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

Supplementary Material

A. Proofs

A.1. Lower Bound of RLHF Objective

In Lemma A.1, we prove that the objective in Equation (6) is a lower bound of the RLHF objective in Equation (5).

Lemma A.1. *Given two diffusion models $\pi_\theta, \pi_{\mathrm{ref}}$, a prompt distribution $p(\mathbf{c})$, a reward function $r(\mathbf{x}_0, \mathbf{c})$, and a constant $\beta > 0$, we have:*

$$\mathbb{E}_{\mathbf{c}\sim p(\mathbf{c})}\left[\mathbb{E}_{\mathbf{x}_0\sim\pi_\theta(\mathbf{x}_0|\mathbf{c})}[r(\mathbf{x}_0,\mathbf{c})] - \beta\mathrm{KL}[\pi_\theta(\mathbf{x}_0|\mathbf{c})||\pi_{\mathrm{ref}}(\mathbf{x}_0|\mathbf{c})]\right] \tag{22}$$

$$\geq \mathbb{E}_{\mathbf{c}\sim p(\mathbf{c})}\left[\mathbb{E}_{\mathbf{x}_0\sim\pi_\theta(\mathbf{x}_0|\mathbf{c})}[r(\mathbf{x}_0,\mathbf{c})] - \beta\mathrm{KL}[\pi_\theta(\bar{\mathbf{x}}|\mathbf{c})||\pi_{\mathrm{ref}}(\bar{\mathbf{x}}|\mathbf{c})]\right], \tag{23}$$

where $\bar{\mathbf{x}} \coloneqq \mathbf{x}_{0:T}$ is the full denoising trajectory, and $\pi_\theta, \pi_{\mathrm{ref}}$ are defined as:

$$\pi(\mathbf{x}_0|\mathbf{c}) = \int \pi(\mathbf{x}_{0:T}|\mathbf{c})\,\mathrm{d}\mathbf{x}_{1:T} = \int p(\mathbf{x}_T)\prod_{t=1}^{T}\pi(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{c})\,\mathrm{d}\mathbf{x}_{1:T}. \tag{24}$$

Proof. It suffices to show that for any $\mathbf{c}$,

$$\mathrm{KL}[\pi_\theta(\bar{\mathbf{x}}|\mathbf{c})||\pi_{\mathrm{ref}}(\bar{\mathbf{x}}|\mathbf{c})] \geq \mathrm{KL}[\pi_\theta(\mathbf{x}_0|\mathbf{c})||\pi_{\mathrm{ref}}(\mathbf{x}_0|\mathbf{c})]. \tag{25}$$

This can be proved similarly as the data processing inequality. We provide the proof below.

$$\begin{aligned}
\mathrm{KL}[\pi_\theta(\bar{\mathbf{x}}|\mathbf{c})||\pi_{\mathrm{ref}}(\bar{\mathbf{x}}|\mathbf{c})] &= \mathbb{E}_{\pi_\theta(\mathbf{x}_{0:T}|\mathbf{c})}\left[\log\frac{\pi_\theta(\mathbf{x}_{0:T}|\mathbf{c})}{\pi_{\mathrm{ref}}(\mathbf{x}_{0:T}|\mathbf{c})}\right] && (26)\\
&= \mathbb{E}_{\pi_\theta(\mathbf{x}_{0:T}|\mathbf{c})}\left[\log\frac{\pi_\theta(\mathbf{x}_0|\mathbf{c})}{\pi_{\mathrm{ref}}(\mathbf{x}_0|\mathbf{c})} + \log\frac{\pi_\theta(\mathbf{x}_{1:T}|\mathbf{x}_0,\mathbf{c})}{\pi_{\mathrm{ref}}(\mathbf{x}_{1:T}|\mathbf{x}_0,\mathbf{c})}\right] && (27)\\
&= \mathbb{E}_{\pi_\theta(\mathbf{x}_0|\mathbf{c})}\left[\log\frac{\pi_\theta(\mathbf{x}_0|\mathbf{c})}{\pi_{\mathrm{ref}}(\mathbf{x}_0|\mathbf{c})}\right] + \mathbb{E}_{\pi_\theta(\mathbf{x}_0|\mathbf{c})}\left[\mathbb{E}_{\pi_\theta(\mathbf{x}_{1:T}|\mathbf{x}_0,\mathbf{c})}\left[\log\frac{\pi_\theta(\mathbf{x}_{1:T}|\mathbf{x}_0,\mathbf{c})}{\pi_{\mathrm{ref}}(\mathbf{x}_{1:T}|\mathbf{x}_0,\mathbf{c})}\right]\right] && (28)\\
&= \mathrm{KL}[\pi_\theta(\mathbf{x}_0|\mathbf{c})||\pi_{\mathrm{ref}}(\mathbf{x}_0|\mathbf{c})] + \mathbb{E}_{\pi_\theta(\mathbf{x}_0|\mathbf{c})}[\mathrm{KL}[\pi_\theta(\mathbf{x}_{1:T}|\mathbf{x}_0,\mathbf{c})||\pi_{\mathrm{ref}}(\mathbf{x}_{1:T}|\mathbf{x}_0,\mathbf{c})]] && (29)\\
&\geq \mathrm{KL}[\pi_\theta(\mathbf{x}_0|\mathbf{c})||\pi_{\mathrm{ref}}(\mathbf{x}_0|\mathbf{c})]. && (30)
\end{aligned}$$

$\square$

A.2. Maximizer of the Lower Bound of RLHF Objective

In Lemma A.2, we prove that Equation (7) maximizes the objective in Equation (6), a lower bound of the RLHF objective.

Lemma A.2. *Define*

$$\pi_{\theta^\star}(\bar{\mathbf{x}}|\mathbf{c}) = \frac{1}{Z(\mathbf{c})}\pi_{\text{ref}}(\bar{\mathbf{x}}|\mathbf{c})\exp\left(\frac{1}{\beta}r(\mathbf{x}_0,\mathbf{c})\right), \tag{31}$$

where

$$Z(\mathbf{c}) = \int \pi_{\text{ref}}(\bar{\mathbf{x}}|\mathbf{c})\exp\left(\frac{1}{\beta}r(\mathbf{x}_0,\mathbf{c})\right)\mathrm{d}\bar{\mathbf{x}} \tag{32}$$

is the partition function. Then $\pi_{\theta^\star}$ is the optimal solution to the following maximization problem:

$$\max_{\pi_\theta}\ \mathbb{E}_{\mathbf{c}\sim p(\mathbf{c})}\left[\mathbb{E}_{\mathbf{x}_0\sim\pi_\theta(\mathbf{x}_0|\mathbf{c})}[r(\mathbf{x}_0,\mathbf{c})] - \beta\mathrm{KL}[\pi_\theta(\bar{\mathbf{x}}|\mathbf{c})||\pi_{\text{ref}}(\bar{\mathbf{x}}|\mathbf{c})]\right]. \tag{33}$$

Proof. We provide the proof below, which is inspired by DPO [35].

$$\begin{aligned}
&\max_{\pi_\theta}\ \mathbb{E}_{\mathbf{c}\sim p(\mathbf{c})}\left[\mathbb{E}_{\mathbf{x}_0\sim\pi_\theta(\mathbf{x}_0|\mathbf{c})}[r(\mathbf{x}_0,\mathbf{c})] - \beta\mathrm{KL}[\pi_\theta(\bar{\mathbf{x}}|\mathbf{c})||\pi_{\text{ref}}(\bar{\mathbf{x}}|\mathbf{c})]\right] && (34)\\
=&\max_{\pi_\theta}\ \mathbb{E}_{\mathbf{c}\sim p(\mathbf{c})}\left[\mathbb{E}_{\bar{\mathbf{x}}\sim\pi_\theta(\bar{\mathbf{x}}|\mathbf{c})}[r(\mathbf{x}_0,\mathbf{c})] - \beta\mathrm{KL}[\pi_\theta(\bar{\mathbf{x}}|\mathbf{c})||\pi_{\text{ref}}(\bar{\mathbf{x}}|\mathbf{c})]\right] && (35)\\
=&\max_{\pi_\theta}\ \mathbb{E}_{\mathbf{c}\sim p(\mathbf{c})}\mathbb{E}_{\bar{\mathbf{x}}\sim\pi_\theta(\bar{\mathbf{x}}|\mathbf{c})}\left[r(\mathbf{x}_0,\mathbf{c}) - \beta\log\frac{\pi_\theta(\bar{\mathbf{x}}|\mathbf{c})}{\pi_{\text{ref}}(\bar{\mathbf{x}}|\mathbf{c})}\right] && (36)\\
=&\min_{\pi_\theta}\ \mathbb{E}_{\mathbf{c}\sim p(\mathbf{c})}\mathbb{E}_{\bar{\mathbf{x}}\sim\pi_\theta(\bar{\mathbf{x}}|\mathbf{c})}\left[\log\frac{\pi_\theta(\bar{\mathbf{x}}|\mathbf{c})}{\pi_{\text{ref}}(\bar{\mathbf{x}}|\mathbf{c})} - \frac{1}{\beta}r(\mathbf{x}_0,\mathbf{c})\right] && (37)\\
=&\min_{\pi_\theta}\ \mathbb{E}_{\mathbf{c}\sim p(\mathbf{c})}\mathbb{E}_{\bar{\mathbf{x}}\sim\pi_\theta(\bar{\mathbf{x}}|\mathbf{c})}\left[\log\frac{\pi_\theta(\bar{\mathbf{x}}|\mathbf{c})}{\pi_{\text{ref}}(\bar{\mathbf{x}}|\mathbf{c})\exp\left(\frac{1}{\beta}r(\mathbf{x}_0,\mathbf{c})\right)}\right] && (38)\\
=&\min_{\pi_\theta}\ \mathbb{E}_{\mathbf{c}\sim p(\mathbf{c})}\mathbb{E}_{\bar{\mathbf{x}}\sim\pi_\theta(\bar{\mathbf{x}}|\mathbf{c})}\left[\log\frac{\pi_\theta(\bar{\mathbf{x}}|\mathbf{c})}{\pi_{\theta^\star}(\bar{\mathbf{x}}|\mathbf{c})Z(\mathbf{c})}\right] && (39)\\
=&\min_{\pi_\theta}\ \mathbb{E}_{\mathbf{c}\sim p(\mathbf{c})}\left[\mathbb{E}_{\bar{\mathbf{x}}\sim\pi_\theta(\bar{\mathbf{x}}|\mathbf{c})}\left[\log\frac{\pi_\theta(\bar{\mathbf{x}}|\mathbf{c})}{\pi_{\theta^\star}(\bar{\mathbf{x}}|\mathbf{c})}\right] - \log Z(\mathbf{c})\right] && (40)\\
=&\min_{\pi_\theta}\ \mathbb{E}_{\mathbf{c}\sim p(\mathbf{c})}[\mathrm{KL}[\pi_\theta(\bar{\mathbf{x}}|\mathbf{c})||\pi_{\theta^\star}(\bar{\mathbf{x}}|\mathbf{c})] - \log Z(\mathbf{c})] && (41)\\
=&\min_{\pi_\theta}\ \mathbb{E}_{\mathbf{c}\sim p(\mathbf{c})}[\mathrm{KL}[\pi_\theta(\bar{\mathbf{x}}|\mathbf{c})||\pi_{\theta^\star}(\bar{\mathbf{x}}|\mathbf{c})]]. && (42)
\end{aligned}$$

Since $\mathrm{KL}[\pi_\theta(\bar{\mathbf{x}}|\mathbf{c})||\pi_{\theta^\star}(\bar{\mathbf{x}}|\mathbf{c})] \geq 0$, and $\mathrm{KL}[\pi_\theta(\bar{\mathbf{x}}|\mathbf{c})||\pi_{\theta^\star}(\bar{\mathbf{x}}|\mathbf{c})] = 0$ if and only if $\pi_\theta(\bar{\mathbf{x}}|\mathbf{c}) = \pi_{\theta^\star}(\bar{\mathbf{x}}|\mathbf{c})$, we conclude that the optimal solution to Equation (33) is $\pi_\theta(\bar{\mathbf{x}}|\mathbf{c}) = \pi_{\theta^\star}(\bar{\mathbf{x}}|\mathbf{c})$ for all $\mathbf{c}$. $\square$

A.3. Necessary and Sufficient Conditions for the Optimal Solution

In Lemma A.3, we provide theoretical justification for our proposed RDP objective in Equation (14).

Lemma A.3.

$$
\begin{aligned}
&\pi_\theta(\bar{\mathbf{x}}|\mathbf{c}) = \pi_{\theta^\star}(\bar{\mathbf{x}}|\mathbf{c}), \quad \forall \bar{\mathbf{x}}, \mathbf{c} && (43)\\
\iff\ &\log \frac{\pi_\theta(\bar{\mathbf{x}}^a|\mathbf{c})}{\pi_{\text{ref}}(\bar{\mathbf{x}}^a|\mathbf{c})} - \log \frac{\pi_\theta(\bar{\mathbf{x}}^b|\mathbf{c})}{\pi_{\text{ref}}(\bar{\mathbf{x}}^b|\mathbf{c})} = \frac{r(\mathbf{x}_0^a, \mathbf{c}) - r(\mathbf{x}_0^b, \mathbf{c})}{\beta}, \quad \forall \bar{\mathbf{x}}^a, \bar{\mathbf{x}}^b, \mathbf{c}. && (44)
\end{aligned}
$$

Proof. We have shown " $\implies$ " in the main text. We provide the proof for " $\impliedby$ " below.

Equation (44) implies that

$$\log \frac{\pi_\theta(\bar{\mathbf{x}}|\mathbf{c})}{\pi_{\text{ref}}(\bar{\mathbf{x}}|\mathbf{c})} - \frac{1}{\beta} r(\mathbf{x}_0, \mathbf{c}) \tag{45}$$

is a constant w.r.t. $\bar{\mathbf{x}}$. Therefore, we can write Equation (45) as a function of $\mathbf{c}$ alone:

$$\log \frac{\pi_\theta(\bar{\mathbf{x}}|\mathbf{c})}{\pi_{\text{ref}}(\bar{\mathbf{x}}|\mathbf{c})} - \frac{1}{\beta} r(\mathbf{x}_0, \mathbf{c}) = f(\mathbf{c}). \tag{46}$$

Hence,

$$\pi_\theta(\bar{\mathbf{x}}|\mathbf{c}) = \pi_{\text{ref}}(\bar{\mathbf{x}}|\mathbf{c}) \exp\left(\frac{1}{\beta} r(\mathbf{x}_0, \mathbf{c})\right) \exp(f(\mathbf{c}))\,. \tag{47}$$

It suffices to show that

$$\exp(f(\mathbf{c})) = \frac{1}{Z(\mathbf{c})}, \quad \forall \mathbf{c}. \tag{48}$$

This follows from the fact that the probability density function $\pi_\theta(\bar{\mathbf{x}}|\mathbf{c})$ must satisfy:

$$
\begin{aligned}
1 &= \int \pi_\theta(\bar{\mathbf{x}}|\mathbf{c})\,\mathrm{d}\bar{\mathbf{x}} && (49)\\
&= \int \pi_{\text{ref}}(\bar{\mathbf{x}}|\mathbf{c}) \exp\left(\frac{1}{\beta} r(\mathbf{x}_0, \mathbf{c})\right) \exp(f(\mathbf{c}))\,\mathrm{d}\bar{\mathbf{x}} && (50)\\
&= \exp(f(\mathbf{c})) \int \pi_{\text{ref}}(\bar{\mathbf{x}}|\mathbf{c}) \exp\left(\frac{1}{\beta} r(\mathbf{x}_0, \mathbf{c})\right) \mathrm{d}\bar{\mathbf{x}} && (51)\\
&= \exp(f(\mathbf{c}))\, Z(\mathbf{c}). && (52)
\end{aligned}
$$

□

B. Instability of DDPO in Large-Scale Reward Finetuning

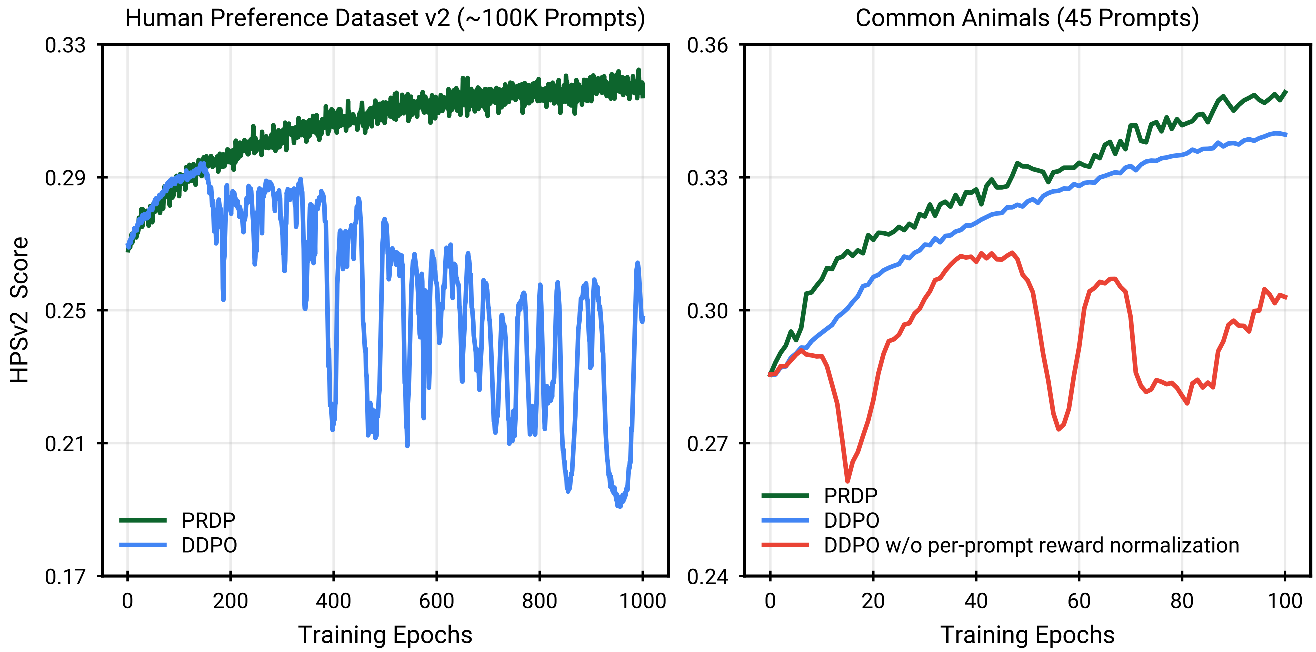

Figure 9. Analysis of the instability of DDPO in large-scale training. We plot the training curves of PRDP and DDPO on the large-scale Human Preference Dataset v2 (Left) and the small-scale Common Animals (Right). PRDP outperforms DDPO in the small-scale setting, and maintains stability in the large-scale setting where DDPO fails. Our ablation study suggests that the per-prompt reward normalization in DDPO is key to its stability, and the inability to perform such normalization in the large-scale setting likely causes its failure.

Figure 9 shows the training curve of PRDP and DDPO [4], where the reward model is HPSv2 [53]. From Figure 9 (Left), we observe that when trained on the large-scale Human Preference Dataset v2 (HPD v2) [53], DDPO fails to stably optimize the reward. We conjecture that this is because the per-prompt reward normalization is rarely enabled in the large-scale setting, since each prompt can only be seen a few times. Specifically, in each epoch, DDPO randomly samples 512 prompts, so on average, each prompt can be seen $512 \times 1000/100\text{K} \approx 5$ times. This is insufficient to obtain a good estimate of the per-prompt expected reward. In this case, DDPO will compute a prompt-agnostic expected reward, by averaging the rewards across all 512 prompts. To verify that such prompt-agnostic reward normalization causes training instability, we conduct an ablation study of DDPO in our small-scale setting with 45 training prompts. As shown in Figure 9 (Right), DDPO without per-prompt reward normalization is unstable even in the small-scale setting, suggesting that the inability to perform per-prompt reward normalization can be a limiting factor in scaling DDPO to large prompt datasets. In contrast to DDPO, PRDP can steadily improve the reward score and maintain stability in both small-scale and large-scale settings.

C. Effect of KL Regularization

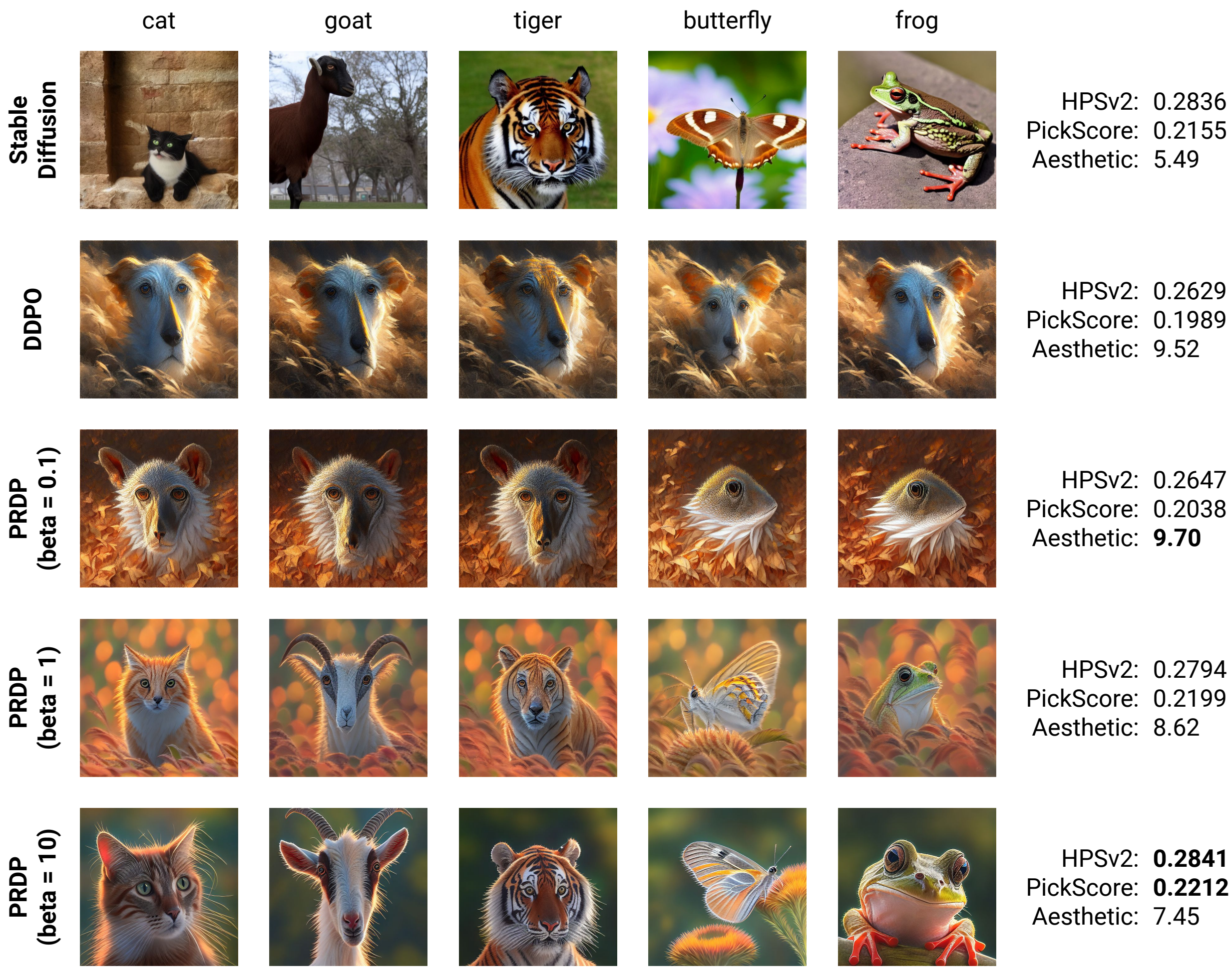

Figure 10. Effect of KL regularization on optimizing aesthetic score. DDPO and PRDP are finetuned from Stable Diffusion v1.4 on 45 prompts of common animal names. Evaluation is performed on the same set of prompts. In addition to aesthetic score, we report HPSv2 and PickScore which reflect text-image alignment but are not used during training. Samples within each column are generated from the prompt shown on top, using the same random seed. PRDP with a large KL weight β can alleviate the reward over-optimization problem encountered by DDPO, significantly improving the aesthetic quality over Stable Diffusion while maintaining text-image alignment.

In contrast to DDPO [4] which only cares about maximizing the reward, PRDP is formulated with a KL regularization, allowing us to alleviate the problem of reward over-optimization by increasing the KL weight β. We demonstrate the effect of KL regularization in Figure 10. Here, the reward used for training is the aesthetic score given by the LAION aesthetic predictor. It only takes images as input, and therefore ignores the text-image alignment. We finetune DDPO and PRDP from Stable Diffusion v1.4 [37] for 250 epochs on 45 training prompts of common animal names as used in DDPO, with 512 reward queries in each epoch. For evaluation, we additionally use HPSv2 [53] and PickScore [22] that reflect text-image alignment. The reported reward scores are averaged over 64 random samples per training prompt, using the same random seed for Stable Diffusion v1.4, DDPO, and PRDP.

We observe that DDPO, without KL regularization, is prone to reward over-optimization. It ignores the text prompt and generates similar images for all prompts. PRDP with a small KL weight (*e.g.*, $\beta = 0.1$) has the same problem, but achieves higher reward scores than DDPO, showing a better reward maximization capability. As the KL weight increases, PRDP is able to better preserve the text-image alignment, indicated by the increase in HPSv2 and PickScore. With $\beta = 10$, PRDP significantly improves the aesthetic score over Stable Diffusion v1.4 without sacrificing text-image alignment.

D. Large-Scale Multi-Reward Finetuning

Table 3. Reward score comparison on unseen prompts. We use a weighted combination of rewards: PickScore = 10, HPSv2 = 2, Aesthetic = 0.05. PRDP is finetuned from Stable Diffusion v1.4 on the training set prompts of Pick-a-Pic v1 dataset.

	Pick-a-Pic v1 Test Set	HPD v2 Animation	HPD v2 Concept Art	HPD v2 Painting	HPD v2 Photo
SD v1.4	2.888	2.927	2.877	2.883	2.984
PRDP	**3.208**	**3.296**	**3.264**	**3.274**	**3.214**

In this section, we provide additional results for our large-scale multi-reward finetuning experiment. Following DRaFT [6], we use a weighted combination of rewards: PickScore = 10, HPSv2 = 2, Aesthetic = 0.05. We finetune Stable Diffusion v1.4 [37] on the training set prompts of Pick-a-Pic v1 dataset [22]. We evaluate our finetuned model on a variety of unseen prompts, including 500 prompts from the Pick-a-Pic v1 test set, and 800 prompts from each of the four benchmark categories of the Human Preference Dataset v2 (HPD v2) [53], namely animation, concept art, painting, and photo. Table 3 reports the reward scores before and after finetuning. The reward scores are averaged over 64 random samples per prompt, using the same random seed for Stable Diffusion v1.4 and PRDP. We further show generation samples for each test prompt set in Figures 11 to 15. As can be seen, PRDP significantly improves generation quality across all five prompt sets.

E. Hyperparameters

Table 4. PRDP training hyperparameters.

Name	Symbol	Small-Scale Finetuning	Large-Scale Finetuning	Large-Scale Multi-Reward Finetuning
Training epochs	E	100	1000	1000
Gradient updates per epoch	K	10	1	1
Prompts per epoch	N	32	64	64
Images per prompt	B	16	8	8
KL weight	β	3×10^{-5}	3×10^{-6}	3×10^{-5}
DDPM steps	T	50	50	50
Stepwise clipping range	ϵ	1×10^{-6}	1×10^{-4}	1×10^{-4}
Classifier-free guidance scale	—	5.0	5.0	5.0
Optimizer	—	AdamW	AdamW	AdamW
Gradient clipping	—	1.0	1.0	1.0
Learning rate	—	1×10^{-5}	7×10^{-6}	1×10^{-5}
Weight decay	—	1×10^{-4}	1×10^{-4}	1×10^{-4}

F. Effect of Clipping

Table 5. **Effect of clipping on training stability.**

	w/o Clipping	w/ Clipping
DDPO	Small scale: Unstable Large scale: Unstable	Small scale: Stable Large scale: Unstable
PRDP	Small scale: Unstable Large scale: Unstable	Small scale: Stable Large scale: Stable

Table 5 summarizes the effect of clipping on the training stability of both DDPO [4] and PRDP. For DDPO, we use PPO-based clipping [42], while for PRDP, we use the proximal updates described in Section 3.3. We observe that clipping is key to stability of small-scale training, whereas using the PRDP objective and clipping are both indispensable for achieving stability in large-scale training.

G. Jax Implementation of PRDP Loss

```python
import jax
import jax.numpy as jnp


def prdp_loss(
    log_probs: jax.Array,         # (B, T)
    log_probs_old: jax.Array,     # (B, T)
    log_probs_ref: jax.Array,     # (B, T)
    rewards: jax.Array,           # (B,)
    clip_range: float,
    kl_weight: float,
) -> jax.Array:
  """Computes PRDP loss for a batch of denoising trajectories with the same text prompt.

  Args:
    log_probs: Log probs of the denoising trajectories under pi_theta.
    log_probs_old: Log probs of the denoising trajectories under pi_theta_old.
    log_probs_ref: Log probs of the denoising trajectories under pi_ref.
    rewards: Rewards of the generated clean images.
    clip_range: Stepwise clipping range (epsilon).
    kl_weight: KL weight (beta).

  Returns:
    loss: The PRDP loss.
  """
  log_ratios = log_probs - log_probs_ref
  log_ratios_old = log_probs_old - log_probs_ref
  clipped_log_ratios = jnp.clip(
      log_ratios, log_ratios_old - clip_range, log_ratios_old + clip_range
  )

  log_ratios = jnp.mean(log_ratios, axis=-1)
  clipped_log_ratios = jnp.mean(clipped_log_ratios, axis=-1)

  log_ratio_diffs = log_ratios[:, None] - log_ratios
  clipped_log_ratio_diffs = clipped_log_ratios[:, None] - clipped_log_ratios
  reward_diffs = rewards[:, None] - rewards

  mse_loss = (log_ratio_diffs - reward_diffs / kl_weight) ** 2
  clipped_mse_loss = (clipped_log_ratio_diffs - reward_diffs / kl_weight) ** 2
  loss = jnp.maximum(mse_loss, clipped_mse_loss)
  loss = jnp.mean(loss, where=reward_diffs > 0)

  return loss
```

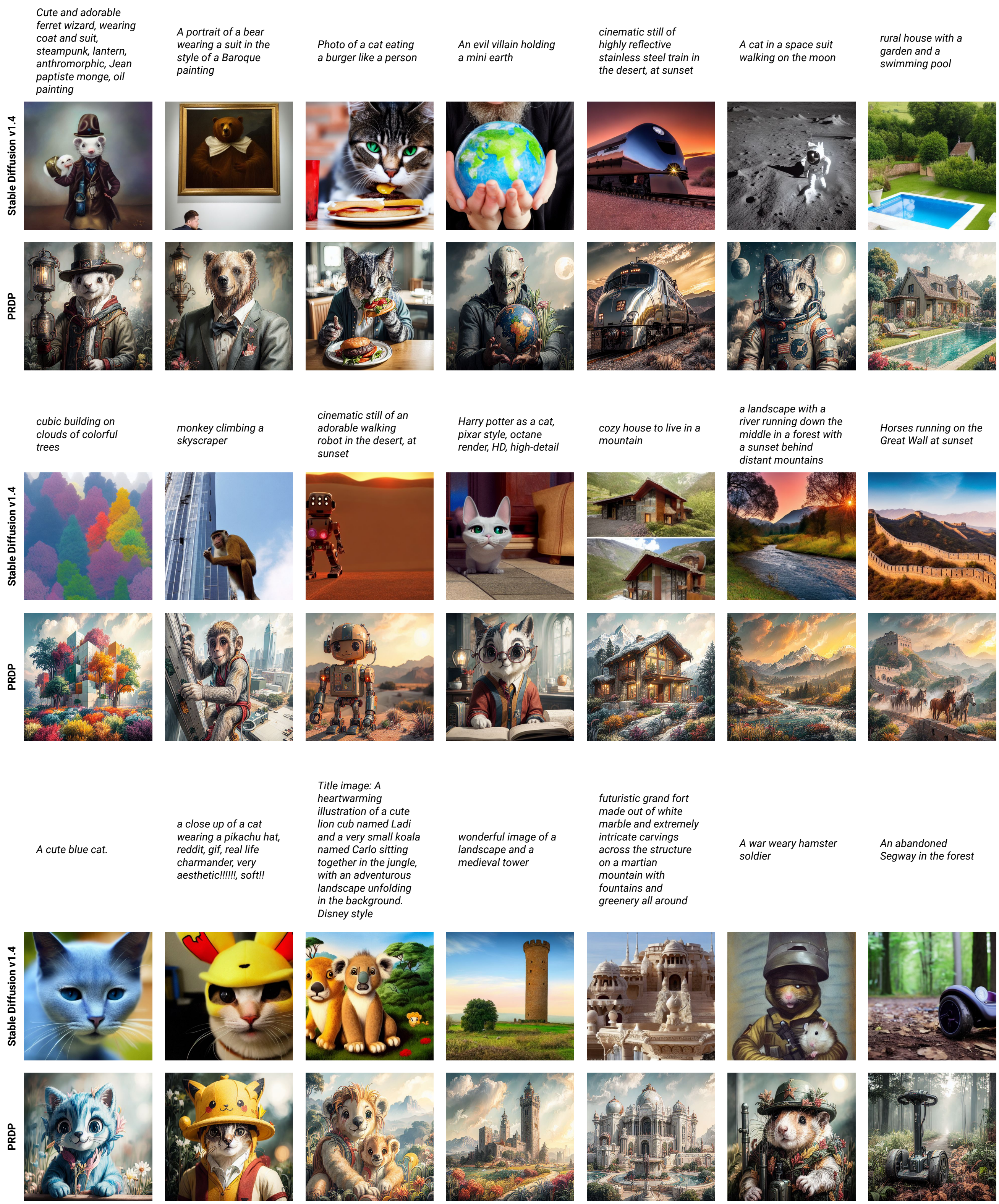

Figure 11. Generation samples on unseen prompts from the Pick-a-Pic v1 test set. PRDP is finetuned from Stable Diffusion v1.4 on the training set prompts of Pick-a-Pic v1 dataset, using a weighted combination of rewards: PickScore = 10, HPSv2 = 2, Aesthetic = 0.05. For each prompt, the generation sample from Stable Diffusion v1.4 and PRDP use the same random seed.

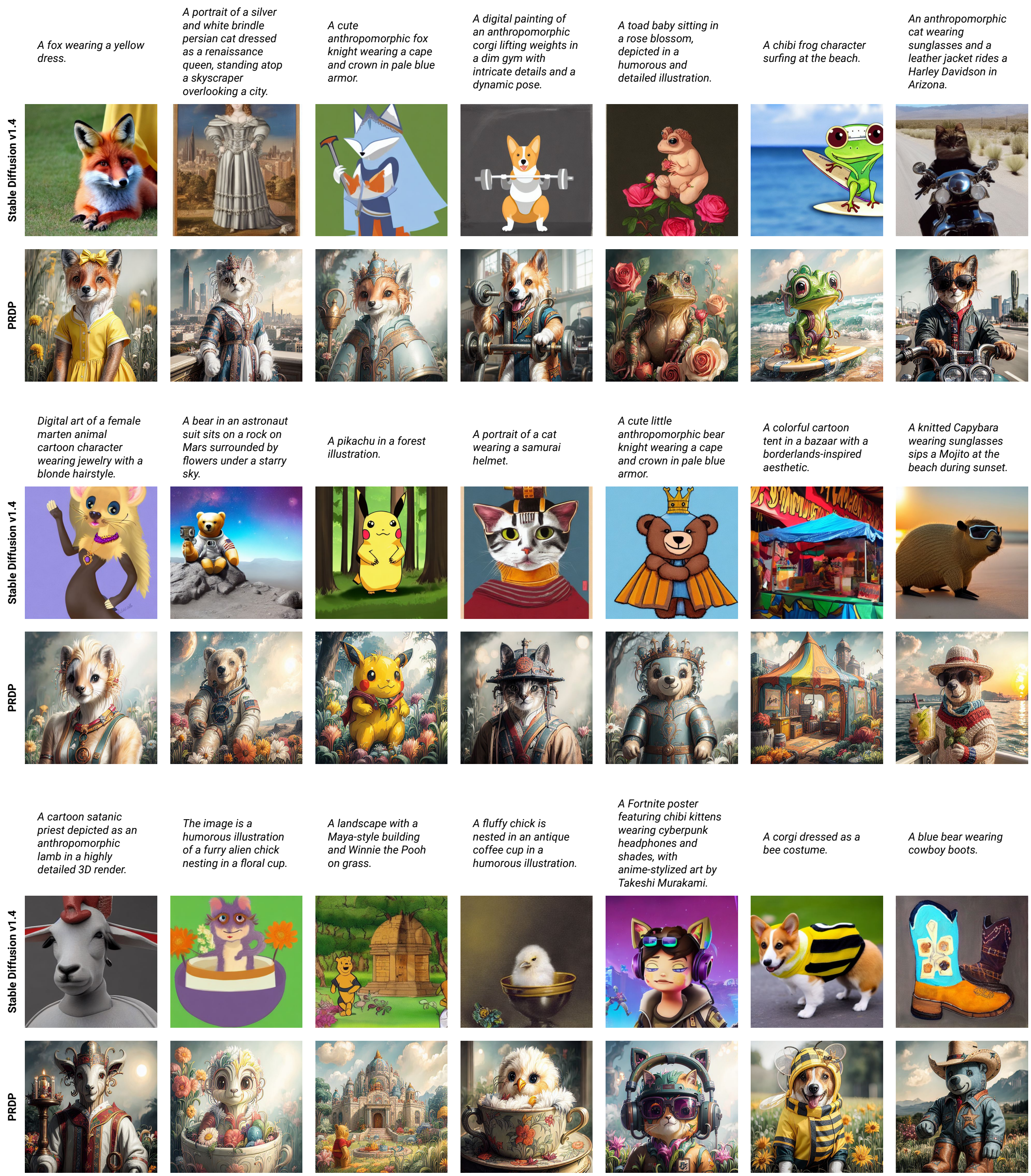

Figure 12. Generation samples on unseen prompts from the HPD v2 animation benchmark. PRDP is finetuned from Stable Diffusion v1.4 on the training set prompts of Pick-a-Pic v1 dataset, using a weighted combination of rewards: PickScore $= 10$, HPSv2 $= 2$, Aesthetic $= 0.05$. For each prompt, the generation sample from Stable Diffusion v1.4 and PRDP use the same random seed.

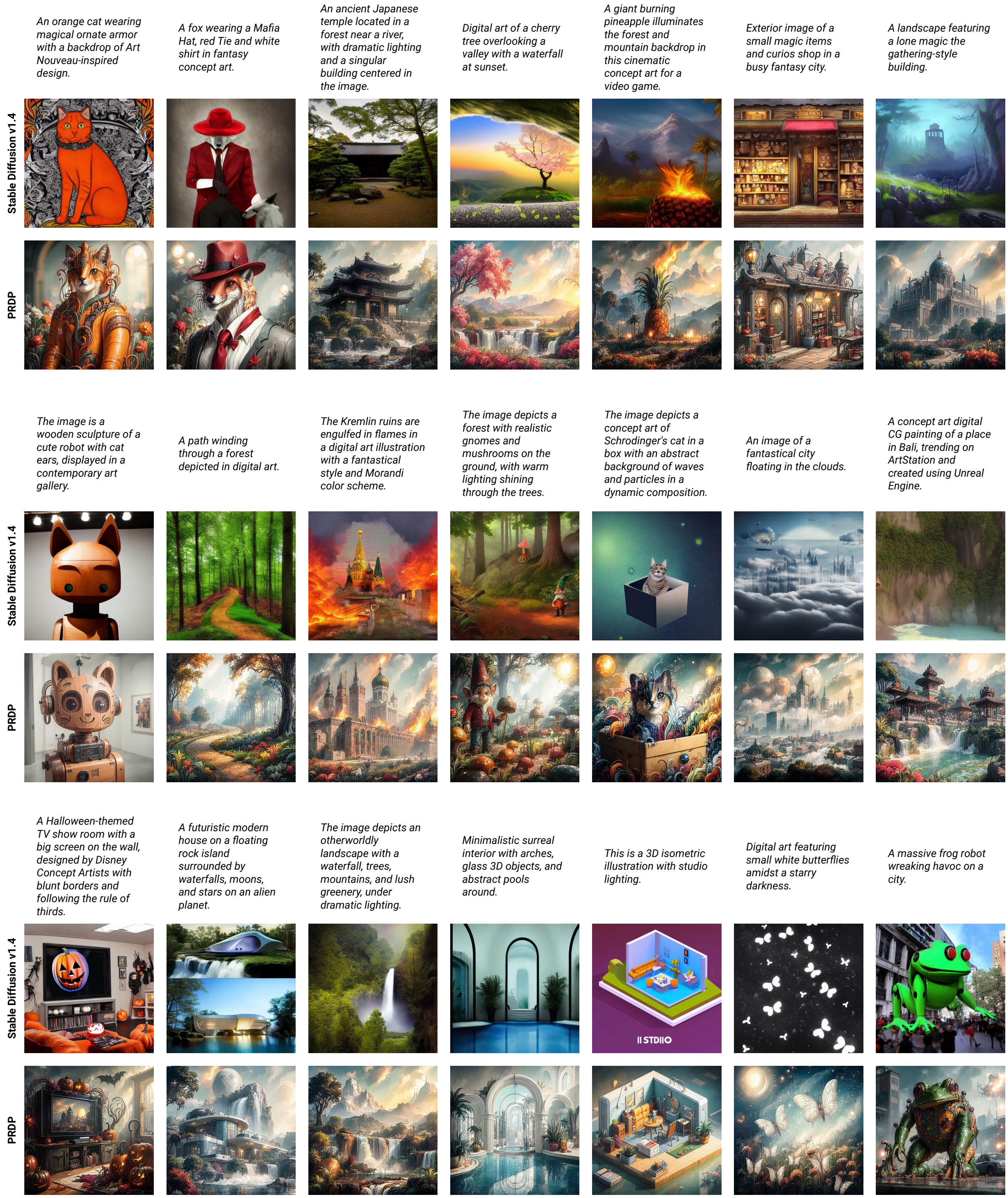

Figure 13. Generation samples on unseen prompts from the HPD v2 concept art benchmark. PRDP is finetuned from Stable Diffusion v1.4 on the training set prompts of Pick-a-Pic v1 dataset, using a weighted combination of rewards: PickScore $= 10$, HPSv2 $= 2$, Aesthetic $= 0.05$. For each prompt, the generation sample from Stable Diffusion v1.4 and PRDP use the same random seed.

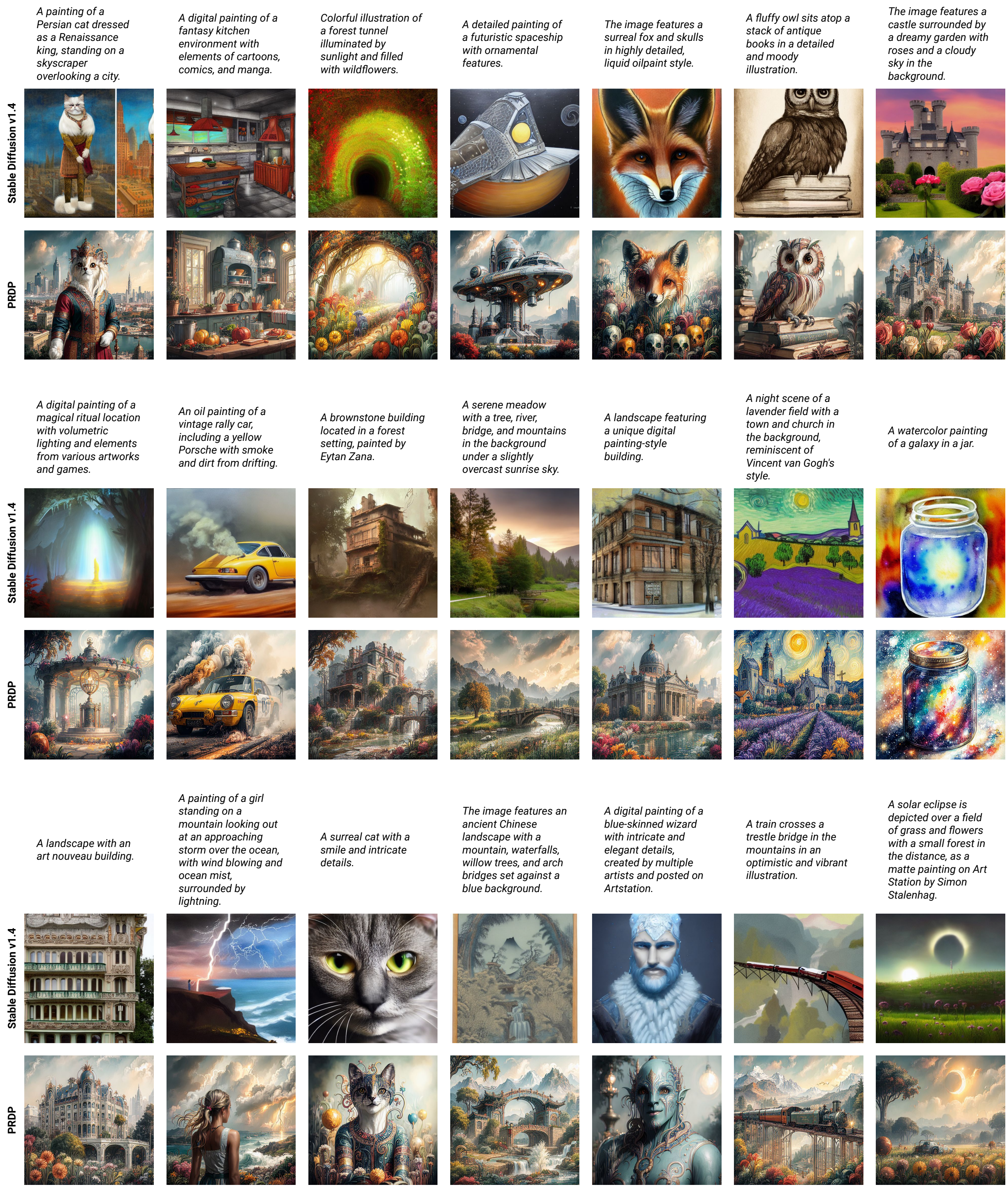

Figure 14. Generation samples on unseen prompts from the HPD v2 painting benchmark. PRDP is finetuned from Stable Diffusion v1.4 on the training set prompts of Pick-a-Pic v1 dataset, using a weighted combination of rewards: PickScore $= 10$, HPSv2 $= 2$, Aesthetic $= 0.05$. For each prompt, the generation sample from Stable Diffusion v1.4 and PRDP use the same random seed.

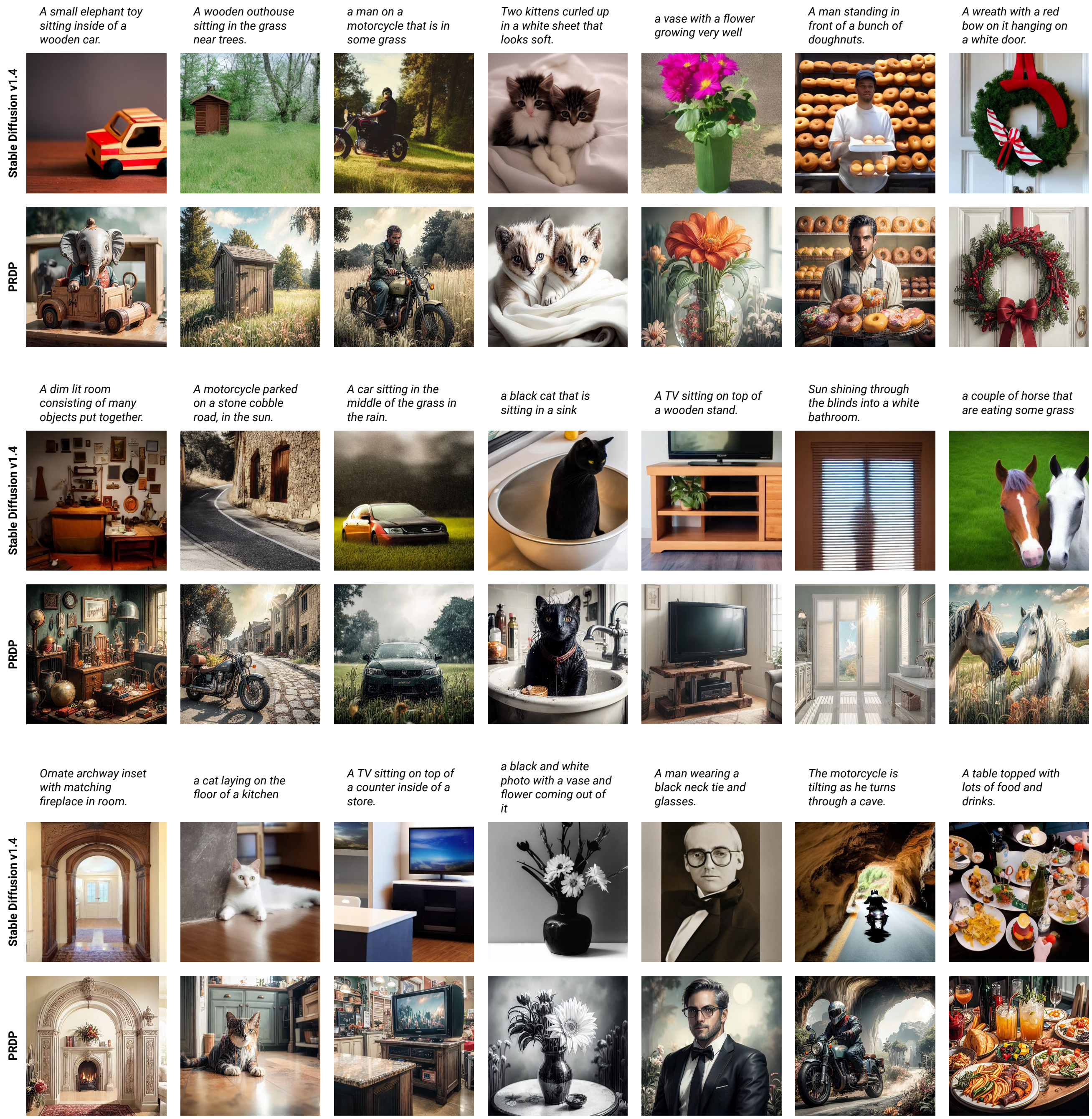

Figure 15. Generation samples on unseen prompts from the HPD v2 photo benchmark. PRDP is finetuned from Stable Diffusion v1.4 on the training set prompts of Pick-a-Pic v1 dataset, using a weighted combination of rewards: PickScore $= 10$, HPSv2 $= 2$, Aesthetic $= 0.05$. For each prompt, the generation sample from Stable Diffusion v1.4 and PRDP use the same random seed.