

DyMO: Training-Free Diffusion Model Alignment with Dynamic Multi-Objective Scheduling

Supplementary Material

7. More Details of the Method

7.1. More Details of Dynamic Scheduling in DyMO

The proposed method includes two alignment objectives (as guidance in inference) and dynamically schedules the usage of the objective (Sec. 4.2), the step size, and the time-travel recurrent steps (Sec. 4.3). We demonstrate the generative denoising process in Fig. 8 with examples. It shows the intermediate images of the baseline model and the proposed method (including the noisy images and the predicted clean images).

To achieve alignment in the generative denoising process, we used the text-aware human preference score $\mathcal{L}_R(\mathbf{x}'_{0|t}, \mathbf{c}, t)$ to guide the denoising process with the gradient computed for the intermediate noisy images. As discussed in the main paper, the samples in the early stage are highly noisy and the predicted clean images obtained through one-step approximation are also blurred, as shown in Fig. 8, demonstrating that the preference model cannot provide effective and accurate guidance. Especially, semantic context is often established during the initial denoising steps, yet it lacks effective supervision at this critical stage. Relying on the proposed semantic alignment objective \mathcal{L}_A depending on the semantic contents reflected in the text-vision attention maps, the proposed method can effectively guide the alignment in the early noisy steps. Some previous works also control the contents by relying on the manipulation on the noise map or attention maps, which are restricted to pre-defined and additionally given layout [40, 54] or oversimplified the semantics [55], limiting them to simple text prompts and restricted usage cases.

We integrate the two objectives \mathcal{L}_R and \mathcal{L}_A into a dynamic scheduling process through the weight w (and the corresponding $1 - w$) in Eq. (10), which are also represented as adaptive weights w_A and w_R for convenience in Algorithm 1. The weights are step- t -dependent and are adjusted according to the relative changes of \mathbf{z} . Considering the strength and the different goals of the two objectives, we let \mathcal{L}_R and \mathcal{L}_A be functional more at the early and later stages, respectively, as shown in Algorithm 1. Additionally, we also dynamically schedule the step size of the update of \mathbf{z} and the number of recurrent steps in the dynamic time-travel strategy, as introduced in Sec. 4.3. The dynamic adjustment operations enable the model to achieve an adaptive alignment process that can be aware of the status of specific steps, leading to both better effectiveness and efficiency.

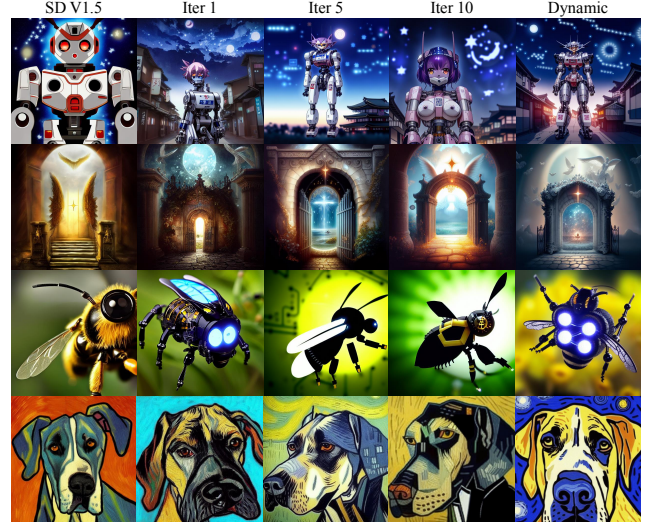


Figure 7. Qualitative comparison of iteration count in time-travel strategy. The prompts from top to bottom are: (1) A portrait of an anime mecha robot with a Japanese town background and a starred night sky. (2) The gate to the eternal kingdom of angels, fantasy, digital painting, HD, detailed. (3) mechanical bee flying in nature, electronics, motors, wires, buttons, lcd, led instead of eyes, antennas instead of feet. (4) A Great dane dog in the style of Vincent Van Gogh.

8. Additional Experimental Details and Results

8.1. Details of Prompts used in Experiments

The text prompts used to generate the images in Fig. 1, Fig. 3, and Fig. 4 are summarized in Tab. 6, Tab. 7, and Tab. 8, respectively, providing a clear reference for the input descriptions corresponding to each figure.

8.2. Constructing Semantic Graph from Input Text Prompts

In Sec. 4.1, we introduce semantic alignment guidance for more effective alignment of the image contents with the semantic contents and intention in the user’s input text prompts, which requires extracting the semantic information (and semantic graph) from the text. It is achieved through a pre-trained large language model (LLM) with some designed instruction prompts, as mentioned in the paper. We provide some exemplar cases of text semantic graph in Fig. 13.



Figure 8. The entire denoising process of SD V1.5 and DyMO, where \mathbf{x}_t and $\mathbf{x}'_{0|t}$ denote the noisy images and one-step predicted clean images at step t , respectively.

8.3. Additional Qualitative Results

We provide more visual results for qualitative evaluation.

In Fig. 3 in the main paper, we provide some exemplar cases of comparing our methods with the baseline model (SD V1.5 [32]), training-free models (*e.g.*, DNO [41], PromptOpt [7], FreeDom [54]) and training-based models (*e.g.*, AlignProp [30], Diffusion-DPO [45], Diffusion-KTO [23], SPO [24]). We provide a more comprehensive comparison with more examples of generated images in Fig. 10 with the corresponding text prompts in Tab. 9. In Fig. 10 and Fig. 3, all the compared methods are based on the same baseline model SD V1.5. It is observed that our method contains semantic information aligning better with the input prompts, such as the frog holding an apple in Fig. 3 as well as the horse shape in the 4th row and the sunglass in the

7th row in Fig. 10. The generated images by our methods contain more visually appealing appearances (*e.g.*, rich details and visual characteristics) that are highly aligned with human preferences, such as color vibrancy (2nd and 3rd rows in Fig. 10), vivid lighting effects (5th and 8th rows in Fig. 10) and detailed textures (1st, 6th and last rows in Fig. 10), etc.

Beyond the baseline SD V1.5, we also validate our method by applying it to other models such as SDXL [29], Diffusion-DPO [45], and SPO [24], and demonstrate the results in Fig. 11 and Fig. 12 (with the corresponding text prompts in Tab. 10). Fig. 4 (with text in Tab. 8) in the main paper also includes a small set of results based on SDXL. And Tab. 1 demonstrates the numerical results of the comparison. The results show that the proposed training-free

Table 5. Geneval Benchmark evaluation based on SD V1.5.

Methods	Overall	Single object	Two object	Counting	Colors	Position	Color attribution
SD V1.5	0.42	1.00	0.38	0.35	0.77	0.04	0.00
DNO	0.43	0.96	0.35	0.35	0.82	0.04	0.05
PromptOpt	0.39	0.96	0.25	0.23	0.80	0.04	0.05
FreeDom	0.52	1.00	0.56	0.64	0.80	0.02	0.13
AlignProp	0.42	1.00	0.23	0.41	0.74	0.04	0.06
Diffusion-DPO	0.47	1.00	0.48	0.47	0.80	0.02	0.05
Diffusion-KTO	0.49	1.00	0.53	0.41	0.82	0.06	0.11
SPO	0.47	0.96	0.48	0.47	0.74	0.11	0.05
SD V1.5+Ours	0.57	1.00	0.72	0.47	0.83	0.07	0.34

alignment method can generally improve the performance after adding it to different pre-trained models. Compared to the baseline models, our approach generates high-quality images more closely aligned with contextual semantics and better cater to human preferences. We also show the latest state-of-the-art generative models (FLUX [21], SD V3.5 [10]) as a reference. The effectiveness of the proposed method DyMO is still obvious in the comparison with them, in terms of visual coherence and detail fidelity.

8.4. Additional Quantitative Results

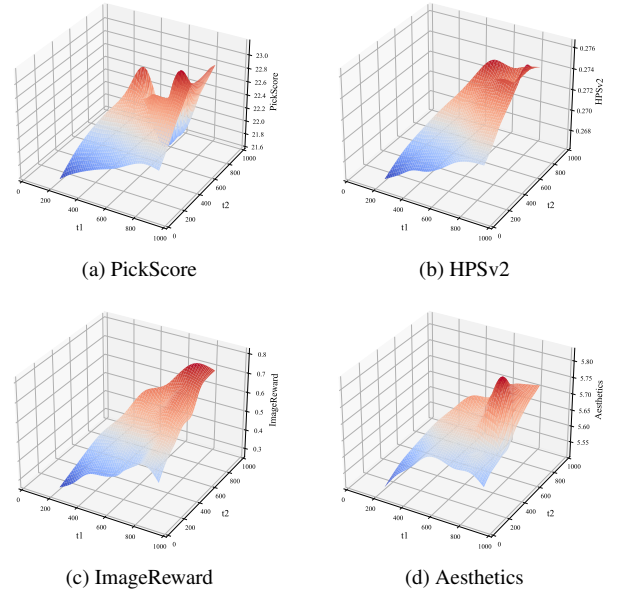
We conduct quantitative evaluation on Geneval benchmark [14] and show comparisons in Tab. 5. Our method performs very well and shows superiority in many aspects, *e.g.*, overall, attribute binding and object synthesis. The proposed semantic alignment can help DyMO on multi-object synthesis. Tab. 5 also shows that DyMO performs well in multiple objects and counting.

8.5. Additional Details on Human Evaluation

We conduct the user study through survey forms, organizing the content into distinct sections based on each prompt. Each section is further divided into three partitions—Q1, Q2, and Q3—corresponding to Fig. 14, Fig. 15, Fig. 16, respectively. In addition, investigators are recruited through an online platform, ensuring their anonymity. Each participant is required to have at least a bachelor’s degree and their privacy and identity are kept confidential throughout the entire process.

8.6. Additional Results on Dynamic Scheduling

The scheduling weights (Eq. (10)) are automatically adjusted based on content changes to balance the roles of \mathcal{L}_A and \mathcal{L}_R in different stages. In the early stage, \mathcal{L}_A dominates due to content instability and noise (where \mathcal{L}_R ’s guidance is weak). As semantics stabilize, \mathcal{L}_R gains more weights,

Figure 9. Insensitivity to scheduling parameter t_1, t_2 .

refining alignment with more detailed visual preferences. Thus, the method prioritizes semantic alignment first, then preference alignment. Based on the observed dynamics of w_R and w_A (trend and smoothness), we set the stage split (t_1, t_2) as (800, 500) for efficiency and simplicity. The model is insensitive to this hyperparameter, as confirmed by the grid-search analysis (Fig. 9).

8.7. Additional Results on Dynamic Time Travel Steps

In Tab. 4 in the main paper, we provide ablation study results of different time-travel steps. By comparison, our proposed dynamic recurrent strategy effectively achieves better performance on different metrics. We provide some visual

comparison examples in Fig. 7, where we can observe that the visual qualities of the generated image are consistent with the numerical results. The proposed method with dynamic recurrent step scheduling is effective in producing better results with less time. With the proposed alignment objectives, the proposed method can also work well with less (and fixed) recurrent step numbers. In addition, compared with the other training-free methods with full-chain backpropagation methods, like DNO [41] and PromptOpt [7], require 370 and 280 seconds, respectively, our method can be more efficient and effective. Compared with the one-step approximation methods, such as FreeDom [54] (with 170 seconds), our method is also efficient and performs better on the results. While many training-free methods take more time to guide the denoising process, our method maintains strong performance in just 40s with fewer iterations as demonstrated in Tab. 4 and Fig. 7. The whole generation process can be further accelerated by incorporating more efficient sampling processes, which is left as future work.

9. Ethical and Social Impacts

The development of DyMO, a training-free alignment framework for text-to-image diffusion models, brings ethical and social implications that require careful consideration to ensure responsible AI deployment. While our method enhances alignment with human preferences and promotes inclusivity, it also raises challenges such as mitigating biases, preserving privacy, and preventing misuse. DyMO relies on pre-trained models and publicly available datasets, which may encode societal biases or reinforce stereotypes. To address this, we emphasize the need for dataset diversity assessment, bias identification, and mechanisms to ensure inclusive and equitable representations. Privacy concerns are mitigated by advocating for anonymization of data and obtaining explicit consent for identifiable imagery. Additionally, we recognize the risks of misuse, such as generating harmful or misleading content, and propose safeguards like content moderation and ethical usage guidelines. Despite these challenges, DyMO holds the potential to advance social equality by improving accessibility and enabling personalized content generation for underrepresented groups. By balancing innovation with responsibility, we aim to democratize advanced generative techniques while upholding fairness, transparency, and inclusivity. Our commitment to responsible AI development underpins our efforts to address these concerns, ensuring that DyMO contributes positively to the field while minimizing potential risks.

Table 6. Detailed prompts used for generated images in Fig. 1.

Image	Prompt
Fig. 1, Row 1, Col 1	Two monkeys are piloting an airplane.
Fig. 1, Row 1, Col 2	Award-winning Kawaii illustration of a cat samurai, holding two swords, background cyberpunk Styles, 4k, golden hour, cinematic light.
Fig. 1, Row 1, Col 3	crop top skinny russian 12 years old teen girl at the water mountain, HDR magazine photo.
Fig. 1, Row 1, Col 4	a tower of cheese.
Fig. 1, Row 1, Col 5	A painting of a koala wearing a princess dress and crown, with a confetti background.
Fig. 1, Row 1, Col 6	Disease Monitoring: Through big data technology, trends in specific diseases can be monitored and predicted, thus improving disease prevention and treatment effectiveness.
Fig. 1, Row 1, Col 7	Gnomes are playing music during Independence Day festivities in a forest near Lake George.
Fig. 1, Row 2, Col 1	paw patrol. 'This is some serious gourmet'. 2 dogs holding mugs.
Fig. 1, Row 2, Col 2	Harry potter as a cat, pixar style, octane render, HD, high-detail.
Fig. 1, Row 2, Col 3	A small green dinosaur toy with orange spots standing on its hind legs and roaring with its mouth open.
Fig. 1, Row 2, Col 3	Two cats watering roses in a greenhouse.
Fig. 1, Row 2, Col 4	Chic Fantasy Compositions, Ultra Detailed Artistic, Midnight Aura, Night Sky, Dreamy, Glowing, Glamour, Glimmer, Shadows, Oil On Canvas, Brush Strokes, Smooth, Ultra High Definition, 8k, Unreal Engine 5, Ultra Sharp Focus, Art By magali villeneuve, rossdraws, Intricate Artwork Masterpiece, Matte Painting Movie Poster.
Fig. 1, Row 2, Col 5	a toy poodle as a rocket scientist.
Fig. 1, Row 2, Col 5	A young woman witch cosplaying with a magic wand and broom, wearing boots, and posing in a full body shot with a detailed face.
Fig. 1, Row 3, Col 1	The image is a portrait of Homer Simpson as a Na'vi from Avatar, created with vibrant colors and highly detailed in a cinematic style reminiscent of romanticism by Eugene de Blaas and Ross Tran, available on Artstation with credits to Greg Rutkowski.
Fig. 1, Row 3, Col 2	Anthropomorphic beagle dog wearing steampunk time traveller outfit, clocks and large round window above, photoreal epic composition, old world deco, tv commercial, sebastian kruger, artem, epic lighting, by Heinz Anger, wow factor, aardman animations, blocking the sun, very artistic pose, alexander abdulov.
Fig. 1, Row 3, Col 3	A happy daffodil with big eyes, multiple leaf arms and vine legs, rendered in 3D Pixar style.
Fig. 1, Row 3, Col 4	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.
Fig. 1, Row 3, Col 5	A slime monster.

Table 7. Detailed prompts used for generated images in Fig. 3.

Image	Prompt
Fig. 3, Row 1	A photo of a frog holding an apple while smiling in the forest.
Fig. 3, Row 2	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.
Fig. 3, Row 3	a gopro snapshot of an anthropomorphic cat dressed as a firefighter putting out a building fire.
Fig. 3, Row 4	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.

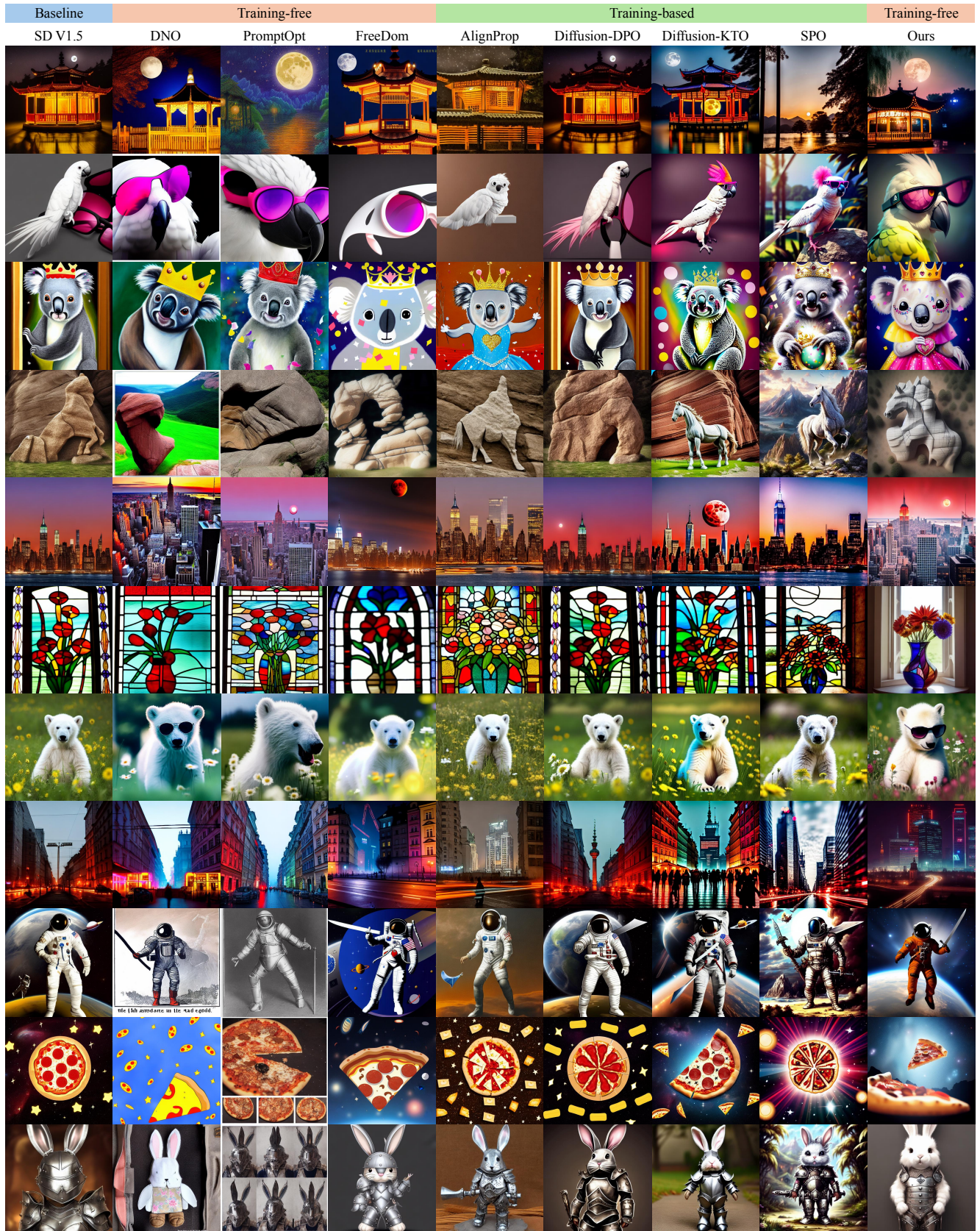


Figure 10. Qualitative comparison based on SD V1.5 backbones. The prompts are provided in the Tab. 9.

Table 8. Detailed prompts used for generated images in Fig. 4.

Image	Prompt
Fig. 4, Row 1	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.
Fig. 4, Row 2	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.
Fig. 4, Row 3	A smiling beautiful sorceress wearing a high necked blue suit surrounded by swirling rainbow aurora, hyper-realistic, cinematic, post-production.
Fig. 4, Row 4	a white polar bear cub wearing sunglasses sits in a meadow with flowers.

Table 9. Detailed prompts used for generated images in Fig. 10.

Image	Prompt
Fig. 10, Row 1	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.
Fig. 10, Row 2	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.
Fig. 10, Row 3	A painting of a koala wearing a princess dress and crown, with a confetti background.
Fig. 10, Row 4	A rock formation in the shape of a horse, insanely detailed.
Fig. 10, Row 5	New York city skyline during the day with a huge red moon in the sky.
Fig. 10, Row 6	A stained glass vase with flowers in front of a window.
Fig. 10, Row 7	A white polar bear cub wearing sunglasses sits in a meadow with flowers.
Fig. 10, Row 8	Warsaw cyberpunk style at night.
Fig. 10, Row 9	astronaut in space with a two handed sword in plate armor in front of the earth.
Fig. 10, Row 10	a slice of pizza floating through space with stars in the background.
Fig. 10, Row 11	a cute bunny wear detailed metal armour.

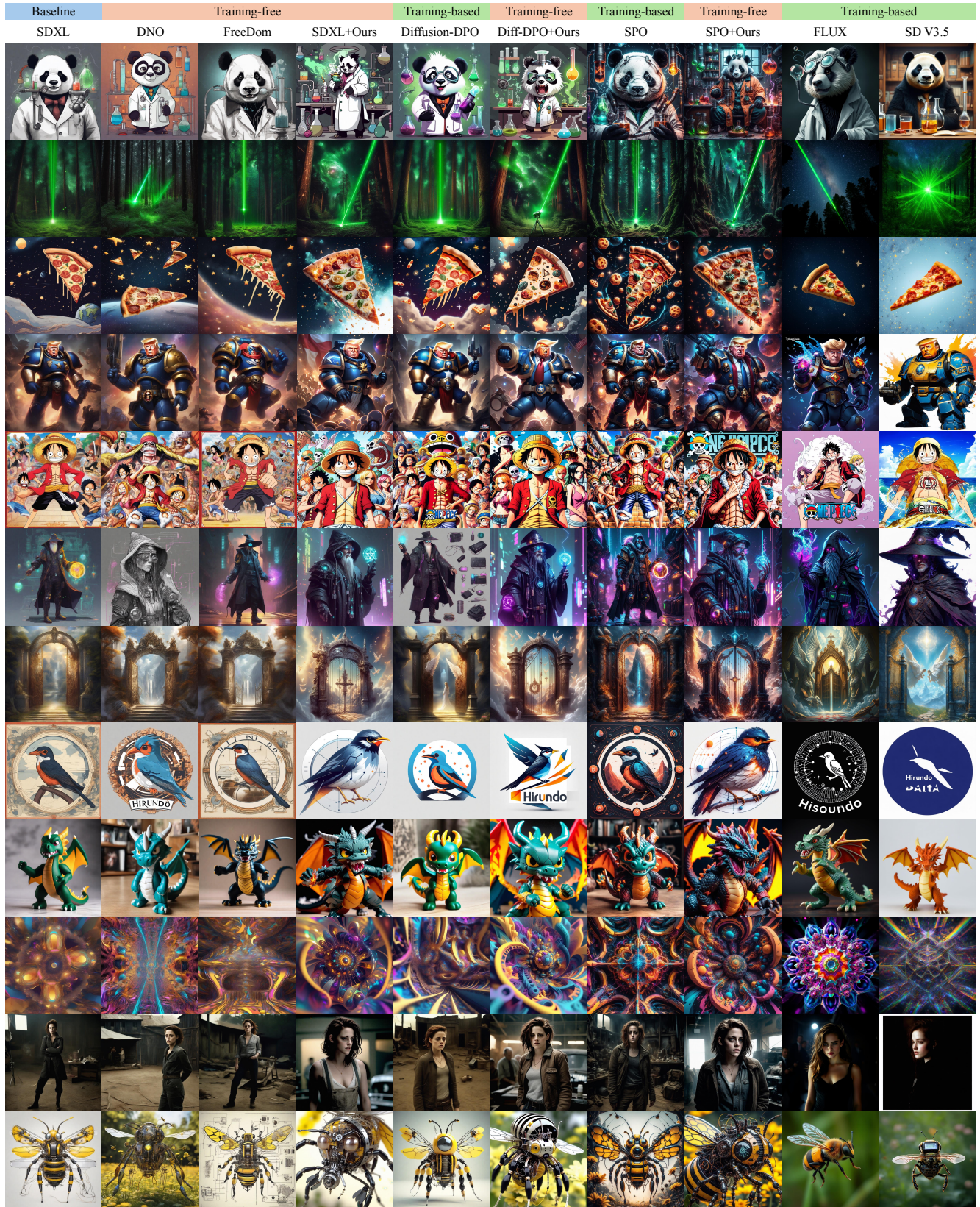


Figure 11. Qualitative comparison based on SDXL backbones. The prompts are provided in the Tab. 10.

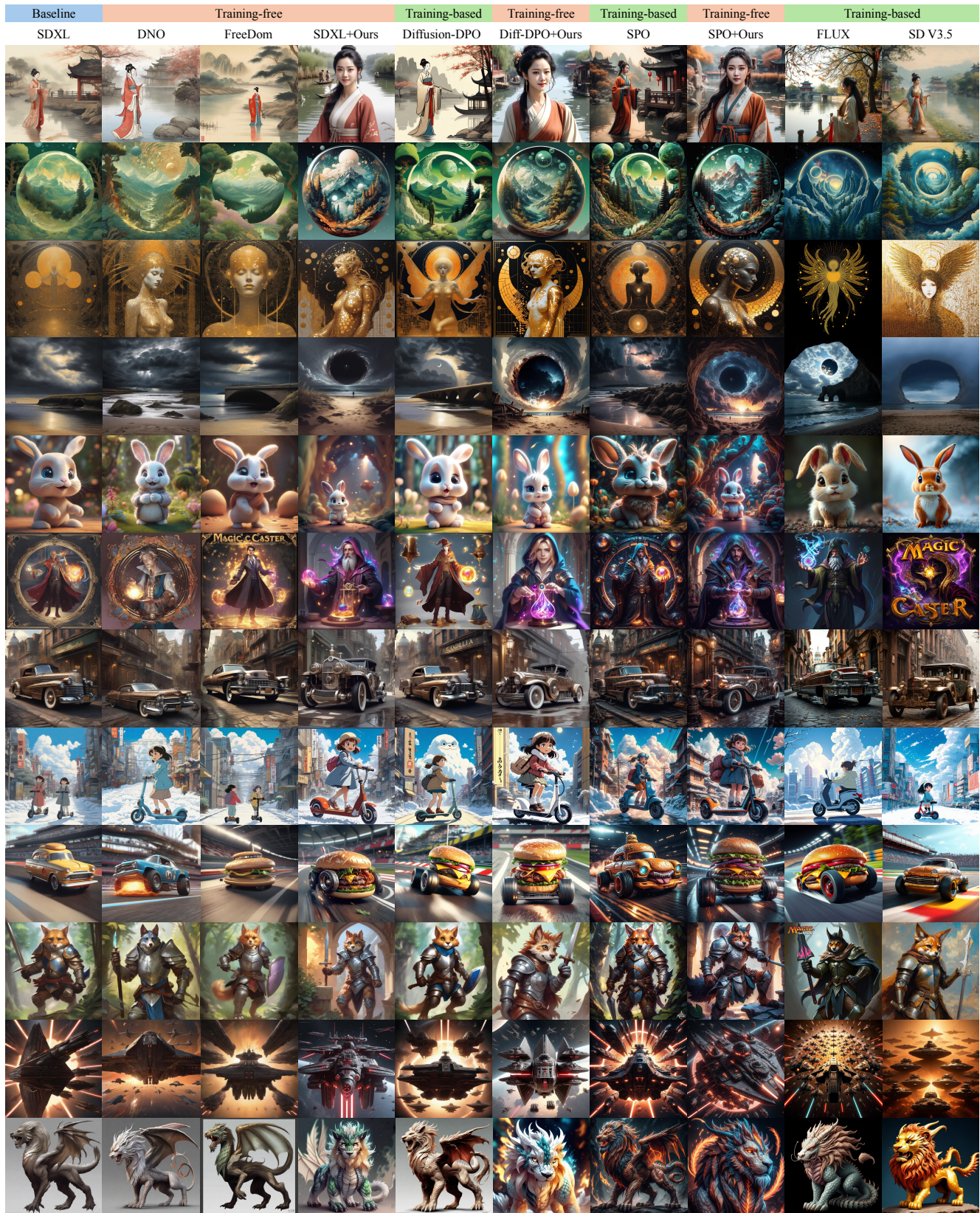


Figure 12. Qualitative comparison based on SDXL backbones. The prompts are provided in the Tab. 10.

Table 10. Detailed prompts used for generated images in Fig. 11 and Fig. 12.

Image	Prompt
Fig. 11, Row 1	A panda bear as a mad scientist.
Fig. 11, Row 2	Green laser in space among galaxies forest.
Fig. 11, Row 3	a slice of pizza floating through space with stars in the background.
Fig. 11, Row 4	disney league of legends splash art of space marine donald trump.
Fig. 11, Row 5	one piece anime cover.
Fig. 11, Row 6	cyberpunk wizard.
Fig. 11, Row 7	The gate to the eternal kingdom of angels, fantasy, digital painting, HD, detailed.
Fig. 11, Row 8	Hirundo a data startup logo.
Fig. 11, Row 9	a dragon vinyl toy in a fighting pose.
Fig. 11, Row 10	Psychedelic synesthesia complex 3d rendered ethereal 8 complex shapes in different sizes fractal futuristic geometry POV replication of future riddim.
Fig. 11, Row 11	masterpiece portrait of Kristen Stewart standing in the movie set, film photography, dark atmosphere, sharp focus, photographed by Annie Leibovitz.
Fig. 11, Row 12	mechanical bee flying in nature, electronics, motors, wires, buttons, lcd, led instead of eyes, antennas instead of feet.
Fig. 12, Row 1	At Song dynasty, a pretty woman in chinese was walking along the river.
Fig. 12, Row 2	spiral mountains, stars, clouds within spheres, green pine trees, chill, calmness, peace, eternity, beauty, ernst haeckel, maria sibylla merian, tristan eaton, victo ngai, artgerm, rhads, ross draws, kaethe butcher, hajime sorayama, greg tocchini, virgil finlay, subtle vignette, volumetric lights, pixiv, by ilya kuvshinov, octane render, 4k, 8k.
Fig. 12, Row 3	Dot matrix, pointillism, seurat, signac, frazetta, brom, Zdzisław Beksiński, Moebius, Egon Schiele, art nouveau, a being made out of pure golden light. angelic and graceful. gold foil inlay with erratic shapes and geometric patterns. black and gold and white and orange. in the style of alphonse mucha and amanda guse and android jones. tarot card style symmetry.
Fig. 12, Row 4	A dark hole in the sky , pont of view from a beach, masterpiece.
Fig. 12, Row 5	Cute and adorable cartoon rabbit baby rhea facing the camera, fantasy, dreamlike, surrealism, super cute, trending on artstationm volumetric light, cinematic, post processing, 8K.
Fig. 12, Row 6	magic caster.
Fig. 12, Row 7	Cadillac El Dorado de style Badass Steampunk dans une vieille rue pavée, trending on artstation, sharp focus, studio photo, intricate details, highly detailed, by greg rutkowski.
Fig. 12, Row 8	character sheet,The little girl riding an electric scooter bike, in a beautiful anime scene by Hayao Miyazaki: a snowy Tokyo city with massive Miyazaki clouds floating in the blue sky, enchanting snowscapes of the city with bright sunlight, Miyazaki's landscape imagery, Japanese art, 16:9.
Fig. 12, Row 9	3D digital illustration, Burger with wheels speeding on the race track, supercharged, detailed, hyperrealistic, 4K.
Fig. 12, Row 10	magic the gathering, anthro furry knight adventurer, showcase promo full art. Painted impressionist style.
Fig. 12, Row 11	hundreds of sith warships in space facing viewer, symmetrical, centered, front view, highly detailed, centered, digital painting, ultradetailed, artstation, digital painting, cgsociety, octane render, sharp focus, illustration, cinematic lighting, 8k hd hyper realistic, intricate, lifelike, golden hour, highly detailed, art by ralph mcquarrie, James Ferdinand Knab, William O'Keefe, Boris Vallejo, Peter Kemp, Joshy Lee, Otto Schmit, and Aja RICKO.
Fig. 12, Row 12	Lion dragon hybrid.


```

{
  "prompt": "a gopro snapshot of an anthropomorphic cat dressed as a
  firefighter putting out a building fire",
  "Graph": [
    {"cat": ["anthropomorphic"]},
    {"firefighter": []},
    {"fire": ["building"]} ]
}

{
  "prompt": "a white polar bear cub wearing sunglasses sits in a
  meadow with flowers.",
  "Graph": [
    {"bear": ["white", "cub"]},
    {"sunglasses": []},
    {"meadow": []},
    {"flowers": []} ]
}

{
  "prompt": "a photo of a frog holding an apple while smiling in the
  forest",
  "Graph": [
    {"frog": ["smiling"]},
    {"apple": []},
    {"forest": []} ]
}

{
  "prompt": "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.",
  "Graph": [
    {"portal": ["swirling", "multicolored", "emerges"]},
    {"ocean": ["depths", "coffee", "rich"]},
    {"waves": ["rippling", "gently", "outward"]},
    {"cup": ["coffee", "gateway"]} ]
}

```

Figure 13. Some examples of text semantic graph.


Human preference investigation

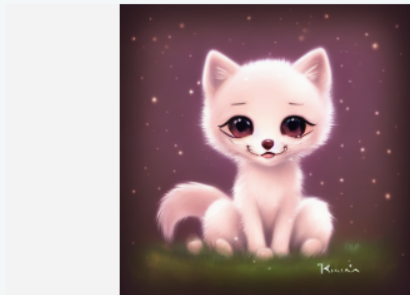
Given a description, please choose the best one according to the question

Test 1

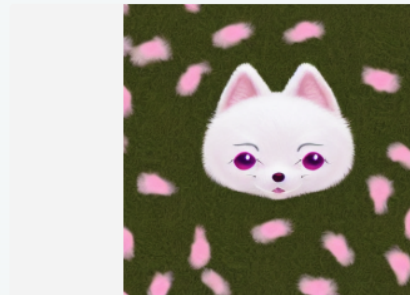
Given a description prompt, select the most appropriate option based on the question.

1. Description: 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

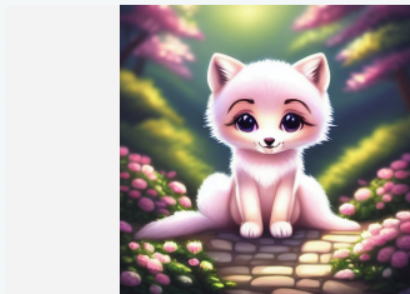
Which image do you prefer given the prompt? 



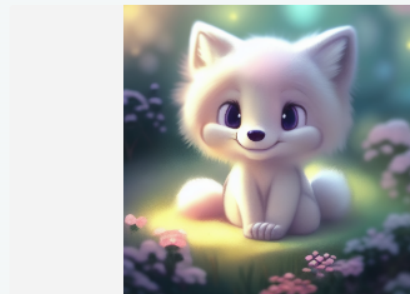
☐ Image 1



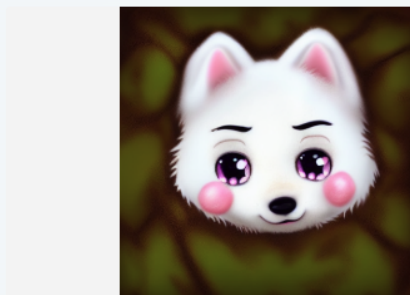
☐ Image 2



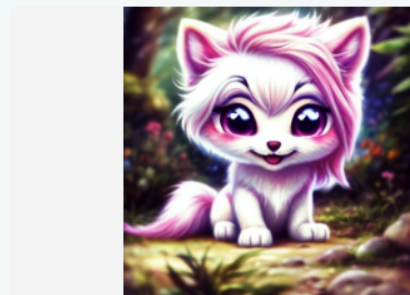
☐ Image 3



☐ Image 4




☐ Image 5

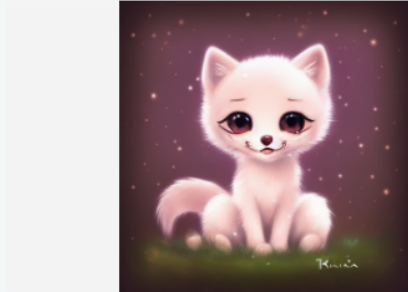


☐ Image 6

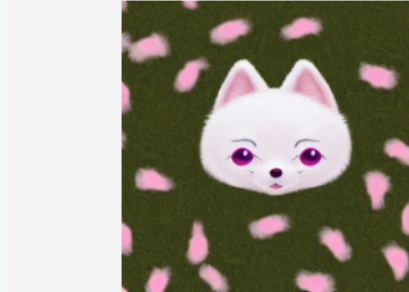
Figure 14. The screenshot of human preference investigation: Which image do you prefer given the prompt?

2. Description: 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

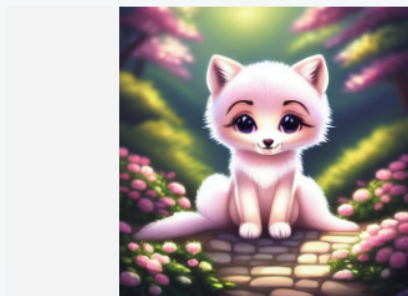
Which image is more visually appealing? 



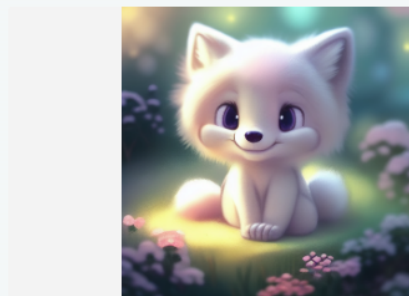
☐ Image 1



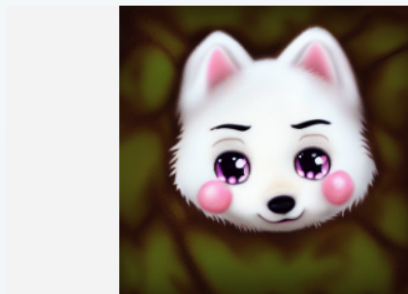
☐ Image 2



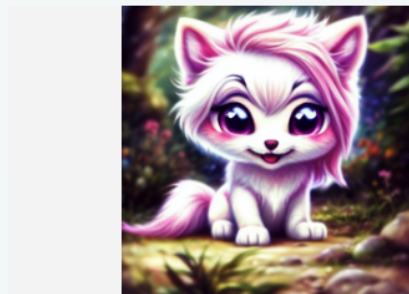
☐ Image 3



☐ Image 4




☐ Image 5

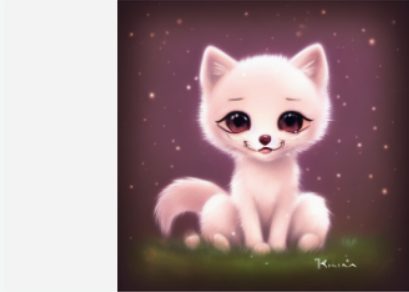


☐ Image 6

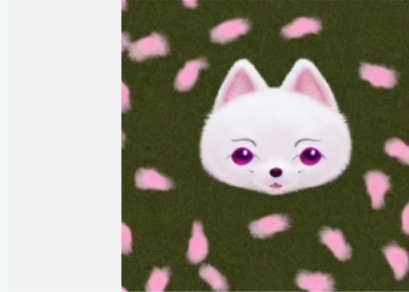
Figure 15. The screenshot of human preference investigation: Which image is more visually appealing?

3. Description: 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

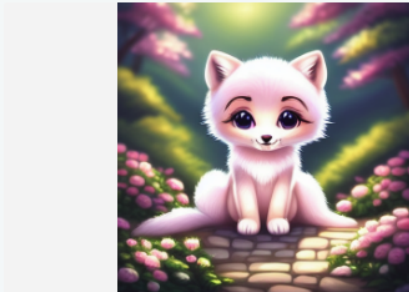
Which image better fits the text description? 



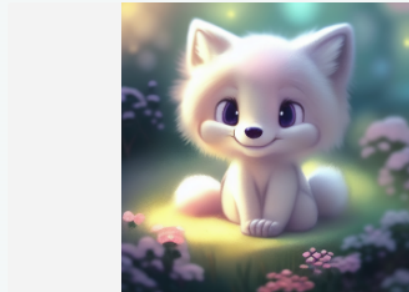
☐ Image 1



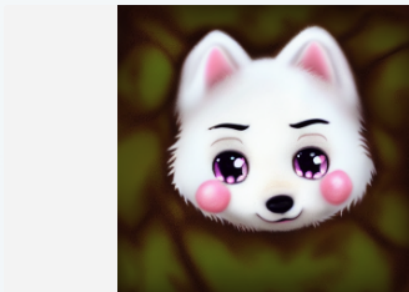
☐ Image 2



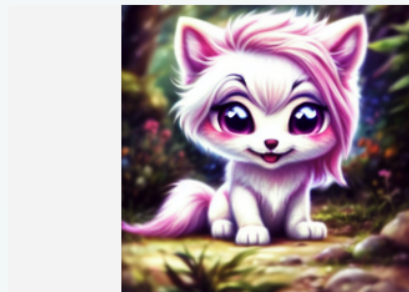
☐ Image 3



☐ Image 4



☐ Image 5



☐ Image 6

Figure 16. The screenshot of human preference investigation: Which image better fits the text description?