Wasserstein Style Distribution Analysis and Transform for Stylized Image Generation (Appendix)

1. More Qualitative Results

In this section, we provide more examples of our WSDT method. In Figures 1 and 2, we show the generated results of WSDT by using SDXL and SD3 base models, respectively . It can be seen that WSDT consistently generates images that closely match the reference style while faithfully adhering to the text prompts, as evident in the richness of details and the stylistic alignment.

2. More Quantitative Results

To provide a more comprehensive evaluation, we compare our method with additional image translation methods. Specifically, we compare our method with AdaAttN [2], StyleFormer [5], StyTr^2 [1], and PnP [4]. We reproduce the results of comparison methods using their official implementations. As shown in Figure 3, the images generated by WSDT exhibit the most consistent style with the style reference images compared to the baselines. Table 1 reports the quantitative results on the evaluation dataset. Our method outperforms all competing approaches in terms of content score, style score, and GM metric, demonstrating its superior capability in stylized image generation.

3. More Ablation Results.

We show more ablation results in Table 2 about the diffusion steps and the classifier-free guidance scale. As shown in the left half of Table 2, when the diffusion step varies from 30 to 70, the content and style scores remain relatively stable across different timesteps. We set 50 diffusion steps as the default configuration. We also investigate the impact of the classifier-free guidance scale, ranging from 2.5 to 12.5. As shown in the right half of Table 2, increasing the guidance scale improves content alignment, as reflected by the rising content score. However, this comes at the cost of degraded style consistency, with the style score showing clear deterioration when the scale exceeds 7.5. We adopt a guidance scale of 7.5 as the default, which provides a good balance between style consistency and content alignment.

4. More Applications

Controllable image generation. Our method can be seamlessly integrated into other diffusion model-based methods. As shown in Figure 4, by combining with other methods, our approach can be extended to more applications. Combined with DreamBooth [3], we can stylize the customized objects. By combining with ControlNet [6], we can perform controllable generation, such as generating stylized images that meet the conditions of depth maps and canny maps.

Style mixing. Our method can use images with multiple styles as references to generate new images that mix these styles. Since our approach achieves stylized image generation by transforming the distribution of the generated image to match the style image distribution, style mixing is achieved by transforming the generated image's distribution into a mixed distribution of multiple style images. Figure 5 shows the results of mixing two styles. We transform the distribution into the mixed distribution of two styles during image generation to generate images with mixed styles. When the mixing coefficient is 0 or 1, the generated image's style matches that of style image 1 or style image 2, respectively. As the mixing coefficient changes from 0 to 1, the style of style image 2 gradually blends in, and the generated image progressively shifts toward style image 2.

5. Computational Cost Comparision

We evaluate the computational cost of $\mathcal{Z}^*$ on an NVIDIA RTX A6000-48G and test the remaining methods, including our WSDT, on a single NVIDIA GeForce RTX 4090-24G. Table 3 shows the computation time and memory costs for different methods. As shown in the table, many methods for stylized image generation rely on either pretraining or test-time optimization. Our method can achieve competitive performance without the need for additional costly processes like pretraining or test-time optimization, enabling efficient stylized image generation.

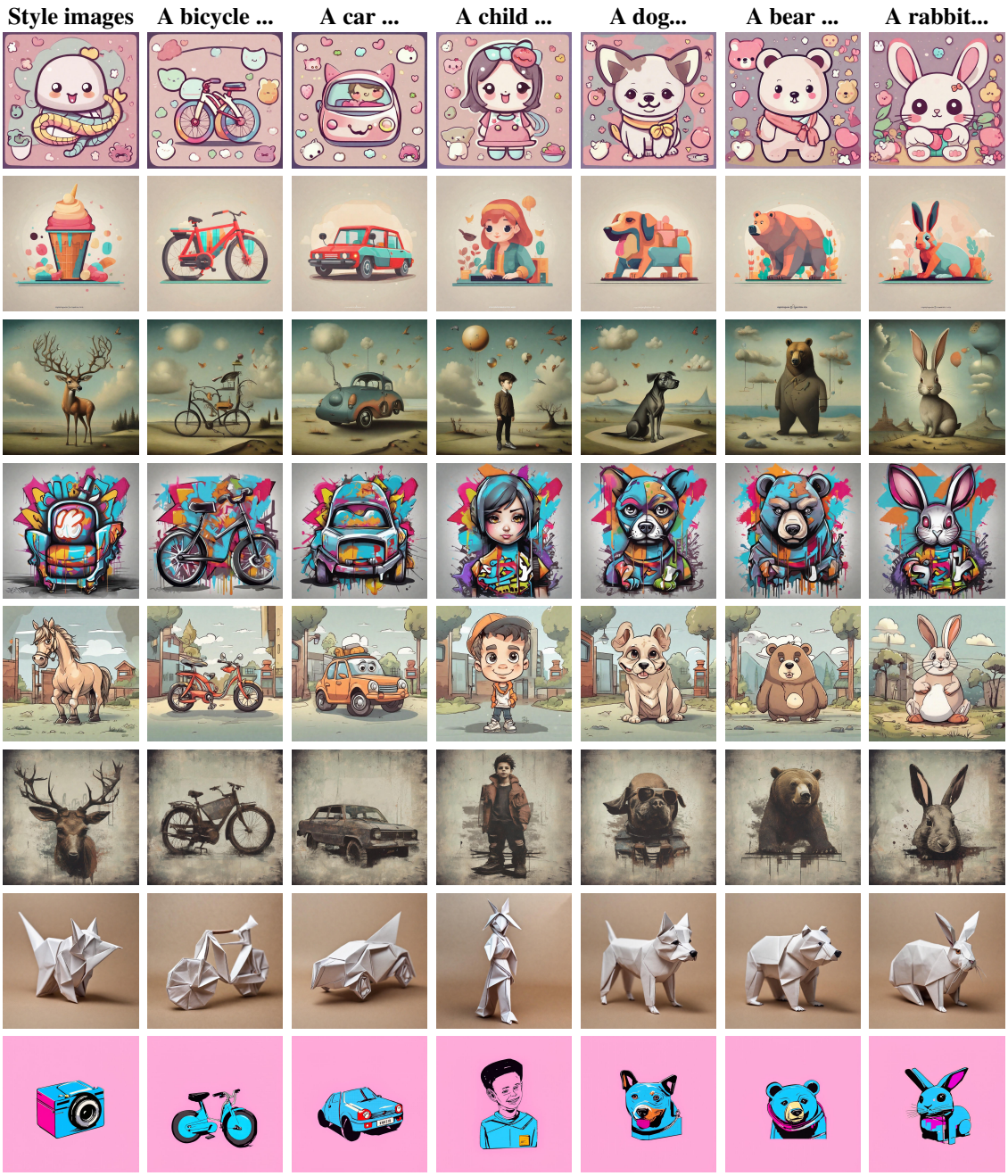

Figure 1. More examples generated by our WSDT method using SDXL base model.

Table 1. Quantitative comparison results on evaluation datasets.

Method	AdaAttN [2]	StyleFormer [5]	StyTr^2 [1]	PnP [4]	WSDT
Content Score ↑	0.266	0.257	0.261	0.257	**0.285**
Style Score ↑	0.416	0.384	0.411	0.459	**0.546**
GM Metric↓	9.95	29.28	19.55	20.00	**5.52**

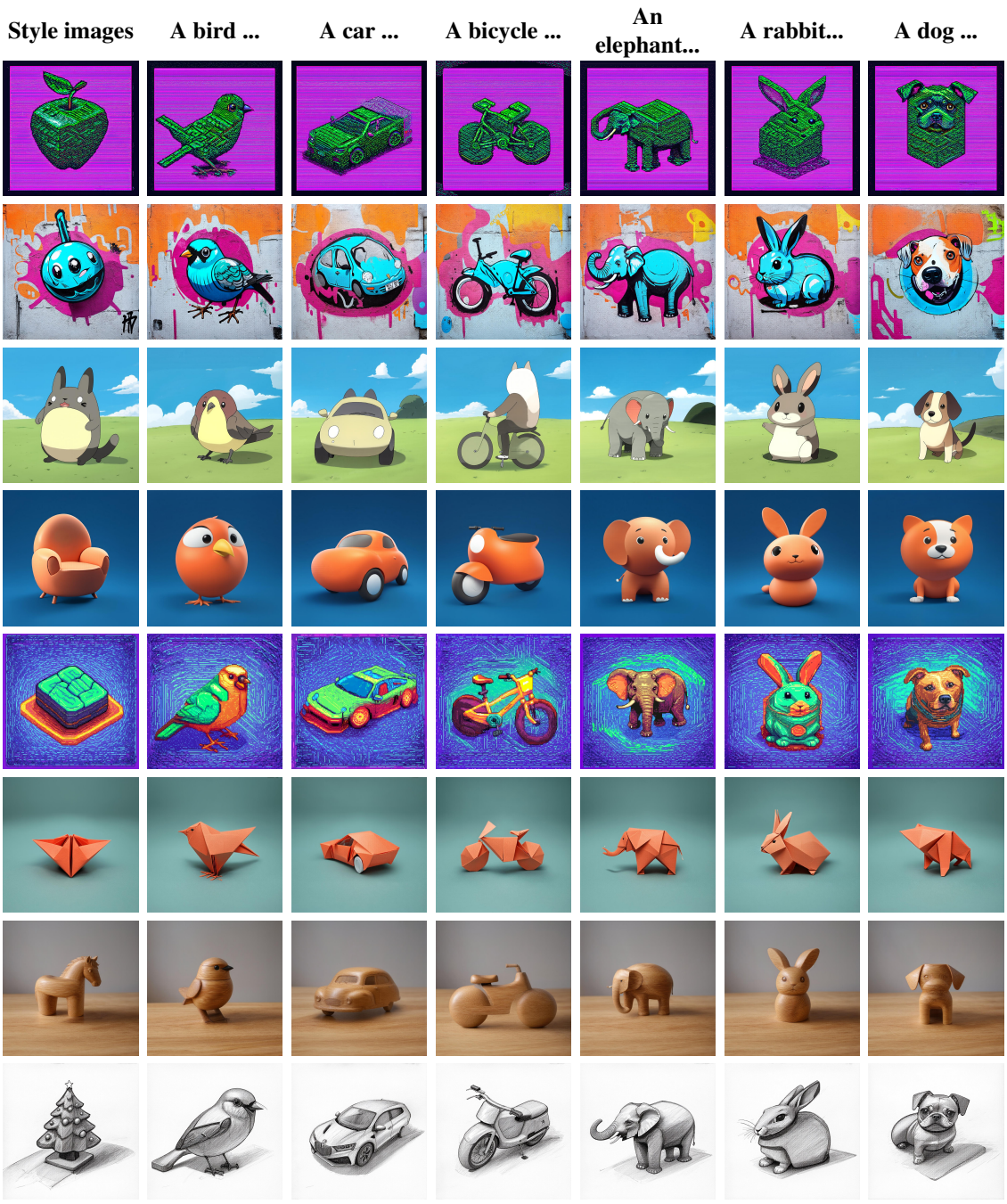

Figure 2. More examples generated by our WSDT method using SD3 base model.

Table 2. More ablation results on evaluation datasets.

Metric	Diffusion steps					Classifier-free guidance scale				
	30	40	50	60	70	2.5	5.0	7.5	10.0	12.5
Content Score ↑	0.284	0.285	0.285	0.284	0.285	0.260	0.278	0.285	0.287	0.288
Style Score ↑	0.543	0.545	0.546	0.547	0.546	0.560	0.549	0.546	0.538	0.534

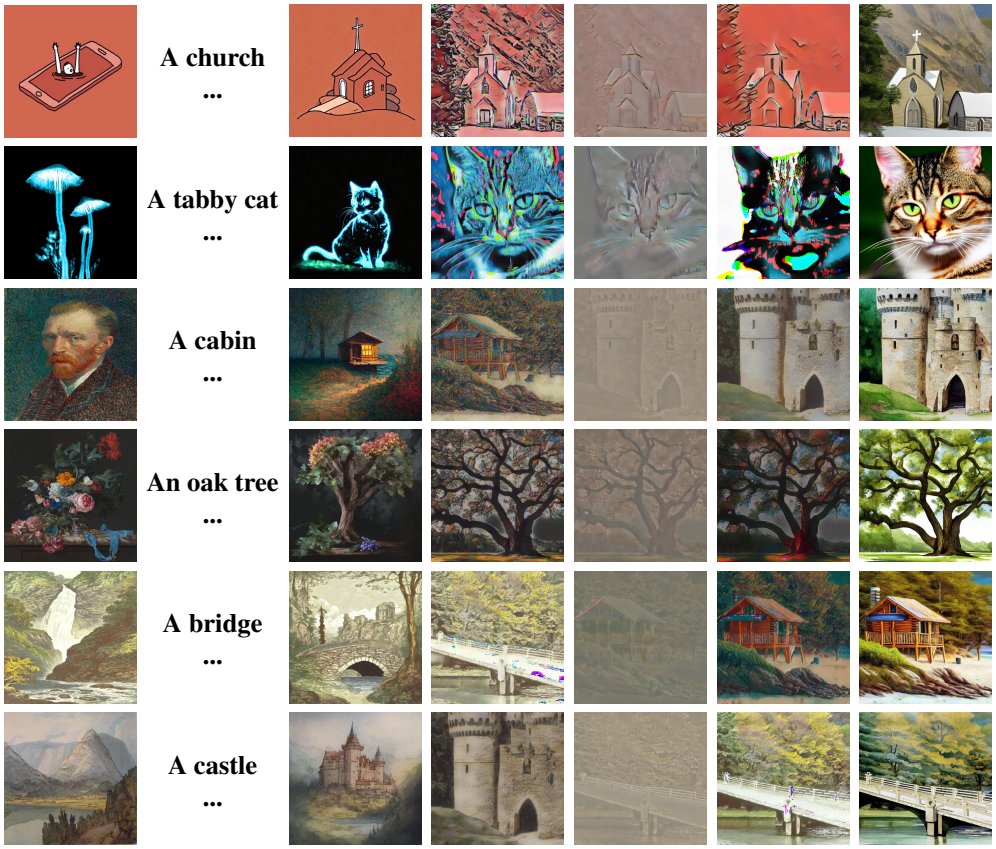

Figure 3. Qualitative comparison results on evaluation datasets.

Table 3. Computational time and memory cost of different methods.

Method	VSP	StyleDrop	StyleAligned	DEADiff	InstaStyle	CSGO	WSDT
Pre-training	✗	✗	✗	✓	✗	✓	✗
Test time optimization	✗	✓	✗	✗	✓	✗	✗
Time per iteration (s) ↓	21.8	372.7	40.0	2.1	400.4	10.2	12.4
Memory(GB) ↓	12.2	21.1	21.1	11.2	9.7	20.6	12.2

References

[1] Yingying Deng, Fan Tang, Weiming Dong, Chongyang Ma, Xingjia Pan, Lei Wang, and Changsheng Xu. Stytr2: Image style transfer with transformers. In *CVPR*, 2022. 1, 2

[2] Songhua Liu, Tianwei Lin, Dongliang He, Fu Li, Meiling Wang, Xin Li, Zhengxing Sun, Qian Li, and Errui Ding. Adaattn: Revisit attention mechanism in arbitrary neural style transfer. In *ICCV*, 2021. 1, 2

[3] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*, 2023. 1, 5

[4] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *CVPR*, 2023. 1, 2

[5] Xiaolei Wu, Zhihao Hu, Lu Sheng, and Dong Xu. Styleformer: Real-time arbitrary style transfer via parametric style composition. In *ICCV*, 2021. 1, 2

[6] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. 1, 5

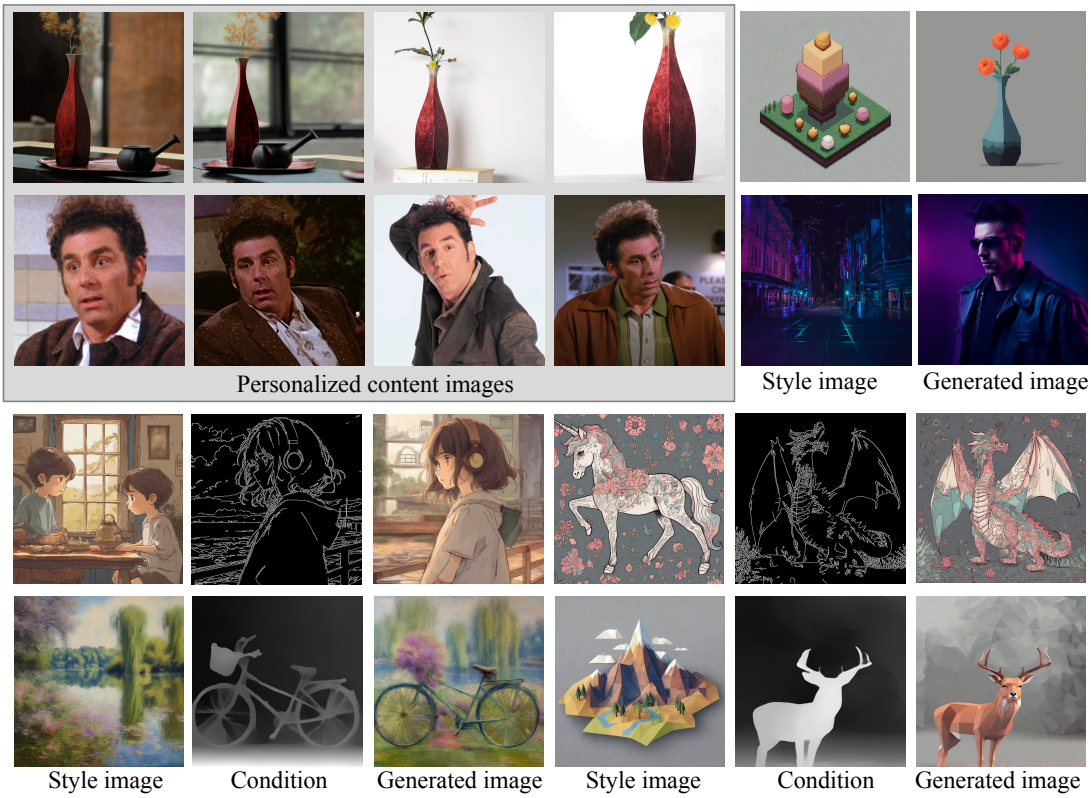

Figure 4. Results obtained by combining WSDT with different methods. The first two rows and the last two rows present the results when combined with DreamBooth [3] and ControlNet [6], respectively. Given personalized content images and style images, WSDT can generate the stylized personalized content. Given style images and control conditions, such as depth or canny maps, WSDT can generate stylized images under the control conditions.

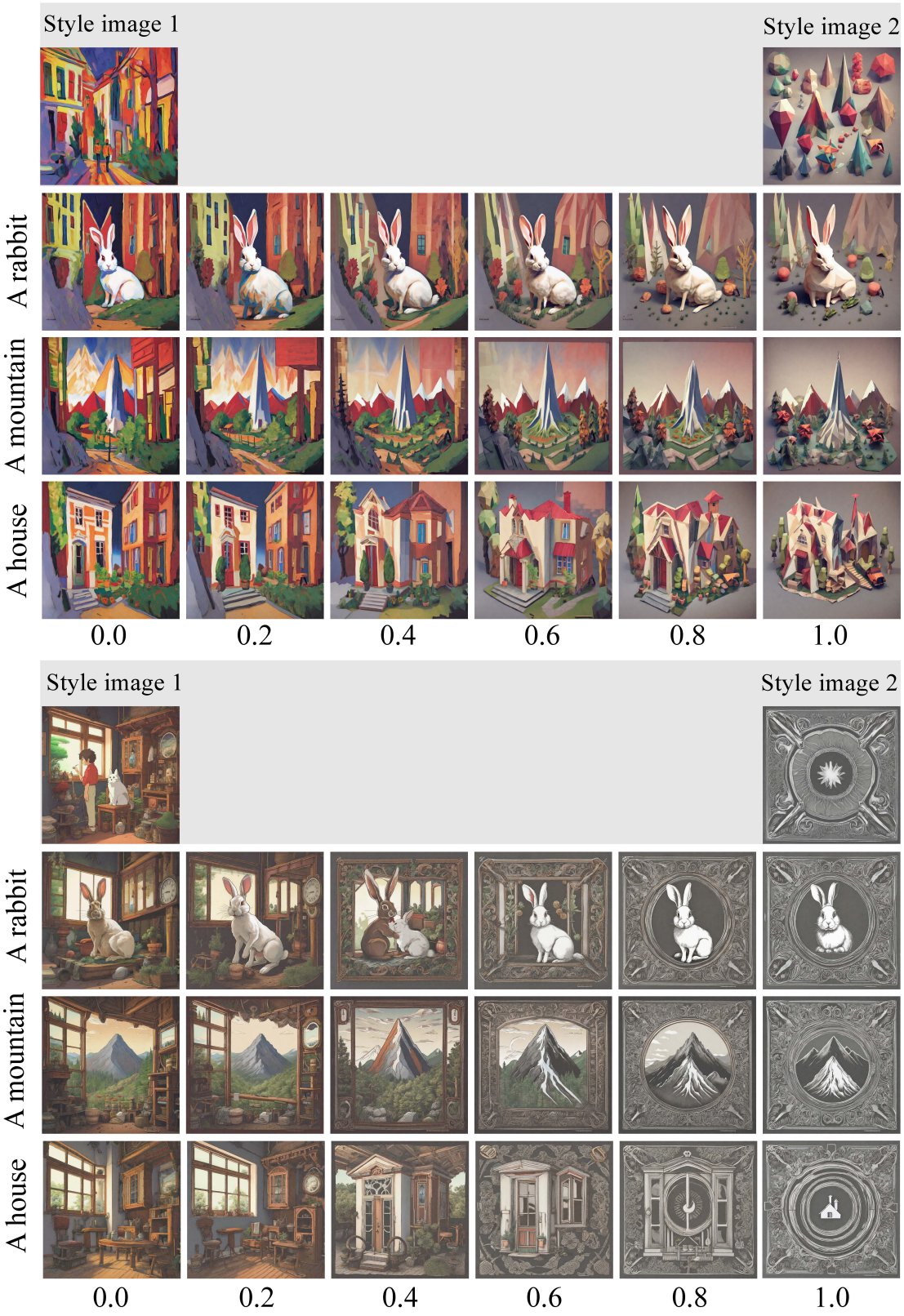

Figure 5. The first/fifth row shows two different style reference images (left: Fauvism/Studio Ghibli style, right: Low Poly/Engraving style). Rows 2–4 and 6–8 are examples of mixed styles, with text prompts 'A rabbit', 'A mountain' and 'A house' respectively. The numbers below rows 4 and 8 indicate the coefficient of style 2 in the mixed styles.