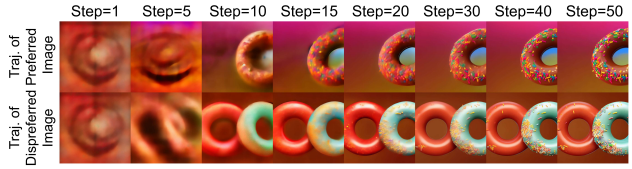
Aesthetic Post-Training Diffusion Models from Generic Preferences with Step-by-step Preference Optimization

Supplementary Material

Case A Prompt: a brown purse abandoned on a green bench



Case B Prompt: a glazed donut with sprinkles, octane render, high quality, hyper realistic, vibrant colors, 4k, soft lighting

Figure 7. Image samples showing disagreement between generic preferences and aesthetic preference. These images are generated by SDXL. The win trajectories in both examples have inferior aesthetics / details, which are detailed in Section 1 of the main text.

1. More Examples of Disagreements Between Generic and Aesthetic Preferences

We show more examples of the disagreement in Fig. 7. Case A of Fig. 7: The output image of the upper trajectory is generally preferred because it aligns more closely with the prompt "a brown purse abandoned on a green bench". However, when considering aesthetic preference, the output image of the upper trajectory is dispreferred because it has some artifacts on the right side of the purse, while the output image of the lower trajectory has clean details. Case B of Fig. 7: The output image of the upper trajectory is generally preferred over the one in the lower trajectory because it has the correct number of donuts as described in the prompt "a glazed donut with sprinkles, octane render, high quality, hyper realistic, vibrant colors, 4k, soft lighting". However, the color in the inner part of the donut of the upper output image is not consistent with the color in the background, making the output image of the lower trajectory aesthetically preferred.

We also examine the preference pairs from the Pick-a-Pic V1 dataset and showcase examples of disagreements between the Pick-a-Pic generic preference labels and aesthetic preference labels in Figure 8. These disagreements in the dataset hinder model's improvement in aesthetics.

Figure 8. More image samples showing disagreements between generic preferences and aesthetic preference in the Pick-a-Pic dataset (best viewed when zoomed in). These images were generated by various diffusion models. Prompt of (a): *"a stuffed animal of a blue fox"*. Prompt of (b): *"girl wearing red lipstick and black leggings"*. Prompt of (c): *"4 cars racing"*. Prompt of (d): *"gangsta clothed chicken"*. Prompt of (e): *"Equity markets were mixed Monday"*. Prompt of (f): *"A Kludde. A mythical monstrous black furry nocturnal dog with bear claws, green glistening scaled wings and glowing crimson eyes. Several heavy chains hang from its body and ankles"*. Prompt of (g): *"**a portrait of a 3D cockroach, wearing a bitcoin shirt, in Hawaii, on the beach, hyper-realistic, ultra-detailed, photography, hyper-realistic, photo-realistic, ultra-photo-realistic, super-detailed, intricate details, 8K, surround lighting, HDR"*. Prompt of (h): *"a suit of armour constructed from meat"*.

2. Timestep-conditional CLIP Vision Encoder

In this section, we present the implementation details of the timestep-conditional CLIP vision encoder introduced in Section 4.2 of the main text. The original CLIP vision encoder is based on a Vision Transformer (ViT) [5]. To incorporate timestep conditioning, we follow the approach in DiT [17], employing time-conditional adaptive layernorm to inject timestep information into the vision encoder of CLIP. We replace all transformer blocks in the CLIP vision

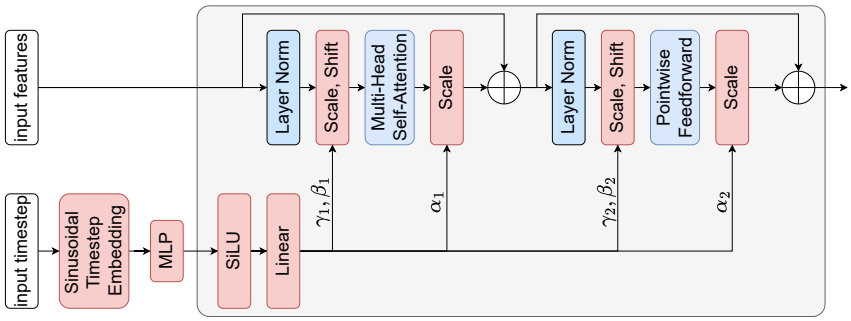


Figure 9. Timestep-conditioned ViT block. The blue components are from the original ViT block and the red componets are newly added. $\oplus$ represents element-wise addition.

Figure 10. Qualitative comparison between Glyph-ByT5-SDXL and Glyph-ByT5-SDXL + SPO in graphic design image generation. SPO consistently improves image aesthetics by creating nuanced textures and vibrant colors without sacrificing image content accuracy.

encoder with timestep-conditioned ViT blocks, the structure of which is illustrated in Figure 9. The blue components in the figure represent the original Transformer block, while the red components denote the newly added components. We embed the input timestep as timestep embeddings using sinusoidal encoding, followed by an MLP with the structure `Linear-SiLU-Linear`. The same timestep embeddings are used as input for all timestep-conditioned ViT blocks.

A linear layer is employed to predict dimension-wise scaling parameters γ and α, as well as the dimension-wise shifting parameter β. Given input features x, the "Scale, Shift" operation modifies x as $x = x \times (1 + \gamma) + \beta$, while the "Scale" operation adjusts x as $x = x \times \alpha$. The "Scale, Shift" operation is applied directly after each layer normalization block, whereas the "Scale" operation is applied immediately after the multi-head self-attention and pointwise feedforward blocks, prior to the residual connections.

We initialize the step-aware preference model (SPM) with PickScore [11] weights. To preserve the pretrained knowledge, the weights of the linear layer responsible for generating γ, β, and α are initialized such that $\gamma = 0$, $\beta = 0$, and $\alpha = 1$, ensuring that the SPM's output matches the pretrained PickScore model's output at the beginning of training.

3. Generalization to Text Generation.

We verify the generalization of SPO by simply marrying the LoRA weights of SPO-SDXL to the Glyph-ByT5-SDXL model [14], which specializes in design image generation. Qualitative examples are shown in Fig. 10, where we observe that SPO consistently improves the visual appeal of Glyph-ByT5-SDXL images, *e.g.*, richer texture of the elephant, flower, robot, and beer mug, while preserving the text generation ability of Glyph-ByT5-SDXL.

4. Detailed Prompts

We summarize the detailed text prompts used in Figure 2 of the main text in Table 10.

5. More Sample Images Generated by SPO-SDXL

Figure 11 presents additional sample images generated by SPO-SDXL, accompanied by the corresponding text prompts listed in Table 11.

Table 10. Detailed prompts used for generated images in Figure 2 of the main text.

Image	Prompt
Col1	Saturn rises on the horizon.
Col2	a watercolor painting of a super cute kitten wearing a hat of flowers
Col3	A galaxy-colored figurine floating over the sea at sunset, photorealistic.
Col4	fireclaw machine mecha animal beast robot of horizon forbidden west horizon zero dawn bioluminiscence, behance hd by jesper ejsing, by rhads, makoto shinkai and lois van baarle, ilya kuvshinov, rossdraws global illumination
Col5	A swirling, multicolored portal emerges from the depths of an ocean of coffee, with waves of the rich liquid gently rippling outward. The portal engulfs a coffee cup, which serves as a gateway to a fantastical dimension. The surrounding digital art landscape reflects the colors of the portal, creating an alluring scene of endless possibilities.
Col6	A profile picture of an anime boy, half robot, brown hair
Col7	Detailed Portrait of a cute woman vibrant pixie hair by Yanjun Cheng and Hsiao-Ron Cheng and Ilya Kuvshinov, medium close up, portrait photography, rim lighting, realistic eyes, photorealism pastel, illustration
Col8	On the Mid-Autumn Festival, the bright full moon hangs in the night sky. A quaint pavilion is illuminated by dim lights, resembling a beautiful scenery in a painting. Camera type: close-up. Camera lens type: telephoto. Time of day: night. Style of lighting: bright. Film type: ancient style. HD.

Table 11. Detailed prompts used for generated images in Figure 11.

Image	Prompt
Row 1, Col1	paw patrol. ”This is some serious gourmet”. 2 dogs holding mugs.
Row 1, Col2	little tiny cub beautiful light color White fox soft fur kawaii chibi Walt Disney style, beautiful smiley face and beautiful eyes sweet and smiling features, snuggled in its soft and soft pastel pink cover, magical light background, style Thomas kinkade Nadja Baxter Anne Stokes Nancy Noel realistic
Row 1, Col3	Full Portrait of Consort Chunhui by Giuseppe Castiglione, symmetrical face, ancient Chinese painting, single face, insanely detailed and intricate, beautiful, elegant, artstation, character concept in the style illustration by Miho Hirano, Giuseppe Castiglione –ar 9:16
Row 1, Col4	Surfer robot dude in the crest of a wave, cinematic, sunny, –ar 16:9
Row 1, Col5	a photo of a frog holding an apple while smiling in the forest
Row 2, Col1	185764, ink art, Calligraphy, bamboo plant :: orange, teal, white, black –ar 2:3 –uplight
Row 2, Col2	large battle Mecha helping with the construction of the Colossus of Rhodos standing above the entry of a harbor, hundreds of ancient workers on scaffolding surrounding the colossus, ancient culture, sunny weather, matte painting, highly detailed, cgsociety, hyperrealistic, –no dof, –ar 16:9
Row 2, Col3	A 3D Rendering of a cockatoo wearing sunglasses. The sunglasses have a deep black frame with bright pink lenses. Fashion photography, volumetric lighting, CG rendering,
Row 2, Col4	a golden retriever dressed like a General in the north army of the American Civil war. Portrait style, looking proud detailed 8k realistic super realistic Ultra HD cinematography photorealistic epic composition Unreal Engine Cinematic Color Grading portrait Photography UltraWide Angle Depth of Field hyperdetailed beautifully colorcoded insane details intricate details beautifully color graded Unreal Engine Editorial Photography Photography Photoshoot DOF Tilt Blur White Balance 32k SuperResolution Megapixel ProPhoto RGB VR Halfrear Lighting Backlight Natural Lighting Incandescent Optical Fiber Moody Lighting Cinematic Lighting Studio Lighting Soft Lighting Volumetric ContreJour Beautiful Lighting Accent Lighting Global Illumination Screen Space Global Illumination Ray Tracing Optics Scattering Glowing Shadows Rough Shimmering Ray Tracing Reflections Lumen Reflections Screen Space Reflections Diffraction Grading Chromatic Aberration GB Displacement Scan Lines Ray Traced Ray Tracing Ambient Occlusion AntiAliasing FKAA TXAA RTX SSAO Shaders
Row 2, Col5	A rock formation in the shape of a horse, insanely detailed,
Row 3, Col1	a gopro snapshot of an anthropomorphic cat dressed as a firefighter putting out a building fire
Row 3, Col2	a desert in a snowglobe, 4k, octane render :: cinematic –ar 2048:858
Row 3, Col3	cat, cute, child, hat
Row 3, Col4	watercolour beaver with tale, white background
Row 3, Col5	corporate office ralph goings – aspect 3:2
Row 4, Col1	there once was a fly on the wall, I wonder, why didn’t it fall, Because its feet stuck, Or was it just luck, Or does gravity miss things so small, high realistic, high detailed, high contrast, unreal render –ar 16:9
Row 4, Col2	lush landscape with mountains with cherry trees by Miyazaki Nausicaa Ghibli, 王ランキング, ranking of kings, spirited away, breath of the wild style, epic composition, clean –w 1024 –h 1792 –no people
Row 4, Col3	what i dream when i close my eyes to sleep
Row 4, Col4	cute cat jumped off plane in parachute, exaggerated expression, photo realism, side angle, epic drama
Row 4, Col5	Full body, a Super cute little girl, wearing cute little giraffe pajamas, Smile and look ahead, ultra detailed sky blue eyes, 8k bright front lighting, fine luster, ultra detail, hyper detailed 3D rendering s750,

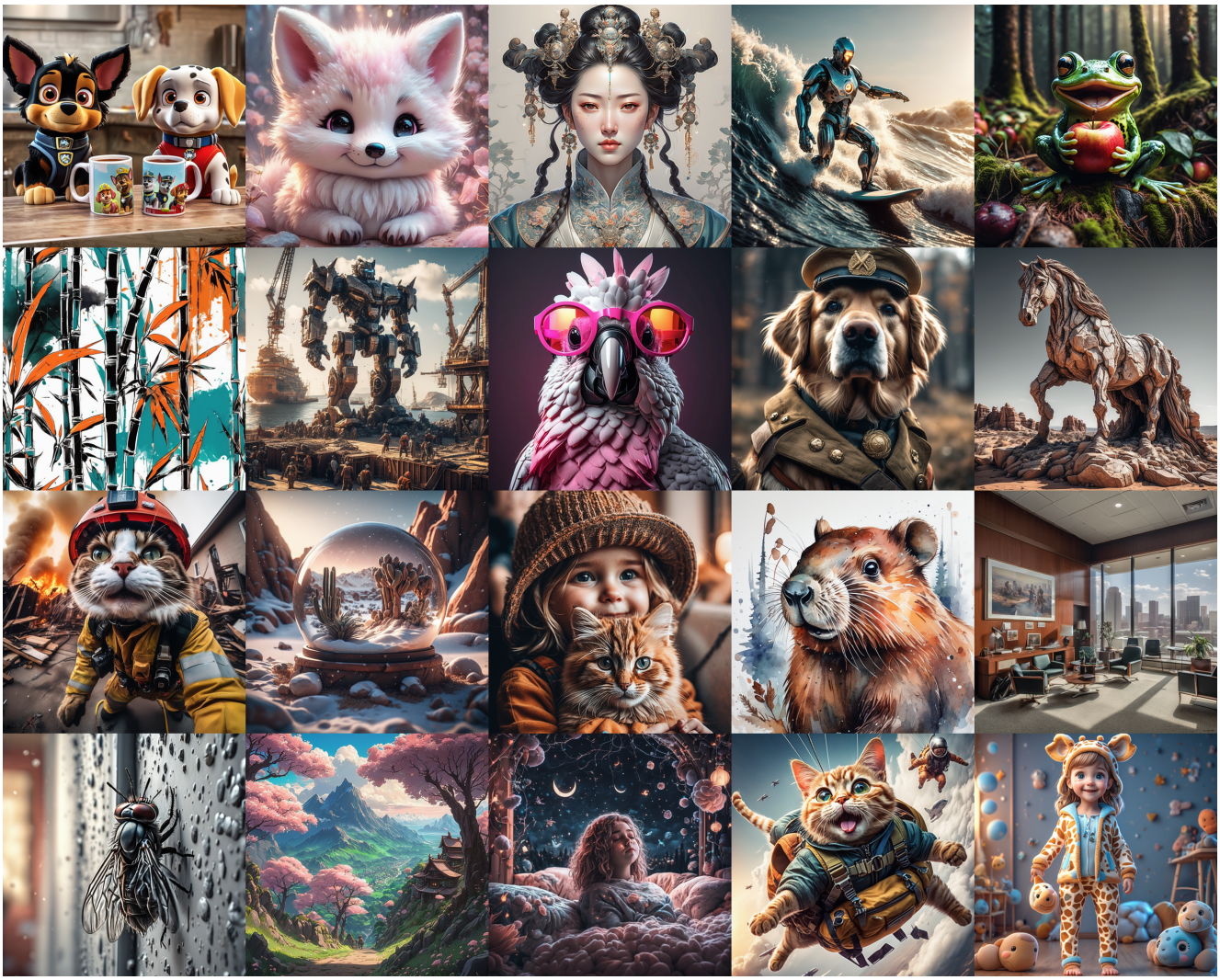


Figure 11. Sample images generated by SPO-SDXL. With the SPO post-training, SPO-SDXL produces high-quality images that are visually attractive and stunning.