Style Aligned Image Generation via Shared Attention:
Supplementary Materials

Amir Hertz* [1], Andrey Voynov* [1], Shlomi Fruchter† [1], and Daniel Cohen-Or† [1,2]
[1] Google Research
[2] Tel Aviv University

Figure 1. **Various remarkable places depicted with the style taken from Bruegels' *"The Tower of Babel".***
Top row: Rome Colosseum, Rio de Janeiro, Seattle Space Needle.

A. Additional Reuslts

Figure 1 shows our techniques being applied for style transfer for the Peter Bruegels' "The Tower of Babel" to multiple places around the world. As for the prompt we always use the places' followed by *"Pieter Bruegel Painting"*, e.g."*Rome Coliseum, Pieter Bruegel Painting*". Even though the original masterpiece is known to model, it fails to reproduce its style with only text guidance. Fig. 2 shows some of the places generated with the direct instruction to resemble the original painting, without self-attention sharing. Notably, the model fails to produce an accurate style alignment with the original picture.

Further examples of style transferring from real examples are presented in Fig.s 10 and 11.

Attention scaling. We also noticed that once the style transfer is performed from an extremely famous image, the default approach may sometimes completely ignore the target prompt, generating an image almost identical to the reference. We suppose that this happens because the outputs of the denoising model for the famous reference image have very high confidence and activations magnitudes. Thus in

*Equal contribution.
†Equal Advising.

Figure 2. **Text-to-image generation with explicit style description.** *Unlike our approach, this fails to produce fine and style-aligned results. See Fig. 1 to inspect our method results.*

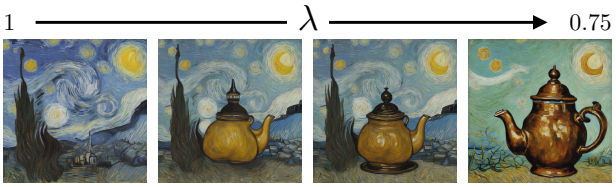

Figure 3. **Reference attention rescaling factor variation used for extremely popular reference image assets.**

the shared self-attention, most of the attention is taken by the reference keys. To compensate for it, we propose the simple trick of the attention scores rescaling. In the self-attention sharing mechanism, for some fixed scale $\lambda < 1$, we rescale the queries and keys products conducting the new scores $\lambda \cdot \langle Q, K_{\text{target}} \rangle$. We apply this only to the reference image keys. First, this suppresses extra-high keys. Also, this makes the attention scores more uniformly distributed, encouraging the generated image to capture style aggregated from the whole reference image. Fig. 3 demonstrates the rescaling factor variation effect for the particularly popular reference *"Starr Night"* by Van Gogh. Notably, without rescaling, the model generates an image almost identical to the reference, while the scale relaxation produces a plausible transfer.

Style prompt selection. For style transfer from an input image, we use an automated method to caption the input image. However, our method can be used with different style descriptions (different columns). As can be seen in Fig. 4, our method produces valid results for different prompts.

Shared self-attention visualization. Figure 5 depicts the self-attention probabilities from a generated target image to the reference style image. In each of the rows, we pick a point on the image and depict the associated probabilities map for the token at this particular point. Notably probabilities mapped on the reference image are semantically close to the query point location. This suggests that the

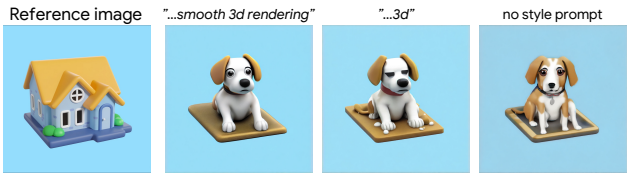

Figure 4. Style transfer results using StyleAligned with different style prompts (different columns) describing the reference image (left column).

self-attention tokens sharing do not perform a global style transfer, but rather match the styles in a semantically meaningful way [4]. In addition, Figure 6 visualizes the three largest components of the average shared attention maps of the rhino image, encoded in RGB channels. Note that the shared attention map is composed of both self-attention and cross-image attention to the giraffe. As can be seen, the components highlight semantically related regions like the bodies, heads, and the background in the images.

B. Integration with Other Methods

Below, we show different examples where our method can provide style aligned image generation capability on top of different diffusion-based image generation methods.

Style aligned subject driven generation. To use our method on top of a personalized diffusion model, first, given a collection of images (3-6) of the personalized content, we follow DreamBooth–LoRA training [8, 12] where the layers of the attention layers are fine-tuned via low-rank adaptation weights (LoRA). Then, during inference, we apply our method by sharing the attention of personalized generated images with a generated reference style image. During this process, the LoRA weights are used only for the generation of personalized content. The results of our method on top of different personalized models are shown in Fig. 9 where in each column we fine-tuned the SDXL model over the image collection on top and generated the personalized content with the reference images on the left. It can be seen that in some cases, like in the backpack photos on the right, the subject in the image remains in the same style as in the original photos. This is a known limitation of training-based personalization methods [13] which we believe can be improved by applying our method over other T2I personalization techniques [6, 9] or more careful search for training hyperparameters that allow better generalization of the personalized model to different styles.

Style aligned MultiDiffusion image generation. Bar et al. [5] presented MultiDiffusion, a method for generating images in any resolution by aggregating diffusion predictions of overlapped squared crops. Our method can be used on top of MultiDiffusion by enabling our shared attention between the crops to a reference image that is gen-

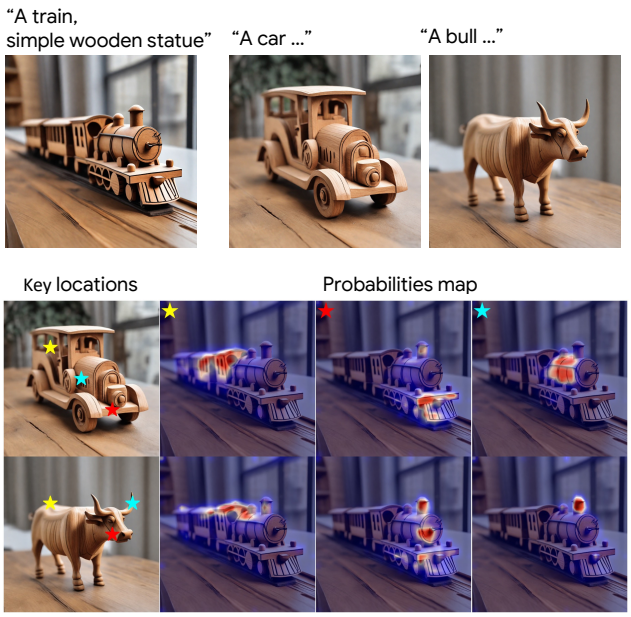

Figure 5. *Self-Attention probabilities maps from different generated image locations (**Key locations** column) to the reference train image with the target style (top-left).*

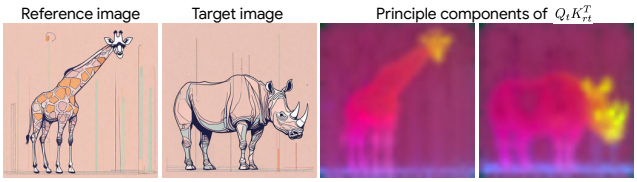

Figure 6. **Principle components of the shared attention map.** *On right, we visualize the principle components of the shared attention map between the reference giraffe and the target rhino generated images. The three largest components of the shared maps are encoded in RGB channels.*

erated in parallel. Fig. 12 shows style aligned panorama images generated with MultiDiffusion in conjunction with our method using the public implementation of MultiDiffusion over Stable Diffusion V2 [3]. Notice that compared to a *vanilla* MultiDiffusion image generation (small images in 12), our method not only enables the generation of style aligned panoramas but also helps to preserve the style within each image.

StyleAligned with additional conditions. Lastly, we show how our method can be combined with ContolNet [17] which enriches the conditioning signals of diffusion text-to-image generation to include additional inputs, like depth map and pose. ContolNet injects the additional information by predicting residual features that are added to the diffusion image features outputs of the down and middle U-Net blocks. Similar to previous modifications, we apply StyleAligned image generation by sharing the attention

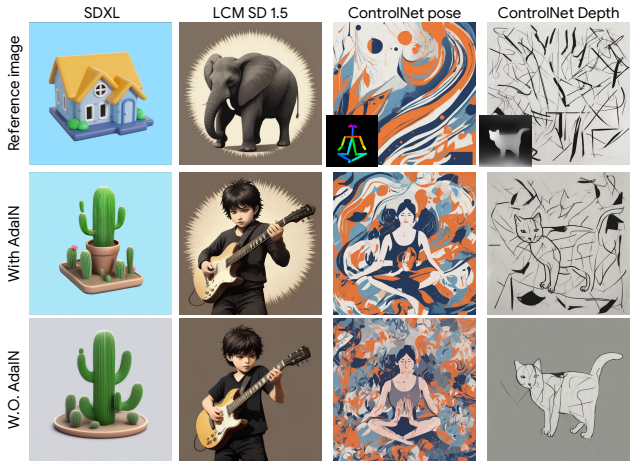

Figure 7. Comparison of StyleAligned applied with (second row) and without (third row) AdaIN applied on keys and queries. Columns represent StyleAligned applied for style transfer from reference image (*col. 1*), StyleAligned applied on Latent Consistency Model (*col. 2*), ControlNet (*col. 3-4*)

of ControlNet conditioned images to a reference image that isn't conditioned on additional input. Fig. 13 shows style aligned image set (different rows) that are conditioned on depth maps (different columns) using ControlNet depth encoder over SDXL [1]. Fig. 14 shows style aligned image set (different rows) that are conditioned on pose estimation obtained by OpenPose [14] (different columns) using ControlNet pose encoder over SDXL [2].

C. Additional Comparisons

We provide additional comparisons of our method to encoder-based text-to-image personalization methods and editing approaches over the evaluation set presented in Section 4 in the main paper. Table 1 summarized the full quantitative results presented in the paper and here.

Ablation comparisons Beyond the quantitative advantage of the use of Query-Key AdaIN, it also qualitatively improves style consistency and makes our method robust across various applications, as demonstrated in Fig. 7. Additionally, we evaluate its importance by performing a user study, similar to the one reported in main paper. The default approach is preferred over No-AdaIN in 75% of cases, and over Full attention sharing in 81% of cases.

Encoder based approaches As reported in the paper, we compare our method to encoder-based text-to-image personalization methods: BLIP-Diffusion [10], ELITE [15], and IP-Adapter [16]. These methods train an image encoder and fine-tune the T2I diffusion model to be conditioned on visual input. Fig. 15 shows a qualitative comparison on the same set shown in the paper (Fig. 7). As can be seen, our image sets are more consistent and aligned to the

Table 1. **Full quantitative comparison for style aligned image generation.** We evaluate the generated image sets in terms of of text alignment (CLIP score) and set consistency (DINO embedding similarity). $\pm X$ denotes the standard deviation of the score across 100 image set results.

Method	Text Alignment (CLIP ↑)	Set Consistency (DINO ↑)
StyleDrop (SDXL)	0.272 ± 0.04	0.529 ± 0.15
StyleDrop (unofficial MUSE)	0.271 ± 0.04	0.301 ± 0.14
DreamBooth-LoRA (SDXL)	0.276 ± 0.03	0.537 ± 0.17
IP-Adapter (SDXL)	0.281 ± 0.03	0.44 ± 0.13
ELITE (SD 1.4)	0.253 ± 0.03	0.481 ± 0.13
BLIP-Diffusion (SD 1.4)	0.245 ± 0.04	0.475 ± 0.12
Prompt-to-Prompt (SDXL)	0.283 ± 0.03	0.454 ± 0.18
SDEdit (SDXL)	0.274 ± 0.03	0.453 ± 0.16
StyleAligned (SDXL)	0.287 ± 0.03	0.51 ± 0.14
StyleAligned (W.O. AdaIN)	0.289 ± 0.03	0.428 ± 0.14
StyleAligned (Full Attn.)	0.28 ± 0.03	0.55 ± 0.15

reference. Notice that, currently, only IP-Adapter provides an encoder model for Stable Diffusion XL (SDXL). Nevertheless, BLIP-Diffusion and ELITE struggle to produce consistent image sets that match the text descriptions.

Zero–shot editing approaches Other baselines that can be used for style aligned image set generation are diffusion-based editing methods when applied over the reference images. However, unlike our method, these methods assume structure preservation of the input image. We report the results of two diffusion-based editing approaches: SDEdit [11] and Prompt-to-Prompt (P2P) [7] in Fig. 8. Notice that similar to our method, these methods provide a level of control that trade-off between alignment to text and alignment to the input image. To get higher text alignment, SDEdit can be applied over an increased percentage of diffusion steps, and P2P can reduce the number of attention injection steps. Our method can achieve higher text alignment, as described in Section 4 in the main paper, by using our shared attention over only a subset of self-attention layers. Fig. 8 presents the trade-off of the results over the different methods. As can be seen, only our method can achieve text alignment while preserving high set consistency.

D. User Study and Evaluation Settings

As described in the main paper, we generate the images for evaluation using a list of 100 text prompts where each prompt describes 4 objects in the same style. The full list is provided at the end of supplementary materials D. We evaluated the results of the different methods using the automatic CLIP and DINO scores and through user evaluation. The format of the user study is provided in Fig. 16 where the user has to select between the results of two methods. For each method from StyleDrop (SDXL), StyleDrop (unofficial Muse), and DreamBooth-LoRA (SDXL), we collected 800 answers compared to our results. In total, we collected 2400 answers from 100 participants.

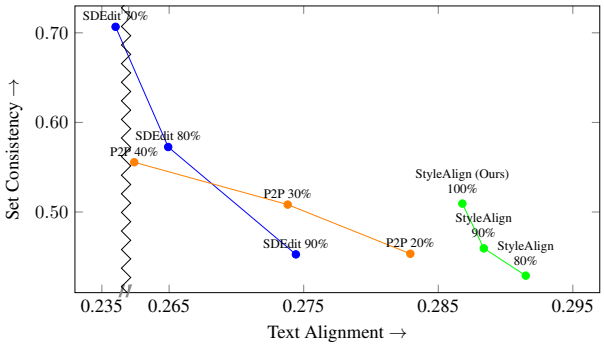

Figure 8. **Quantitative Comparison to zero shot editing approaces.** *We compare the results of the different methods in terms of text alignment (CLIP score) and set consistency (DINO embedding similarity).*

References

[1] Diffusers: controlnet-depth-sdxl-1.0. https://huggingface.co/diffusers/controlnet-depth-sdxl-1.0, 2023. 3

[2] Diffusers: controlnet-openpose-sdxl-1.0. https://huggingface.co/thibaud/controlnet-openpose-sdxl-1.0, 2023. 3

[3] Diffusers: Multidiffusion piepeline. https://huggingface.co/docs/diffusers/api/pipelines/panorama, 2023. 3

[4] Yuval Alaluf, Daniel Garibi, Or Patashnik, Hadar Averbuch-Elor, and Daniel Cohen-Or. Cross-image attention for zero-shot appearance transfer, 2023. 2

[5] Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. Multidiffusion: Fusing diffusion paths for controlled image generation. 2023. 2, 8

[6] Ligong Han, Yinxiao Li, Han Zhang, Peyman Milanfar, Dimitris N. Metaxas, and Feng Yang. Svdiff: Compact parameter space for diffusion fine-tuning. *ArXiv*, abs/2303.11305, 2023. 2

[7] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-or. Prompt-to-prompt image editing with cross-attention control. In *The Eleventh International Conference on Learning Representations*, 2022. 4

[8] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2021. 2

[9] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1931–1941, 2023. 2

[10] Dongxu Li, Junnan Li, and Steven C. H. Hoi. BLIP-Diffusion: Pre-trained subject representation for con-

Figure 9. **Personalized T2I diffusion with StyleAligned.** *Each row shows style aligned image st using the reference image on the left, applied on different personalized diffusion models, fine-tuned over the personalized content on top. The top two rows were generated using the prompt "[my subject] in the style of a beautiful papercut art." The bottom two rows were generated using the prompt "[my subject] in beautiful flat design." where [my subject] is replaced with the subject name.*

trollable text-to-image generation and editing. *ArXiv*, abs/2305.14720, 2023. 3

[11] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. In *ICLR*, 2021. 4

[12] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Wei Wei, Tingbo Hou, Yael Pritch, Neal Wadhwa, Michael Rubinstein, and Kfir Aberman. Hyperdreambooth: Hypernetworks for fast personalization of text-to-image models. *arXiv preprint arXiv:2307.06949*, 2023. 2

[13] Yoad Tewel, Rinon Gal, Gal Chechik, and Yuval Atzmon. Key-locked rank one editing for text-to-image personalization. *SIGGRAPH 2023 Conference Proceedings*, 2023. 2

[14] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *CVPR*, 2016. 3

[15] Yuxiang Wei. Official implementation of ELITE. https://github.com/csyxwei/ELITE, 2023. 3

[16] Hu Ye, Jun Zhang, Siyi Liu, Xiao Han, and Wei Yang. IP-Adapter: Text compatible image prompt adapter for text-to-image diffusion models. *ArXiv*, abs/2308.06721, 2023. 3

[17] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. 3

Figure 10. **Samples of the proposed style transfer techniques applied for a variety of different images and target prompts.**

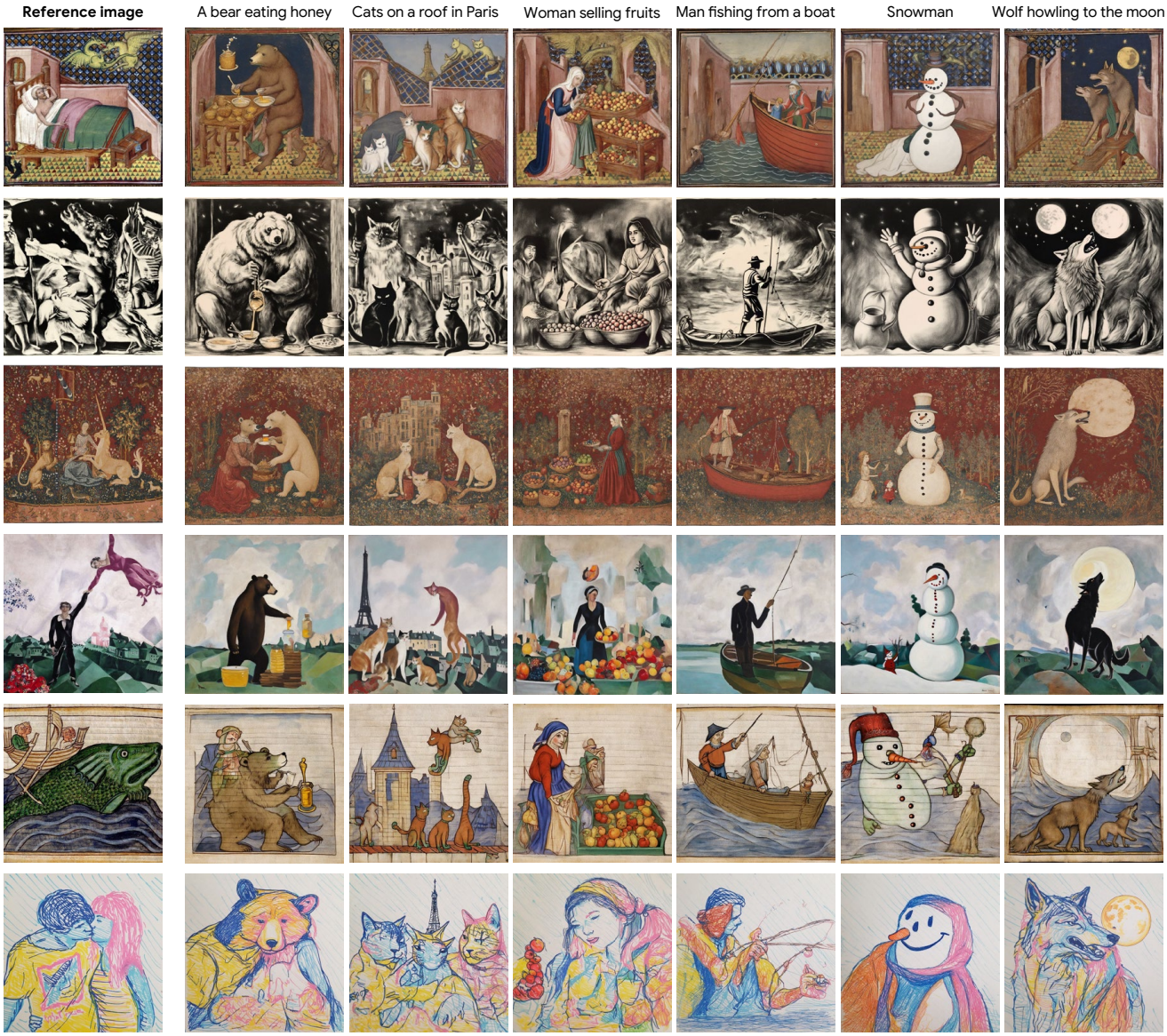

Figure 11. **Samples of the proposed style transfer techniques applied for a variety of different images and target prompts.**

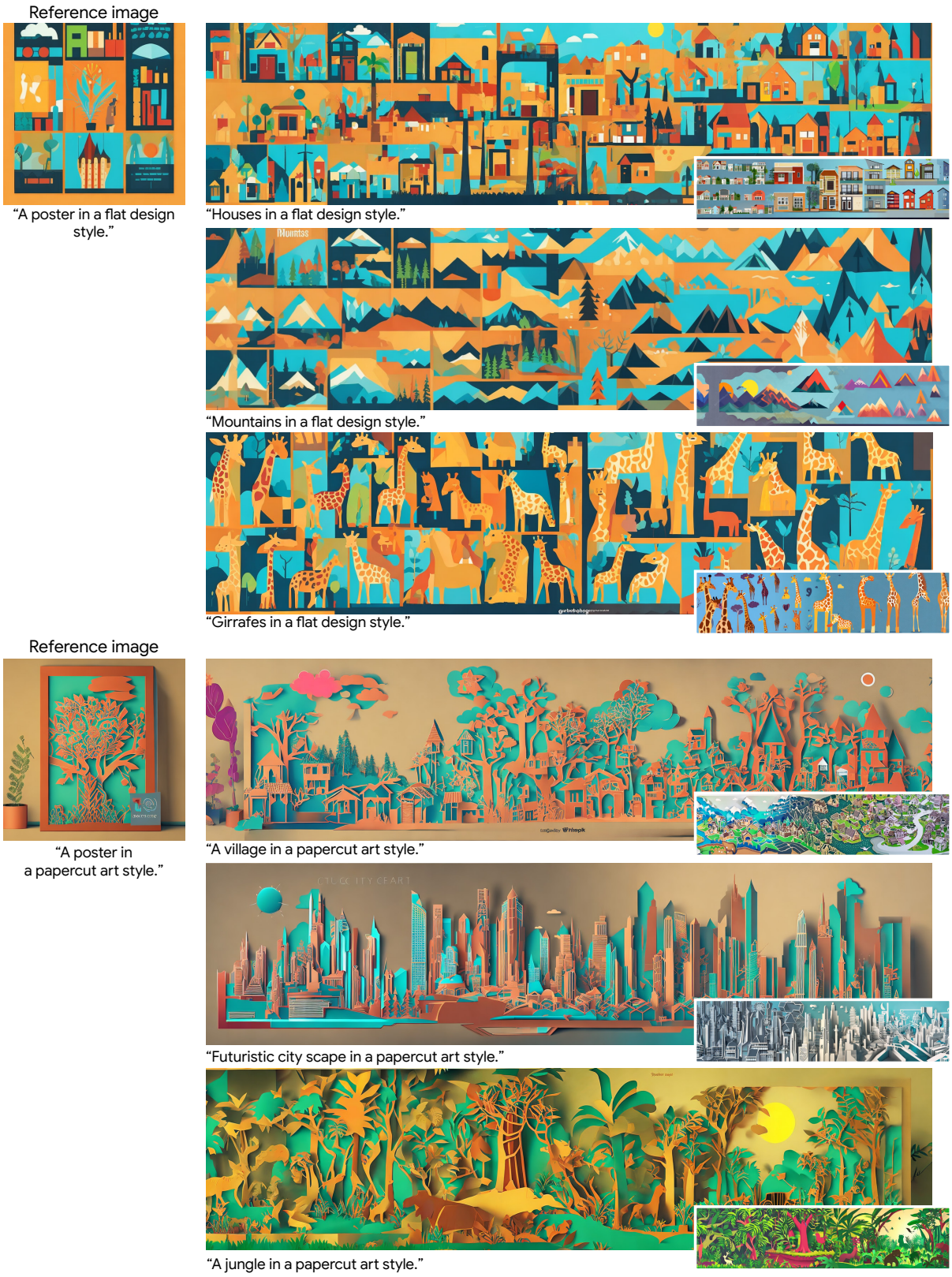

Figure 12. **MultiDiffusion with StyleAligned.** *The panoramas were generated with MultiDiffusion [5] using the text prompt beneath and the left image as reference. The small images in the bottom right corners are the results of MultiDiffusion results without our method.*

Figure 13. **ControlNet Depth with StyleAligned.**

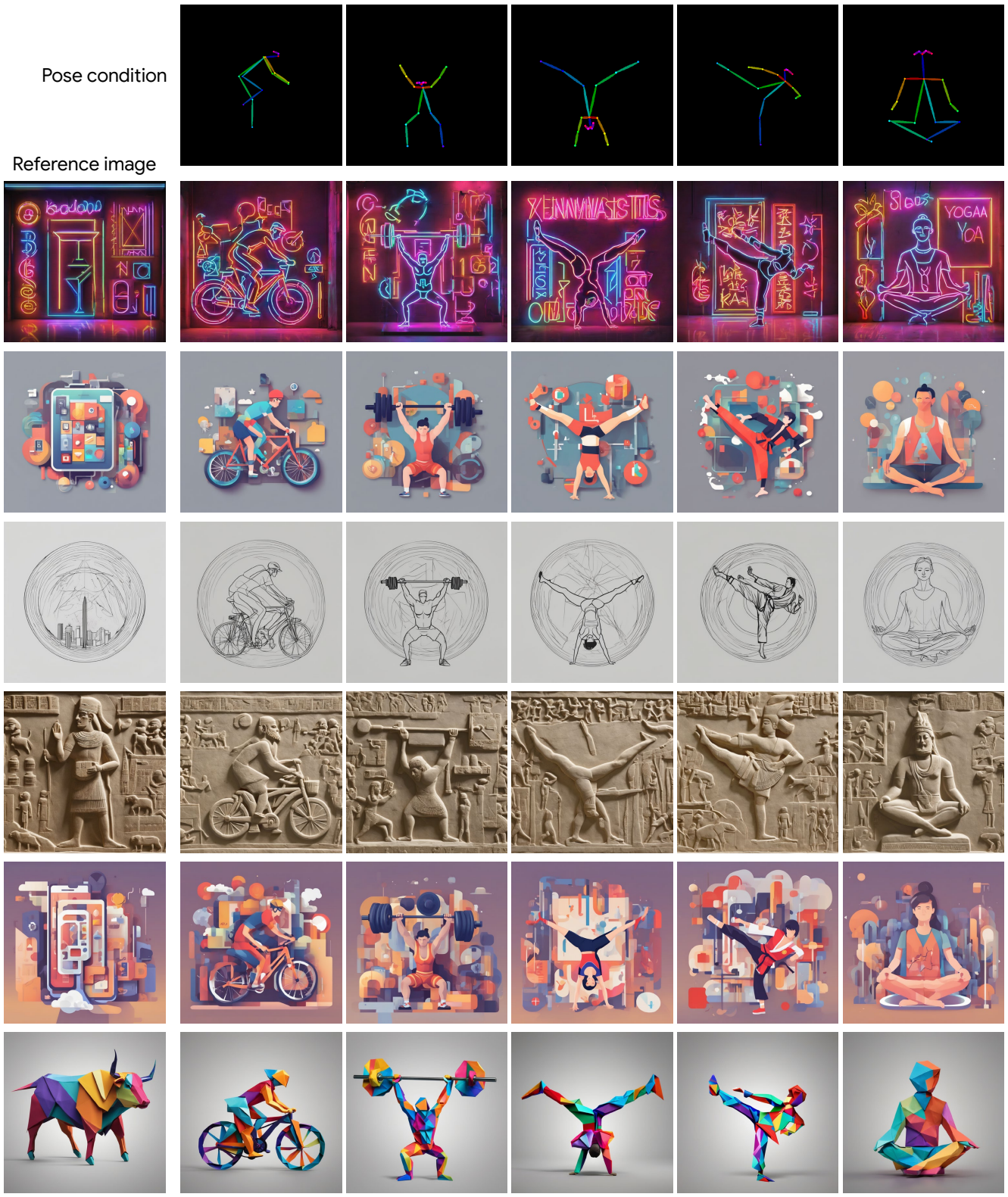

Figure 14. **ControlNet pose with StyleAligned.**

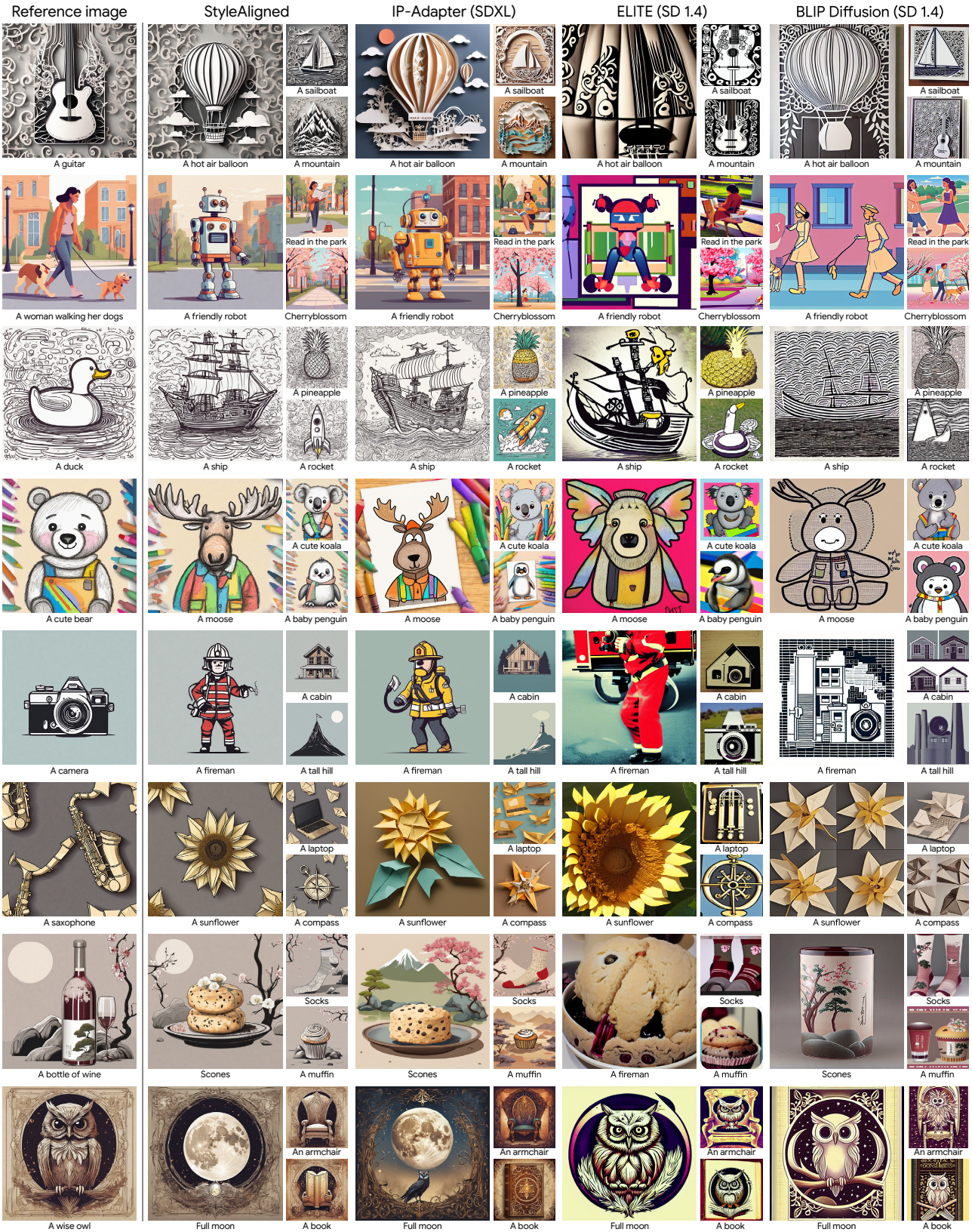

Figure 15. **Qualitative comparison to encoders based personalization methods.**

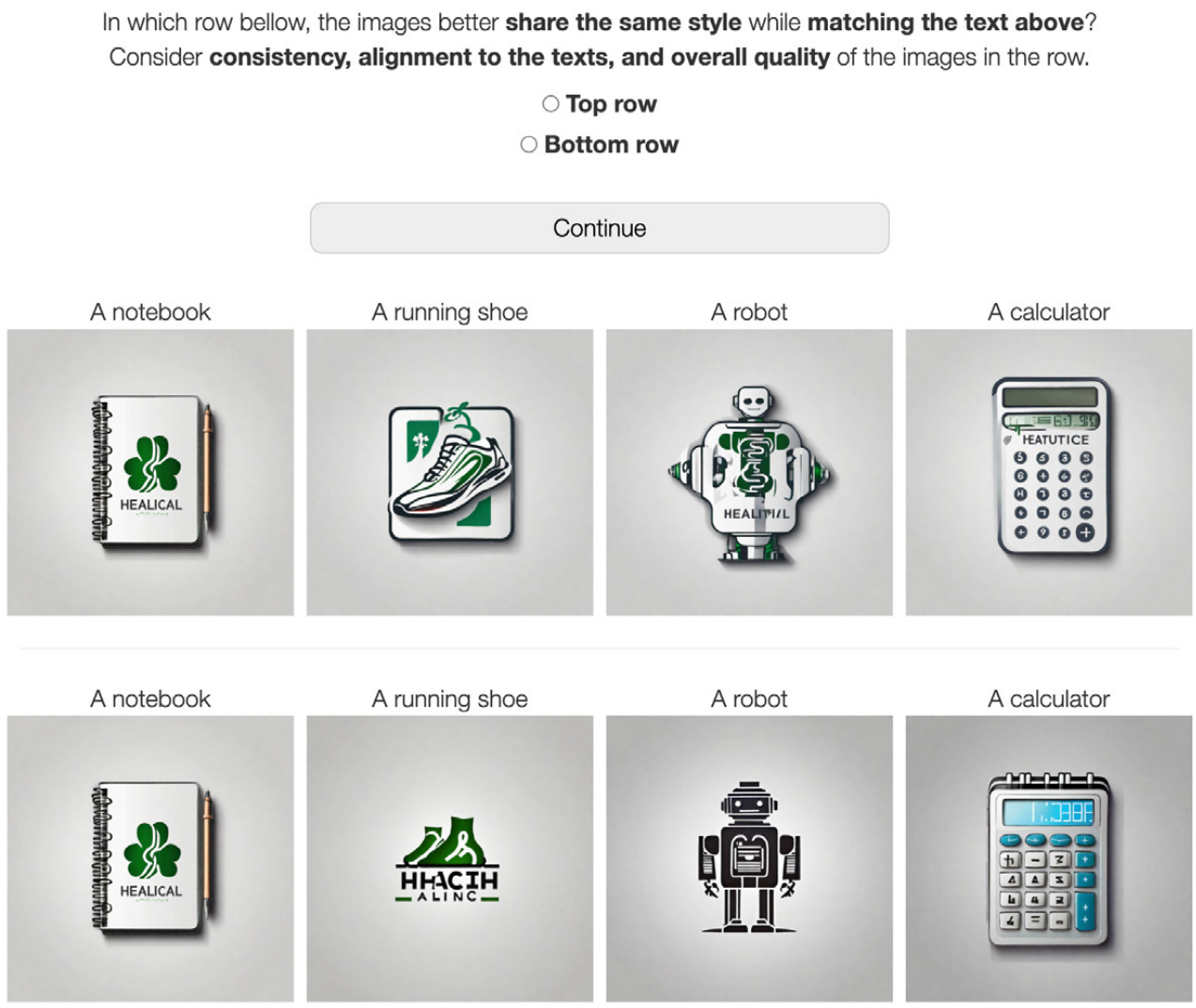

Figure 16. **Screenshot from the user study.** *Each row of images represents the result obtained by different method. The user had to assess which row is better in terms of style alignment and text alignment.*

List of prompts for our evaluation set generation:

1. {A house, A temple, A dog, A lion} in sticker style.
2. {Flowers, Golden Gate bridge, A chair, Trees, An airplane} in watercolor painting style.
3. {A village, A building, A child running in the park, A racing car} in line drawing style.
4. {A phone, A knight on a horse, A train passing a village, A tomato in a bowl} in cartoon line drawing style.
5. {Slices of watermelon and clouds in the background, A fox, A bowl with cornflakes, A model of a truck} in 3d rendering style.
6. {A mushroom, An Elf, A dragon, A dwarf} in glowing style.
7. {A thumbs up, A crown, An avocado, A big smiley face} in glowing 3d rendering style.
8. {A bear, A moose, A cute koala, A baby penguin} in kid crayon drawing style.
9. {An orchid, A Viking face with beard, A bird, An elephant} in wooden sculpture.
10. {A portrait of a person wearing a hat, A portrait of a woman with a long hair, A person dancing, A person fishing} in oil painting style.
11. {A woman walking a dog, A friendly robot, A woman reading in the park, Cherryblossom} in flat cartoon illustration style.
12. {A bithday cake, The letter A, An espresso machine, A Car} in abstract rainbow colored flowing smoke wave design.
13. {A flower, A piano, A butterfly, A guiter} in melting golden 3d rendering style.
14. {A train, A car, A bicycle, An airplane} in minimalist round BW logo.
15. {A rocket, An astronaut, A man riding a snowboard, A pair of rings} in neon graffiti style.
16. {A teapot, A teacup, A stack of books, A cozy armchair} in vintage poster style.
17. {A mountain range, A bear, A campfire, A pine forest} in woodblock print style.
18. {A surfboard, A beach shack, A wave, A seagull} in retro surf art style.
19. {A paintbrush, A sunflower field, A scarecrow, A rustic barn} in a minimal origami style.
20. {A cityscape, Hovering vehicles, Dragons, Boats} in cyberpunk art style.
21. {A treasure box, A pirate ship, A parrot, A skull} in tattoo art style.
22. {Music stand, A vintage microphone, A turtle, A saxophone} in art deco style.
23. {A tropical island, A mushroom, A palm tree, A cocktail} in vintage travel poster style.
24. {A carousel, Cotton candy, A ferris wheel, Balloons} in retro amusement park style.
25. {A serene river, A rowboat, A bridge, A willow tree} in 3D render, animation studio style.
26. {A retro guitar, A jukebox, A chess piece, A milkshake} in 1950s diner art style.
27. {A snowy cabin, A sleigh, A snowman, A winter forest} in Scandinavian folk art style.
28. {A bowl with apples, A pencil, A big armor, A magical sunglasses} in fantasy poison book style.
29. {A kiwi fruit, A set of d0rums, A hammer, A tree} in Hawaiian sunset painting style.
30. {A guitar, A hot air balloon, A sailboat, A mountain} in papercut art style.
31. {A coffee cup, A typewriter, A pair of glasses, A vintage camera} in retro hipster style.
32. {A board of backgammon, A shirt and pants, Shoes, A cocktail} in vintage postcard style.
33. {A roaring lion, A soaring eagle, A dolphin, A galloping horse} in tribal tattoo style.
34. {A pizza, Candles and roses, A bottle, A chef} in Japanese ukiyo-e style.
35. {A wise owl, A full moon, A magical chair, A book of spells} in fantasy book cover style.
36. {A cozy cabin, Snow-covered trees, A warming fireplace, A steaming cup of cocoa} in hygge style.
37. {A bottle of wine, A scone, A muffin, Pair of socks} in Zen garden style.
38. {A diver, Bowl of frutis, An astronaut, A carousel} in celestial artwork style.
39. {A horse,A castle, A cow, An old phone} in medieval fantasy illustration style.
40. {A mysterious forest, Bioluminescent plants, A graveyard, A train station} in enchanted 3D rendering style.
41. {A globe, An airplane, A suitcase, A compass} in travel agency logo style.
42. {A persian cat playing with a ball of wool, A man skiing down the hill, A train at the station, A bear eating honey} in cafe logo style.
43. {A book, A quill pen, An inkwell, An umbrella} in educational institution logo style.
44. {A hat, A strawberry, A screw, A giraffe} in mechanical repair shop logo style.
45. {A notebook, A running shoe, A robot, A calculator} in healthcare and medical clinic logo style.
46. {A rubber duck, A pirate ship, A rocket, A pineapple} in doodle art style.
47. {A trumpet, A fishbowl, A palm tree, A bicycle} in abstract geometric style.
48. {A teapot, A kangaroo, A skyscraper, A lighthouse} in mosaic art style.
49. {A ninja, A hot air balloon, A submarine, A watermelon} in paper collage style.
50. {A saxophone, A sunflower, A compass, A laptop} in origami style.
51. {A penguin, A bicycle, A tornado, A pineapple} in abstract graffiti style.
52. {A magician's hat, A UFO, A roller coaster, A beach ball} in street art style.
53. {A cactus, A shopping cart, A child playing with cubes, A camera} in mixed media art style.
54. {A snowman, A surfboard, A helicopter, A cappuccino} in abstract expressionism style.
55. {A robot, A cupcake, A woman playing bascketball, A sunflower} in digital glitch art style.
56. {A treehouse, A disco ball, A sailing boat, A cocktail} in psychedelic art style.
57. {A football helmet, A playmobil, A truck, A watch} in street art graffiti style.
58. {A cabin, A leopard, A squirrel, A rose} in pop art style.
59. {A bus, A drum, A rabbit, A shopping mall} in minimalist surrealism style.
60. {A frisbee, A monkey, A snake, skates} in abstract cubism style.
61. {A piano, A villa, A snowboard, A rubber duck} in abstract impressionism style.
62. {A laptop, A man playing soccer, A woman playing tennis, A rolling chair} in post-modern art style.
63. {A cute puppet, A glass of beer, A violin, A child playing with a kite} in neo-futurism style.
64. {A dog, A brick house, A lollipop, A woman playing on a guiter} in abstract constructivism style.
65. {A kite surfing, A pizza, A child doing homework, A person doing yoga} in fluid art style.
66. {Ice cream, A vintage typewriter, A pair of reading glasses, A handwritten letter} in macro photography style.
67. {A gourmet burger, A sushi, A milkshake, A pizza} in professional food photography style for a menu.
68. {A crystal vase, A pocket watch, A compass, A leather-bound journal} in vintage still life photography style.
69. {A sake set, A stack of books, A cozy blanket, A cup of hot cocoa} in miniature model style.
70. {A retro bicycle, A sunhat, A picnic basket, A kite} in outdoor lifestyle photography style.
71. {A group of hikers on a mountain trail, A winter evening by the fire, A hen, A person enjoying music} in realistic 3D render.
72. {A tent, A person knitting, A rural farm scene, A basket of fresh eggs} in retro music and vinyl photography style.

73. {A giraffe, A blanket, A fork and knife, A pile of candies} in cozy winter lifestyle photography style.
74. {A wildflower, A ladybug, An igloo in antarctica, A person running} in bokeh photography style.
75. {A coffee machine, A laptop, A person working, A plant on the desk} in minimal flat design style.
76. {A camera, A fireman, A wooden house, A tall hill} in minimal vector art style.
77. {A person texting, A person scrawling, A cozy chair, A lamp} in minimal pastel colors style.
78. {A smartphone, A book, A dinner table, A glass of wine} in minimal digital art style.
79. {A brush, An artist painting, A girl holding umbrella, a pool table} in minimal abstract illustration style.
80. {A pair of running shoes, A motorcycle, Keys, A fitness machine} in minimal monochromatic style.
81. {A compass rose, A cactus, A zebra, A blizzard} in woodcut print style.
82. {A lantern, A tricycle, A seashell, A swan} in chalk art style.
83. {Magnifying glass, Gorilla, Airplane, Swing} in pixel art style.
84. {Hiking boots, Kangaroo, Ice cream cone, Hammock} in comic book style.
85. {Horseshoe, Vintage typewriter, Snail, Tornado} in vector illustration style.
86. {A lighthouse, A hot air balloon, A cat, A cityscape} in isometric illustration style.
87. {A compass, A violin, A palm tree, A koala} in wireframe 3D style.
88. {Beach umbrella, Rocket ship, Fox, Waterfall} in paper cutout style.
89. {Tree stump, Harp, Chameleon, Canyon} in blueprint style.
90. {Elephant, UFO toy, Flamingo, Lightning bolt} in retro comic book style.
91. {Robot, Temple, Jellyfish, Sofa} in infographic style.
92. {Microscope, Giraffe, Laptop, Rainbow} in geometric shapes style.
93. {Teapot, Dragon toy, Skateboard, Storm cloud} in cartoon line drawing style.
94. {Crystal ball, Carousel horse, Hummingbird, Glacier} in watercolor and ink wash style.
95. {Feather quill, Satellite dish, Deer, Desert scene} in dreamy surreal style.
96. {Map, Saxophone, Mushroom, Dolphin} in steampunk mechanical style.
97. {Anchor, Clock, Globe, Bicycle} in 3D realism style.
98. {Clock, Helicopter, Whale, Starfish} in retro poster style.
99. {Binoculars, Bus, Pillow, Cloud} in bohemian hand-drawn style.
100. {Rhino, Telescope, Stool, Panda} in vintage stamp style.