Implicit Style-Content Separation using B-LoRA

Yarden Frenkel[1], Yael Vinker[1], Ariel Shamir[2], and Daniel Cohen-Or[1]

[1] Tel Aviv University
[2] Reichman University

Table of Contents

1 Comparisons

User Study As described in the main paper, we conducted a user study to further validate our findings. We constructed an evaluation set comprising 50 unique pairs of style and content images, randomly sampled from a diverse pool of 23 objects and 25 style references. From this evaluation set, we selected 10 representative pairs for each of the competing methods: ZipLoRA, StyleDrop, and StyleAligned. For each pair, we generated images using both the respective method and our approach, presenting them alongside the original style and content references, as illustrated in Figure 1 The generated images were displayed in a randomized order to avoid bias. Participants were asked to choose the result that "better transfers the style from the style image while preserving the content of the content image." In total, we gathered 1020 responses from 34 participants, ensuring a comprehensive evaluation of our method against alternative approaches.

Qualitative Comparisons In Section 5 of the main paper, we conducted a comparison of our B-LoRA method against four state-of-the-art baselines for image stylization incorporating personalization [2,3,8,9]. In this section, we delve deeper into the implementation details and present additional qualitative results. To begin, we employed DreamBooth-LoRA [7] fine-tuning to obtain both style and content LoRAs, utilizing the same parameter configuration as ZipLoRA [8]. For content images, we conducted fine-tuning across a set of images of the same object, except for the experiment involving a single image. However, for style LoRAs, we conducted fine-tuning using a single style image. We utilized the

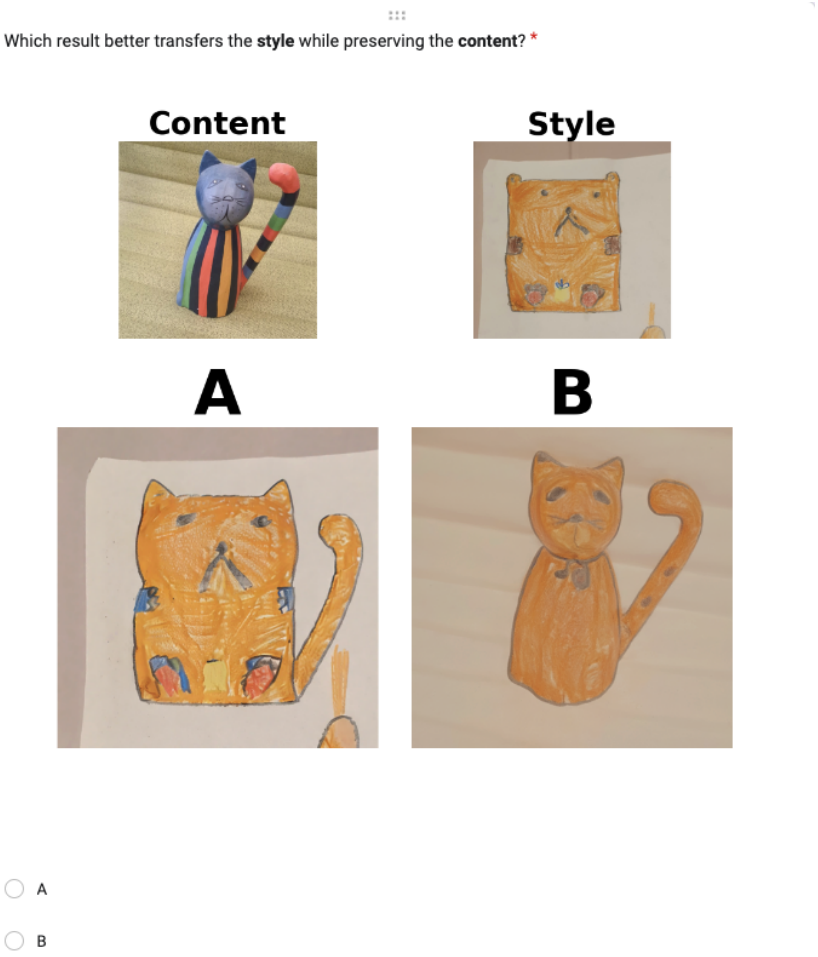

Fig. 1: Screenshot from the user study. Each of the two images, labeled A and B, represents a result obtained from a different method. Participants were tasked with selecting the image they believe is better in terms of both style adaptation and content preservation.

prompts provided in DreamBooth [6] and StyleDrop [9], specifically "A [v] <object>" or "A <object> in [s] style" for content and style, respectively. Subsequently, for ControlNet combined with DreamBooth-LoRA, we leveraged the publicly available implementation of ControlNet on SDXL from huggingface [3]. this approach involved utilizing the style LoRAs we trained for style transfer while employing CannyEdge with thresholds of 100 and 200 for content guidance in ControlNet. For StyleDrop [9], we followed the methodology outlined in StyleAligned [2] for fine-tuning the model over the style images, followed by fus-

ing the content LoRAs with the SDXL weights. Similarly, for StyleAligned [2], we utilized the authors' implementation for subject-driven generation alongside our content LoRAs. Lastly, for ZipLoRA [8], we use the unofficial implementation [5] with default parameters. We provide additional comparisons of our B-LoRA method against the aforementioned approaches using the same evaluation set presented in Section 5 of the main paper. These additional comparisons are illustrated in Figure 5. Furthermore, we provide comparisons with challenging content inputs, such as stylized images, presented in Figure 6. We also showcase comparisons with challenging style inputs, such as object images, in Figure 7. These examples demonstrate the robustness of our method in handling diverse and complex content and style references.

Comparisons to Baselines Beyond SDXL-Based Approaches We provide additional comparisons of our method with three other image stylization techniques that do not rely on SDXL: StyTr2 [1], AdaAttn [4], and SWAG [11]. We evaluated the results using the same quantitative metrics described in the main paper. Figure 2 presents a qualitative comparison of the same set shown in the main paper, and Table 1 contains the quantitative results.

Table 1: Quantitative comparison: We measure the average cosine similarity between the DINO features of the output image and the reference style and content. In this experiment, we use a single input image for evaluation.

	StyTr2	AdaAttn	SWAG	Ours
Style Transfer	0.83	0.818	0.883	0.881
Content	0.854	0.828	0.788	0.790

Comparisons to InstantStyle InstantStyle [10] is a concurrent work to ours. Aimed at performing image stylization tasks based on a style image reference. InstantStyle achieves this by injecting the CLIP embedding of the style image into style-specific blocks within SDXL, similar to our method, where the fifth block is selected for the style condition. Notably, InstantStyle uses a trained IP-Adapter model and does not require fine-tuning, which is its main advantage over our method. Both approaches provide compelling results in consistent style generation, as presented in Figure 3. For content conditioning, InstantStyle utilizes ControlNet, while our method separates content from style and extracts both. This allows for better content preservation in scenarios where ControlNet may not capture the content well enough or may override the style, as shown in Figure 4. Additionally, InstantStyle requires the content component to be explicitly defined to subtract its CLIP embedding from the style embedding, whereas our approach learns the content and style implicitly. For a fair comparison, we trained our method using the style images from InstantStyle.

Fig. 2: Comparison with alternative approaches that do not rely on SDXL. The input style and content references are shown on the left.

2 Limitations

While our work enables robust image stylization across various complex input images, it does have limitations. First, in our style-content separation procedure, the object's color is often included in the style component. However, in some cases, color plays a crucial role in preserving identity. Therefore, when stylizing the content component, the results may not properly preserve the object's identity, as illustrated in Figure 8(a). Second, since we use a single reference image, our learned style component may encompass background elements rather than focusing solely on the central object, as demonstrated in Figure 8(b). Lastly, while our method effectively stylizes scene images, it may encounter challenges with complex scenes containing numerous elements. Consequently, it may struggle to accurately capture the scene structure, potentially compromising content preservation, as depicted in Figure 8(c).

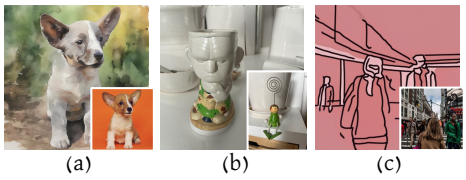

Fig. 8: Method limitations. (a) Suboptimal identity preservation due to color separation. (b) Style leakage from background objects. (c) Inability to adequately capture content in complex scenes.

Here, we expand upon this section and propose potential approaches to mitigate these limitations.

The first limitation we aim to address is the sub-optimal identity preservation due to color separation. To overcome this issue, we propose applying a scaling factor of alpha between 0.4-0.5 to the style adapter ΔW^5. This adjustment allows for preserving the original colors of the subject while minimizing interference with other style B-LoRA injections, as illustrated in Figure 9.

To mitigate style leakage from background objects in the style reference image, we suggest preprocessing the training data by center cropping the desired style reference image. This approach increases precision by focusing on the central object during the B-LoRA training process.

Addressing the final limitation of adequately capturing content in complex scenes, we conducted an ablation study to explore the effect of injecting different prompts into different blocks of the network. Specifically, we conducted five experiments:

(1) Injecting our method's prompt "A [c] in [s] style", into all transformer blocks of the UNet. **(2)** Injecting "A [c]" into the content block W^4 while injecting "A [s]" into all other blocks. **(3)** The complement of (2), injecting "A [s]" into the style block W^5 and "A [c]" into all other blocks. **(4)** Similar to (2), but injecting "A [c]" into W^4 while other blocks receive "A [c] in [s] style". **(5)** Similar to (3), but injecting "A [s]" into W^5 while other blocks receive our method's prompt "A [c] in [s] style".

We present the results of these experiments in Figure 10. Our findings indicate that injecting the prompt "A [c]" into W^4 while other blocks receive the prompt "A [c] in [s] style" often leads to improved generation results, particularly for complex scenes containing numerous elements.

3 Analysis and Ablation

Layers Optimization As detailed in Section 4 of the main paper, the SDXL UNet comprises 11 transformer blocks, with the high-resolution blocks containing 2 attention layers each and the middle 6 blocks containing 10 attention layers each (see Figure 3 in the main paper). To explore the impact of different block combinations on the resulting image, we divided the UNet into 8 blocks $\{W_0^0 \dots W_0^7\}$, where $\{W_0^1 \dots W_0^6\}$ represent the bottleneck blocks, as discussed in Section 4, and designated W_0^0 and W_0^7 for the high-resolution blocks at the edges. We aimed to assess the effects of optimizing various block combinations $\{\Delta W^i, \Delta W^j\}$ by jointly training the LoRA weights of the corresponding blocks. Qualitative results are depicted in Figures 11 and 12, where each cell (i, j) represents the reconstruction image for the prompt "A [v]" after training the LoRAs solely for the i-th and j-th blocks of the SDXL Unet. The diagonal entries represent output generated by training a single block. Upon examination, we observed that optimizing $\{\Delta W^4, \Delta W^5\}$ consistently produced the most satisfactory results for content and style, respectively, outperforming other combinations. Notably, the reconstruction in cell (4, 5) yielded the best results achievable among all com-

binations, supporting our findings in the main paper. Furthermore, we noted that the combination of blocks 2 and 5 also achieved satisfactory reconstruction. However, employing this combination may lead to less disentanglement of style from content, as ΔW^5 needs to "cover" ΔW^2 by learning content details instead of focusing primarily on style, as intended. This observation further solidifies our choice of optimizing $\{\Delta W^4, \Delta W^5\}$ for effective style-content separation.

Prompt Selection To validate our choice of the prompt "A [v]" during optimization, we conducted an ablation study regarding the prompt used during training. As described in the DreamBooth [6] paper, the authors suggest that the most efficient way to conduct the fine-tuning process is by using the prompt "A [v] <class-name>", where [v] is the token dedicated for personalization, and <class-name> is the class of the object depicted in the input image. We compare our method of optimizing ΔW^4 and ΔW^5 with the prompt "A [v]" against using the suggested "A [v] <class-name>" prompt.

In Figure 13, we demonstrate the impact of different prompts on style transfer between objects by fusing ΔW^4_{c1} and ΔW^5_{c2} to transfer the style of object1 to object2. We use four different prompts: **(1)** "A [c1] in [c2] style" (our method), **(2)** "A [c1] <obj1> in [c2] style", **(3)** "A [c1] <obj1> in [c2] <obj2> style", and **(4)** our method optimized without the class name.

As can be seen, the first column, using "A [c1] in [c2] style", fails to reconstruct the object's structure correctly. The second column, with "A [c1] <obj1> in [c2] style", successfully reconstructs the content but struggles to transfer the style. In the third column, using "A [c1] <obj1> in [c2] <obj2> style", the structure of the resulting image is affected by the obj2 class name.

In contrast, our method in the fourth column, optimized without the class name, is able to preserve the content image's structure and effectively transfer the style from the other object. This demonstrates the effectiveness of our approach using the prompt "A [v]" during optimization.

Alpha Effect As mentioned in the main paper, by the end of the training, we can obtain the tuned model weights using $W = W_0 + \Delta W$, where ΔW is our trained B-LoRA update. The strength of the fine-tuning merge equation can be adjusted and controlled by the alpha scalar: $W = W_0 + \alpha \Delta W$. (in our experiments $\alpha = 1$). We demonstrate alpha's effect on style and content components in Figure 14. As can be seen, when the alpha value is small, both the style and the content may be lost.

4 B-LoRA for Personalization

Throughout the paper, our method has been implemented using a single image for decoupling style and content. However, by training our method using multiple images for content, we can recontextualize reference objects while preserving stylization quality. In Figures 15 and 16, we showcase the versatility of our method by combining various B-LoRAs for style and content with text prompts. Note that the style can be derived from the reference style or from other objects.

5 Additional Results

Our B-LoRA method focuses on three main applications: image stylization based on image style references, text-based image stylization, and consistent style generation. In Figures 17 and 18, we present additional results generated by our approach for image stylization based on image style references. The columns represent the style reference images, while the rows correspond to the content reference images. As discussed, our method demonstrates proficiency in extracting content from style images (Figure 19) and extracting style from objects for object mixing tasks (Figure 20). In Figures 21 and 22, we provide qualitative results showcasing our method's performance on randomly selected objects and styles from our evaluation set. These examples further highlight the robustness of our approach to handling diverse content and style references. In Figure 23 we present additional qualitative results for text-based image stylization. As discussed in the paper, by utilizing only the learned B-LoRA weights capturing the content, our method enables text-guided style manipulation while effectively preserving the input object's content and structure. These results demonstrate the flexibility of our approach in allowing challenging style manipulations through textual guidance.

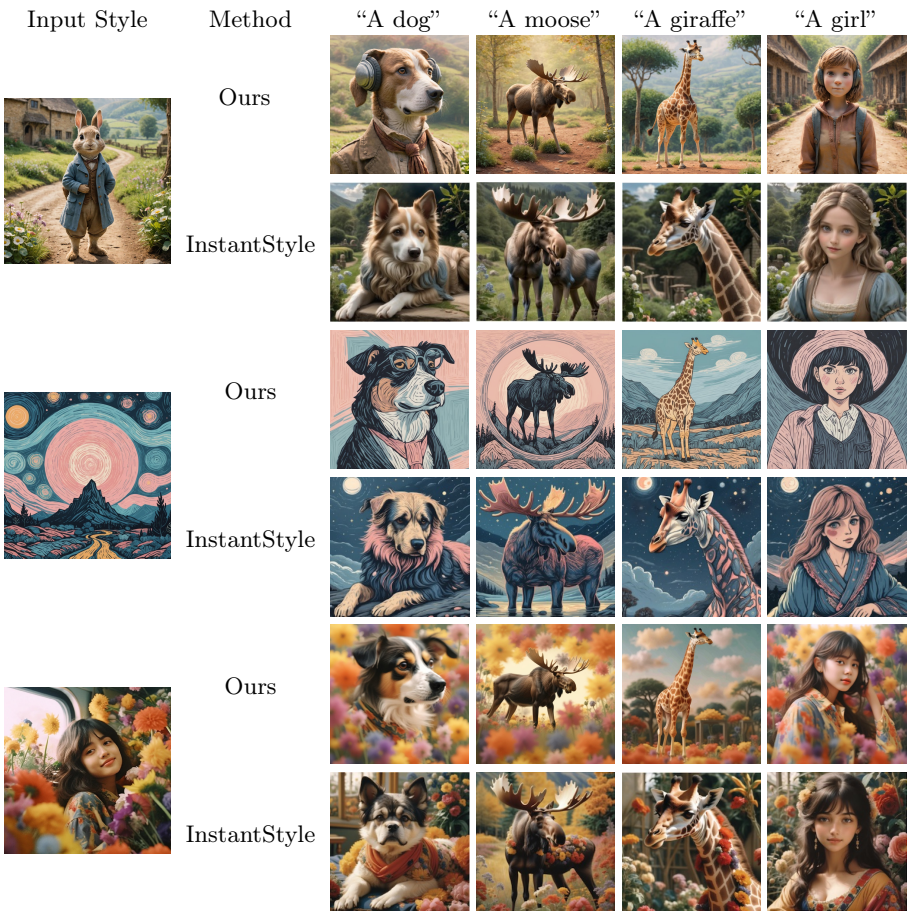

Fig. 3: Comparison of stylization results between our method and InstantStyle. The input style image is shown in the first column, followed by results generated by our method and InstantStyle for different prompts: “A dog”, “A moose”, “A giraffe”, and “A girl”. The images of InstantStyle are taken from the original paper. Both approaches achieve consistent style generation, demonstrating the effectiveness of style transfer.

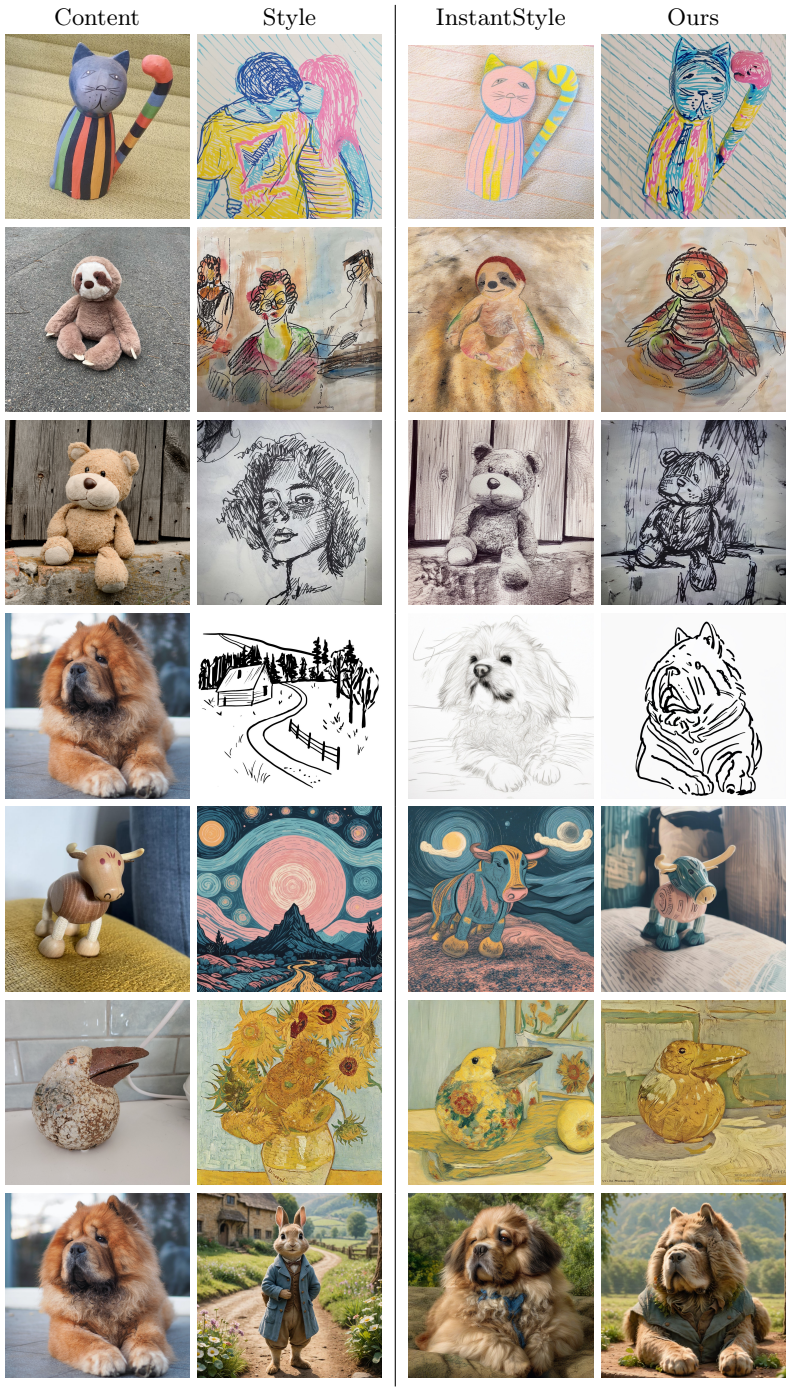

Fig. 4: Comparison of style and content mixing between our method and InstantStyle. The results illustrate cases where ControlNet, used by InstantStyle, may fail to adequately capture the content or may override the style. For example, in the fourth row, we can see that ControlNet failed to extract the shape of the dog, leading to unsatisfactory results, While our method demonstrates better content preservation. The images showcase the stylization applied to various content images, highlighting differences in how each approach handles content and style integration.

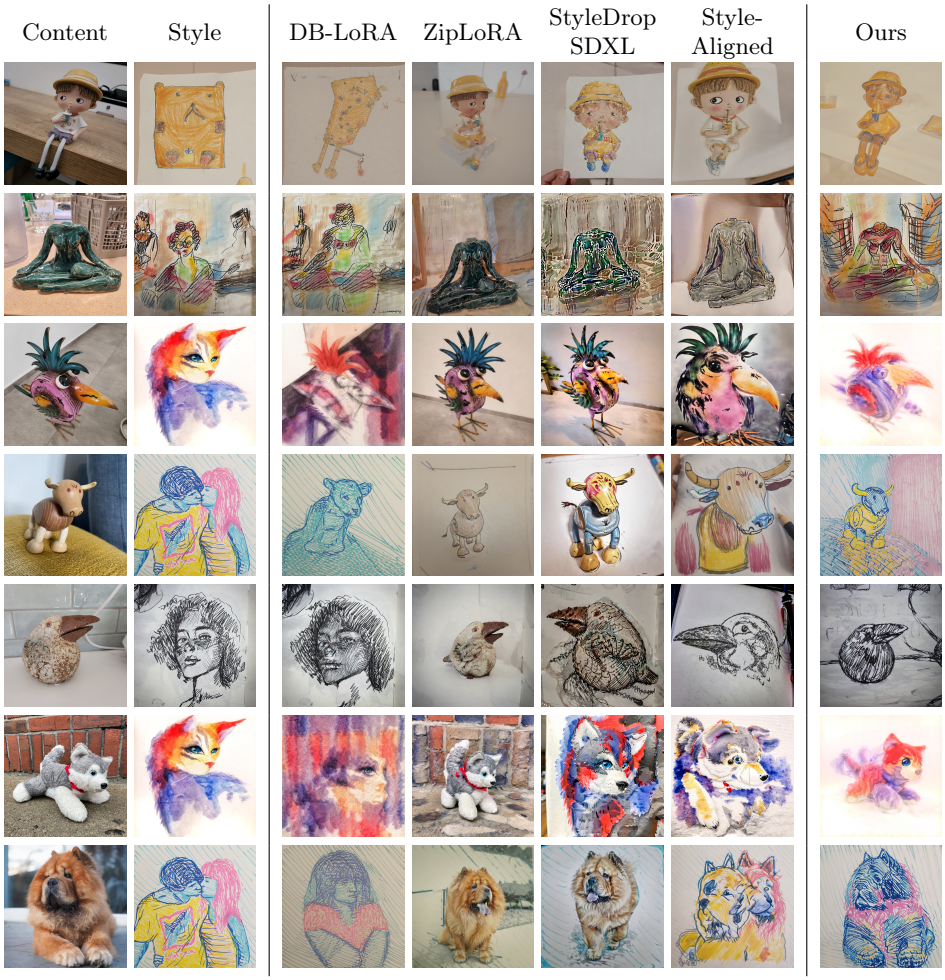

Fig. 5: Additional comparisons for image stylization based on reference image.

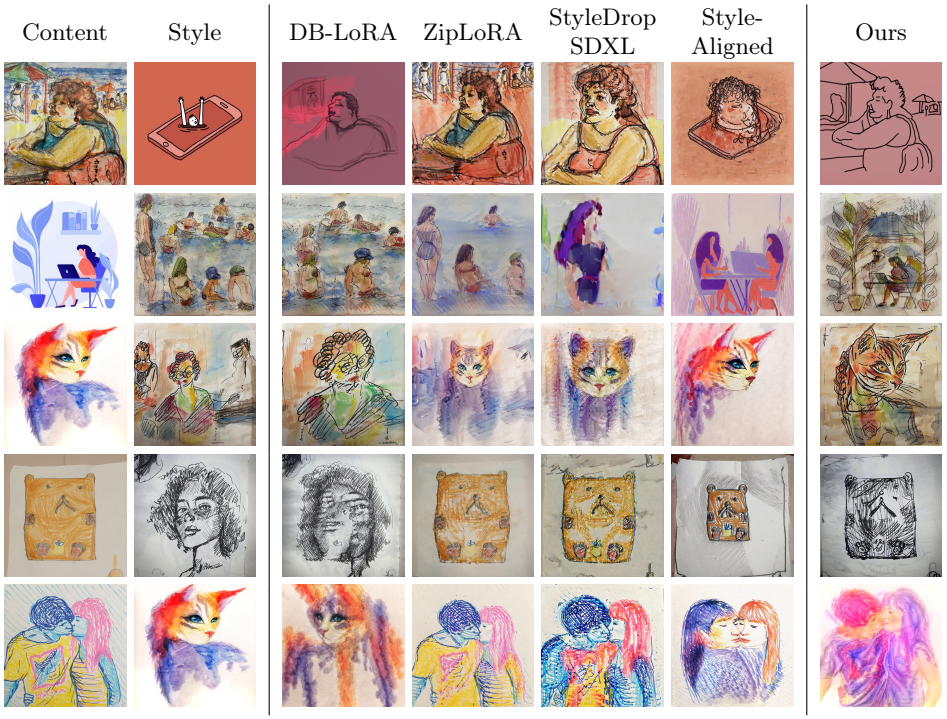

Fig. 6: Additional comparisons using challenging stylized images as content input. As can be seen, other methods encounter difficulties in disentangling the style and content from these images, consequently struggling to effectively transfer the style from one stylized image to another. ©The paintings in the first three rows are by Judith Kondor Mochary

Fig. 7: Additional comparisons using challenging subject images as style reference. As can be seen, other methods encounter difficulties in disentangling the style and content from these images, consequently struggling to effectively transfer the style from one object to another.

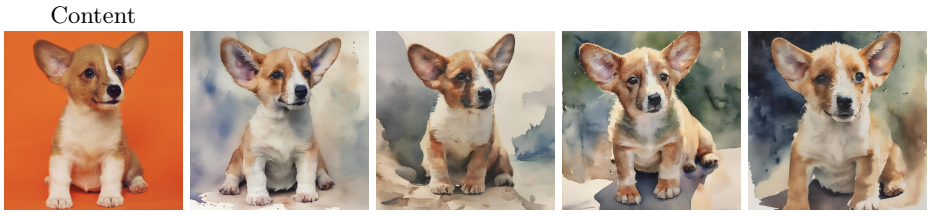

Fig. 9: To mitigate the limitation of sub-optimal identity preservation due to color separation, we propose combining adapters $\{\Delta W^4, \Delta W^5\}$, with ΔW^5 assigned a coefficient α within the range of [0.4, 0.5]. This method preserves the original colors of the subject while allowing stylizations using text prompts. The generated contents depicted in the figure are based on the prompt "Watercolor painting of [c]".

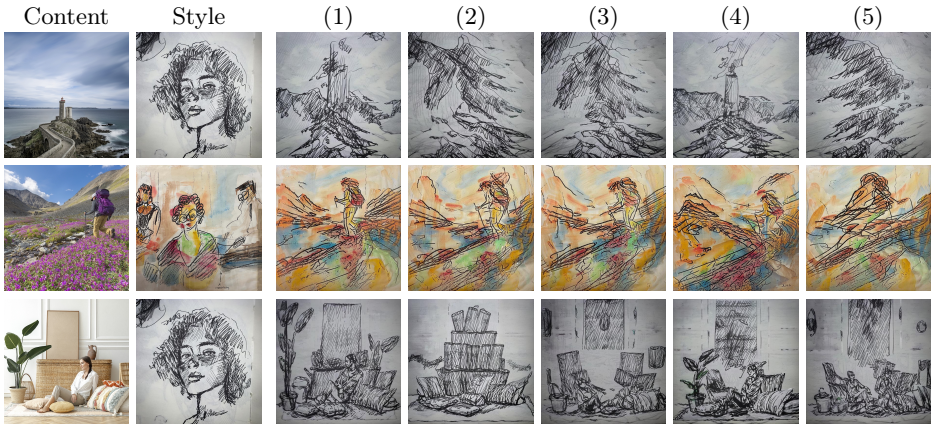

Fig. 10: Qualitative results of an ablation study investigating the effect of injecting different prompts into different blocks of the network to address the limitation of capturing content in complex scenes. Five experiments were conducted presented in the five columns [1-5]: **(1)** Injecting our method prompt, denoted as p_1 = "A [c] in [s] style", into the entire Unet. **(2)** Injecting "A [c]" into the content block W^4 while all other blocks receive "A [s]". **(3)** The complement, injecting "A [s]" into the style block W^5 and "A [c]" into all other blocks. **(4)** Similar to 2, but injecting "A [c]" into W^4 while other blocks receive "A [c] in [s] style". **(5)** Similar to 3, but injecting "A [s]" into W^5 while other blocks receive our method's prompt "A [c] in [s] style". As can be seen the (4) columns contains the best results.

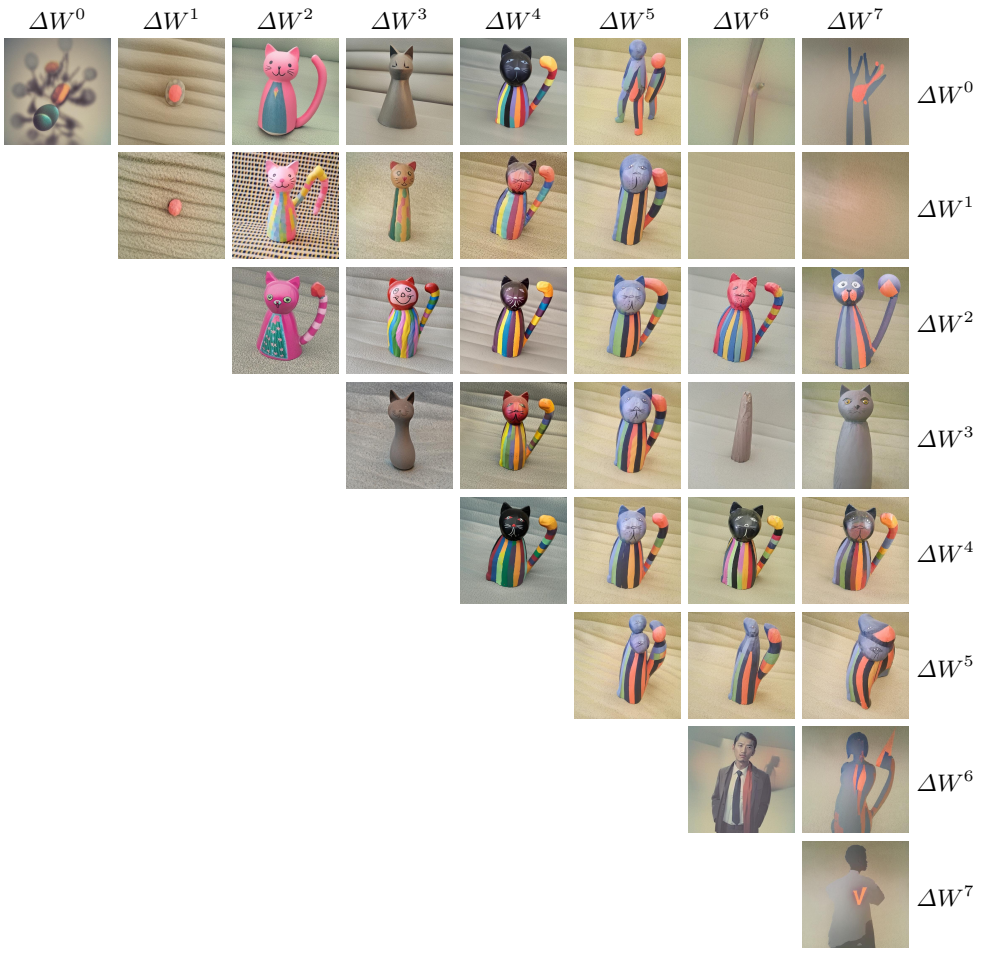

Fig. 11: Qualitative results of the ablation study showcasing the reconstruction images for prompt "A [v]" after training LoRAs for different block combinations of the SDXL Unet. Each cell (i, j) represents a specific block combination, with the diagonal representing output generated by training a single block. Notably, cells (4, 5) demonstrate the most consistent and optimal reconstruction for content and style, respectively

Fig. 12: Qualitative results of the ablation study showcasing the reconstruction images for prompt "A [v]" after training LoRAs for different block combinations of the SDXL Unet. Each cell (i, j) represents a specific block combination, with the diagonal representing output generated by training a single block. Notably, cells (4, 5) demonstrate the most consistent and optimal reconstruction for content and style, respectively.

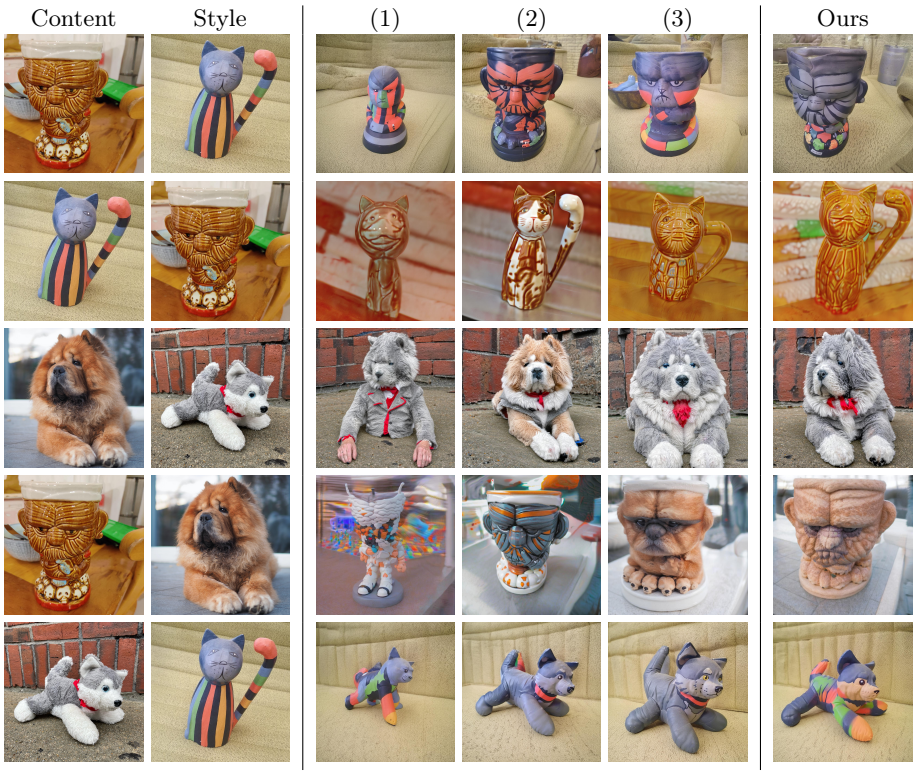

Fig. 13: Ablation study on the impact of different prompts for style transfer between objects. The first three columns use the prompts: (1) "A [c1] in [c2] style", (2) "A [c1] <obj1> in [c2] style", (3) "A [c1] <obj1> in [c2] <obj2> style", respectively. The fourth column shows our method using the prompt "A [v]" without class names during optimization of ΔW^4 and ΔW^5. Our approach in the fourth column better preserves the content object's structure while effectively transferring the style from the other object.

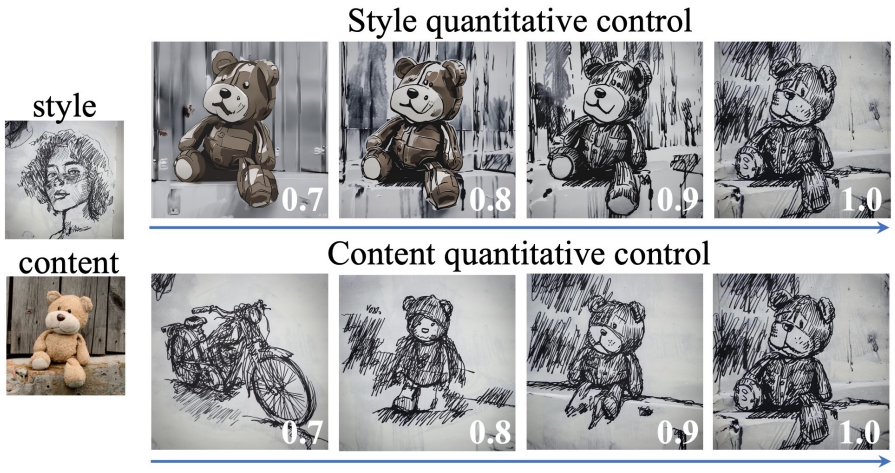

Fig. 14: On the left is the style-content input pair. On the right is quantitative control over style and content by altering the α parameter, shown in white.

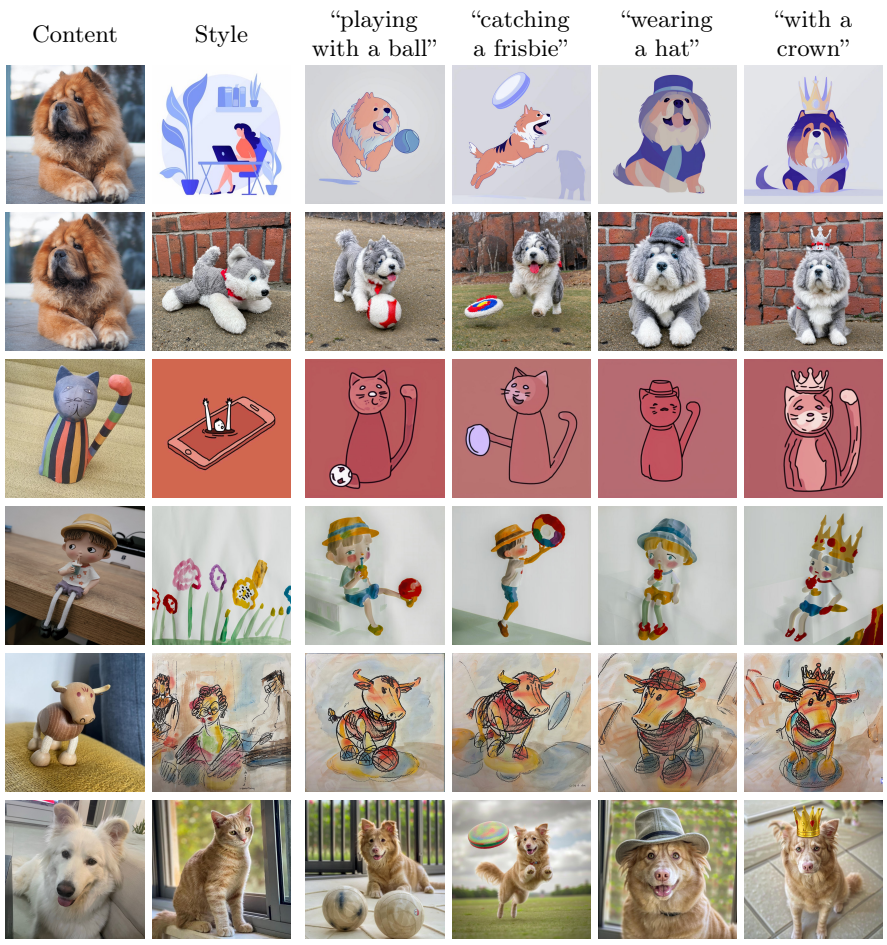

Fig. 15: While maintaining the stylistic characteristics of the style, our method effectively re-contextualizes the content object. Note that our approach is capable of transferring the style from either a style or object reference image.

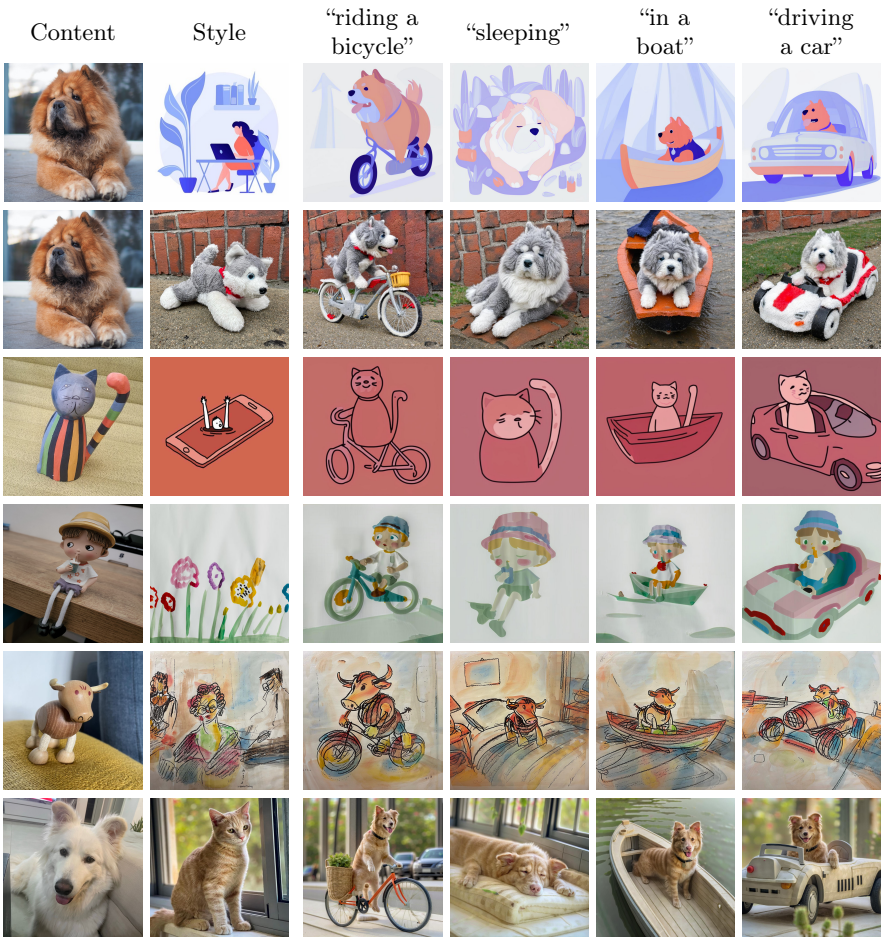

Fig. 16: While maintaining the stylistic characteristics of the style, our method effectively re-contextualizes the content object. Note that our approach is capable of transferring the style from either a style or object reference image.

Fig. 17: Image stylization based on image style reference using B-LoRA, illustrating the performance on challenging content image references. ©The paintings in the first three columns are by Judith Kondor Mochary.

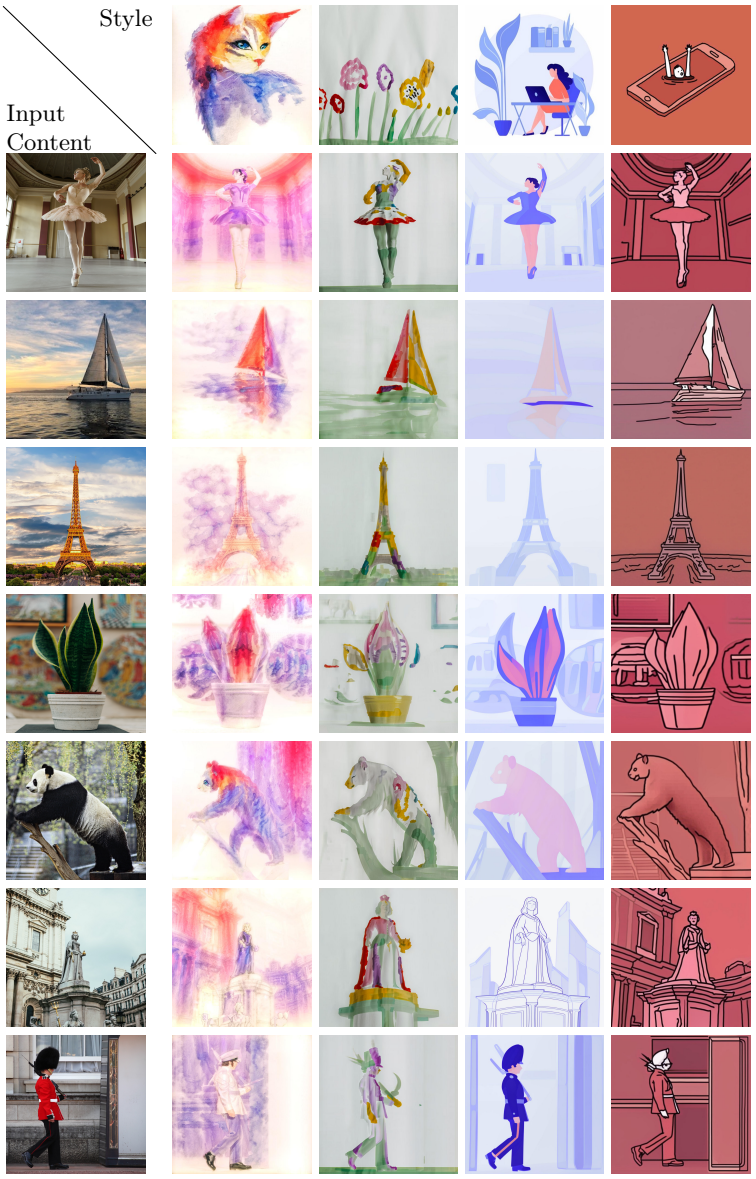

Fig. 18: Image stylization based on image style reference using B-LoRA, illustrating the performance on challenging content image references.

Fig. 19: Additional results generated using B-LoRA. Our method able to blend content and styles across different style images. Each object in the (i, j) cell is created by combining the ΔW^4 of the i-th row with the ΔW^5 of the j-th column, while the diagonal represents the reconstruction image. ©The paintings in the second and third columns (and rows) are by Judith Kondor Mochary.

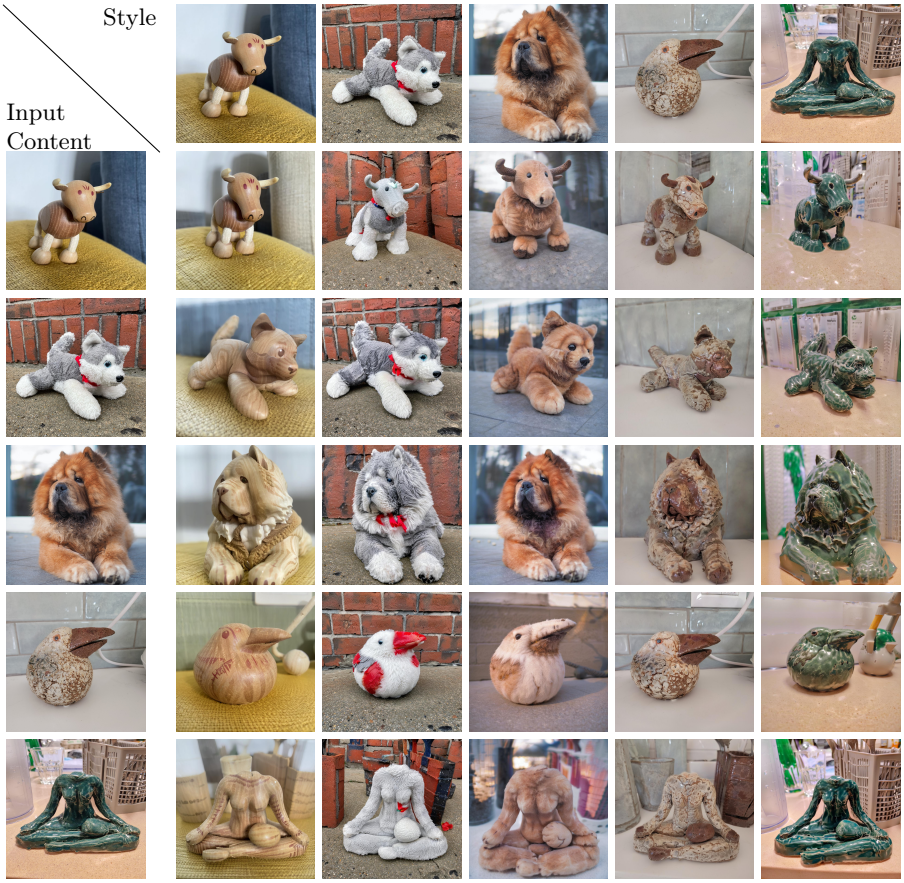

Fig. 20: Additional results generated using B-LoRA. Our method able to blend content and styles across different objects. Each object in the (i, j) cell is created by combining the ΔW^4 of the i-th row with the ΔW^5 of the j-th column, while the diagonal represents the reconstruction image.

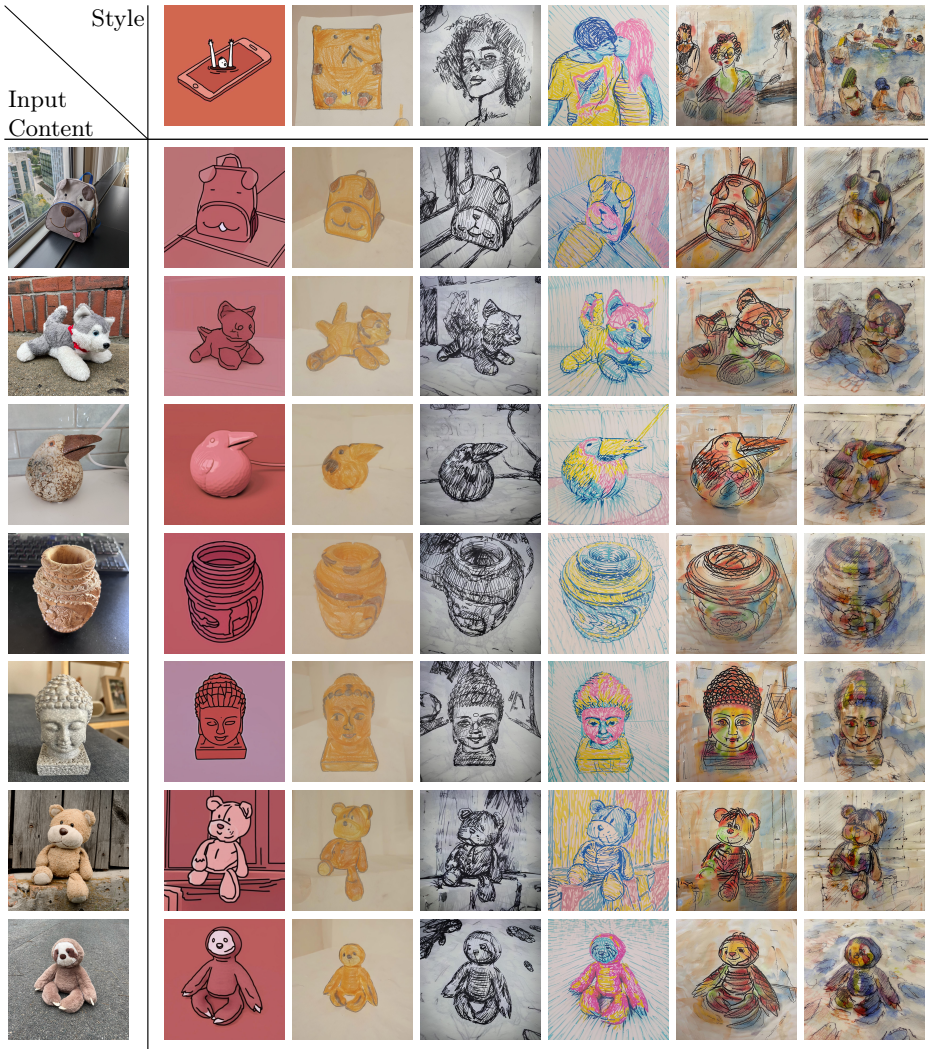

Fig. 21: Image stylization based on image style reference using B-LoRA for randomly selected objects and styles. ©The paintings in the last two columns are by Judith Kondor Mochary.

Fig. 22: Image stylization based on image style reference using B-LoRA for randomly selected objects and styles.

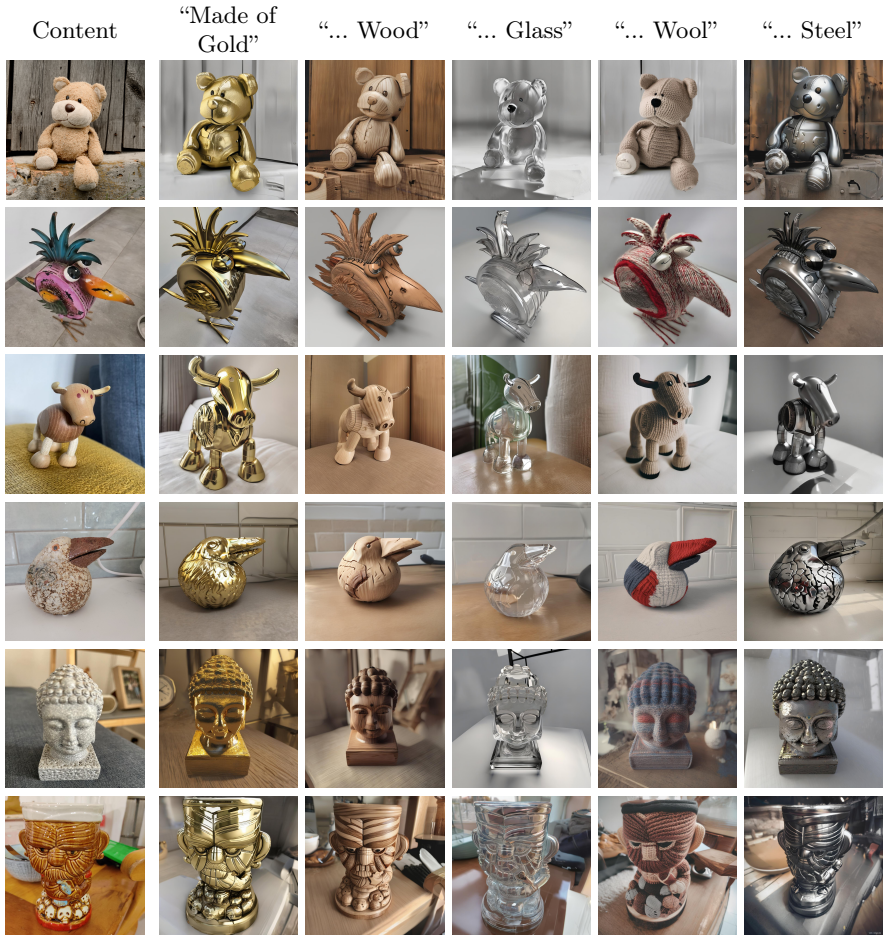

Fig. 23: Text-based Image stylization using B-LoRA, generated using the prompt "A [v] made of ...".

References

1. Deng, Y., Tang, F., Dong, W., Ma, C., Pan, X., Wang, L., Xu, C.: Stytr2: Image style transfer with transformers. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 11316–11326 (2021), https://api.semanticscholar.org/CorpusID:247922246 3
2. Hertz, A., Voynov, A., Fruchter, S., Cohen-Or, D.: Style aligned image generation via shared attention. arXiv preprint arXiv:2312.02133 (2023) 1, 2, 3
3. huggingface: Controlnet with stable diffusion xl 1, 2
4. Liu, S., Lin, T., He, D., Li, F., Wang, M., Li, X., Sun, Z., Li, Q., Ding, E.: Adaattn: Revisit attention mechanism in arbitrary neural style transfer. 2021 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 6629–6638 (2021), https://api.semanticscholar.org/CorpusID:236956663 3
5. mkshing: Ziplora-pytorch. https://github.com/mkshing/ziplora-pytorch 3
6. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: 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. pp. 22500–22510 (2023) 2, 6
7. Ryu, S.: Low-rank adaptation for fast text-to-image diffusion fine-tuning. https://github.com/cloneofsimo/lora 1
8. Shah, V., Ruiz, N., Cole, F., Lu, E., Lazebnik, S., Li, Y., Jampani, V.: Ziplora: Any subject in any style by effectively merging loras. arXiv preprint arXiv:2311.13600 (2023) 1, 3
9. Sohn, K., Ruiz, N., Lee, K., Chin, D.C., Blok, I., Chang, H., Barber, J., Jiang, L., Entis, G., Li, Y., Hao, Y., Essa, I., Rubinstein, M., Krishnan, D.: Styledrop: Text-to-image generation in any style (2023) 1, 2
10. Wang, H., Spinelli, M., Wang, Q., Bai, X., Qin, Z., Chen, A.: Instantstyle: Free lunch towards style-preserving in text-to-image generation. ArXiv **abs/2404.02733** (2024), https://api.semanticscholar.org/CorpusID:268876474 3
11. Wang, P., Li, Y., Vasconcelos, N.: Rethinking and improving the robustness of image style transfer. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 124–133 (2021), https://api.semanticscholar.org/CorpusID:233209896 3