Enhancing Diffusion Models with Text-Encoder Reinforcement Learning

Chaofeng Chen[1⋆], Annan Wang[1⋆], Haoning Wu[1], Liang Liao[1], Wenxiu Sun[3], Qiong Yan[3], and Weisi Lin[2] ✉

[1] S-Lab, Nanyang Technological University
[2] CCDS, Nanyang Technological University [3] Sensetime Research
chaofeng.chen@ntu.edu.sg c190190@e.ntu.edu.sg wslin@ntu.edu.sg
https://github.com/chaofengc/TexForce

(a) SDv1.4 (b) TexForce (SDv1.4) (c) SDv1.4 (d) TexForce

Fig. 1: By refining the text encoder through reinforcement learning, the proposed ***TexForce*** with Stable Diffusion v1.4 can generate images that align better with human quality preference. *The compared images are generated with the same seed and prompts. (a)(b): "Impressionist painting of a cat, high quality"; (c)(d): "A photo of a hand" & "A complete face of a man".*

Abstract. Text-to-image diffusion models are typically trained to optimize the log-likelihood objective, which presents challenges in meeting specific requirements for downstream tasks, such as image aesthetics and image-text alignment. Recent research addresses this issue by refining the diffusion U-Net using human rewards through reinforcement learning or direct backpropagation. However, many of them overlook the importance of the text encoder, which is typically pretrained and fixed during training. In this paper, we demonstrate that by finetuning the text encoder through reinforcement learning, we can enhance the text-image alignment of the results, thereby improving the visual quality. Our primary motivation comes from the observation that the current text encoder is suboptimal, often requiring careful prompt adjustment. While fine-tuning the U-Net can partially improve performance, it remains suffering from the suboptimal text encoder. Therefore, we propose to use reinforcement learning with low-rank adaptation to finetune the text encoder based on task-specific rewards, referred as **TexForce**. We first

⋆ These authors contributed equally to this work.

show that finetuning the text encoder can improve the performance of diffusion models. Then, we illustrate that TexForce can be simply combined with existing U-Net finetuned models to get much better results without additional training. Finally, we showcase the adaptability of our method in diverse applications, including the generation of high-quality face and hand images.

Keywords: Diffusion Models · Text Encoder · Reinforcement Learning

Generative models have witnessed notable advancements in recent years, transitioning from earlier Generative Adversarial Networks (GAN) [3, 13, 20] to the more recent diffusion models [9, 16, 43]. Text-to-image models like Stable Diffusion [36], DALLE [34, 35], and Imagen [38], trained on extensive datasets, have demonstrated impressive capabilities in producing high-quality images from textual prompts. However, these diffusion models primarily optimize the log-likelihood objective, which, although effective for generative tasks, may not consistently fulfill specific requirements for downstream applications. Key challenges include achieving desirable image aesthetics and aligning generated images with text descriptions, both of which are critical for applications in areas such as content generation and multimedia synthesis.

Although prompt engineering is helpful in some cases, such as Fig. 2, these techniques [15, 47, 48] have inherent limitations, including a lack of precise control, limited generalization across various models, and inadequacy in addressing complex demands. For instance, current models encounter difficulties in generating visual texts [27], and comprehending object counts [19, 23]. Therefore, recent efforts draw inspiration from the success of *reinforcement learning from human feedback (RLHF)* employed in large language models [30] and adopt similar strategies to enhance the alignment capabilities of diffusion models [2, 11, 23, 32]. While these methods show promise, they all finetune the U-Net conditioned on the fixed suboptimal text encoder, which constrains their efficacy.

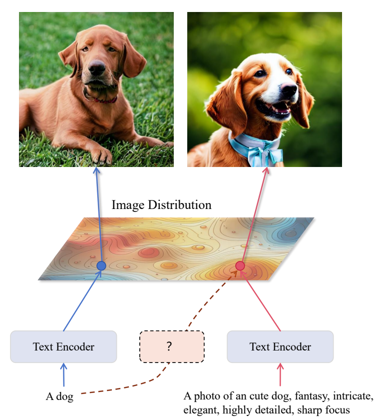


Fig. 2: The qualities of outputs from pretrained diffusion models vary a lot with different prompts. Through the reinforcement learning, we can finetune text encoder to better align with images.

In this paper, we introduce **TexForce**, an innovative method that applies reinforcement learning combined with low-rank adaptation to enhance the text encoder using task-specific rewards. We utilize the DDPO (denoising diffusion policy optimization) [2] algorithm to update the text encoder, which is based on PPO (proximal policy optimization [41]) in the iterative denoising process. Unlike direct backpropagation, this RL algorithm does not require differentiable rewards, offering greater flexibility. By finetuning with LoRA [18], TexForce can adapt to diverse tasks by simply switching the LoRA weights, and also allows for the fusion of different LoRA weights to combine the capabilities learned from different rewards. Most importantly, TexForce can be seamlessly integrated with existing finetuned U-Net models from previous methods and achieves much better performance without additional training.

As illustrated in Fig. 1, our approach significantly enhances the result quality of Stable Diffusion 1.4 (SDv1.4) [36] when finetuned to align with different rewards. We validate our approach through extensive experiments on both single-prompt and multi-prompt tasks across various reward functions. Moreover, we showcase the adaptability of our method across various applications, including the generation of high-quality face and hand images. Our contributions can be summarized as follows:

- We observe that when optimizing reinforcement learning (RL) rewards, finetuning the U-Net component of diffusion models carries the risk of compromising image appearance, whereas finetuning the text encoder mitigates such concerns and preserves semantics better.
- We find that it is possible to directly combine the LoRA weights from text encoder and U-Net *without extra training*. This straightforward fusion addresses the challenges associated with finetuned U-Net while upholding the merits of finetuned text encoders.

1 Related Works

1.1 Text-to-Image Diffusion Models

Denoising diffusion models [16, 42, 44, 45] have become the de facto standard for generative tasks, owing to their remarkable capabilities in generating diverse multimedia content, including images [9, 36], videos [14, 21, 51], 3D content [28, 31], and more. Text-to-image models, particularly those creating images based on textual prompts, have gained significant traction attributable to the availability of powerful models such as StableDiffusion [36], DALLE [34, 35] and Imagen [38]. Several approaches have emerged to enhance control over texture details in the generated outputs. Noteworthy methods like DreamBooth [37] and Texture Inversion [12] offer tailored solutions for specific image requirements. To improve the generalization capabilities, ControlNet [54], T2I-Adapter [29] and $\mathcal{W}_+$-Adapter [25] introduce additional image encoders to control the structure and details. Nonetheless, they still require a large number of paired images to train, and may struggle to meet the diverse demands of various tasks. Prompt engineering [15] is another popular approach aimed at enhancing the quality of generated images. However, this method is constrained by the expressiveness of text prompts and pretrained models and may not be straightforward when addressing complex tasks such as aesthetic quality and object composition [11, 23].

1.2 Learning from Feedback in Diffusion Models

Recent efforts have aimed to optimize diffusion models using human rewards or task objectives, typically categorized into three main approaches: reward-weighted regression (RWR), reinforcement learning (RL), and direct backpropagation. RWR methods like RAFT [10], Lee et al. [23], and Emu [8] start by assessing image quality with human feedback and then re-weight or select high-quality examples to enhance performance. RL methods, exemplified by DDPO [2] and DPOK [11], treat the denoising process as a Markov decision process and optimize the model using RL algorithms, such as PPO [41]. Direct backpropagation methods, including AlignProp [32], ReFL [53], and DRaFT [7], propagate gradients directly from the reward function to the model. Because these models only finetune U-Net conditioned on the suboptimal text encoder, their effectiveness in aligning outputs with text prompts is often limited. A concurrent work, TextCraftor [26], also explores fine-tuning the text encoder; however, it relies on direct backpropagation and is incompatible with non-differentiable rewards.

1.3 Quality Metrics for Generative Models

With the increasing popularity of generative models, several benchmarks [22, 24, 49, 52, 53] have been developed to evaluate the quality of generated images. Notably, ImageReward [53], PickScore [22], and HPS [52] are among the more frequently employed benchmarks. Additionally, image aesthetic metrics, particularly the LAION Aesthetics Predictor [6], find widespread application in data filtering [36] and results assessment.

2 Method

2.1 Preliminaries on Diffusion Models

Diffusion models [16, 36] belong to the class of generative models that leverage noise-driven processes to progressively transform data distributions. This process contains a controlled noise addition phase (forward diffusion) and a noise removal phase (reverse diffusion). Given image samples $\mathbf{x}_0$ originating from the data distribution $q(\mathbf{x}_0)$, the forward diffusion process generates a sequence of images $\{\mathbf{x}_t\}_{t=1}^{T}$ by iteratively introducing noise via a Markov chain with a predefined noise schedule. Then, the reverse diffusion process is to learn a denoising U-Net ϵ_θ to estimate the cleaner $\mathbf{x}_{t-1}$ with the noisy $\mathbf{x}_t$:

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \mathbf{\Sigma}_\theta(\mathbf{x}_t, t)), \tag{1}$$

where θ is the learnable parameter. For text-to-image diffusion models [36], this process is conditioned on a text input s, encoded with a text encoder $\mathbf{z} = \tau_\phi(s)$. Then the network ϵ_θ is trained with the following objective:

$$\mathcal{L}(\theta) = \mathbb{E}_{\mathbf{x}_t, s, t, \epsilon \sim \mathcal{N}(0, \mathbf{I})} \left[\|\epsilon - \epsilon_\theta(\mathbf{x}_t, t, \mathbf{z})\|_2^2 \right], \tag{2}$$

which aims to optimize the variational lower bound on the log-likelihood of the data distribution $q(\mathbf{x}_0)$. It is worth noting that the text encoder τ_ϕ is usually a pretrained model, such as CLIP [33], and is fixed during training.

2.2 Reinforcement Learning with LoRA

According to the above formulation, diffusion models are learnt to optimize the log-likelihood objective Eq. (2), which is not directly related to the task requirements. To address this issue, we propose to finetune the text encoder τ_ϕ with reinforcement learning (RL). Details come as follows.

RL in Diffusion Models. In our setting, the RL framework optimizes the policy defined by the diffusion model conditioned on the text embeddings. The text encoder τ_ϕ acts as the policy network that maps text descriptions to actions (text embeddings), which then influences the generative process of the diffusion model. Let R be the reward function that evaluates the quality of the generated images, which could encapsulate various aspects, such as image-text alignment and image quality, and adherence to specific attributes desired in the output. Then the objective of RL is to maximize the expected reward:

$$J(\phi) = \mathbb{E}\left[R(\mathbf{x}_0, s)\right]. \tag{3}$$

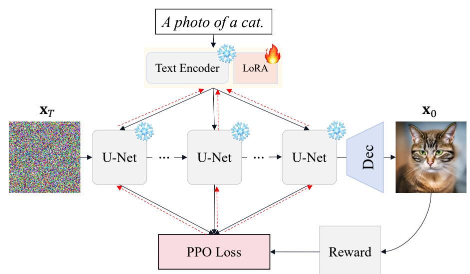


Fig. 3: Illustration of text encoder finetune with PPO algorithm.

Since the denoising process can be formulated as a Markov decision process [2], *i.e.*, $p_\theta(\mathbf{x}_0|\mathbf{z}) = p(\mathbf{x}_T)\prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{z})$, the policy gradient of Eq. (3) can be computed as:

$$\nabla_\phi J = \mathbb{E}\left[\sum_{t=0}^{T} \nabla_\phi \log p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t, \tau_\phi(s)) R(\mathbf{x}_0, s)\right]. \tag{4}$$

Following the DDPO algorithm, we use the Proximal Policy Optimization (PPO) [41] to keep stable learning dynamics. It applies importance sampling with clipped probability ratio to Eq. (3) which becomes:

$$J = \mathbb{E}\left[\min(r_t(\phi)A, \text{clip}(r_t(\phi), 1-\lambda, 1+\lambda)A)\right], \tag{5}$$

where the advantage value A is the normalized rewards R over a buffer set of $\mathbf{x}_0$, and r_t is the probability ratio between the new policy and the old policy for the denoise step $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t, \tau_\phi(s))$. Since the policy is an isotropic Gaussian, the probability can be easily calculated. Then, we can calculate the gradient for Eq. (5) similar to Eq. (4) to update the policy network τ_ϕ. More details are provided in supplementary materials.

Low-Rank Adaptation (LoRA) [18] is a technique that allows for the modification of large pre-trained models without the need for extensive re-training. It achieves this by inserting trainable low-rank matrices into the original feed-forward layer as $W' = W + \alpha\Delta W$, where ΔW is the learnable weights initialized to zero and α is a scale factor. Such low-rank weight matrices are shown to be helpful in preventing the model from overfitting to the training data [18].

2.3 Discussion of Finetuning for Diffusion Model

In this part, we briefly discuss the advantages of finetuning the text encoder with reinforcement learning to improve the performance of diffusion models.

Finetune of Diffusion Model. As discussed in Sec. 2.1, given the text s and $\mathbf{x}_0$, the denoising network ϵ_θ is learned by maximizing the following lower bound:

$$\mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|s)}\left[\log(p_\theta(\mathbf{x}_{0:T}|\mathbf{z}))\right] - D_{KL}(q_\phi(\mathbf{z}|s)||p(\mathbf{z})). \tag{6}$$

In the training stage of diffusion models, ϕ is usually fixed and $p_\theta(\mathbf{x}_0|\mathbf{z})$ are learned through classifier free guidance [17]. With an extremely large amount of s in datasets such as LAION-400M [39, 40], it is reasonable to assume that q_ϕ is a good estimation of $p(\mathbf{z})$ even when ϕ is fixed. However, in the finetuning stage, we expect to use a small amount of s to optimize Eq. (6) for specific tasks. In such cases, the q_ϕ is likely to be a suboptimal estimation of $p(\mathbf{z})$, and thus largely increasing the second KL term. Therefore, we believe that it is necessary to finetune the text encoder τ_ϕ to minimize the second term when the finetune dataset is limited.

RL v.s. Direct Backpropagation. Besides reinforcement learning, recent approaches also directly backpropagate the gradients through the denoising steps [7, 32, 53]: $\nabla_\theta\mathcal{L} = \sum_m^n \frac{\partial R}{\partial \mathbf{x}_t}\frac{\partial \mathbf{x}_t}{\partial \theta}$, where $m \leq n \in [0, T]$. However, this approach is more likely to overfit the reward function and lead to mode collapse. For instance, in DRaFT [7], the model may collapse to generate a single image to achieve high aesthetic rewards. Besides, RL does not require differentiable quality rewards, and is much more flexible than direct backpropagation. For example, current applications can collect human feedbacks and use them as rewards to directly finetune the model.

3 Experiments

3.1 Implementation Details

Prompt Datasets. We follow previous works [2, 11, 32, 52, 53] and use three types of prompt datasets with their corresponding experimental settings:

- **Simple animal prompts [2].** A simple dataset with a curated list of 45 common animals for training.
- **Single phrases.** Four single phrases from DPOK [11] to test the model capabilities under different scenarios.
- **Complex long prompts.** Subsets from ImageReward [53] and HPSv2 [52]. The former contains 20,000 prompts for training and 100 for testing. The latter contains 750 prompts for training, 50 for testing.
- **Specific task prompts.** Example task prompts for face and hand images.

Reward Functions. We conduct experiments with different kinds of reward functions as below:

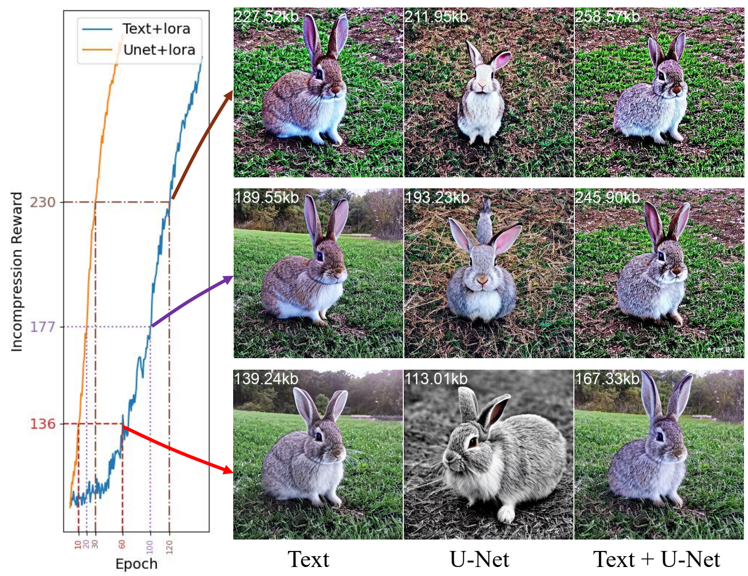


Fig. 4: Comparison of training progress between finetuning text encoder and U-Net with LoRA. The image size after JPEG compression is marked on the top-left corner as "kb".

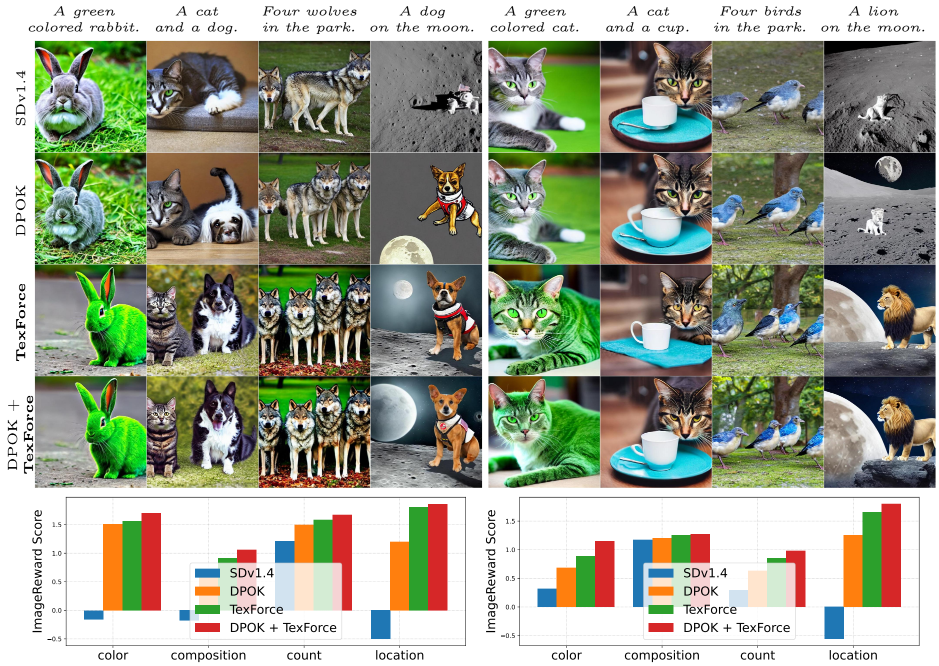
- **Text-to-Image Rewards.** These rewards are trained on text-to-image datasets, such as ImageReward [53] and HPSv2 [52].
- **Specific task rewards.** Following [2], we evaluate model performance for the compression and incompression. Besides, we also design specific rewards for face and hand.

Please refer to the supplementary materials for more training details and hyper-parameter settings.

3.2 Finetuning Text Encoder v.s. U-Net

In this section, we will empirically analyze the difference between finetuning the text encoder and U-Net, through the incompression task as introduced in DDPO [2]. It aims to enhance the complexity of generated images through reinforcement learning (RL). The reward for this task is assessed based on the image size after JPEG compression with a quality factor of 95. Given its objective nature and the ambiguity of possible solutions, this task is suitable for analyzing behaviors of different models when optimized for the reward. We finetune both the text encoder and the U-Net with LoRA using the simple animal prompts.

Figure 4 shows the results comparison between finetuning text encoder and U-Net with LoRA. By comparing models with the same incompression score, we can have the following observations:



(a) Results on seen text prompts (b) Results on unseen text prompts

Fig. 5: Qualitative and quantitative comparisons with SDv1.4 and DPOK on individual scenarios. Images for comparison are generated with the same random seed. The results show that TexForce can generate more consistent images with better quality than SDv1.4 and DPOK, and simple combination of DPOK and TexForce gives even better performance without any additional training.

– U-Net tends to change the visual appearance to increase the reward, whereas the text encoder introduces novel visual concepts to attain the same objective. As shown in Fig. 4, despite having comparable incompression scores, the outcomes from the text encoder are more coherent than those from the U-Net. However, this also makes the optimization of the text encoder more challenging and time-consuming.
– We can directly combine the LoRA weights from TexForce and U-Net to achieve even better results. As shown in Fig. 4, the results in the third column achieve the highest incompression score and still maintain a similar visual structure from the first column, successfully combining advantages from the LoRA weights of both the text encoder and U-Net. *It is worth noting that this is achieved without additional training.*

3.3 Comparison with Existing Works

Results on Different Individual Prompts. As demonstrated in [11, 23], existing stable diffusion models exhibit misalignment with simple text prompts,

Fig. 6: Visual comparison with ReFL on ImageReward dataset using real user prompts.

such as color consistency (*e.g.*, *A green colored dog.*) and combination of objects (*e.g.*, *A cat and a dog*). Therefore, we start with these simple individual scenarios to highlight the advantages of our method. We follow the experimental settings of DPOK [11], and conduct our experiments with four different capabilities: color consistency, object composition, object count, and object location, as shown in Fig. 5. Both DPOK and our TexForce are trained with the ImageReward feedback function and 20K samples. All models are evaluated with ImageReward scores and averaged over 50 samples with the same random seeds.

Figure 5 presents a comprehensive overview of both quantitative and qualitative results for seen and unseen prompts. The results show that our method can generate more consistent images with better quality than the original SDv1.4 and DPOK. For instance, we can see that the results of TexForce are more consistent with the prompts, such as the color of the rabbit and cat, the number of birds, and the location of the dog. Besides, the results of TexForce are more realistic than DPOK and SDv1.4. Quantitatively, TexForce attains better average ImageReward scores for both seen and unseen prompts. This underscores the overall superiority of TexForce over DPOK and SDv1.4 across multiple samples. Moreover, the combination of TexForce and DPOK demonstrates the best performance in terms of both ImageReward scores and visual quality. This demonstrates the flexibility of our method, which can be seamlessly integrated with existing methods to achieve best performance without additional training.

Results on Complex Long Prompts Next, we conduct experiments using larger dataset with complex long prompts, *i.e.*, the ImageReward dataset [53] and HPS dataset [52]. We retrained with ReFL and AlignProb with official codes,

Fig. 7: Visual comparison with AlignProb on HPSv2 dataset using real user prompts.

Table 1: Quantitative results on ImageReward and HPSv2. Results are tested with the same seed and prompts.

Method	ImageReward
SDv1.4	0.2154
ReFL	0.4485
TexForce	**0.4556**
ReFL + TexForce	**0.6553**

Method	HPSv2
SDv1.4	0.2752
AlignProb	**0.2821**
TexForce	0.2767
AlignProb + TexForce	**0.2914**

and the results are shown in Figs. 6 and 7. As we can observe, since ReFL is trained with a single step backward to update the U-Net, it is less effective than the proposed TexForce to align the input prompts with generated images, such as the white shirt of the fox. The improvement of ReFL is mainly the appearance of the generated images, such as the color of the man and the texture of the fox. Meanwhile, the proposed TexForce is better at aligning the text prompts with generated images, which makes TexForce better in Tab. 1. Furthermore, when merging the strengths of TexForce and ReFL, we observe a notable improvement in both quantitative results and visual appearance.

Similarly, we compared our method with AlignProb using the HPSv2 reward. The results from AlignProb appeared overly optimized towards the rewards, evident in the abundance of yellow spotlights and clouds, leading to disrupted semantics. In contrast, our proposed TexForce primarily aims to improve the text-image alignment. While our reward score was slightly lower than AlignProb due to limited changes in color style, our results better match the textual prompts in terms of semantics. Additionally, our combined results maintain the visual styles preferred by the HPSv2 while preserving the meaning of the text prompts, resulting in significant improvement over AlignProb.

Table 2: Quantitative results with SDv1.5 and SDv2.1 on the ImageReward dataset.

Backbone	Original	ReFL	TexForce	ReFL+TexForce
SDv1.5	0.2140	0.5484	0.4086	**0.6703**
SDv2.1	0.3891	0.5223	0.5084	**0.6158**

3.4 Experiments with Different Backbones

To show the robustness of our method, we conduct experiments with more different backbones, including SDv1.5[3] and SDv2.1[4]. We use the ImageReward score and prompts dataset to train all the models, and the results are shown in Fig. 8 and Tab. 2. Notably, our method consistently improves the text-image alignment of the original models. For instance, the results generated by TexForce exhibit enhanced visual appeal and better consistency with the prompts, such as the *victorian lady*, *old king*, and *atom model*. It is also worth noting that although SDv2.1 is already much better than SDv1.5, TexForce continues to augment the performance. This demonstrates the adaptability and robustness of our method when employed with different backbones. Although ReFL achieves higher ImageReward scores, our observations indicate that it primarily enhances color and fine details and is less effective than TexForce in aligning images with text prompts. For both SDv1.5 and SDv2.1, the combined model yields the best performance, which clearly affirms the effectiveness of TexForce.

3.5 GPT-4V Evaluation

As GPT-4V has recently shown to be comparable with human-level performance in evaluating image quality [49, 50], we decide to rely on GPT-4V evaluations instead of traditional user studies, which may be inconsistent and hard to reproduce. Our approach involves using GPT-4V to rank image quality based on aesthetic quality and coherence with text. In Fig. 9, we present the average scores from three rounds of evaluations using the ImageReward test dataset. We can see that our TexForce method does a better job at aligning text with images in diffusion models, while ReFL improves the appearance of the images. Combining both approaches successfully takes advantage of both of them and yields the best results. Please refer to the supplementary material to reproduce the results.

3.6 Ablation Study about Joint Finetune

Given the efficacy of straightforward fusion, it prompts us to inquire whether a joint finetuning approach yields good results. We conducted such experiments using the SDv1.4 backbone, and the results are illustrated in Tab. 3 and Fig. 10. It is evident that, although the quantitative performance of joint finetuning

[3] `https://huggingface.co/runwayml/stable-diffusion-v1-5runwayml/stable-diffusion-v1-5`

[4] `https://huggingface.co/stabilityai/stable-diffusion-2-1`

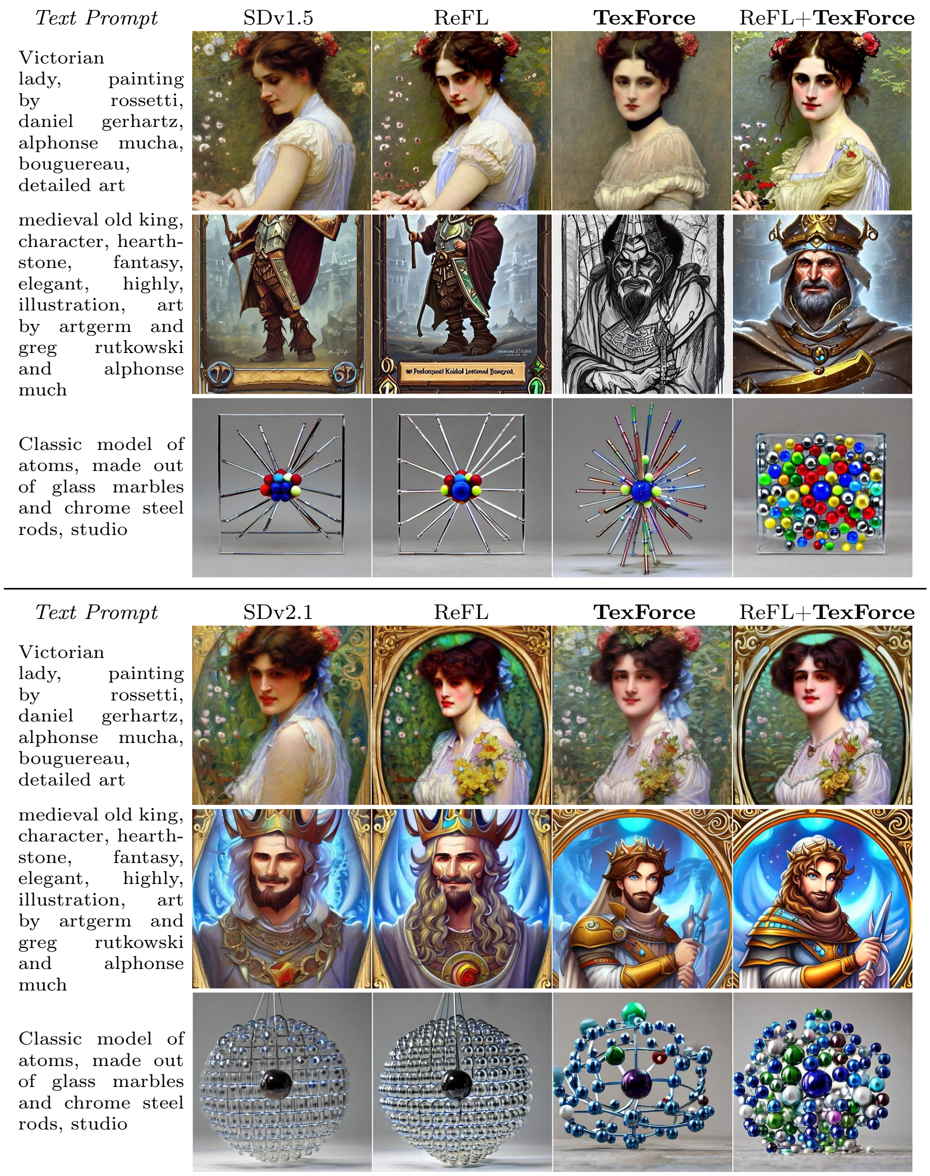


Fig. 8: Results with SDv1.5 and SDv2.1 backbones on ImageReward test dataset.

surpasses that of ReFL and TexForce, it still falls short of the performance achieved by their simple combination. We hypothesize that this disparity arises because the fixed U-Net can serve as a prior for pixel generation during the finetuning of the text encoder. Consequently, joint fine-tuning complicates the optimization process for the text encoder, thereby leading to inferior results.

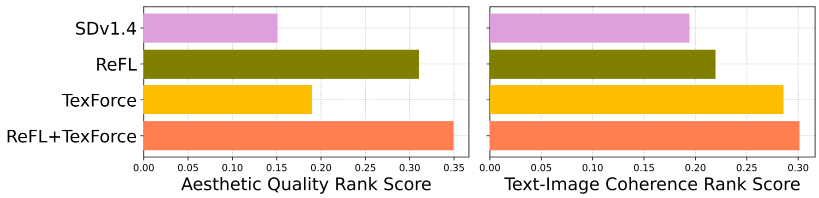


Fig. 9: GPT4V evaluation for aesthetic quality and text-image coherence with ImageReward testset and SDv1.4. Refer to supplementary for SDv1.5 and SDv2.1 results.

Table 3: Quantitative comparison between simple fusion and joint training.

Methods	SDv1.4	ReFL	TexForce	ReFL + TexForce	Joint
Score	0.2154	0.4485	0.4556	**0.6553**	**0.5009**

3.7 Applications

TexForce demonstrates remarkable adaptability to diverse tasks, as it does not require differentiable rewards. In this section, we showcase its capabilities in enhancing the quality of generated face and hand images.

Face reward. We employ the face quality evaluation metric from [4], which is based on an image quality evaluation network [5] trained using the face quality dataset [46].

Hand reward. Regarding the hand quality evaluation, we recognize the absence of specific hand quality metrics. Instead, we employ a straightforward hand detection confidence score as a reward function and observe its utility. The hand detection model from [1] is used to calculate the confidence score.

Figure 11 illustrates the progressive improvement in the quality of generated face and hand images over the course of training. These results illustrate the capacity of TexForce to enhance image quality, utilizing either direct quality metrics or a simple confidence score.

Moreover, by utilizing LoRA weights for fine-tuning the text encoder, we find that it is feasible to blend specific LoRA weights to enhance the quality of specific objects. Suppose the LoRA weight θ_i from i-th task, we can simply fuse them via $\sum_i \alpha_i\theta_i$. In Fig. 12, we demonstrate how the fusion of ImageReward LoRA weights and face quality LoRA weights can produce high-quality face images. This flexibility significantly broadens the range of potential applications for TexForce.

4 Conclusion

In this paper, we introduce a new method called **TexForce** for enhancing the text encoder of diffusion models using reinforcement learning. Our research demonstrates that refining the text encoder can enhance the overall performance of diffusion models, specifically in terms of aligning text and images as well as

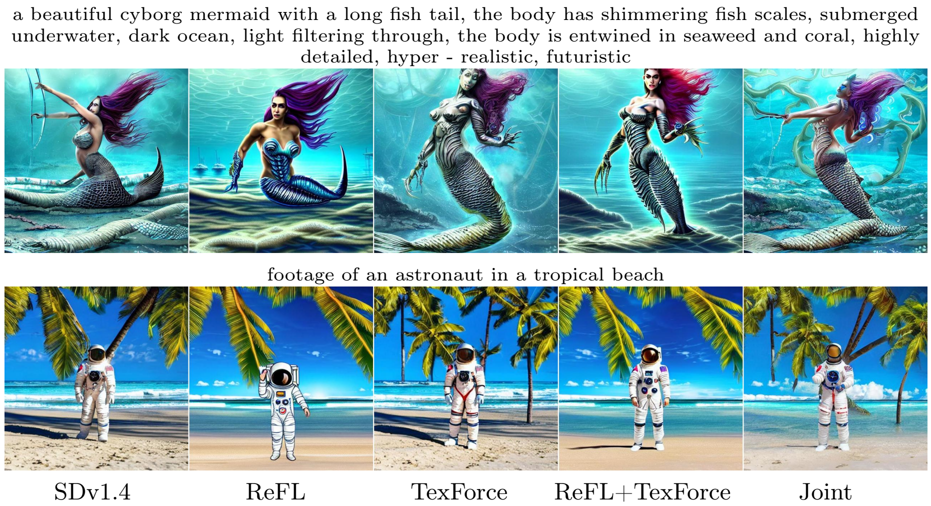


Fig. 10: Visual comparison between simple fusion and joint training.

Fig. 11: Two example applications of TexForce: high-quality face and hand generation.

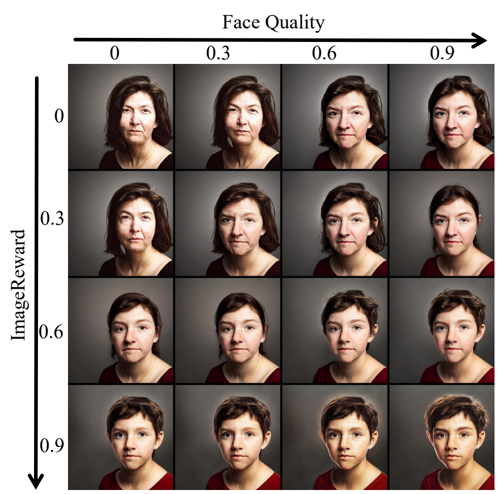


Fig. 12: Fusion of ImageReward LoRA weights and face quality LoRA weights. Prompt: *A realistic portrait photo.*

improving visual quality. Furthermore, we illustrate that TexForce can be seamlessly integrated with existing U-Net models that have undergone fine-tuning, without the need for extra training, resulting in significant performance improvements. Lastly, we showcase the versatility of our approach across various applications, including the generation of high-quality images of faces and hands. We also provide evidence that the finetuned LoRA weights with different tasks can be combined to enhance the specific quality of image generation.

Acknowledgments

This study is supported under the RIE2020 Industry Alignment Fund – Industry Collaboration Projects (IAF-ICP) Funding Initiative, as well as cash and in-kind contribution from the industry partner(s).

References

1. Alam, M.M., Islam, M.T., Rahman, S.M.M.: Unified learning approach for egocentric hand gesture recognition and fingertip detection. Pattern Recognition (2021). https://doi.org/10.1016/j.patcog.2021.108200
2. Black, K., Janner, M., Du, Y., Kostrikov, I., Levine, S.: Training diffusion models with reinforcement learning. Int. Conf. Learn. Represent. (2024)
3. Brock, A., Donahue, J., Simonyan, K.: Large scale gan training for high fidelity natural image synthesis. Int. Conf. Learn. Represent. (2019)
4. Chen, C., Mo, J.: IQA-PyTorch: Pytorch toolbox for image quality assessment. [Online]. Available: https://github.com/chaofengc/IQA-PyTorch (2022)
5. Chen, C., Mo, J., Hou, J., Wu, H., Liao, L., Sun, W., Yan, Q., Lin, W.: Topiq: A top-down approach from semantics to distortions for image quality assessment. In: IEEE Trans. Image Process. (2024)
6. Christoph, S., Romain, B.: Laion-aesthetics. (2022)
7. Clark, K., Vicol, P., Swersky, K., Fleet, D.J.: Directly fine-tuning diffusion models on differentiable rewards. Int. Conf. Learn. Represent. (2024)
8. Dai, X., Hou, J., Ma, C.Y., Tsai, S., Wang, J., Wang, R., Zhang, P., Vandenhende, S., Wang, X., Dubey, A., et al.: Emu: Enhancing image generation models using photogenic needles in a haystack. arXiv preprint arXiv:2309.15807 (2023)
9. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. In: Adv. Neural Inform. Process. Syst. (2021)
10. Dong, H., Xiong, W., Goyal, D., Pan, R., Diao, S., Zhang, J., Shum, K., Zhang, T.: Raft: Reward ranked finetuning for generative foundation model alignment. arXiv preprint arXiv:2304.06767 (2023)
11. Fan, Y., Watkins, O., Du, Y., Liu, H., Ryu, M., Boutilier, C., Abbeel, P., Ghavamzadeh, M., Lee, K., Lee, K.: Dpok: Reinforcement learning for fine-tuning text-to-image diffusion models. Adv. Neural Inform. Process. Syst. (2024)
12. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-Or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. Int. Conf. Learn. Represent. (2023)
13. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. Communications of the ACM **63**(11), 139–144 (2020)
14. Guo, Y., Yang, C., Rao, A., Wang, Y., Qiao, Y., Lin, D., Dai, B.: Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. Int. Conf. Learn. Represent. (2024)
15. Hao, Y., Chi, Z., Dong, L., Wei, F.: Optimizing prompts for text-to-image generation. Adv. Neural Inform. Process. Syst. (2023)
16. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Adv. Neural Inform. Process. Syst. (2020)
17. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)

18. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. Int. Conf. Learn. Represent. (2022)
19. Hu, Y., Liu, B., Kasai, J., Wang, Y., Ostendorf, M., Krishna, R., Smith, N.A.: Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. Int. Conf. Comput. Vis. (2023)
20. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 4401–4410 (2019)
21. Khachatryan, L., Movsisyan, A., Tadevosyan, V., Henschel, R., Wang, Z., Navasardyan, S., Shi, H.: Text2video-zero: Text-to-image diffusion models are zero-shot video generators. Int. Conf. Comput. Vis. (2023)
22. Kirstain, Y., Polyak, A., Singer, U., Matiana, S., Penna, J., Levy, O.: Pick-a-pic: An open dataset of user preferences for text-to-image generation. Adv. Neural Inform. Process. Syst. (2024)
23. Lee, K., Liu, H., Ryu, M., Watkins, O., Du, Y., Boutilier, C., Abbeel, P., Ghavamzadeh, M., Gu, S.S.: Aligning text-to-image models using human feedback. arXiv preprint arXiv:2302.12192 (2023)
24. Li, C., Zhang, Z., Wu, H., Sun, W., Min, X., Liu, X., Zhai, G., Lin, W.: Agiqa-3k: An open database for ai-generated image quality assessment. IEEE Trans. Circuit Syst. Video Technol. pp. 1–1 (2023). https://doi.org/10.1109/TCSVT.2023.3319020
25. Li, X., Hou, X., Loy, C.C.: When stylegan meets stable diffusion: a $\mathcal{W}_+$ adapter for personalized image generation. IEEE Conf. Comput. Vis. Pattern Recog. (2024)
26. Li, Y., Liu, X., Kag, A., Hu, J., Idelbayev, Y., Sagar, D., Wang, Y., Tulyakov, S., Ren, J.: Textcraftor: Your text encoder can be image quality controller. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 7985–7995 (2024)
27. Liu, R., Garrette, D., Saharia, C., Chan, W., Roberts, A., Narang, S., Blok, I., Mical, R., Norouzi, M., Constant, N.: Character-aware models improve visual text rendering. arXiv preprint arXiv:2212.10562 (2022)
28. Liu, R., Wu, R., Van Hoorick, B., Tokmakov, P., Zakharov, S., Vondrick, C.: Zero-1-to-3: Zero-shot one image to 3d object. In: Int. Conf. Comput. Vis. pp. 9298–9309 (2023)
29. Mou, C., Wang, X., Xie, L., Wu, Y., Zhang, J., Qi, Z., Shan, Y., Qie, X.: T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. AAAI (2024)
30. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al.: Training language models to follow instructions with human feedback. In: Adv. Neural Inform. Process. Syst. (2022)
31. Poole, B., Jain, A., Barron, J.T., Mildenhall, B.: Dreamfusion: Text-to-3d using 2d diffusion. Int. Conf. Learn. Represent. (2023)
32. Prabhudesai, M., Goyal, A., Pathak, D., Fragkiadaki, K.: Aligning text-to-image diffusion models with reward backpropagation. arXiv preprint arXiv:2310.03739 (2023)
33. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: Proc. Int. Conf. Mach. Learn. pp. 8748–8763 (2021)
34. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 (2022)

35. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: Proc. Int. Conf. Mach. Learn. (2021)
36. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: IEEE Conf. Comput. Vis. Pattern Recog. (2022)
37. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 22500–22510 (2023)
38. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S.K.S., Ayan, B.K., Mahdavi, S.S., Lopes, R.G., et al.: Photorealistic text-to-image diffusion models with deep language understanding. In: Adv. Neural Inform. Process. Syst. (2022)
39. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al.: Laion-5b: An open large-scale dataset for training next generation image-text models. arXiv preprint arXiv:2210.08402 (2022)
40. Schuhmann, C., Vencu, R., Beaumont, R., Kaczmarczyk, R., Mullis, C., Katta, A., Coombes, T., Jitsev, J., Komatsuzaki, A.: Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. arXiv preprint arXiv:2111.02114 (2021)
41. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O.: Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347 (2017)
42. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: Proc. Int. Conf. Mach. Learn. (2015)
43. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. Int. Conf. Learn. Represent. (2021)
44. Song, Y., Durkan, C., Murray, I., Ermon, S.: Maximum likelihood training of score-based diffusion models. Adv. Neural Inform. Process. Syst. **34**, 1415–1428 (2021)
45. Song, Y., Ermon, S.: Improved techniques for training score-based generative models. In: Adv. Neural Inform. Process. Syst. (2020)
46. Su, S., Lin, H., Hosu, V., Wiedemann, O., Sun, J., Zhu, Y., Liu, H., Zhang, Y., Saupe, D.: Going the extra mile in face image quality assessment: A novel database and model. IEEE Trans. Multimedia (2023)
47. Wang, Z.J., Montoya, E., Munechika, D., Yang, H., Hoover, B., Chau, D.H.: Diffusiondb: A large-scale prompt gallery dataset for text-to-image generative models. arXiv preprint arXiv:2210.14896 (2022)
48. Witteveen, S., Andrews, M.: Investigating prompt engineering in diffusion models. arXiv preprint arXiv:2211.15462 (2022)
49. Wu, H., Zhang, Z., Zhang, E., Chen, C., Liao, L., Wang, A., Li, C., Sun, W., Yan, Q., Zhai, G., et al.: Q-bench: A benchmark for general-purpose foundation models on low-level vision. Int. Conf. Learn. Represent. (2024)
50. Wu, H., Zhang, Z., Zhang, E., Chen, C., Liao, L., Wang, A., Xu, K., Li, C., Hou, J., Zhai, G., Xue, G., Sun, W., Yan, Q., Lin, W.: Q-instruct: Improving low-level visual abilities for multi-modality foundation models. IEEE Conf. Comput. Vis. Pattern Recog. (2024)
51. Wu, J.Z., Ge, Y., Wang, X., Lei, S.W., Gu, Y., Shi, Y., Hsu, W., Shan, Y., Qie, X., Shou, M.Z.: Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In: Int. Conf. Comput. Vis. pp. 7623–7633 (2023)
52. Wu, X., Sun, K., Zhu, F., Zhao, R., Li, H.: Better aligning text-to-image models with human preference. Int. Conf. Comput. Vis. (2023)

53. Xu, J., Liu, X., Wu, Y., Tong, Y., Li, Q., Ding, M., Tang, J., Dong, Y.: Imagereward: Learning and evaluating human preferences for text-to-image generation. Adv. Neural Inform. Process. Syst. (2023)
54. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Int. Conf. Comput. Vis. (2023)