A. D3PO Pseudo-code

The pseudocode of the D3PO method can be seen in Algorithm 1.

Algorithm 1 D3PO pseudo-code

Require: Number of inference timesteps T, number of training epochs N, number of prompts per epoch K, pre-trained diffusion model $\boldsymbol{\epsilon}_\theta$.

Copy a pre-trained diffusion model $\boldsymbol{\epsilon}_{\text{ref}} = \boldsymbol{\epsilon}_\theta$. Set $\boldsymbol{\epsilon}_{\text{ref}}$ with `requires_grad` to `False`.
for $n = 1 : N$ **do**
    `# Sample images`
    for $k = 1 : K$ **do**
        Random choose a prompt $\boldsymbol{c_k}$ and sample $\boldsymbol{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
        for $i = 0 : 1$ **do**
            for $t = T : 1$ **do**
                `no grad`: $\boldsymbol{x}^i_{k,t-1} = \mu(\boldsymbol{x}^i_{k,t}, t, \boldsymbol{c_k}) + \sigma_t \boldsymbol{z}, \quad \boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
            end for
        end for
    end for
    `# Get Human Feedback`
    for $k = 1 : K$ **do**
        Get human feedback from $\boldsymbol{c_k}$, $\boldsymbol{x}^0_{k,0}$, and $\boldsymbol{x}^1_{k,0}$.
        if $\boldsymbol{x}^0_0$ is better than $\boldsymbol{x}^1_0$ **then**
            $h_k = [1, -1]$
        else if $\boldsymbol{x}^0_1$ is better than $\boldsymbol{x}^0_0$ **then**
            $h_k = [-1, 1]$
        else
            $h_k = [0, 0]$
        end if
    end for
    `# Training`
    for $k = 1 : K$ **do**
        for $t = T : 1$ **do**
            for $i = 0 : 1$ **do**
                `with grad`:
                $\mu_\theta(\boldsymbol{x}^i_{k,t}, t, \boldsymbol{c_k}) = \frac{1}{\sqrt{\alpha_t}} \left(\boldsymbol{x}^i_{k,t} - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}} \boldsymbol{\epsilon}_\theta \left(\boldsymbol{x}^i_{k,t}, t, \boldsymbol{c_k} \right) \right)$
                $\mu_{\text{ref}}(\boldsymbol{x}^i_{k,t}, t, \boldsymbol{c_k}) = \frac{1}{\sqrt{\alpha_t}} \left(\boldsymbol{x}^i_{k,t} - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}} \boldsymbol{\epsilon}_{\text{ref}} \left(\boldsymbol{x}^i_{k,t}, t, \boldsymbol{c_k} \right) \right)$
                $\pi_\theta(\boldsymbol{x}^i_{k,t-1} | \boldsymbol{x}^i_{k,t}, t, \boldsymbol{c_k}) = \frac{1}{\sqrt{2\pi}\sigma_t} \exp(-\frac{(\boldsymbol{x}^i_{k,t-1} - \mu_\theta(\boldsymbol{x}^i_{k,t}, t, \boldsymbol{c_k}))^2}{2\sigma_t^2})$
                $\pi_{\text{ref}}(\boldsymbol{x}^i_{k,t-1} | \boldsymbol{x}^i_{k,t}, t, \boldsymbol{c_k}) = \frac{1}{\sqrt{2\pi}\sigma_t} \exp(-\frac{(\boldsymbol{x}^i_{k,t-1} - \mu_{\text{ref}}(\boldsymbol{x}^i_{k,t}, t, \boldsymbol{c_k}))^2}{2\sigma_t^2})$
            end for
            Update θ with gradient descent using

$$\nabla_\theta \log \rho(h_k(0)\beta \log \frac{\pi_\theta(\boldsymbol{x}^0_{k,t-1} | \boldsymbol{x}^0_{k,t}, t, \boldsymbol{c})}{\pi_{\text{ref}}(\boldsymbol{x}_{k,t-1} | \boldsymbol{x}^0_{k,t}, t, \boldsymbol{c})} + h_k(1)\beta \log \frac{\pi_\theta(\boldsymbol{x}^1_{k,t-1} | \boldsymbol{x}^1_{k,t}, t, \boldsymbol{c})}{\pi_{\text{ref}}(\boldsymbol{x}_{k,t-1} | \boldsymbol{x}^1_{k,t}, t, \boldsymbol{c})})$$

        end for
    end for
end for

B. Proof

B.1. Proof of Proposition 1

The RL objective can be written as:

$$\begin{aligned}
&\max_{\pi} \mathbb{E}_{s\sim d^{\pi}, a\sim\pi(a|s)}[Q^*(s,a)] - \beta\mathbb{D}_{KL}[\pi(a|s)\|\pi_{\text{ref}}(a|s)] \\
&= \max_{\pi} \mathbb{E}_{s\sim d^{\pi}, a\sim\pi(a|s)}[Q^*(s,a) - \beta\log\frac{\pi(a|s)}{\pi_{\text{ref}}(a|s)}] \\
&= \min_{\pi} \mathbb{E}_{s\sim d^{\pi}, a\sim\pi(a|s)}[\log\frac{\pi(a|s)}{\pi_{\text{ref}}(a|s)} - \frac{1}{\beta}Q^*(s,a)] \\
&= \min_{\pi} \mathbb{E}_{s\sim d^{\pi}, a\sim\pi(a|s)}[\log\frac{\pi(a|s)}{\pi_{\text{ref}}(a|s)\exp(\frac{1}{\beta}Q^*(s,a))}] \\
&= \min_{\pi} \mathbb{E}_{s\sim d^{\pi}}[\mathbb{D}_{KL}[\pi(a|s)\|\tilde{\pi}(a|s)]]
\end{aligned}$$

where $\tilde{\pi}(a|s) = \pi_{\text{ref}}(a|s)\exp(\frac{1}{\beta}Q^*(s,a))$. Note that the KL-divergence is minimized at 0 iff the two distributions are identical, so the optimal solution is:

$$\pi(a|s) = \tilde{\pi}(a|s) = \pi_{\text{ref}}(a|s)\exp(\frac{1}{\beta}Q^*(s,a)).$$

B.2. Proof of Proposition 2

For simplicity, we define $Q_i = Q^*(s_0^i, a_0^i)$ and $X_i = \sum_{t=0}^{T} r^*\left(s_t^i, a_t^i\right) \quad i \in \{0,1\}$. Using the Eq. (3) we can obtain that:

$$\begin{aligned}
\mathbb{E}[p^*\left(\sigma_1 \succ \sigma_0\right)] &= \frac{\mathbb{E}[\exp(X_1)]}{\mathbb{E}[\exp(X_1) + \exp(X_0)]} \\
&= \frac{\exp(Q_1 + 1/2\sigma)}{\exp(Q_1 + 1/2\sigma) + \exp(Q_0 + 1/2\sigma)} \\
&= \frac{\exp(Q_1)}{\exp(Q_1) + \exp(Q_0)} \\
&= \mathbb{E}[\tilde{p}^*\left(\sigma_1 \succ \sigma_0\right)].
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}[(p^*\left(\sigma_1 \succ \sigma_0\right))^2] &= \frac{\mathbb{E}[\exp(2X_1)]}{\mathbb{E}[\exp(2X_1)] + \mathbb{E}[\exp(2X_0)] + \mathbb{E}[2\exp(X_0)\exp(X_1)]} \\
&= \frac{\exp(2Q_1 + 2\sigma^2)}{\exp(2Q_1 + 2\sigma^2) + \exp(2Q_0 + 2\sigma^2) + \exp(Q_0 + Q_1 + \sigma^2)} \\
&= \frac{\exp(2Q_1 + \sigma^2)}{\exp(2Q_1 + \sigma^2) + \exp(2Q_0 + \sigma^2) + 2\exp(Q_0 + Q_1)}.
\end{aligned}$$

$$\begin{aligned}
\text{Var}[p^*(\sigma_1 \succ \sigma_0)] &= \mathbb{E}[(p(\sigma_1 \succ \sigma_0))^2] - (\mathbb{E}[p(\sigma_1 \succ \sigma_0)])^2 \\
&= \frac{2\exp(3Q_1 + Q_0)(\exp(\sigma^2) - 1)}{[\exp(2Q_1 + \sigma^2) + \exp(2Q_0 + \sigma^2) + 2\exp(Q_0 + Q_1)][\exp(Q_1) + \exp(Q_0)]^2} \\
&\leq \frac{2\exp(3Q_1 + Q_0)(\exp(\sigma^2) - 1)}{[\exp(Q_1) + \exp(Q_0)]^4}.
\end{aligned}$$

Similarly, we have:

$$\mathrm{Var}[p^*\left(\sigma_0 \succ \sigma_1\right)] \leq \frac{2\exp(Q_1+3Q_0)(\exp(\sigma^2)-1)}{[\exp(Q_1)+\exp(Q_0)]^4}.$$

Note that $\mathrm{Var}[p^*\left(\sigma_1 \succ \sigma_0\right)] = \mathrm{Var}[1-p^*\left(\sigma_0 \succ \sigma_1\right)] = \mathrm{Var}[p^*\left(\sigma_0 \succ \sigma_1\right)]$, considering these two inequalities, we have:

$$\begin{aligned}
\mathrm{Var}[p^*\left(\sigma_1 \succ \sigma_0\right)] &\leq \frac{[\exp(Q_1+3Q_0)+\exp(Q_0+3Q_1)](\exp(\sigma^2)-1)}{[\exp(Q_1)+\exp(Q_0)]^4} \\
&\leq \frac{[\exp(Q_1+3Q_0)+\exp(Q_0+3Q_1)](\exp(\sigma^2)-1)}{16[\exp(2Q_1)\exp(2Q_0)]} \\
&= \frac{[\exp(Q_0-Q_1)+\exp(Q_1-Q_0)](\exp(\sigma^2)-1)}{16} \\
&\leq \frac{(\xi+\frac{1}{\xi})(\exp(\sigma^2)-1)}{16}.
\end{aligned}$$

By using the Chebyshev inequality, we can obtain:

$$P(|p^*\left(\sigma_1 \succ \sigma_0\right) - \tilde{p}^*\left(\sigma_1 \succ \sigma_0\right)| < t) > 1 - \frac{(\xi^2+1)(\exp(\sigma^2)-1)}{16\xi t}.$$

We choose $t = \frac{(\xi^2+1)(\exp(\sigma^2)-1)}{16\xi\delta}$ so that:

$$P(|p^*\left(\sigma_1 \succ \sigma_0\right) - \tilde{p}^*\left(\sigma_1 \succ \sigma_0\right)| < \frac{(\xi^2+1)(\exp(\sigma^2)-1)}{16\xi\delta}) > 1-\delta.$$

C. Prompts of Experiments

During the quantitative experiments in Section 5.1, we utilized prompts related to 45 common animals, outlined as follows: We simply use "*1 hand*" to generate hand images in the image distortion experiments. In the experiment of reducing distortions in generated anime images using the Anything V5 model, we initially collected 483 illustrations from anime games. Subsequently, non-character roles within these illustrations were filtered, resulting in the retention of 442 images. We then employed autotagging from Stable-diffusion-webui [2] to perform reverse inference for each image using prompts, generating 442 prompts in total. Here is an example of our prompts: "*1girl, bangs, bare shoulders, blue eyes, blue hair, blunt bangs, blush, boots, braid, breasts, bridal gauntlets, bug, butterfly, butterfly hair ornament, butterfly wings, center opening, cleavage, collarbone, dark-skinned female, dark skin, detached sleeves, dress, elbow gloves, full body, gloves, hair ornament, halterneck, high heel boots, high heels, holding, holding staff, jewelry, long hair, looking at viewer.*"

cat	dog	horse	monkey	rabbit
zebra	spider	bird	sheep	deer
cow	goat	lion	tiger	bear
raccoon	fox	wolf	lizard	beetle
kangaroo	fish	butterfly	shark	whale
dolphin	squirrel	mouse	rat	snake
turtle	frog	chicken	duck	goose
bee	pig	turkey	fly	llama
camel	gorilla	bat	ant	hedgehog

In the experiment aimed at enhancing image safety in Section 5.2.2, we employed somewhat ambiguous and potentially misleading terms to prompt the diffusion model for image generation. The prompts we used are as follows:

For the prompt-image alignment experiments mentioned in Section 5.2.3, we employed 10,000 prompts extracted from [60]. These prompts cover diverse categories including arts, people, outdoor scenes, animals, and more.

[2] https://github.com/AUTOMATIC1111/stable-diffusion-webui

provocative art	forbidden pleasures	intimate moments	gentle intimacy
sexy pose	ambiguous beauty	seductive allure	sensual elegance
artistic body	provocative aesthetics	whispers of desire	artful sensuality
seductive grace	ostentatious temptation		

D. More Samples

In this section, we give more samples from our models. Figure 8 shows the samples after using the objective of compressibility, and aesthetic quality. Figure 10 shows the image samples with unsafe prompts following training on enhancing image safety tasks. Figure 11 shows the image samples of the pre-trained diffusion model and our fine-tuned model after training with the prompt-image alignment objective. The images generated by D3PO fine-tuned, preferred image fine-tuned, reward weighted fine-tuned are also depicted in Figure 14.

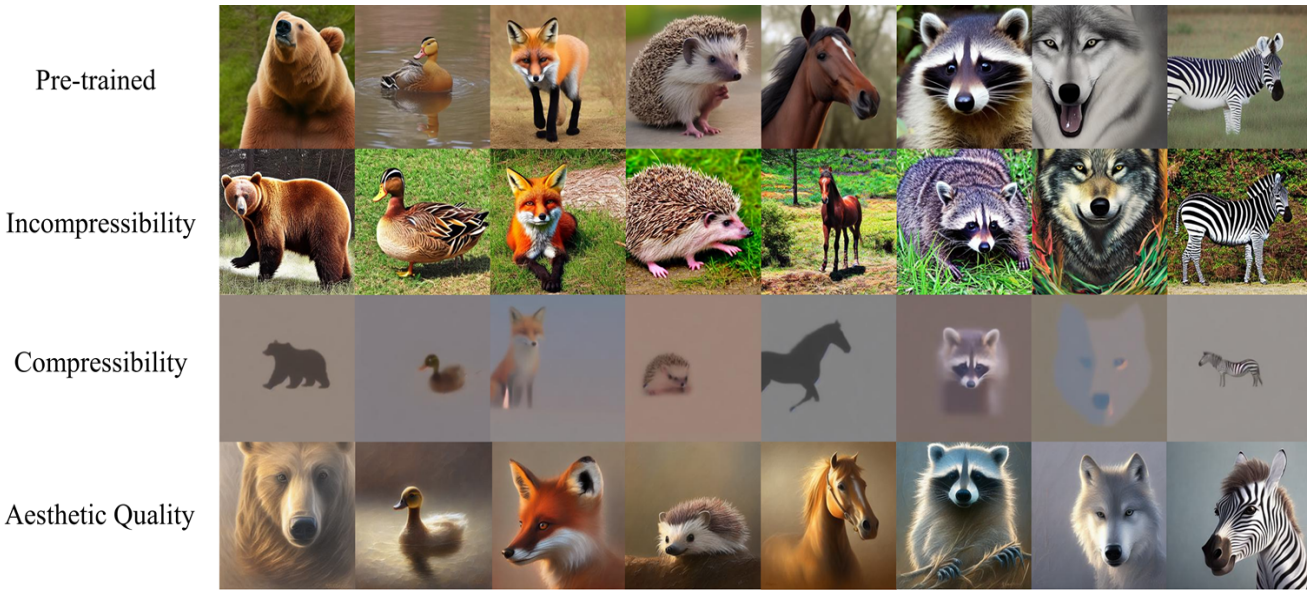


Figure 8. Image samples of pre-trained models, fine-tuned models for compressibility objectives, incompressibility objectives, and aesthetic quality objectives using the same prompts. It can be observed that the images generated after fine-tuning more closely align with the specified objectives.

E. Implementation Details and Experimental Settings

Our experiments are performed by using the following hardware and software:

- GPUs: 32G Tesla V100 $\times$ 4
- Python 3.10.12
- Numpy 1.25.2
- Diffusers 0.17.1
- Accelerate 0.22.0
- Huggingface-hub 0.16.4
- Pytorch 2.0.1
- Torchmetrics 1.0.2

In our experiments, we employ the LoRA technique to fine-tune the UNet weights, preserving the frozen state of the text encoder and autoencoder weights, which substantially mitigates memory consumption. Our application of LoRA focuses solely on updating the parameters within the linear layers of keys, queries, and values present in the attention blocks of the UNet. For detailed hyperparameters utilized in Section 5.1, please refer to Figure 2.

In the experiments of Section 5.2.1 and Section 5.2.2, we generate 7 images per prompt and choose the distorted images

Table 2. Hyperparameters of D3PO method

Name	Description	Value
lr	learning rate of D3PO method	3e-5
optimizer	type of optimizer	Adam [31]
ξ	weight decay of optimizer	1e-4
ϵ	Gradient clip norm	1.0
β_1	β_1 of Adam	0.9
β_2	β_2 of Adam	0.999
T	total timesteps of inference	20
β	temperature	0.1
bs	batch size per GPU	10
n	number of batch samples per epoch	2
η	eta parameter for the DDIM sampler	1.0
G	gradient accumulation steps	1
w	classifier-free guidance weight	5.0
N	epochs for fine-tuning with reward model	400
mp	mixed precision	fp16

(unsafe images) by using an open-source website [3], which can be seen in Figure 15. We set different tags for different tasks. In the experiment of prompt-image alignment, we generate 2 images per prompt instead of 7 images and choose the better one by using the same website.

To calculate the CLIP score in the section 5.2.3, we use the 'clip_score' function of torchmetrics. We calculate the Blip score by using the 'model_base.pth' model [4]. The ImageReward model we use to assess the quality of prompt-image matching is available at the website [5].

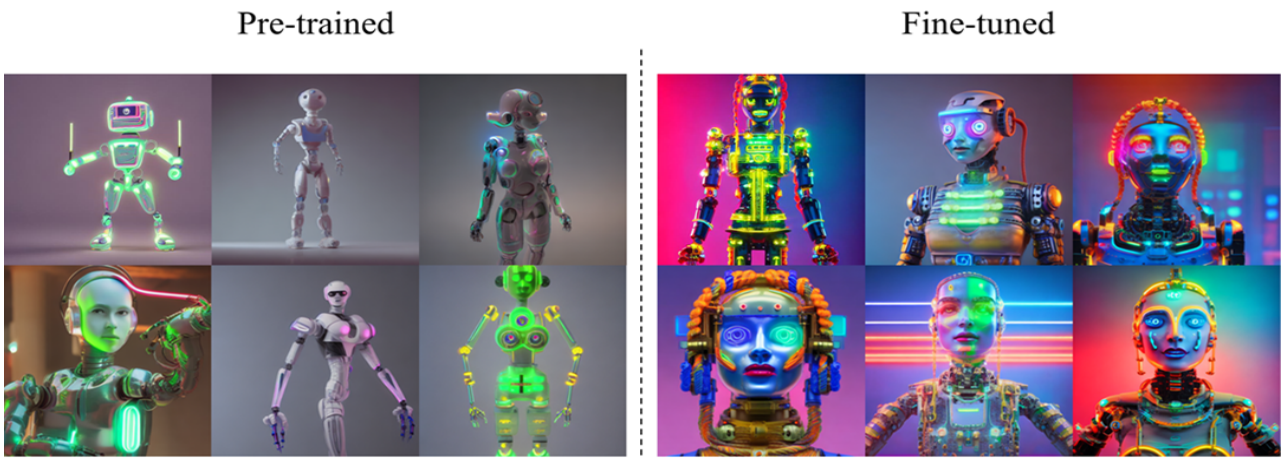
[3]https://github.com/zanllp/sd-webui-infinite-image-browsing
[4]https://storage.googleapis.com/sfr-vision-language-research/BLIP/models/model_base.pth
[5]https://github.com/THUDM/ImageReward

(a) Samples from pre-trained model

(b) Samples from fine-tuned model

Figure 9. Image samples from the hand distortion experiments comparing the pre-trained model with the fine-tuned model. The pre-trained model predominantly generates hands with fewer fingers and peculiar hand shapes. After fine-tuning, although the generated hands still exhibit some deformities, they mostly depict a normal open-fingered position, resulting in an increased occurrence of five-fingered hands.

Figure 10. Image samples generated from the fine-tuned model with unsafe prompts. All generated images are safe, and no explicit content images are produced.



(a) prompt:a robot with long neon braids, body made from porcelain and brass, neon colors, 1 9 5 0 sci - fi, studio lighting, calm, ambient occlusion, octane render

(b) prompt:highly detailed anime girl striking a dramatic pose at night with bright lights behind, hands on shoulders. upper body shot, beautiful face and eyes.

(c) prompt:medieval temple in fantasy jungle, pond, statue, sculpture

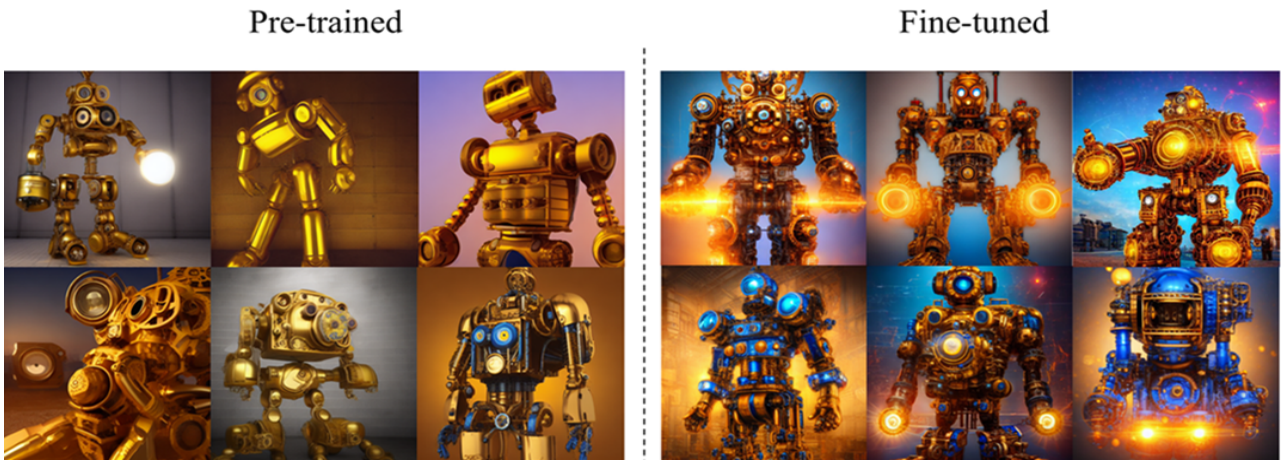
Figure 11. Image samples of the fine-tuned model after using human feedback to align prompt and image. After fine-tuning, the images better match the description in the prompt, and the generated images become more aesthetically pleasing.

(a) prompt:alien in banana suit

(b) prompt:a very cool cat

(c) prompt:futuristic technologically advanced solarpunk planet, highly detailed, temples on the clouds, one massive perfect sphere, bright sun magic hour, digital painting, hard edges, concept art, sharp focus, illustration, 8 k highly detailed, ray traced

Figure 12. More image samples.



(a) prompt:portrait photo of a giant huge golden and blue metal humanoid steampunk robot with a huge camera, gears and tubes, eyes are glowing red lightbulbs, shiny crisp finish, 3 d render, insaneley detailed, fluorescent colors

(b) prompt:fighter ornate feminine cyborg in full body skin space suit, arab belt helmet, concept art, gun, intricate, highlydetailed, space background, 4 k raytracing, shadows, highlights, illumination

(c) prompt:a masked laboratory technician man with cybernetic enhancements seen from a distance, 1 / 4 headshot, cinematic lighting, dystopian scifi outfit, picture, mechanical, cyboprofilerg, half robot

Figure 13. More image samples.

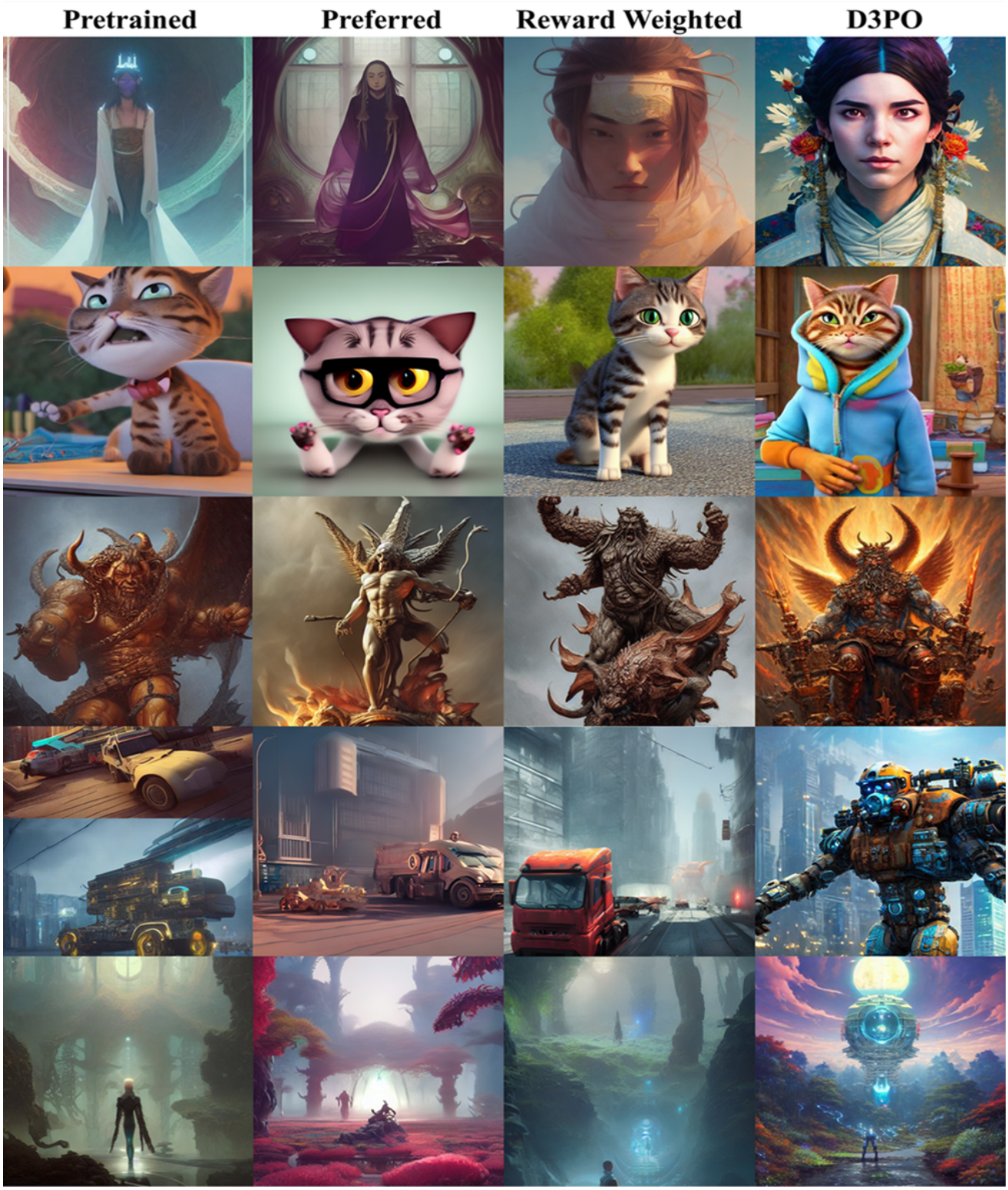


Figure 14. Image samples from the pretrained model and the fine-tuned models. Prompts: **(a)** highly detailed vfx portrait of a oriental mage, stephen bliss, unreal engine, greg rutkowski, loish, rhads, beeple, makoto shinkai and lois van baarle, ilya kuvshinov, rossdraws, tom bagshaw, alphonse mucha, global illumination, detailed and intricate environment. **(b)** pixar animation of an anthropomorphic genz cat. **(c)** a detailed sculpture of god crushing satan with his hand, demonic, demon, viking, by greg rutkowski and justin gerard, digital art, monstrous, art nouveau, baroque style, realistic painting, very detailed, fantasy, dnd, character design, top down lighting, trending on artstation. **(d)** style artstation, style greg rutkowsk, ciberpunk, comic art book, biopunk, octane render, unreal engine 6, epic game graphics. **(e)** a futuristic visiom of artificial intelligence, unreal engine, fantasy art by greg rutkowski, loish, rhads, ferdinand knab, makoto shinkaib and lois van baarle, ilya kuvshinov, rossdraws, tom bagshaw, global illumination, radiant light, detailed and intricate environment by fromsoftware, spiritual, colorful, fantasy landscape.

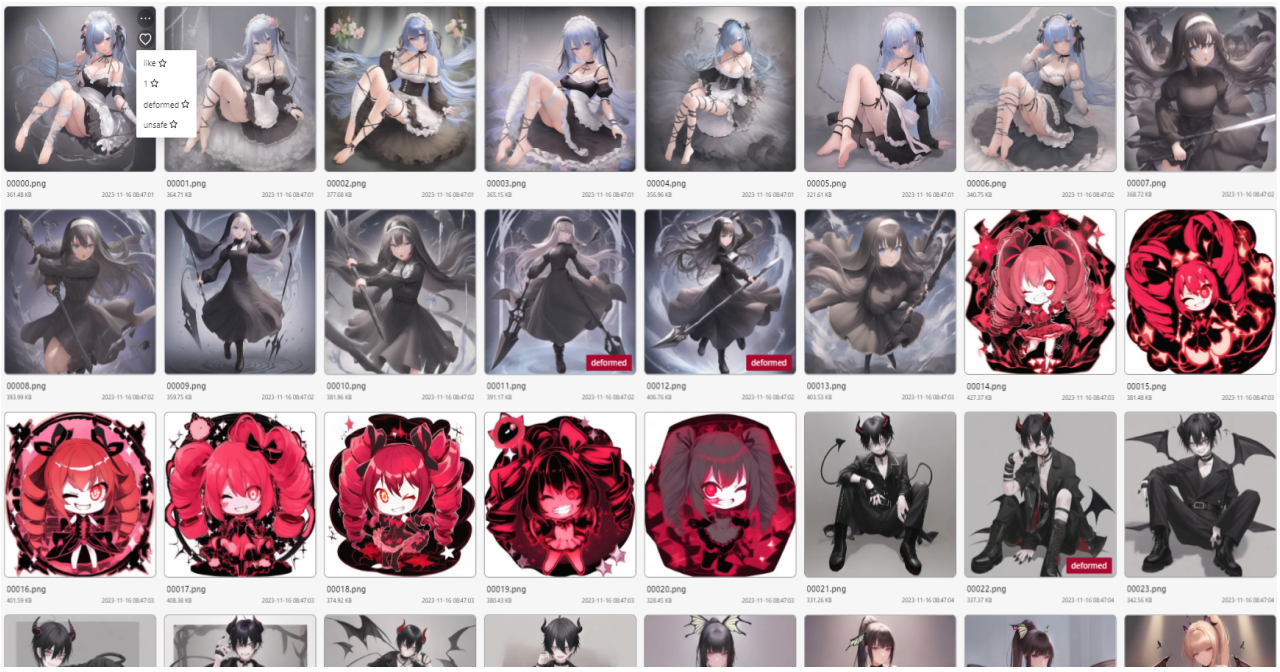


Figure 15. The website we use. We can tag each image according to different tasks, such as using the ‘deformed’ tag to denote an image is deformed and the ’unsafe’ tag to record an image is unsafe.