Fine-Tuning Diffusion Models Using Reinforcement Learning

Shaocong Ma

October 9, 2025

Training Diffusion Models with Reinforcement Learning

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, Sergey Levine

UC Berkeley, MIT

ICLR 2024

http://rl-diffusion.github.io

Motivation

- Traditional diffusion models are trained to maximize data likelihood.
- However, real-world goals are often different:
  - Aesthetic quality (e.g., human preference)
  - Prompt-image alignment
  - Image compressibility
- These objectives are hard to specify via prompts or likelihoods.
- Key idea: **formulate denoising as a multi-step decision process.**
- This enables reinforcement learning to directly optimize black-box rewards.

Why Not Supervised Fine-Tuning?

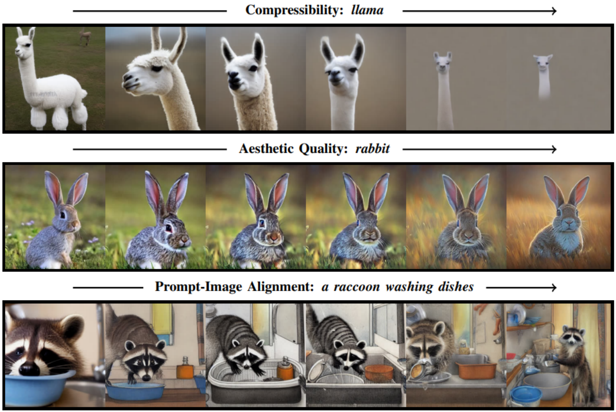
- **Reward is often non-differentiable or black-box:**
  - Human preference scores (e.g., aesthetics)
  - VLM-based feedback (e.g., LLaVA descriptions)
  - File size after compression
- **No supervision during denoising steps:**
  - Reward only applies to the final output x_0
  - No ground-truth data for intermediate steps
- **Sampling-based generation = sequential decision process**
  - Each step modifies the sample: $x_T \to \cdots \to x_0$
  - RL can assign credit across this trajectory
- **Conclusion:** RL is better suited for optimizing non-differentiable, sample-level objectives.

Goal of the Paper

- **Goal:** Train diffusion models to satisfy arbitrary downstream objectives.
- **Key insight:** Treat the denoising process as a multi-step Markov Decision Process (MDP).
- **Method:** Propose **DDPO** (Denoising Diffusion Policy Optimization), a policy gradient algorithm for optimizing diffusion models with black-box rewards.
- **Result:** Enables fine-tuning for tasks like aesthetic quality, alignment, and compressibility without additional human labels or prompt engineering.

Key Applications

- Improve image compressibility
- Increase aesthetic quality
- Enhance prompt-image alignment



Preliminaries: Diffusion Models

- Goal: model data distribution via a denoising process.
- Forward process: gradually add noise to data $x_0 \to x_T$.
- Reverse process: learn to recover data by removing noise $x_T \to x_0$.
- Trained to minimize denoising loss (variational lower bound on log-likelihood):

$$\mathcal{L}_{\text{DDPM}} = \mathbb{E}_{x_0,t,\epsilon}\left[\left\|\epsilon - \epsilon_\theta(x_t, t, c)\right\|^2\right]$$

- Common in text-to-image generation (e.g., Stable Diffusion).

Preliminaries: Reinforcement Learning

- RL solves sequential decision-making problems.
- Key elements:
  - States s, actions a, reward $r(s,a)$, transitions $P(s'|s,a)$
- Agent learns a policy $\pi(a|s)$ to maximize expected cumulative reward:

$$J(\pi) = \mathbb{E}_\pi \left[\sum_t r(s_t, a_t) \right]$$

- In this paper: **denoising steps** $\sim$ **actions**, final image $\sim$ reward.

Core Idea: Denoising as a Markov Decision Process (MDP)

- Key idea: model the diffusion sampling process as a multi-step decision process.
- Each denoising step becomes an RL timestep:

$$\text{State: } s_t = (x_t, t, c) \qquad \text{Action: } a_t = x_{t-1}$$

- Policy: $\pi_\theta(a_t|s_t) = p_\theta(x_{t-1}|x_t, c)$
- Transition: deterministic, x_{t-1} becomes new state at $t-1$
- Reward: only at final step:

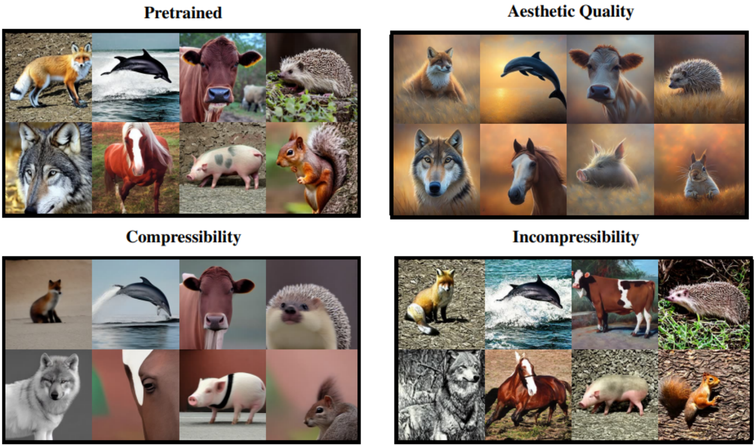
$$r(s_t, a_t) = \begin{cases} r(x_0, c), & \text{if } t = 0 \\ 0, & \text{otherwise} \end{cases}$$

- This formulation allows applying policy gradient methods to train diffusion models.

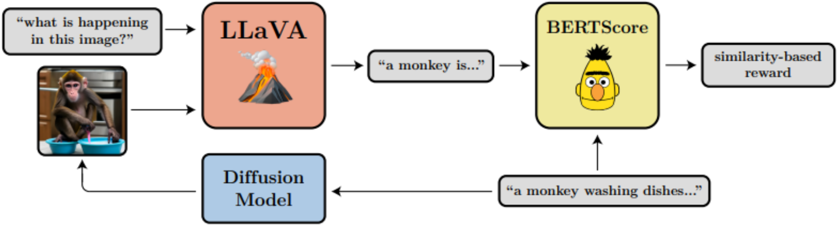
Reward Functions

- Compressibility: JPEG file size
- Aesthetics: LAION-predicted score
- Prompt alignment: LLaVA + BERTScore

DDPO Samples



Prompt-Image Alignment via VLMs



Sample Improvements

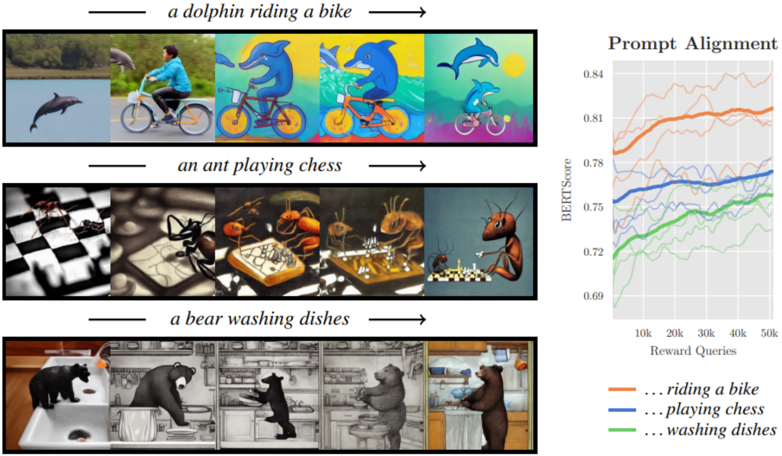


Figure: Visual effects of DDPO finetuning (Prompt alignment)

Overoptimization Risks

- Overfitting reward functions can degrade output quality
- Add KL-penalty or use early stopping

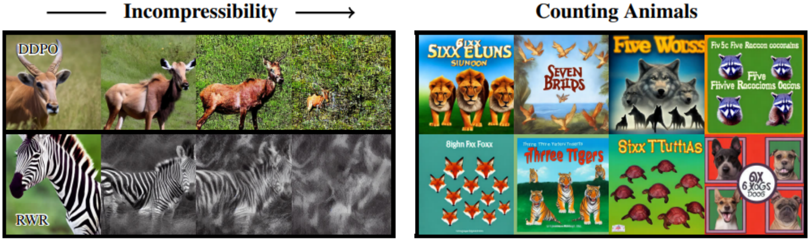


Figure 7 (Reward model overoptimization) Examples of RL overoptimizing reward functions. **(L)** The diffusion model eventually loses all recognizable semantic content and produces noise when optimizing for incompressibility. **(R)** When optimized for prompts of the form "n *animals*", the diffusion model exploits the VLM with a typographic attack (Goh et al., 2021), writing text that is interpreted as the specified number n instead of generating the correct number of animals.

Conclusion

- DDPO is a powerful framework for RL-trained diffusion
- Enables optimization of diverse, meaningful reward functions
- Generalizes across prompts without new data

DPOK: Reinforcement Learning for Fine-tuning Text-to-Image Diffusion Models

Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, Kimin Lee

Google Research, UC Berkeley, Amazon, KAIST, UW-Madison

NeurIPS 2023

https://arxiv.org/abs/2305.16381

Similarity to DDPO

Aspect	DDPO (Black et al., 2024)	DPOK (Fan et al., 2023)
MDP Formulation (same)	Multi-step MDP: State: (c, t, x_t) Action: x_{t-1} Reward: only at $t = 0$	Multi-step MDP: State: (z, x_t) Action: x_{t-1} Reward: only at $t = T - 1$
Policy Definition (same)	$\pi(a_t \mid s_t) = p_\theta(x_{t-1} \mid x_t, c)$	$\pi_\theta(a_t \mid s_t) = p_\theta(x_{t-1} \mid x_t, z)$
Reward Source	JPEG compressibility, Aesthetic score (LAION), VLM + BERTScore	ImageReward model trained from human feedback

Table: Comparison of DDPO and DPOK: MDP formulation, policy structure, and reward signal.

Difference

Policy Gradient estimation:

- (same, policy gradient theorem) $r \cdot \nabla \log p_\theta$
- (difference) DPOK contains the gradient of KL divergence.

Practical implementation: PPO.

Other Differences to DDPO

Aspect	DDPO (Black et al., 2024)	DPOK (Fan et al., 2023)
KL Regularization	Optional; PPO-style clipping used for stability	Essential; KL divergence between fine-tuned and pre-trained model is a core regularizer
Tasks	Compressibility / Incompressibility, Aesthetic optimization, Prompt-image alignment	Color, count, composition, location, Bias correction (e.g., "Four roses" as flower not whiskey)
Model Used	Stable Diffusion v1.4	Stable Diffusion v1.5 with LoRA

Table: Comparison of DDPO and DPOK: KL regularization, tasks, and base model.

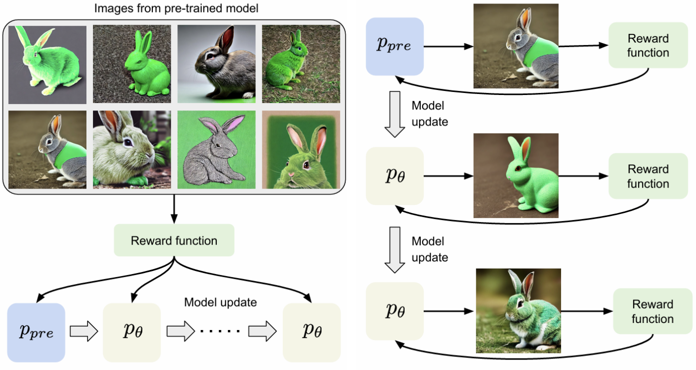
Motivation

- Text-to-image diffusion models often fail on fine-grained details:
  - Object count, color, spatial composition
- Learning from human feedback (LHF) improves alignment with user intent
- Supervised fine-tuning struggles with data quality and overfitting
- Key idea: **Use online reinforcement learning to optimize feedback-trained reward functions**

Goal of the Paper

- **Goal:** Fine-tune diffusion models using online RL to optimize human preference-based rewards
- **Method:** Propose **DPOK** – Policy gradient with KL regularization
- **Result:** Outperforms supervised methods on alignment and image quality

Supervised vs. RL Fine-Tuning (Figure 1)



(a) Supervised fine-tuning (b) RL fine-tuning

Figure 1: Illustration of (a) reward-weighted supervised fine-tuning and (b) RL fine-tuning. Both start with the same pre-trained model (the blue rectangle). In supervised fine-tuning, the model is updated on a fixed dataset generated by the pre-trained model. In contrast, the model is updated using new samples from the previously trained model during online RL fine-tuning.

MDP Formulation of Denoising

- Treat diffusion process as Markov Decision Process (MDP):

$$s_t = (z, x_{T-t}), \quad a_t = x_{T-t-1}$$

- Policy: $\pi_\theta(a_t|s_t) = p_\theta(x_{t-1}|x_t, z)$
- Deterministic transition, reward only at $t = 0$: $r(x_0, z)$
- RL objective:

$$\min_\theta \mathbb{E}_z \left[\mathbb{E}_{p_\theta(x_0|z)}[-r(x_0, z)]\right]$$

KL Regularization (Main difference)

- Prevent overfitting to reward by penalizing deviation from pre-trained model
- Use upper bound of KL divergence (Lemma 4.2):

$$\mathrm{KL}(p_\theta(x_0|z) \| p_{\text{pre}}(x_0|z)) \leq \sum_t \mathrm{KL}(p_\theta(x_{t-1}|x_t, z) \| p_{\text{pre}}(x_{t-1}|x_t, z))$$

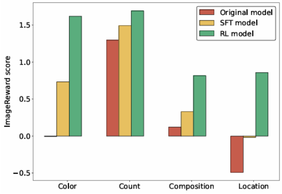
- Final RL objective:

$$\mathbb{E}_{p_\theta(x_{0:T}|z)}[-\alpha r(x_0, z) + \beta \sum_t \mathrm{KL}]$$

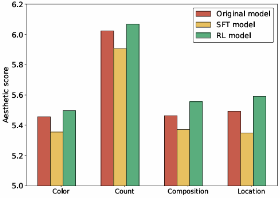
Qualitative Comparison (Figure 2)

Figure 2: Comparison of images generated by the original Stable Diffusion model, supervised fine-tuned (SFT) model, and RL fine-tuned model. Images in the same column are generated with the same random seed. Images from seen text prompts: "A green colored rabbit" (color), "A cat and a dog" (composition), "Four wolves in the park" (count), and "A dog on the moon" (location).

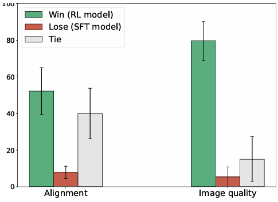
Quantitative Evaluation (Figure 3)



(a) ImageReward score



(b) Aesthetic score



(c) Human evalution

Figure 3: (a) ImageReward scores and (b) Aesthetic scores of three models: the original model, supervised fine-tuned (SFT) model, and RL fine-tuned model. ImageReward and Aesthetic scores are averaged over 50 samples from each model. (c) Human preference rates between RL model and SFT model in terms for image-text alignment and image quality. The results show the mean and standard deviation averaged over eight independent human raters.

Conclusion

- DPOK is a robust framework for RL fine-tuning of diffusion models
- Combines reward feedback and KL regularization
- Outperforms supervised methods on alignment and quality
- Demonstrates potential of RLHF in generative modeling

Thank You!