Supplementary Material

The supplementary material is structured as follows:

1. We first evaluated our results on extensive datasets and community LoRAs on different models to validate the effectiveness of our approach in section A.
2. We compared our method with the other methods in section B.
3. We assessed the influence of complex prompts on the model's performance in section C.
4. We experimented with a new scale and tested its comparative effects in section D.
5. We utilized Community LoRA in combination with local LoRA to conduct integrated performance evaluations and examined random seeds on model performance through comprehensive testing in section E.
6. We tested the choice of different parameters in scale factors in section F.

A. Visual Results

We employ datasets from StyleDrop [4] and DreamBooth [2] with Stable Diffusion (SD), as depicted in Fig. 4 and Fig. 5, we also evaluated our method on FLUX using LoRAs from Hugging Face, as shown in Fig. 2 and Fig. 3. By systematically combining these object and style LoRAs, we obtained a sequence of images that demonstrates the effectiveness of our approach in seamlessly integrating both object and style, yielding consistent and high-quality visual outputs.

B. Additional Comparisons

We have added a comparison with StyleID [1], as shown in Fig. 6. It can be observed that StyleID [1] effectively achieves style transfer while preserving texture quality. However, the generated objects might be slightly blurred or the style generated may not be distinct. Additionally, compared to our method, their approach is based on the fixed layout of original image, which may not generalize well to backgrounds and actions.

C. Prompt Control

We conduct experiments to evaluate whether our method can modify the object's actions, the surrounding environment, or introduce new elements through prompt adjustments. As illustrated in Fig. 9 and Fig. 10, after modifying the prompts, our method effectively retains the original object's features and stylistic attributes, while also integrating new elements or scene details seamlessly.

D. New Scale

In the main text of our paper, we employ the scale as follows:

$$S = \alpha \cdot \frac{t_{now}}{t_{all}} + \beta. \tag{1}$$

Inspired by [5], we also introduce an alternative scale factor:

$$S^{*} = \left(\alpha' \cdot \frac{t_{now}}{t_{all}} + \beta' \right) \% \, \alpha. \tag{2}$$

In this equation, we set $\alpha' = 1.5$ and $\beta' = 1.3$, which means that the style information is enhanced to some extent at the beginning of the generation process, allowing the model to capture certain block information from the style LoRA. Fig. 1 below illustrates the primary differences between the two scales.

For S^{*} results, since the style information is enhanced during the early diffusion steps, the generated images capture the background and color block information from the style LoRA. However, this approach results in a weakened learning effect for the texture and brushstrokes information in the style LoRA. This represents a trade-off, and users can select different scale factors based on their preferences.

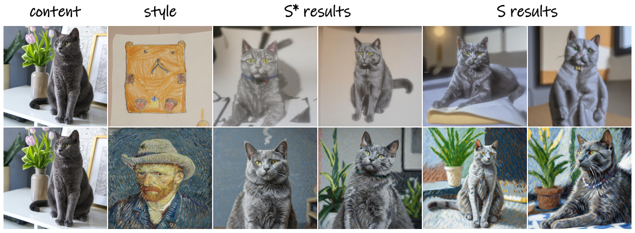

Figure 1. **Results of different scaling factors.** Corresponding generation results of K-LoRA with differernt scaling factor and for each object-style pair, two seeds are randomly selected.

E. Robustness Analysis

We evaluate LoRA models from various sources, where the object LoRA is sourced from the community, while the style LoRA is trained locally. We also compare DirectMerge [3], Multi-LoRA composition [6], and our proposed Fixed Selection approach. As shown in Fig. 7, our method demonstrates superior performance in learning both object and style characteristics, surpassing other approaches. Furthermore, we test the robustness of our approach by selecting random seeds to assess stability. The results, presented in Fig. 8, indicate that our method consistently achieves stable fusion across a broad range of seed selections, ensuring reliable integration.

F. Additional Ablations

In the main text, we employe a scale with two hyperparameters, α and β. Specifically, we set α to 1.5 and β to 0.5, enabling objects and styles to exert varying levels of influence at different positions. To validate the suitability of the selected parameters, we compute the CLIP similarity scores between 18 randomly chosen sets of generated images and their corresponding original object/style references. The results shown in the table below represent the summation of CLIP similarity scores.

$\beta\backslash\alpha$	1.0	1.5	2.0
0.25	125.3%	126.7%	127.0%
0.50	126.5%	**128.1%**	126.2%
0.75	124.5%	125.8%	125.3%

We can see that the optimal setting for α and β is 1.5 and 0.5, respectively. This weight configuration satisfies almost all content-style pairs according to our experiments, and users do not need to make further adjustments.

References

[1] Jiwoo Chung, Sangeek Hyun, and Jae-Pil Heo. Style Injection in Diffusion: A Training-free Approach for Adapting Large-scale Diffusion Models for Style Transfer. *arXiv e-prints*, art. arXiv:2312.09008, 2023. 1, 7

[2] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 1

[3] Simo Ryu. Merging loras. https://github.com/cloneofsimo/lora, 2023. 1

[4] Kihyuk Sohn, Nataniel Ruiz, Kimin Lee, Daniel Castro Chin, Irina Blok, Huiwen Chang, Jarred Barber, Lu Jiang, Glenn Entis, Yuanzhen Li, et al. Styledrop: Text-to-image generation in any style. *arXiv preprint arXiv:2306.00983*, 2023. 1

[5] Yu xin Zhang, Weiming Dong, Fan Tang, Nisha Huang, Haibin Huang, Chongyang Ma, Tong-Yee Lee, Oliver Deussen, and Changsheng Xu. Prospect: Prompt spectrum for attribute-aware personalization of diffusion models. *ACM Transactions on Graphics (TOG)*, 42:1 – 14, 2023. 1

[6] Ming Zhong, Yelong Shen, Shuohang Wang, Yadong Lu, Yizhu Jiao, Siru Ouyang, Donghan Yu, Jiawei Han, and Weizhu Chen. Multi-lora composition for image generation. *arXiv preprint arXiv:2402.16843*, 2024. 1, 8

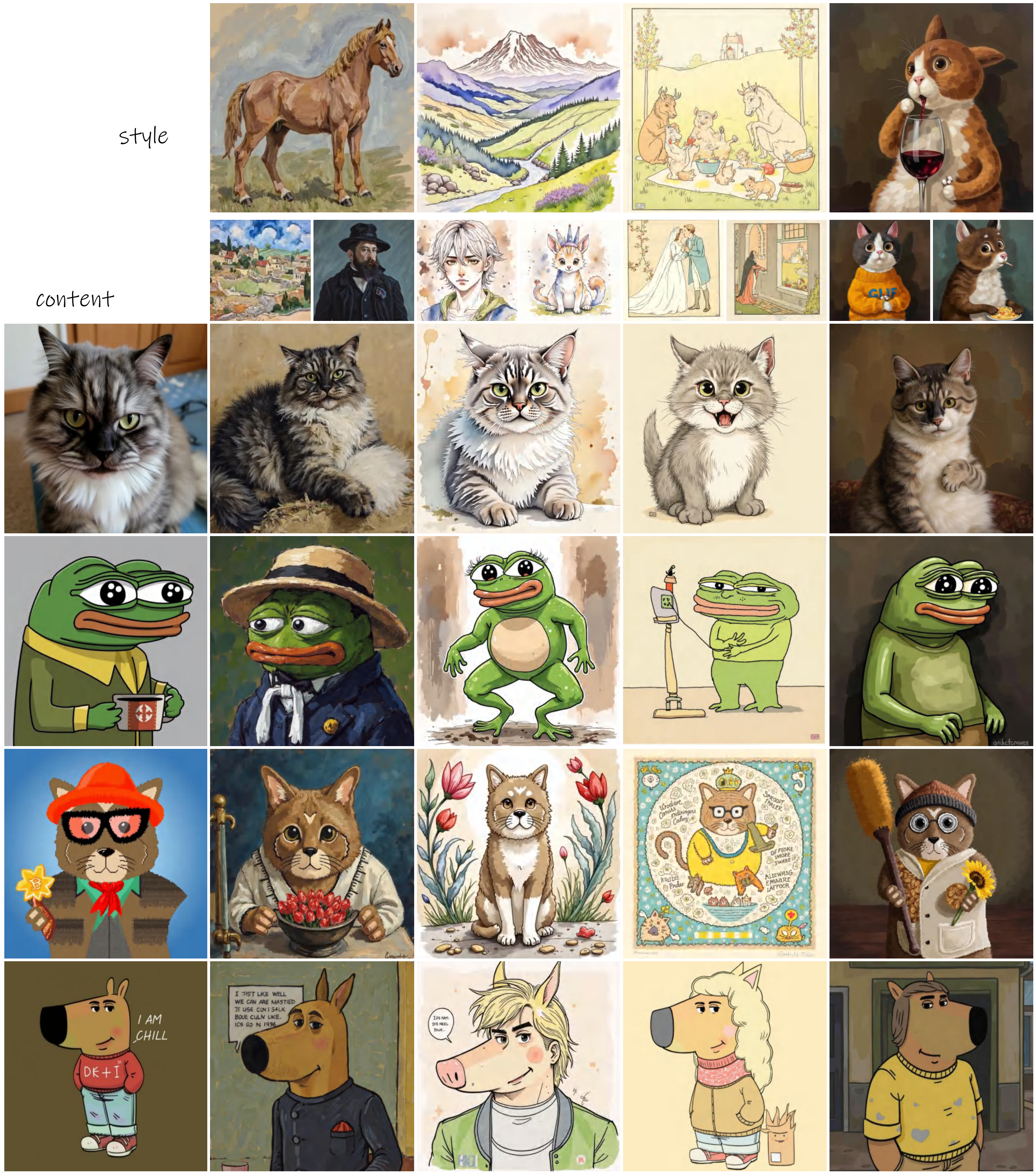

Figure 2. **Additional Generated Results using FLUX.** The images in each position correspond to the object above and the style on the left, showing the results generated by applying the different LoRAs with our method.

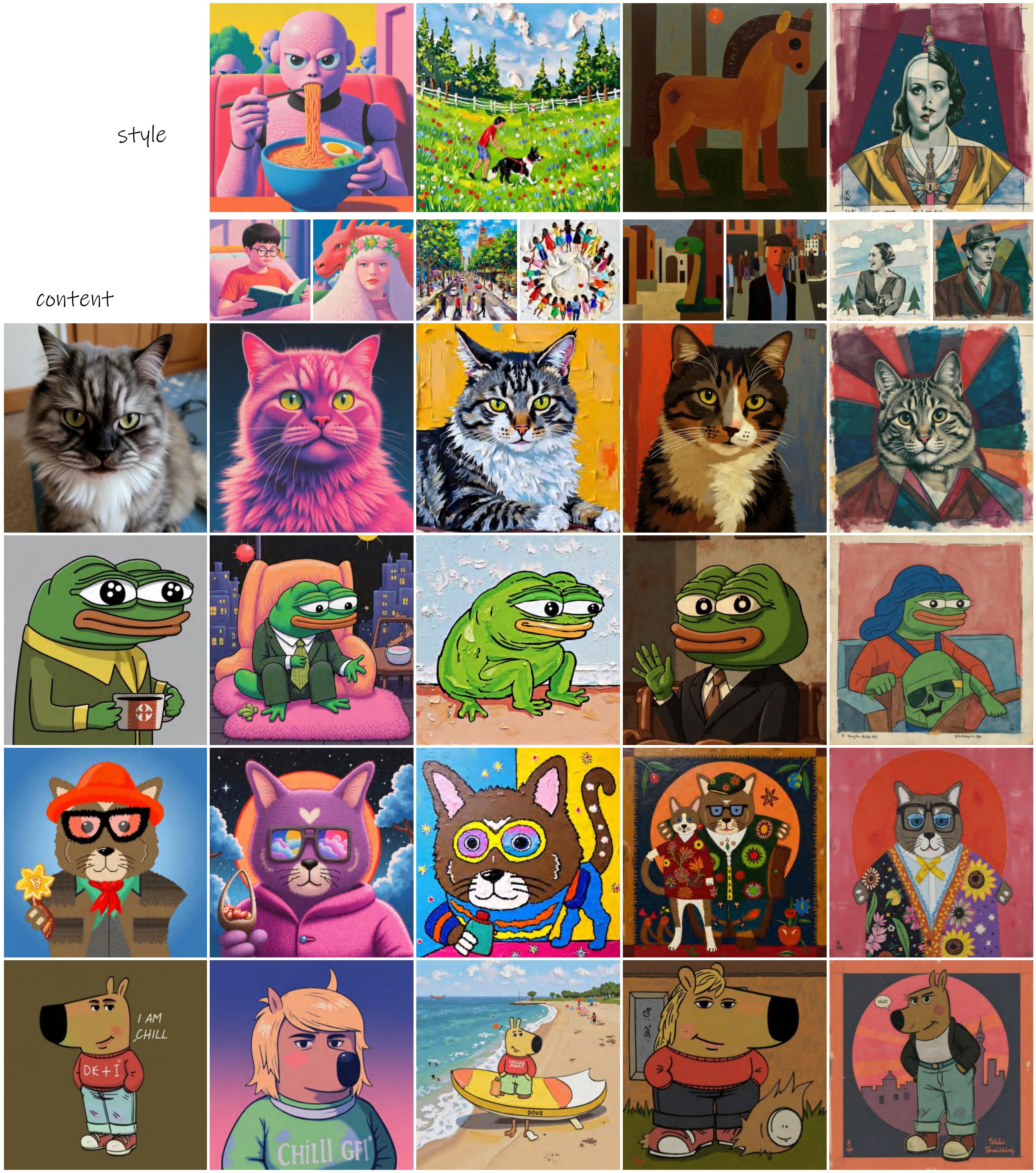

Figure 3. **Additional Generated Results using FLUX.** The images in each position correspond to the object above and the style on the left, showing the results generated by applying the different LoRAs with our method.

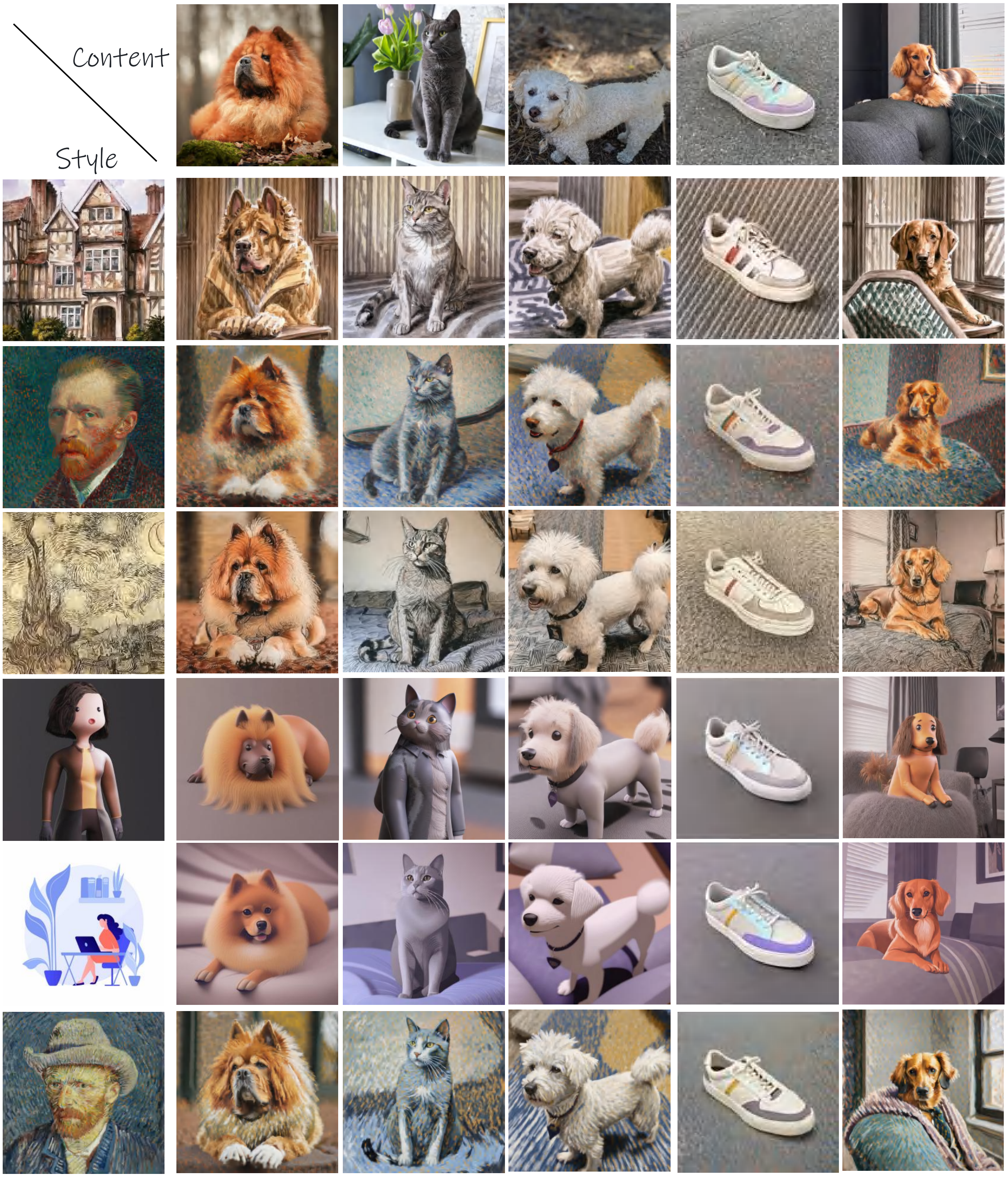

Figure 4. **Additional Generated Results using SD.** The images in each position correspond to the object above and the style on the left, showing the results generated by applying the different LoRAs with our method.

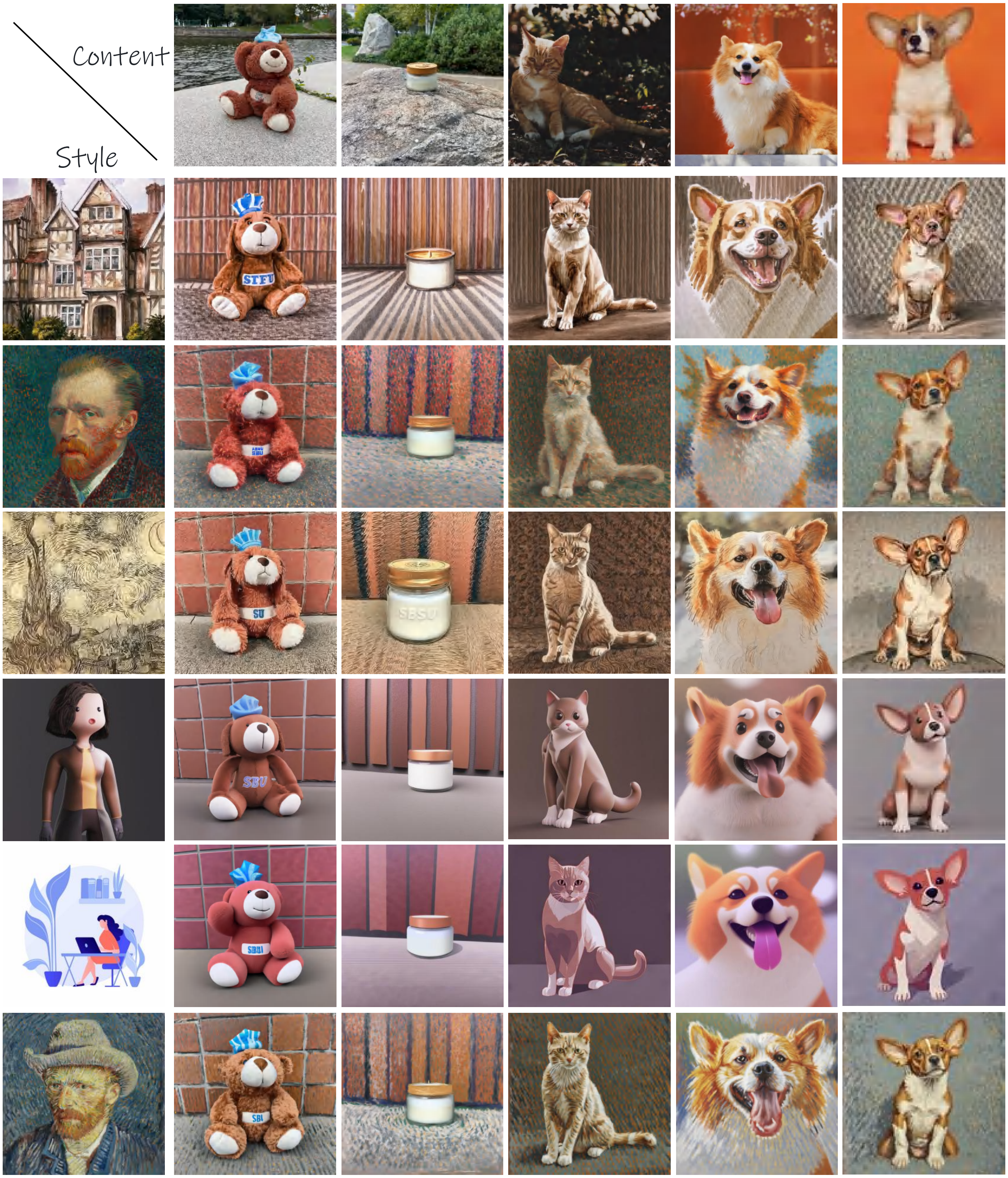

Figure 5. **Additional Generated Results using SD.** The images in each position correspond to the object above and the style on the left, showing the results generated by applying the different LoRAs with our method.

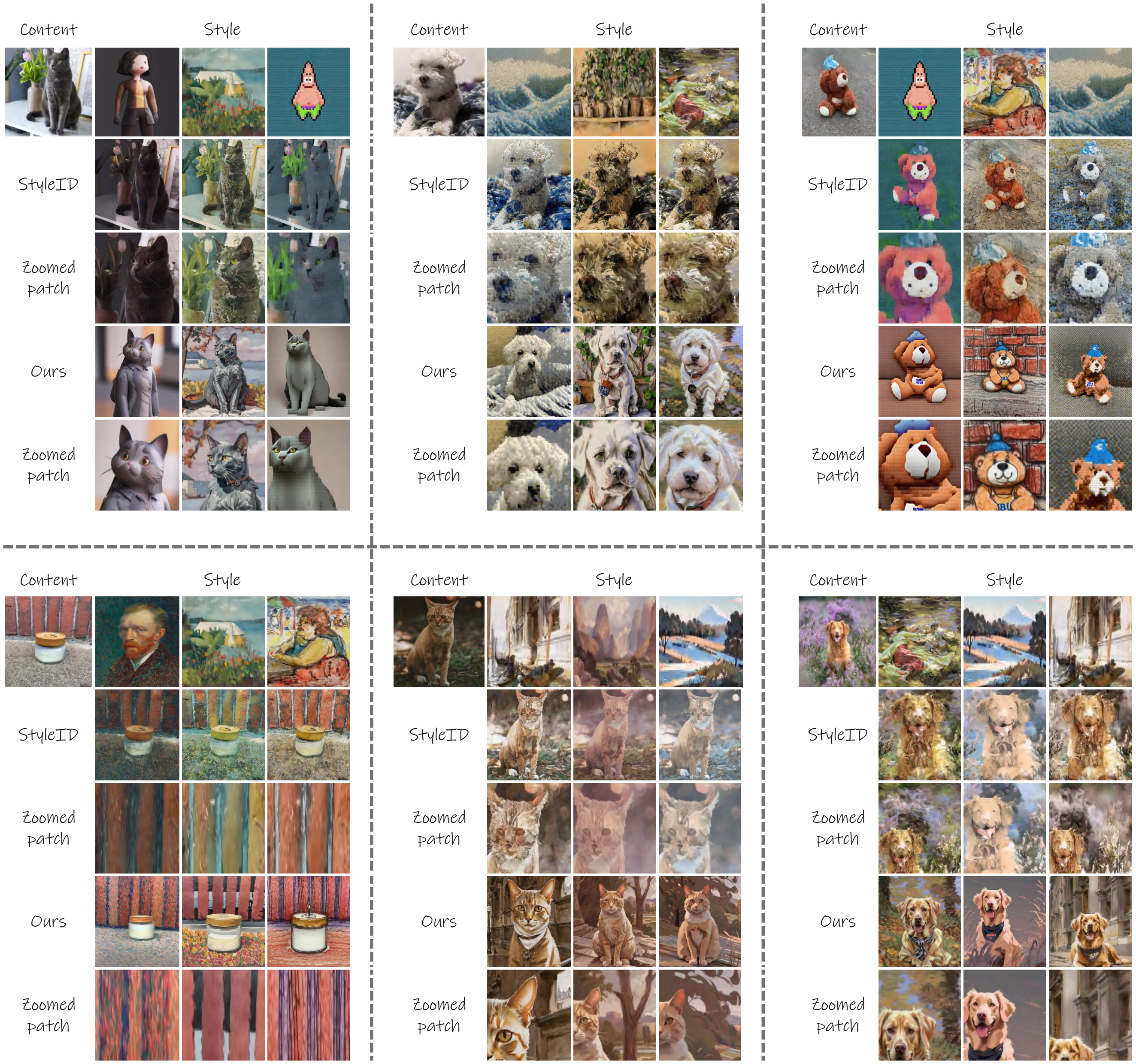

Figure 6. **Additional Comparisons.** We compare the StyleID [1] method and then capture zoomed patches in the output image to observe detailed texture information and stylistic features. Within each block, the second and third rows represent StyleID results along with its corresponding zoomed patch, while the subsequent two rows illustrate the result of our method and the associated zoomed patch.

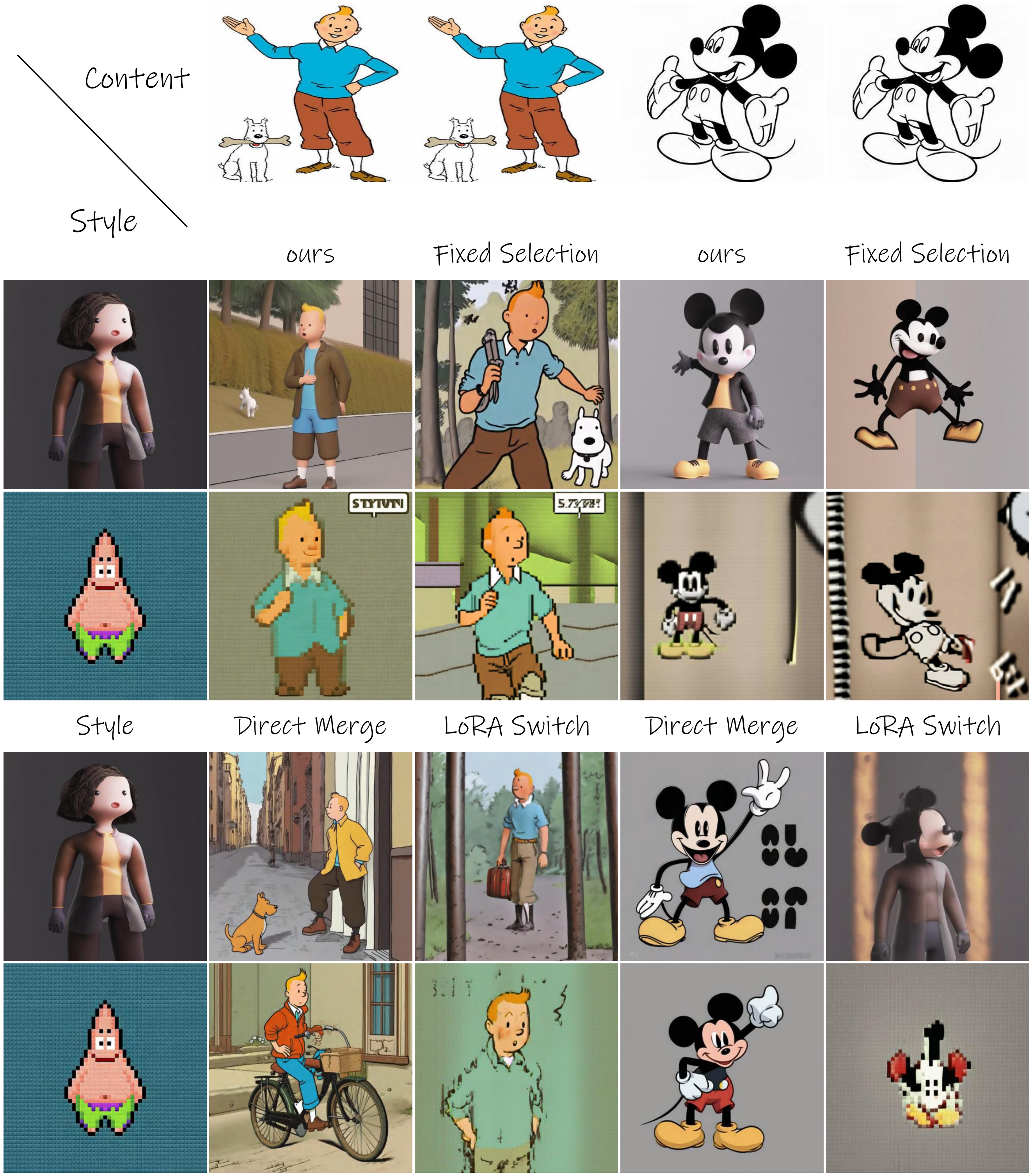

Figure 7. **Robustness Validation.** We utilize community LoRAs and locally trained LoRAs to compare the Fixed Selection proposed in the main text, direct arithmetic merging LoRA as a baseline comparison, Multi-LoRA Composition [6] methods, in order to validate generalizability and robustness.

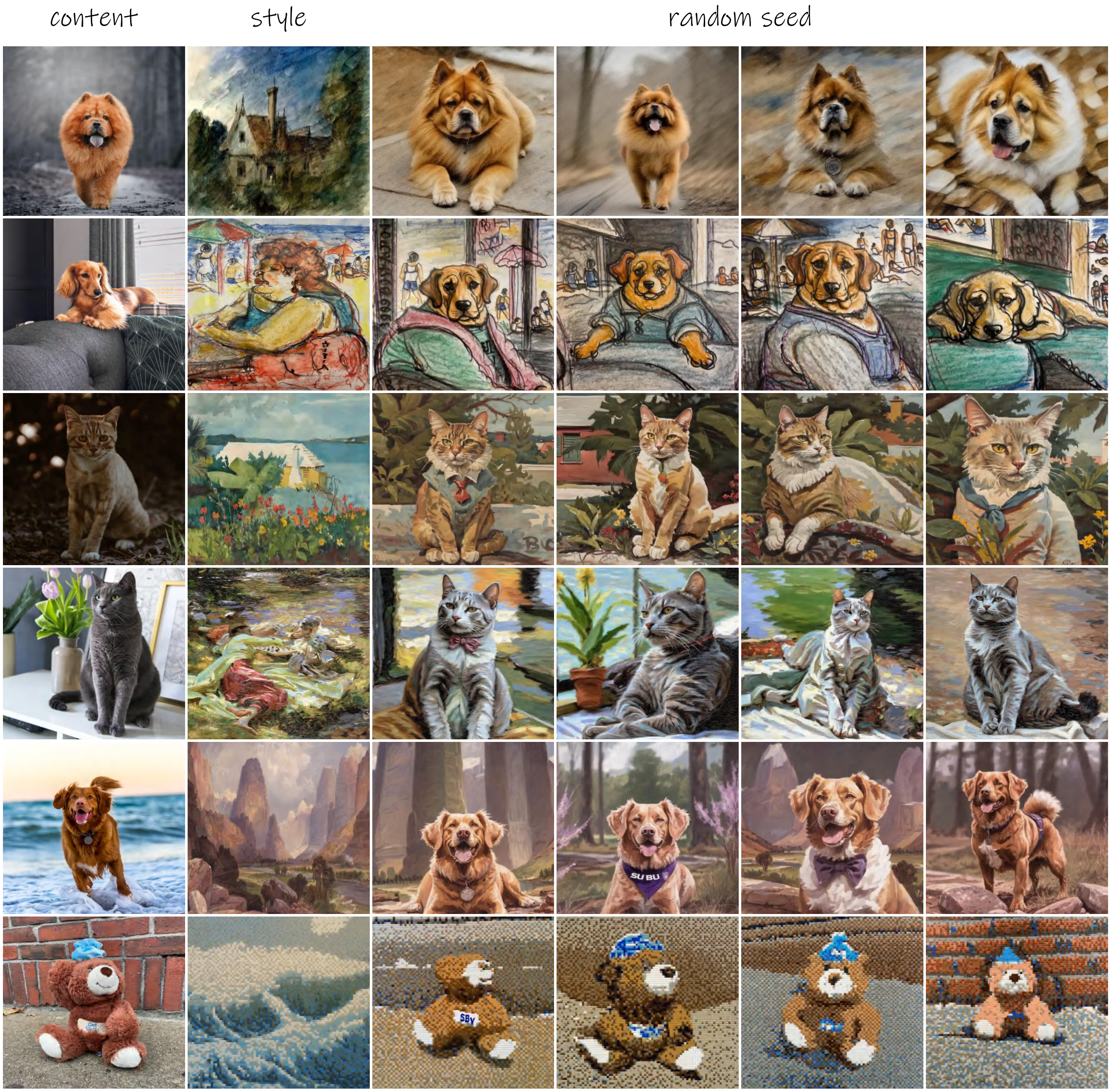

Figure 8. **Robustness Validation.** We randomly select seeds to further validate stability.

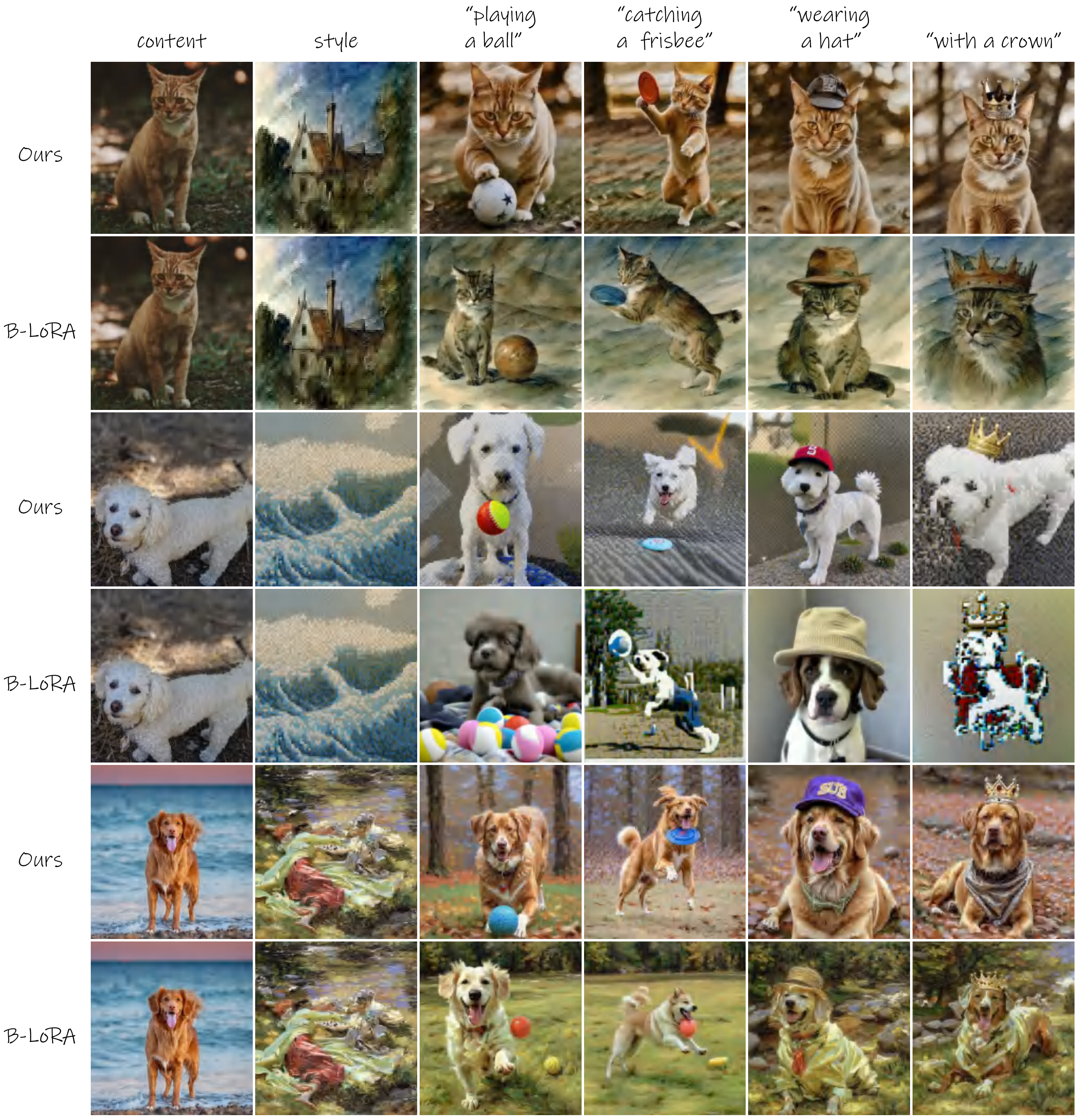

Figure 9. **Prompt Control.** We introduce prompts for new scenes, new actions, and new objects to validate our method's ability to recontextualize content and maintain stylistic consistency.

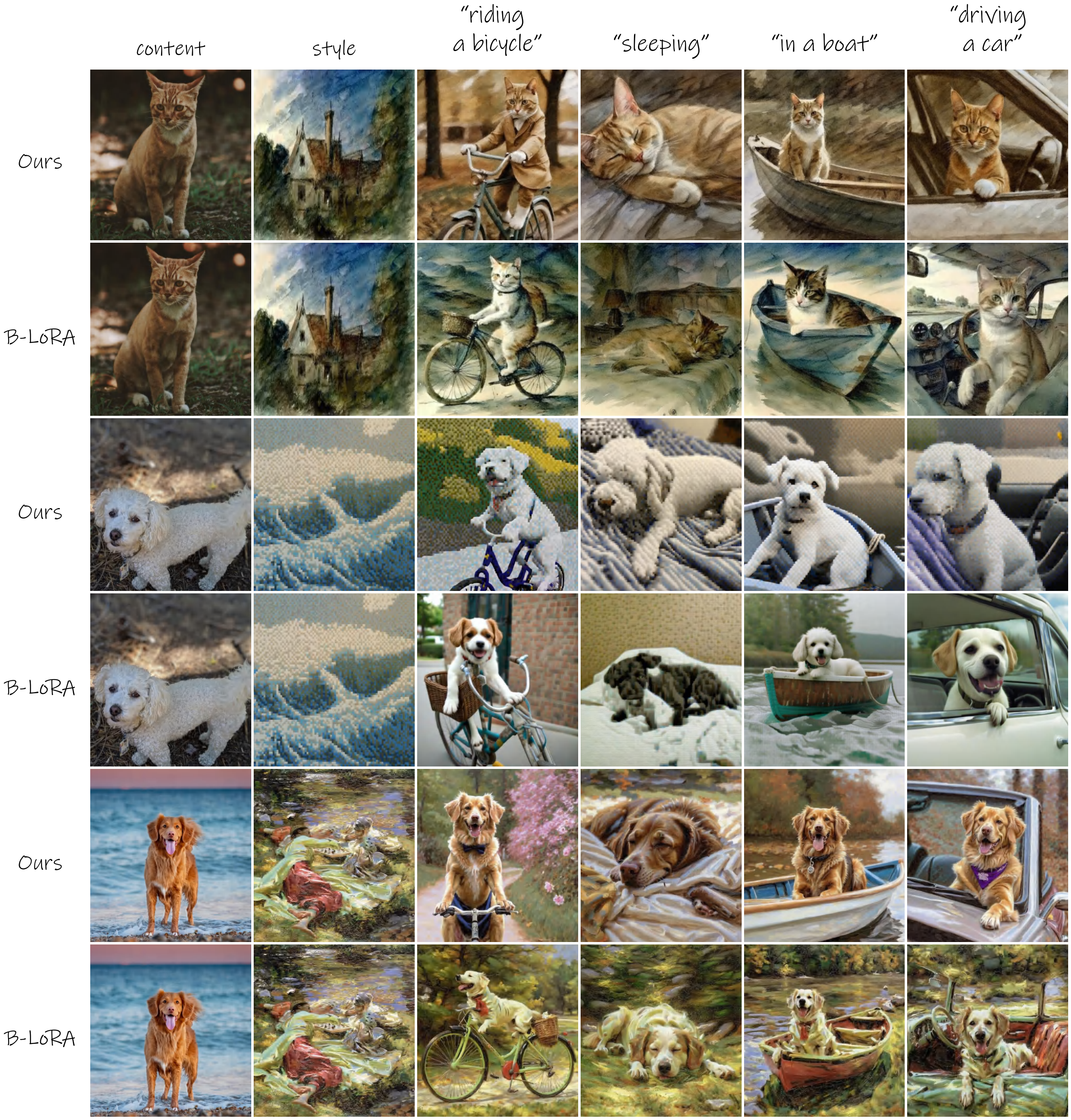

Figure 10. **Prompt Control.** We introduce prompts for new scenes, new actions, and new objects to validate our method's ability to re-contextualize content and maintain stylistic consistency.