Training Diffusion Models with Reinforcement Learning

Kevin Black Michael Janner Yilun Du Ilya Kostrikov Sergey Levine

https://arxiv.org/abs/2305.13301

Presenter: Yulai Zhao

Compressibility: *llama*

Incompressibility: *bird*

Aesthetic Quality: *rabbit*

Prompt-Image Alignment: *a raccoon washing dishes*

RL training

Introduction

- Diffusion models are powerful generative models that can generate high-quality images, videos, 3D shapes, etc.
- They work by adding noise to data through a chain, then training a model to reverse the process.
- Conventional training maximizes a variational lower bound on data log-likelihood.

Motivation

- However, most use cases of diffusion models are not **directly** concerned with matching the training data, but instead with a downstream objective. We don't just want an image that looks like existing images, but one that has a **specific** type of appearance; we don't just want a drug molecule that is physically plausible, but one that is as effective as possible.

How diffusion models can be trained on these downstream objectives directly using reinforcement learning (RL)?

TL;DR

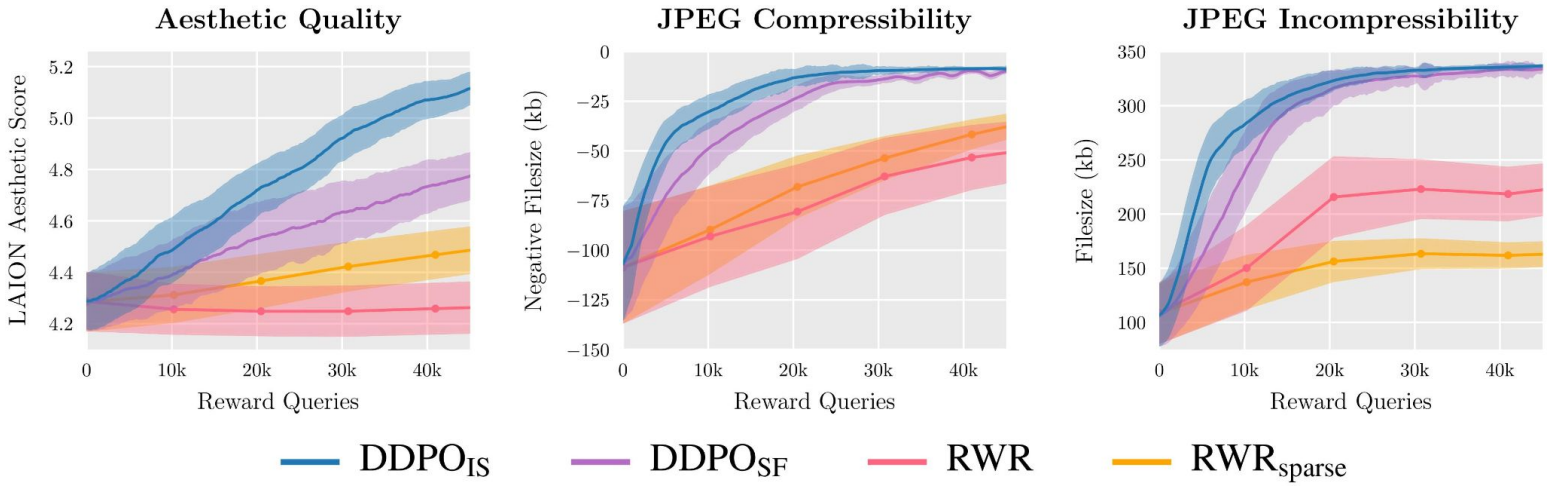
- We describe how posing denoising as a **multi-step decision-making** problem enables a class of policy gradient algorithms, which we refer to as denoising diffusion policy optimization (DDPO)
- This paper finetuned Stable Diffusion on a variety of objectives, including **image compressibility**, **human-perceived aesthetic quality**, and **prompt-image alignment**.

Using RL for Diffusion Models: Formulations

- The key insight is we can better maximize the reward of the final sample: each denoising step is an action, and the agent only gets a reward on the final step of each denoising trajectory. (i.e., multi-step decision making)
- Each denoising step is an action, states are noise levels and contexts.
- Reward is given only for final denoised sample based on downstream objective.
- Our goal is for the diffusion model to generate samples that maximize this **reward function**.
- This allows applying policy gradient algorithms to directly optimize arbitrary reward functions.
- The proposed method is called Denoising Diffusion Policy Optimization (DDPO).

Experiments

- Tested on image compressibility, aesthetics, and text-to-image alignment tasks.
- DDPO variants significantly outperformed baselines like Reward Weighted Regression.
- Qualitative results showed DDPO can effectively optimize obscure objectives.



Finetuning Stable Diffusion Using DDPO

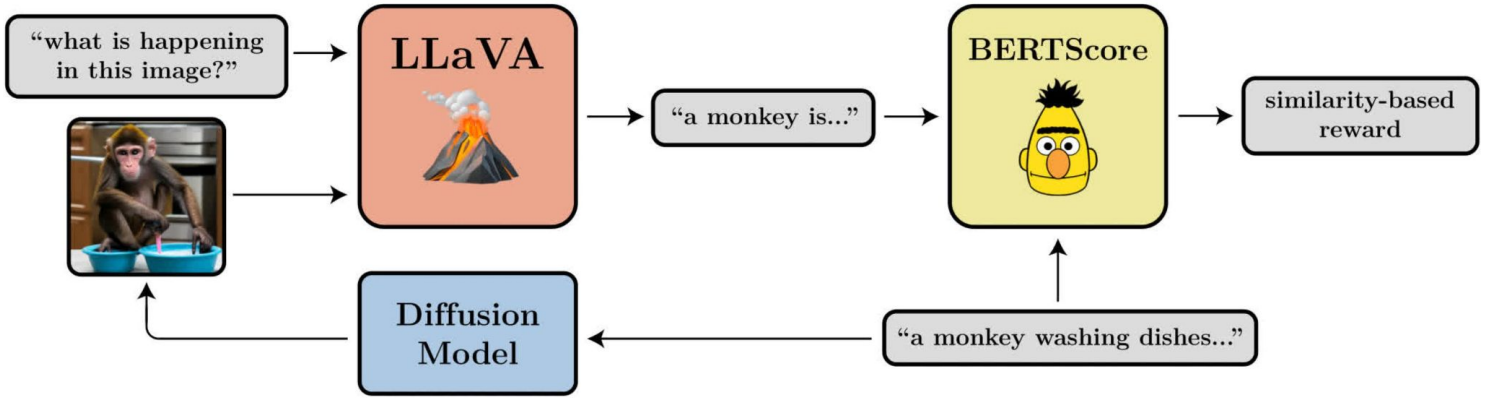
We have four tasks, each defined by a different reward function:

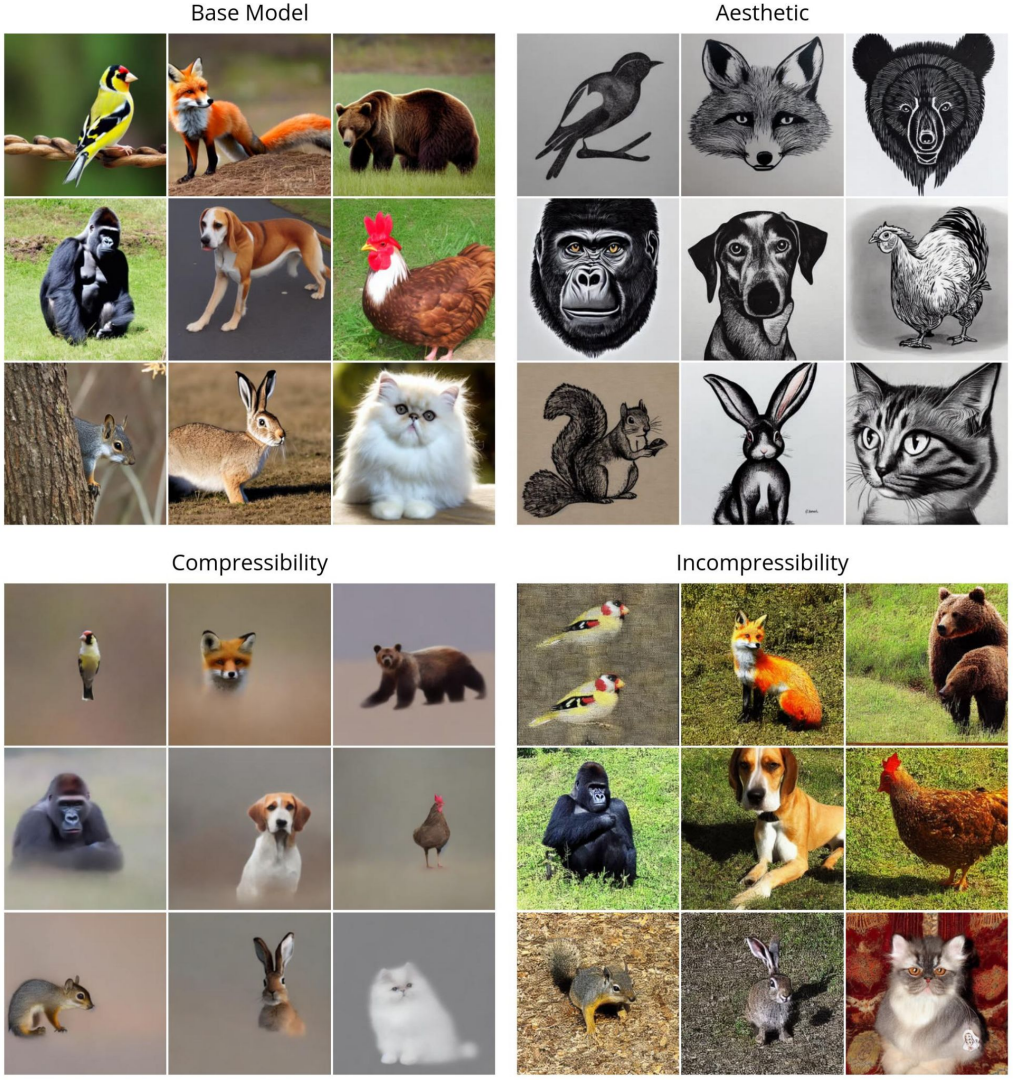
- **Compressibility**: How easy is the image to compress using the JPEG algorithm? The reward is the negative file size of the image (in kB) when saved as a JPEG.
- **Incompressibility**: Very similar. How hard is the image to compress using the JPEG algorithm?
- **Aesthetic Quality**: How aesthetically appealing is the image to human? The reward is the output of the LAION aesthetic predictor, which is a NNs trained on human preferences.

Finetuning Stable Diffusion Using DDPO

We have four tasks, each defined by a different reward function:

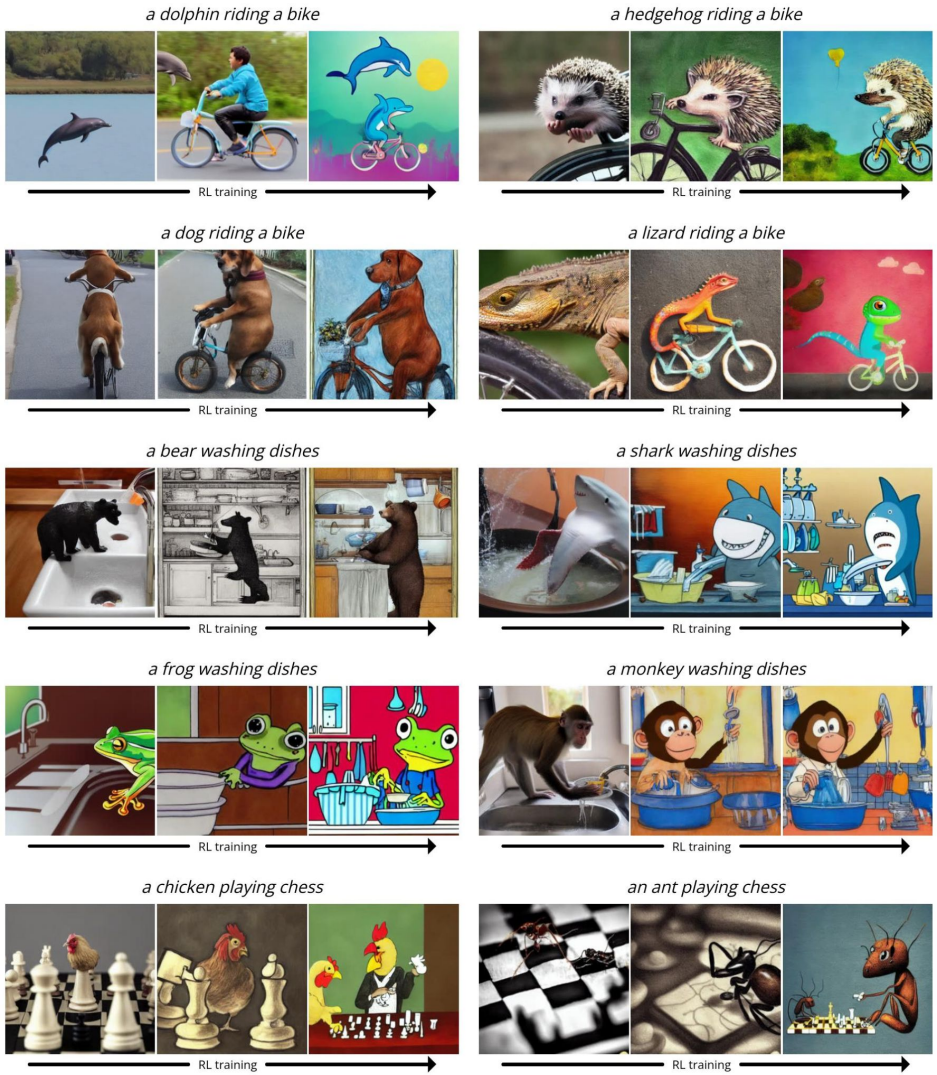
- **Prompt-Image Alignment**: How well does the image represent what was asked for in the prompt? Feed the image into LLaVA, ask it to describe the image, and then compute the similarity between that description and the original prompt using BERTScore. Thus, no human-labelling needed.





Performance of DDPO on the simple rewards

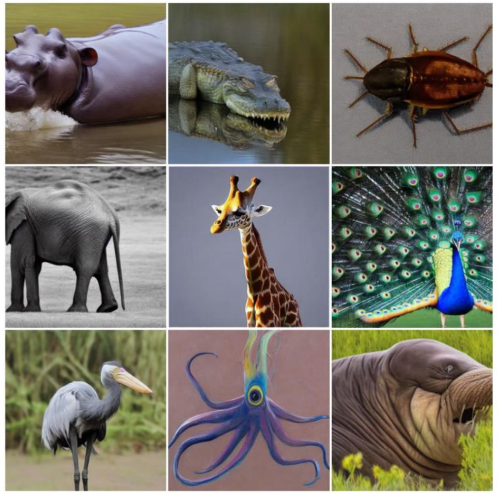
- The first image illustrates what “vanilla” Stable Diffusion generates for nine different animals;
- All of the RL-finetuned models show a clear qualitative difference. Interestingly, the aesthetic quality model (top right) tends towards minimalist black-and-white line drawings, revealing the kinds of images that the LAION aesthetic predictor considers “more aesthetic”



Performances on more complex prompt-image alignment task

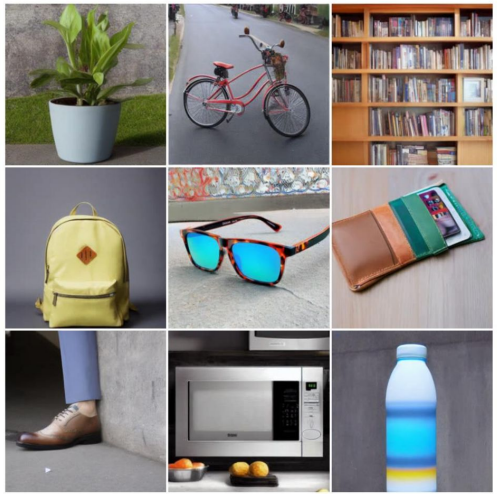
- Each series of three images shows samples for the same prompt and random seed over time, with the first sample coming from vanilla Stable Diffusion.
- Interestingly, the model shifts towards a more cartoon-like style
- Why?

New Animals: Base Model

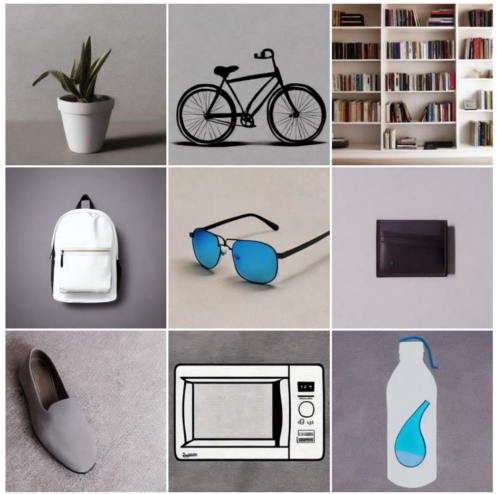


New Animals: Aesthetic

Non-Animals: Base Model

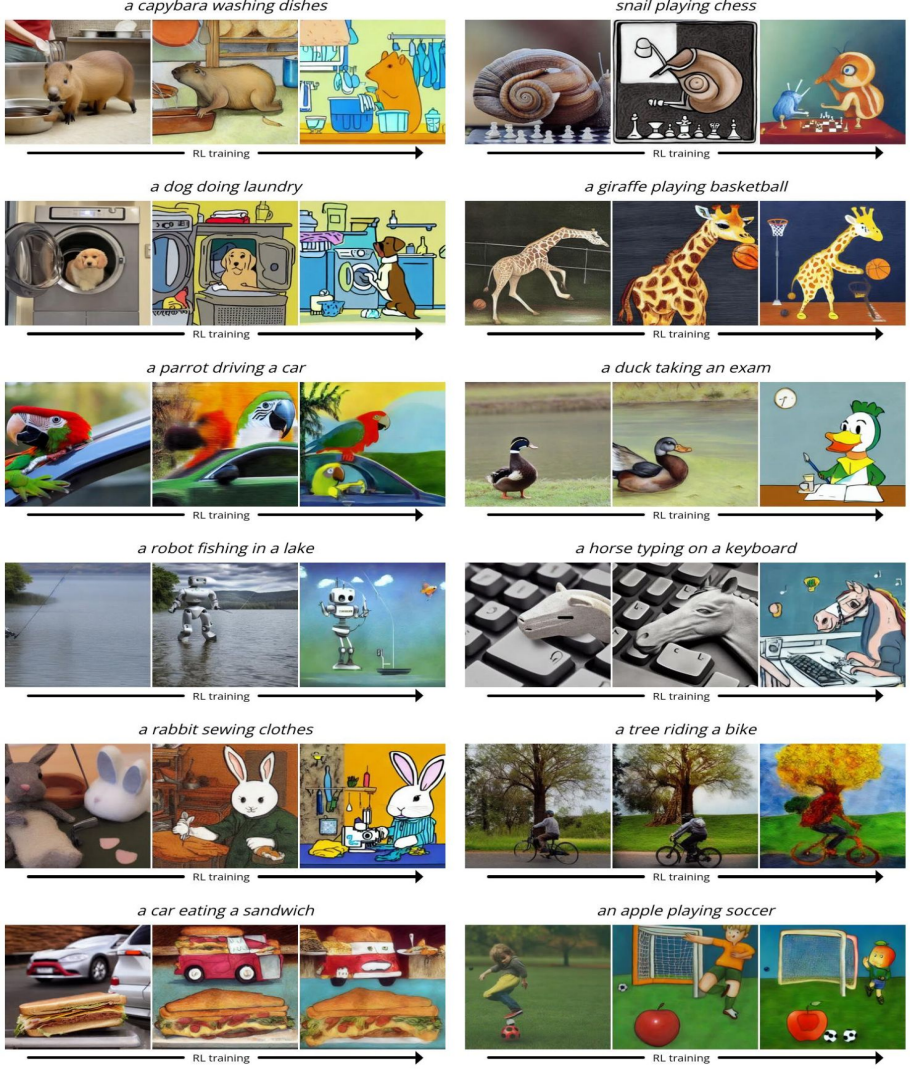


Non-Animals: Aesthetic



Unexpected Generalization

- The aesthetic quality model was finetuned using prompts that were selected from a list of 45 common animals.
- We find that it generalizes not only to unseen animals but also to everyday objects.

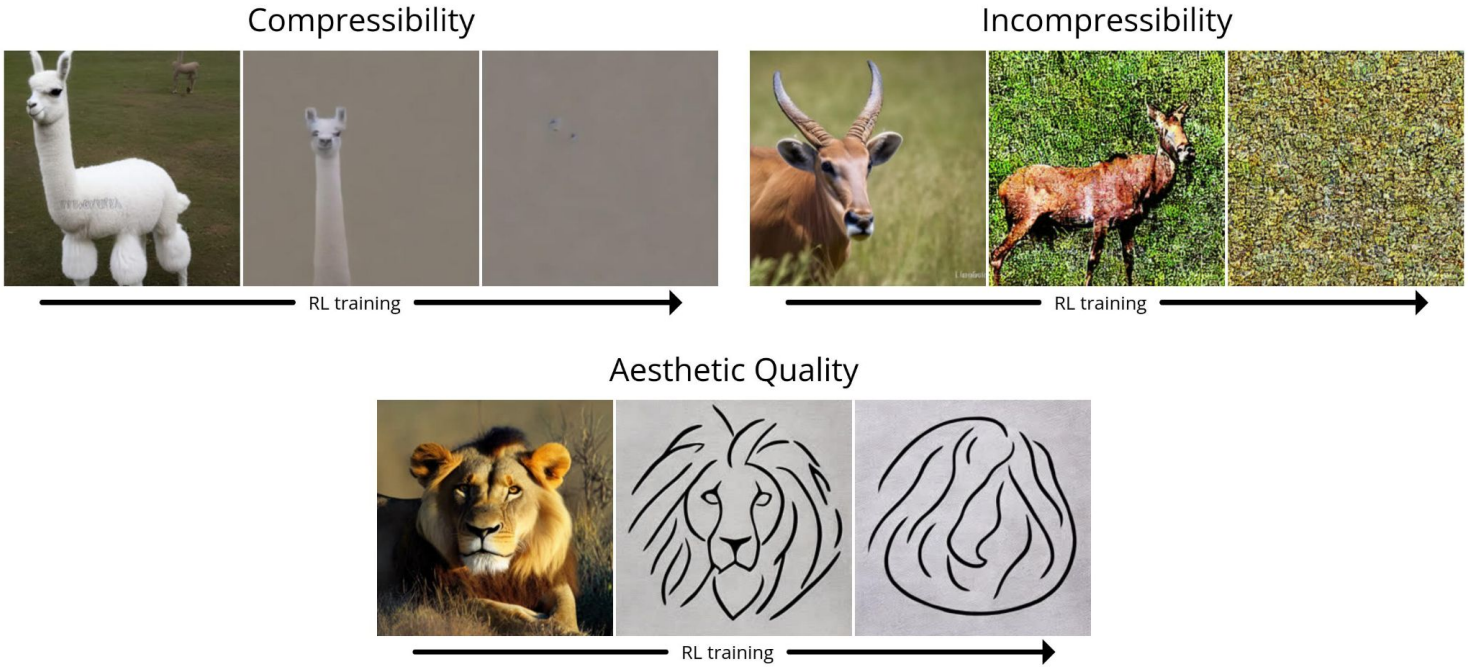


Unexpected Generalization

- The prompt-image alignment model used the same list of 45 common animals during training, and only three activities.
- We find that it generalizes not only to unseen animals but also to unseen activities, and even novel combinations of the two.

Limitations

- Like other RL fine-tuning methods, can overfit to reward and degrade image quality, aka “reward overoptimization”



Conclusions

- Diffusion models are good at producing complex, high-dimensional outputs. However, so far they've mostly been successful in applications where the goal is to learn patterns from lots of data
- DDPO enables directly optimizing diffusion models for user-specified goals.
- It is effective on diverse objectives that are hard to specify via likelihoods.
- Automated rewards from VLMs enable optimizing obscure objectives.
- The approach improves controllability without additional human labeling.