PRDP: Proximal Reward Difference Prediction for Large-Scale Reward Finetuning of Diffusion Models

Fei Deng[1,2], Qifei Wang[1], Wei Wei[1,3], Matthias Grundmann[1], Tingbo Hou[1,4]

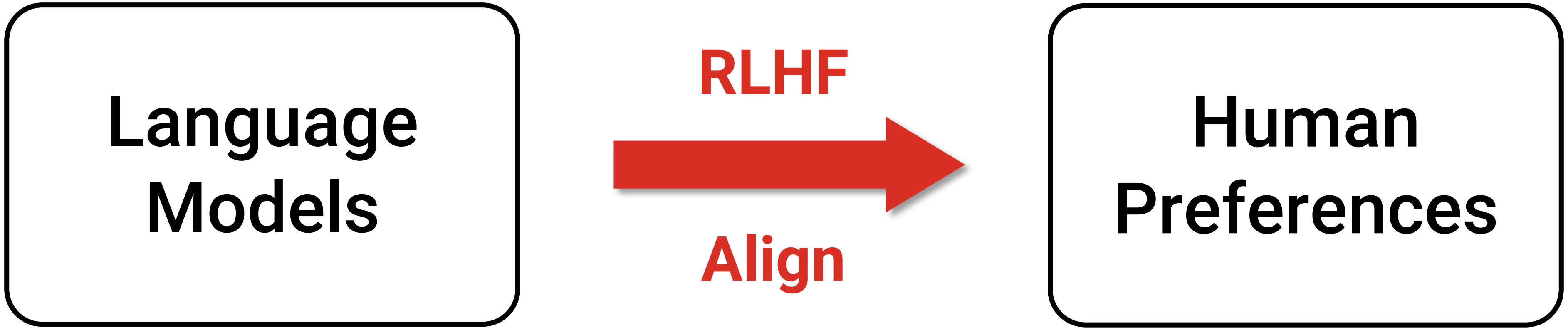
[1]Google, [2]Rutgers University, [3]Accenture, [4]Meta

A painting of a girl standing on a mountain looking out at an approaching storm over the ocean, with wind blowing and ocean mist, surrounded by lightning.

PRDP Training

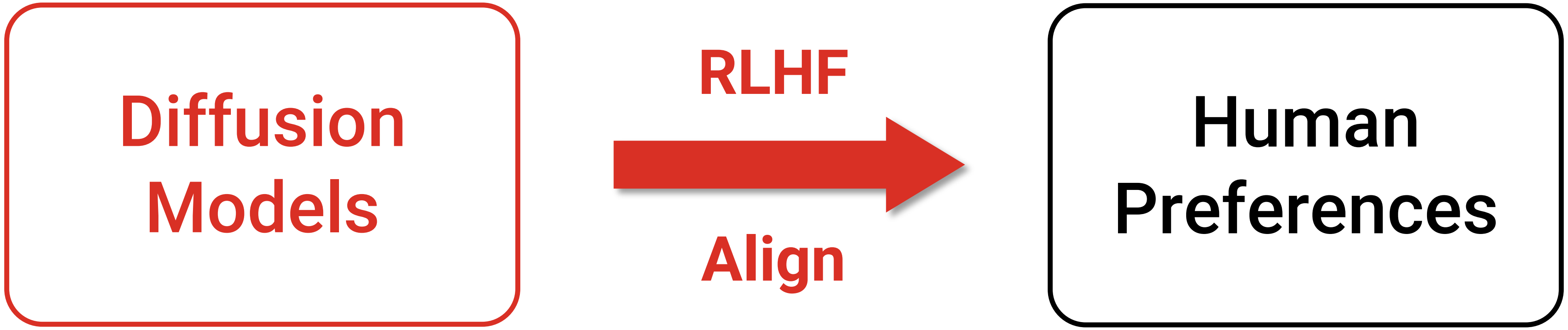
Aligning with Human Preferences

- Numerous products

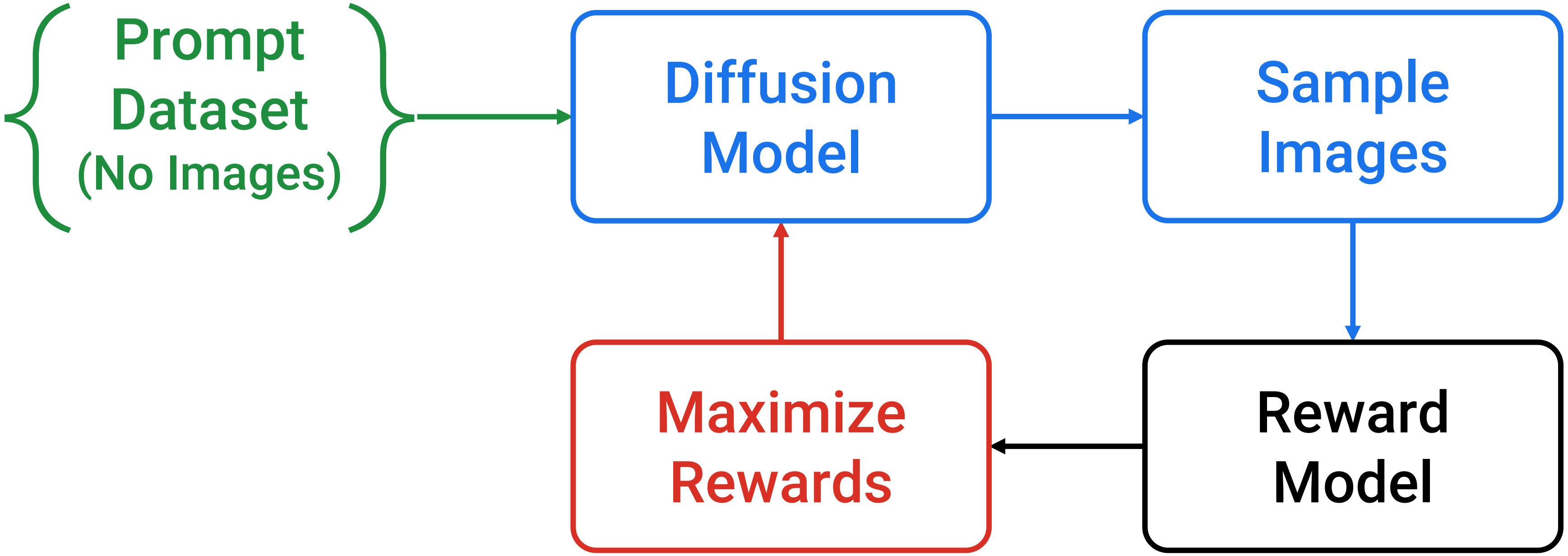


Aligning with Human Preferences

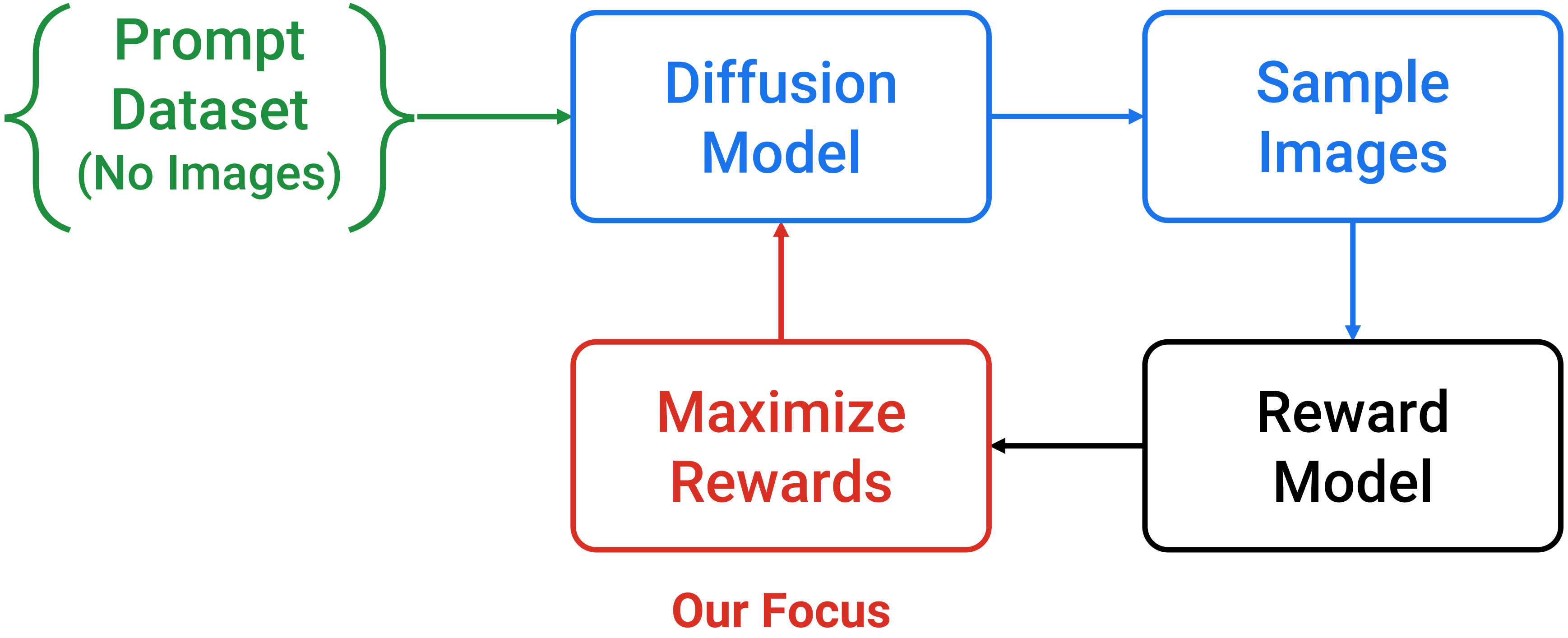
- Our paper



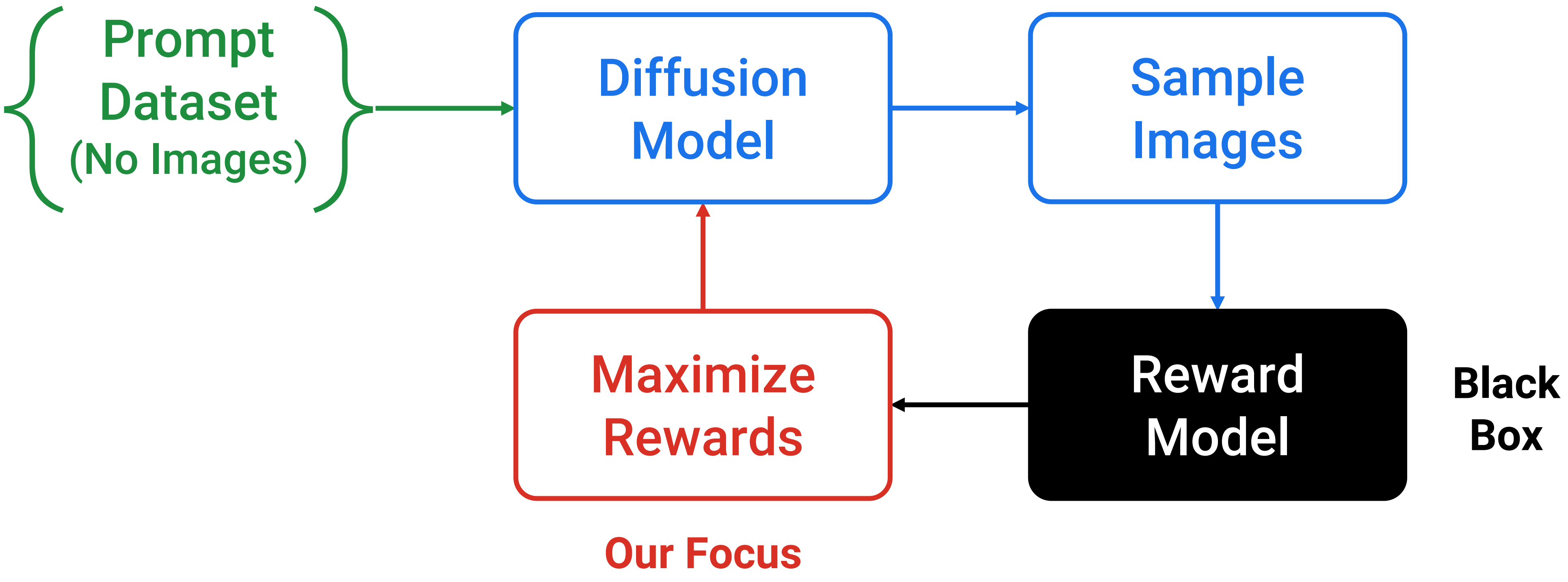
RLHF Pipeline



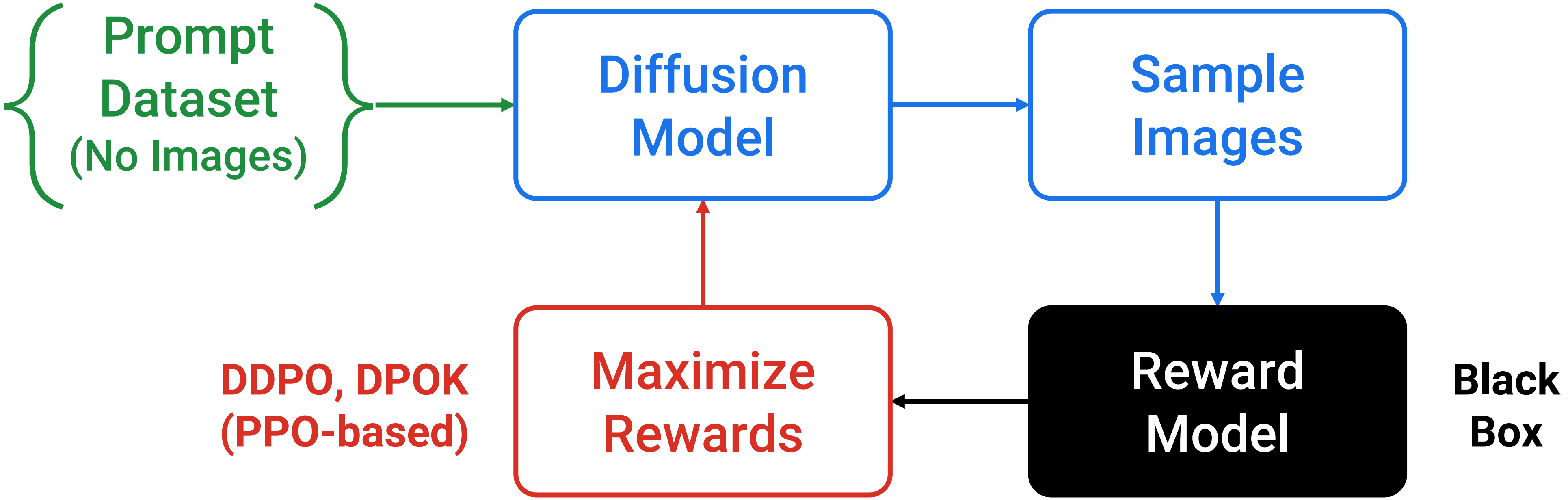
RLHF Pipeline



RLHF Pipeline

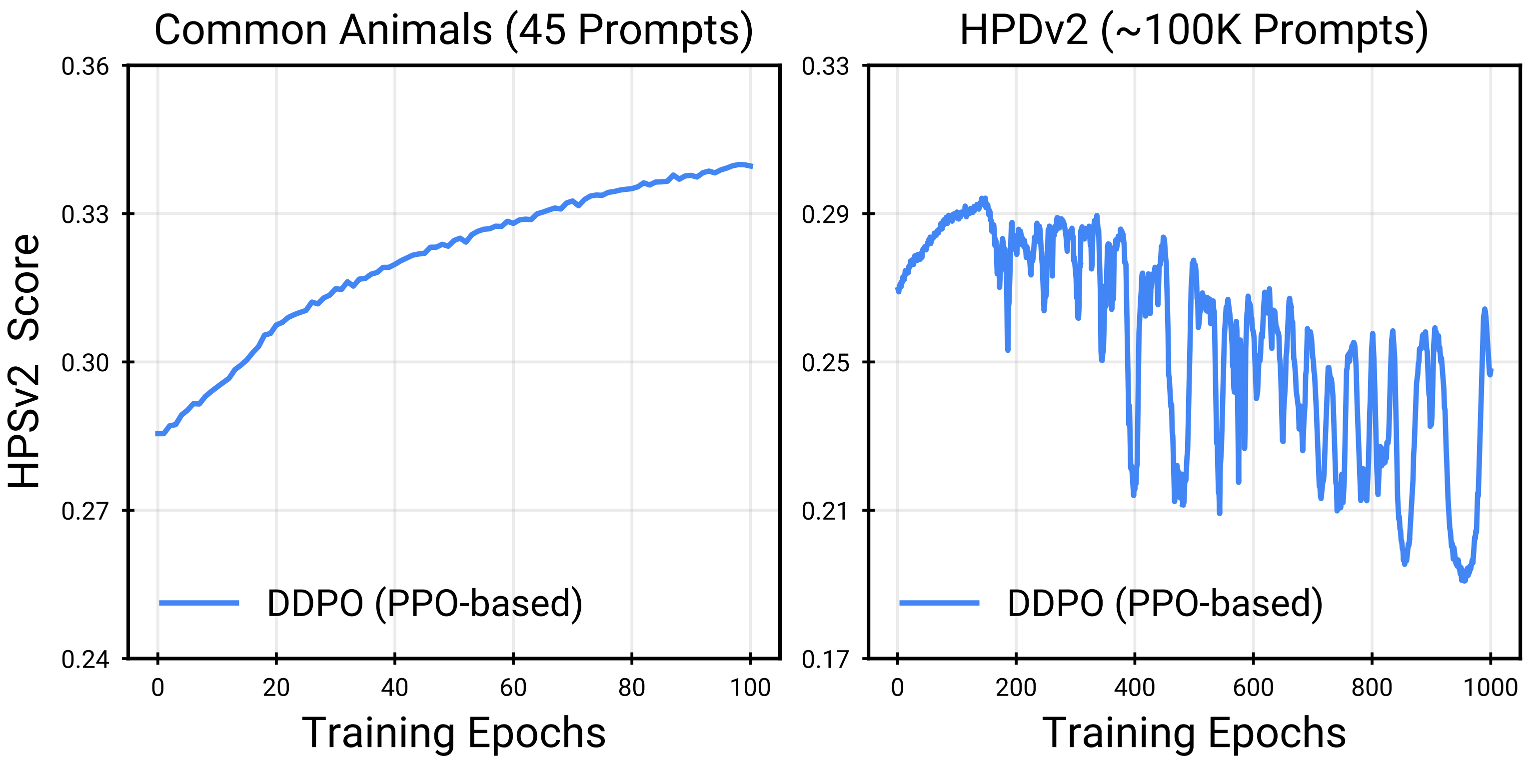


RLHF Pipeline

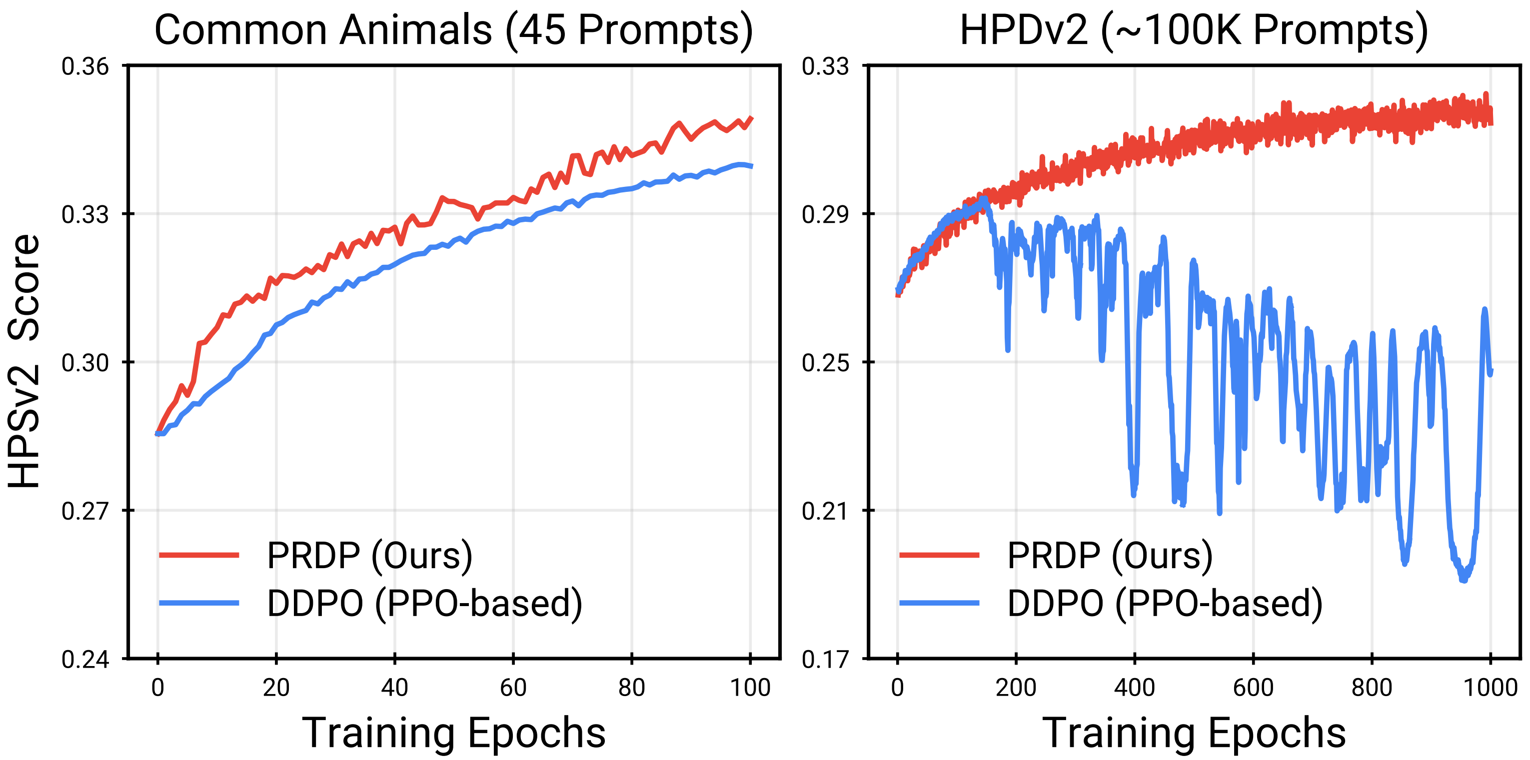


Black *et al*. Training Diffusion Models with Reinforcement Learning. *ICLR 2024*.
Fan *et al*. DPOK: Reinforcement Learning for Fine-tuning Text-to-Image Diffusion Models. *NeurIPS 2023*.
Schulman *et al*. Proximal Policy Optimization Algorithms.

Previous Methods: Unstable at Large Scale



Our Contribution: Stable Large-Scale Training



Method Overview: Novel Training Objective

RLHF Objective

$$\max_{\pi_\theta} \ \mathbb{E}\Big[r - \beta \mathrm{KL}\big[\pi_\theta \parallel \pi_{\mathrm{ref}}\big]\Big]$$

Reward Maximization

Equivalent

More Stable

PRDP Objective

$$\min_{\theta} \ \mathbb{E}\Big\| \Delta \hat{r}_\theta - \Delta r / \beta \Big\|^2$$

Reward Difference Prediction

RLHF Objective

$$\pi_{\theta^\star} = \arg\max_{\underbrace{\pi_\theta}_{\text{Diffusion Model}}} \mathbb{E}_{\mathbf{c}\sim \underbrace{p(\mathbf{c})}_{\text{Prompt Dataset}}} \Bigg[\mathbb{E}_{\overbrace{\mathbf{x}_{0:T}}^{\text{Denoising Trajectory}} \sim \underbrace{\pi_\theta}_{\text{Diffusion Model}}(\mathbf{x}_{0:T}|\mathbf{c})} \Big[\overbrace{r(\mathbf{x}_0, \mathbf{c})}^{\text{Reward}} \Big] - \beta \mathrm{KL} \Big[\underbrace{\pi_\theta}_{\text{Diffusion Model}}(\mathbf{x}_{0:T}|\mathbf{c}) \,\|\, \underbrace{\pi_{\text{ref}}}_{\text{Pretrained Diffusion Model}}(\mathbf{x}_{0:T}|\mathbf{c}) \Big] \Bigg]$$

Optimal Solution

$$\pi_{\theta^\star}(\mathbf{x}_{0:T}|\mathbf{c}) = \underbrace{\frac{1}{Z(\mathbf{c})}}_{\text{Intractable}}\pi_{\text{ref}}(\mathbf{x}_{0:T}|\mathbf{c})\exp\left(\frac{1}{\beta}r(\mathbf{x}_0,\mathbf{c})\right)$$

Optimal Solution

$$\pi_{\theta^\star}(\mathbf{x}_{0:T}|\mathbf{c}) = \frac{1}{\underline{Z(\mathbf{c})}}\pi_{\text{ref}}(\mathbf{x}_{0:T}|\mathbf{c})\exp\left(\frac{1}{\beta}r(\mathbf{x}_0,\mathbf{c})\right)$$

Intractable

Equivalent Condition

$$\log\frac{\pi_{\theta^\star}(\mathbf{x}_{0:T}|\mathbf{c})}{\pi_{\text{ref}}(\mathbf{x}_{0:T}|\mathbf{c})} = \frac{1}{\beta}r(\mathbf{x}_0,\mathbf{c}) - \log Z(\mathbf{c})$$

Rafailov *et al*. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. *NeurIPS 2023*.

Optimal Solution

$$\pi_{\theta^\star}(\mathbf{x}_{0:T}|\mathbf{c}) = \frac{1}{\underline{Z(\mathbf{c})}}\pi_{\text{ref}}(\mathbf{x}_{0:T}|\mathbf{c})\exp\left(\frac{1}{\beta}r(\mathbf{x}_0,\mathbf{c})\right)$$

Intractable

Equivalent Condition

$$\log\frac{\pi_{\theta^\star}(\mathbf{x}_{0:T}|\mathbf{c})}{\pi_{\text{ref}}(\mathbf{x}_{0:T}|\mathbf{c})} = \frac{1}{\beta}r(\mathbf{x}_0,\mathbf{c}) - \log Z(\mathbf{c})$$

Cancel logZ by considering two denoising trajectories

Rafailov *et al*. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. *NeurIPS 2023*.

Equivalent Condition

$$\log \frac{\pi_{\theta^\star}(\mathbf{x}_{0:T}|\mathbf{c})}{\pi_{\text{ref}}(\mathbf{x}_{0:T}|\mathbf{c})} = \frac{1}{\beta} r(\mathbf{x}_0, \mathbf{c}) - \log Z(\mathbf{c})$$

Cancel logZ by considering two denoising trajectories

Equivalent Condition

$$\log \frac{\pi_{\theta^\star}(\mathbf{x}^a_{0:T}|\mathbf{c})}{\pi_{\text{ref}}(\mathbf{x}^a_{0:T}|\mathbf{c})} - \log \frac{\pi_{\theta^\star}(\mathbf{x}^b_{0:T}|\mathbf{c})}{\pi_{\text{ref}}(\mathbf{x}^b_{0:T}|\mathbf{c})} = \frac{r(\mathbf{x}^a_0, \mathbf{c}) - r(\mathbf{x}^b_0, \mathbf{c})}{\beta}$$

Rafailov *et al*. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. *NeurIPS 2023*.

Equivalent Condition

$$\log \frac{\pi_{\theta^\star}(\mathbf{x}^a_{0:T}|\mathbf{c})}{\pi_{\text{ref}}(\mathbf{x}^a_{0:T}|\mathbf{c})} - \log \frac{\pi_{\theta^\star}(\mathbf{x}^b_{0:T}|\mathbf{c})}{\pi_{\text{ref}}(\mathbf{x}^b_{0:T}|\mathbf{c})} = \frac{r(\mathbf{x}^a_0, \mathbf{c}) - r(\mathbf{x}^b_0, \mathbf{c})}{\beta}$$

$$\hat{r}_\theta(\mathbf{x}_{0:T}, \mathbf{c}) := \log \frac{\pi_\theta(\mathbf{x}_{0:T}|\mathbf{c})}{\pi_{\text{ref}}(\mathbf{x}_{0:T}|\mathbf{c})}$$

Equivalent Condition

$$\underbrace{\hat{r}_{\theta^\star}(\mathbf{x}^a_{0:T}, \mathbf{c}) - \hat{r}_{\theta^\star}(\mathbf{x}^b_{0:T}, \mathbf{c})}_{\text{Reward Difference Prediction}} = \frac{r(\mathbf{x}^a_0, \mathbf{c}) - r(\mathbf{x}^b_0, \mathbf{c})}{\beta}$$

CVPR SEATTLE, WA JUNE 17-21, 2024

Equivalent Condition

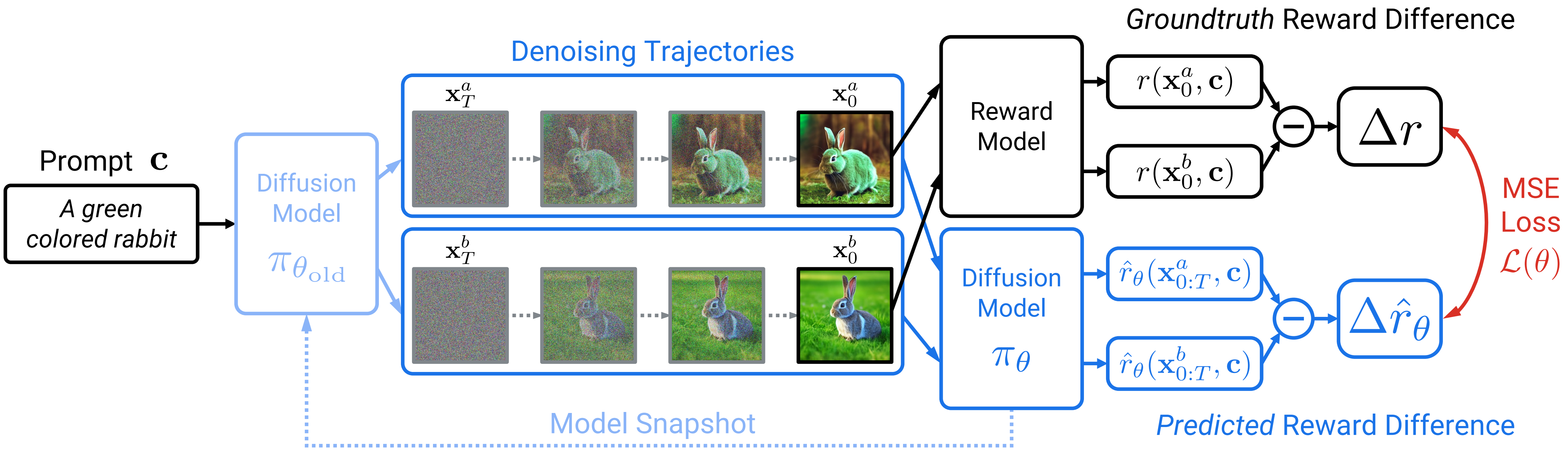
$$\underbrace{\hat{r}_{\theta^\star}(\mathbf{x}^a_{0:T}, \mathbf{c}) - \hat{r}_{\theta^\star}(\mathbf{x}^b_{0:T}, \mathbf{c})}_{\text{Reward Difference Prediction}} = \frac{r(\mathbf{x}^a_0, \mathbf{c}) - r(\mathbf{x}^b_0, \mathbf{c})}{\beta}$$

$$\pi_\theta = \pi_{\theta^\star} \iff \mathcal{L}(\theta) = 0$$

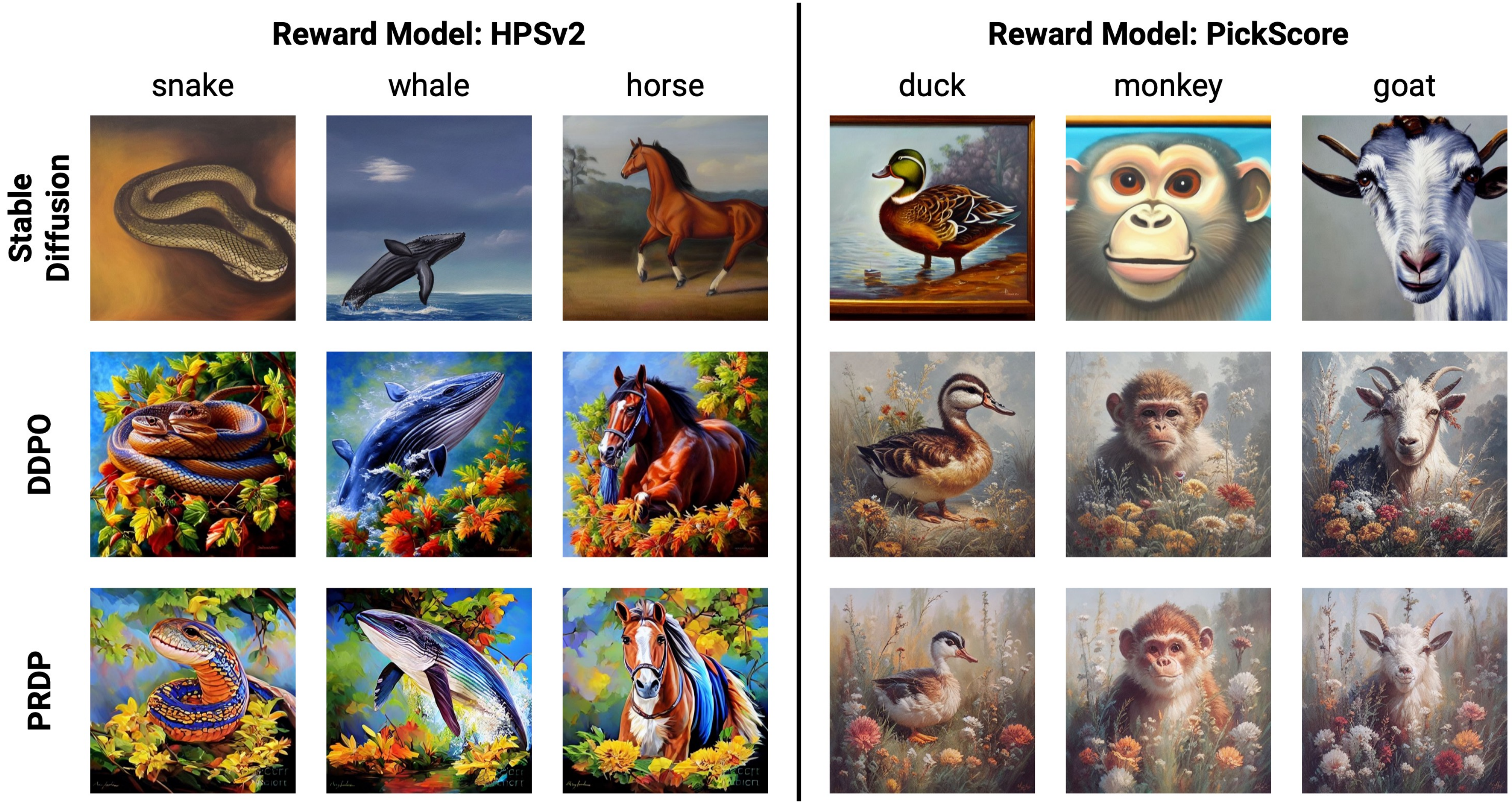
PRDP Objective

$$\mathcal{L}(\theta) = \mathbb{E}_{\mathbf{x}^a_{0:T}, \mathbf{x}^b_{0:T}, \mathbf{c}} \left\| \overbrace{\left[\hat{r}_\theta(\mathbf{x}^a_{0:T}, \mathbf{c}) - \hat{r}_\theta(\mathbf{x}^b_{0:T}, \mathbf{c})\right]}^{\text{Predicted Reward Difference}} - \overbrace{\left[r(\mathbf{x}^a_0, \mathbf{c}) - r(\mathbf{x}^b_0, \mathbf{c})\right]}^{\text{Groundtruth Reward Difference}} / \beta \right\|^2$$

Online Training Pipeline



Small-Scale Training (45 Prompts)



Wu *et al*. Human Preference Score v2: A Solid Benchmark for Evaluating Human Preferences of Text-to-Image Synthesis.
Kirstain *et al*. Pick-a-Pic: An Open Dataset of User Preferences for Text-to-Image Generation. *NeurIPS 2023*.

Large-Scale Training (~100K Prompts)

Reward Model: HPSv2

cinematic still of highly reflective stainless steel train in the desert, at sunset

The image is a wooden sculpture of a cute robot with cat ears, displayed in a contemporary art gallery.

A chibi frog character surfing at the beach.

Reward Model: PickScore

An anthropomorphic frog wizard wearing a cape and holding a wand.

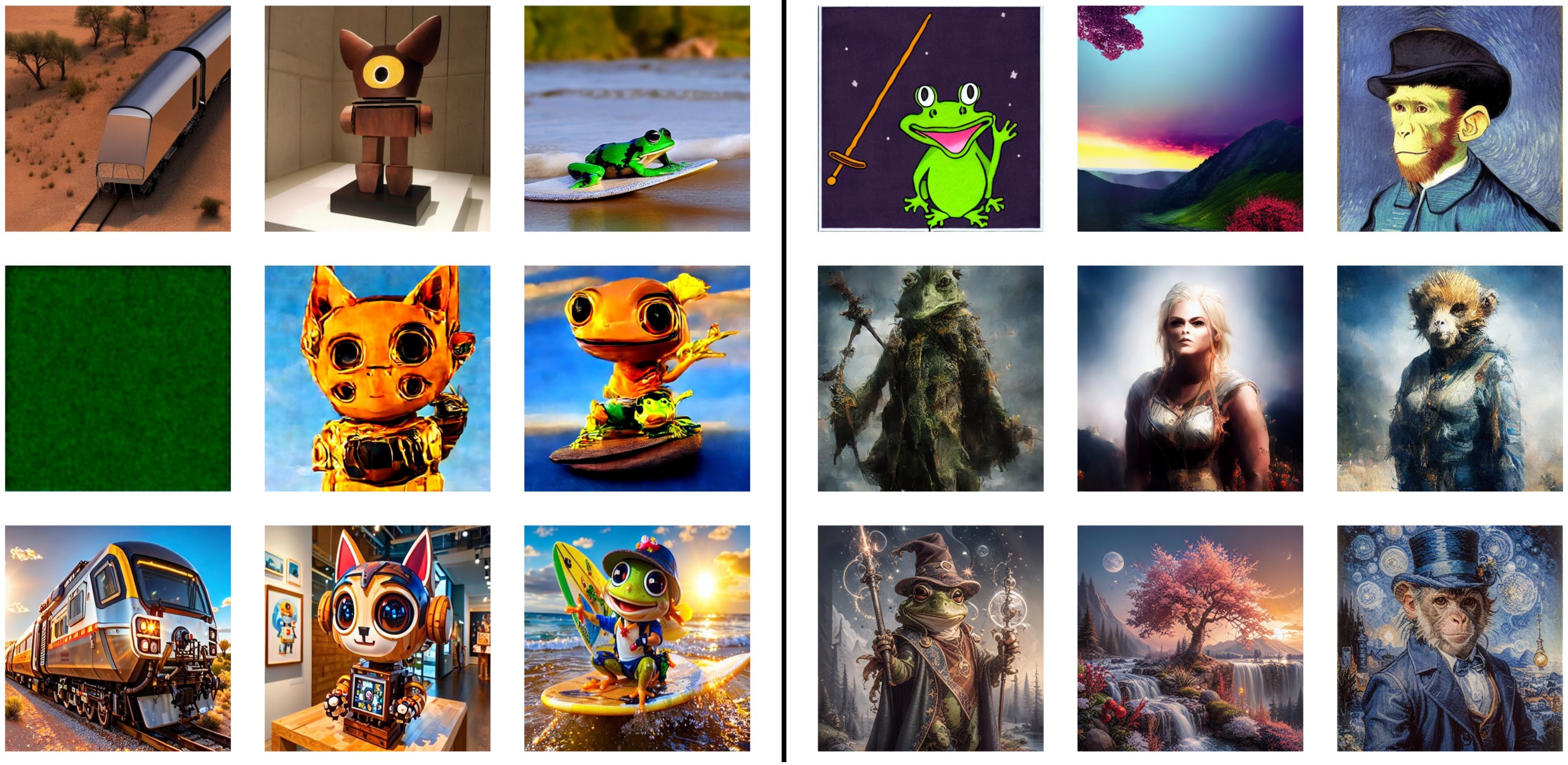
Digital art of a cherry tree overlooking a valley with a waterfall at sunset.

A monkey in a blue top hat painted in oil by Vincent van Gogh in the 1800s.

Stable Diffusion

DDPO

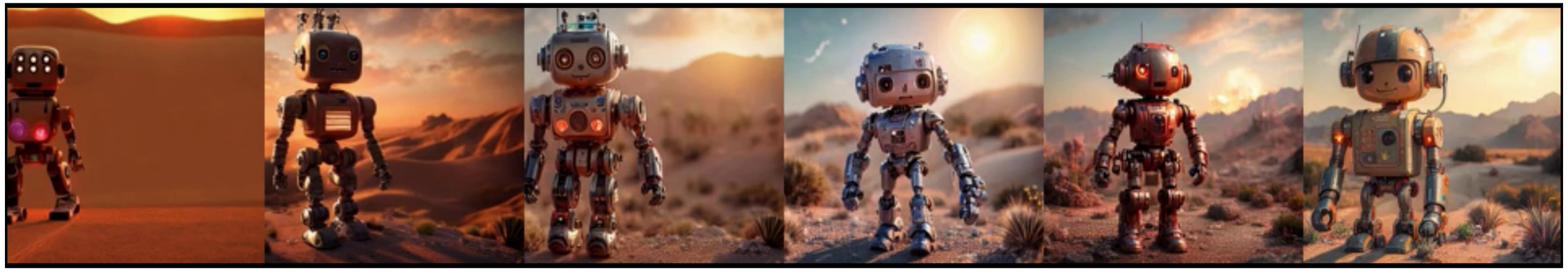
PRDP



Steady Improvement in Large-Scale Training

rural house with a garden and a swimming pool

cinematic still of an adorable walking robot in the desert, at sunset



PRDP Training

CVPR SEATTLE, WA JUNE 17-21, 2024

Summary

- PRDP: Scalable diffusion model alignment
- Superior generation quality
- Generalization to unseen prompts

Project Page

fei.deng@rutgers.edu