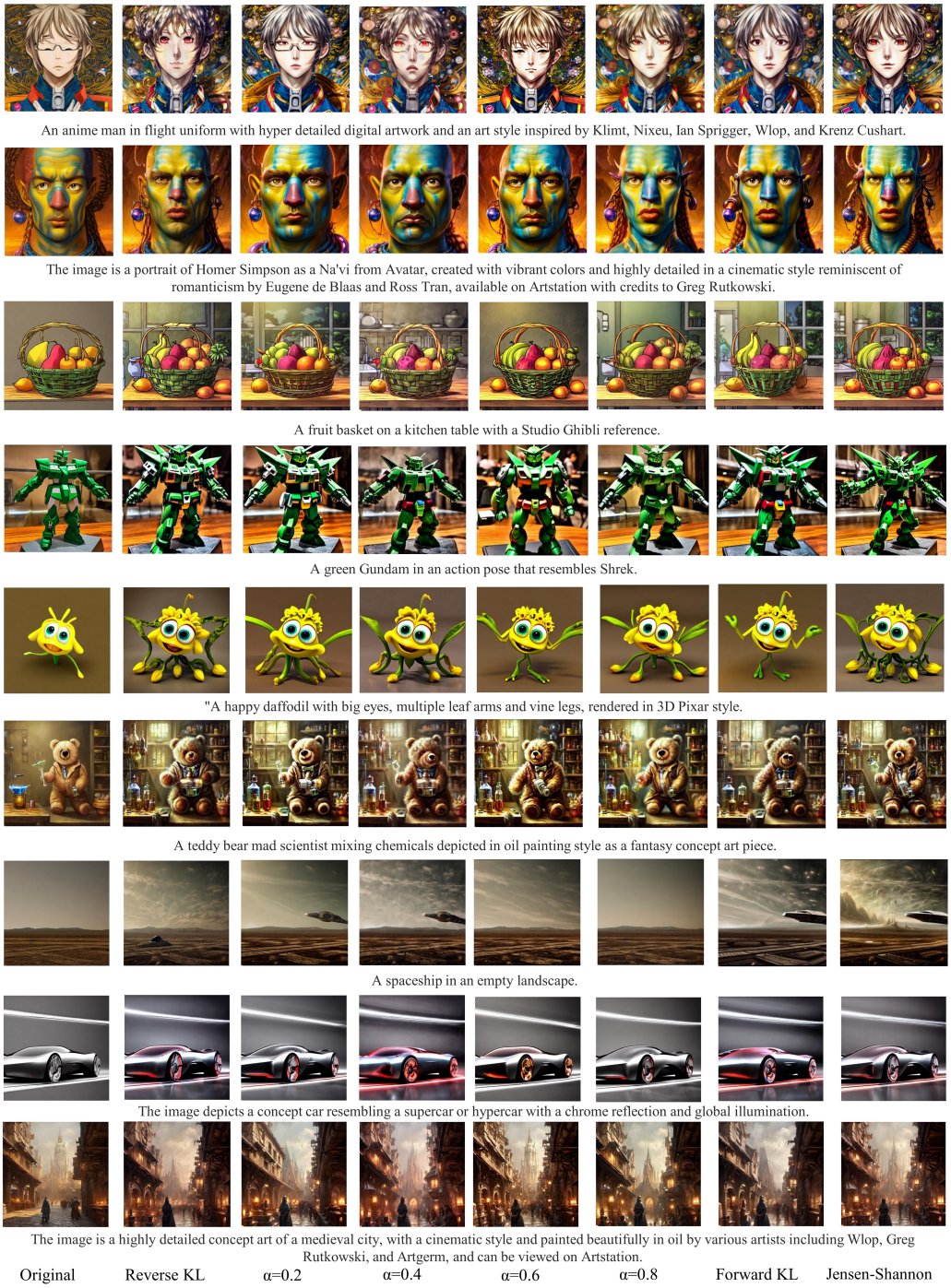
Table 6: Evaluations of the alignment performance with larger-scale (parti-prompts and HPS-V2 training set, **total 4832 prompts**) experiments, where the CLIPScore and VQAScore evaluates image-text alignment performance, and the remaining four metrics evaluate human value alignment performance.

Model		Image-Image CLIP score ↓	Entropy 1D ↑	Entropy 2D ↑	LPIPS ↑
Original Model		0.8096 ± 0.0982	3.8279 ± 0.2675	7.5427 ± 0.5896	0.3002 ± 0.0404
Reverse KL Divergence		0.8491 ± 0.0891	3.9519 ± 0.1578	7.7858 ± 0.3904	0.2957 ± 0.0349
α-Divergence	α =0.2	0.8420 ± 0.0887	3.9231 ± 0.1743	**7.8985 ± 0.3214**	0.3096 ± 0.0321
	α =0.4	**0.8471 ± 0.0892**	**3.9689 ± 0.1564**	7.7860 ± 0.3721	0.2995 ± 0.0348
	α =0.6	**0.8472 ± 0.0881**	3.9216 ± 0.1932	7.7592 ± 0.4400	**0.3132 ± 0.0140**
	α =0.8	0.8423 ± 0.0795	3.9448 ± 0.1851	7.7059 ± 0.4409	0.3030 ± 0.0363
Forward KL Divergence		0.8505 ± 0.0876	3.9409 ± 0.1639	7.7315 ± 0.3868	0.3004 ± 0.0346
Jensen-Shannon Divergence		0.8503 ± 0.0885	3.9532 ± 0.1569	7.8239 ± 0.3928	0.3036 ± 0.0358

Model		RMSE ↑	PSNR ↓	SSIM ↓	FSIM ↓
Original Model		0.0133 ± 0.0028	37.660 ± 1.879	0.8819 ± 0.0380	0.3779 ± 0.0227
Reverse KL Divergence		0.0154 ± 0.0027	36.360 ± 1.587	0.8521 ± 0.0374	0.3800 ± 0.0185
α-Divergence	α =0.2	**0.0176 ± 0.0025**	**35.192 ± 1.280**	**0.8354 ± 0.0350**	**0.3757 ± 0.0182**
	α =0.4	0.0155 ± 0.0026	36.317 ± 1.489	0.8535 ± 0.0357	0.3806 ± 0.0184
	α =0.6	0.0166 ± 0.0027	35.708 ± 1.488	0.8371 ± 0.0348	0.3765 ± 0.0212
	α =0.8	0.0155 ± 0.0028	36.327 ± 1.646	0.8528 ± 0.0383	0.3789 ± 0.0210
Forward KL Divergence		0.0157 ± 0.0026	36.133 ± 1.536	0.8475 ± 0.0366	0.3763 ± 0.0194
Jensen-Shannon Divergence		0.0158 ± 0.0026	36.054 ± 1.511	0.8450 ± 0.0363	0.3798 ± 0.0202

Table 7: Evaluations of the generation diversity with larger-scale (parti-prompts and HPS-V2 training set, **total 4832 prompts**) experiments. The metrics originally utilized for evaluating image similarity exhibit an opposite property when evaluating generation diversity.

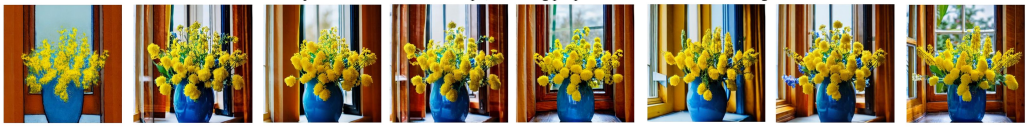
I Qualitative Comparison of Alignment with Diverse Divergence



Side-view blue-ice sneaker inspired by Spiderman created by Weta FX.

An abstract painting depicting the balance between dark and light in nature with rough brushstrokes and fine details in natural colors.

Michal Karcz has painted a beautiful landscape, featuring purple trees with intricate and elegant details.



A blue vase filled with yellow flowers on a window sill.

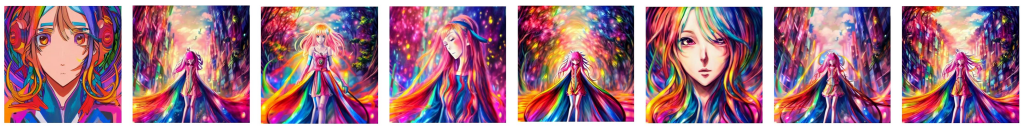
A blue airplane in a blue, cloudless sky

Three cows eating in a field with sea in background.

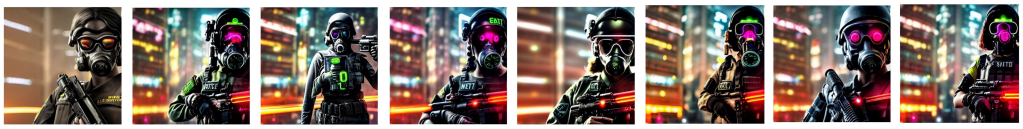
a girl a white bear a camera a tea pot and a cup

A street light and other road signs.

Original | Reverse KL | α=0.2 | α=0.4 | α=0.6 | α=0.8 | Forward KL | Jensen-Shannon



A colorful digital painting with a front view and anime-inspired vibes featuring a magical composition.



A key visual of a young female swat officer with a neon futuristic gas mask in a cyberpunk setting.

A person wearing a lab coat holding a green apple and standing in front of a whiteboard filled with equations and diagrams.

16-year-old teenager wearing a white bear-ear hat with a smirk on their face.

Original	Reverse KL	α=0.2	α=0.4	α=0.6	α=0.8	Forward KL	Jensen-Shannon