Supplementary Material for ZipLoRA: Any Subject in Any Style by Effectively Merging LoRAs

Viraj Shah[1,2], Nataniel Ruiz[1], Forrester Cole[1], Erika Lu[1], Svetlana Lazebnik[2], Yuanzhen Li[1], and Varun Jampani[1]

[1] Google Research
[2] UIUC

Table of Contents

1 Additional Implementation Details

In this section, we provide additional implementation details for our algorithm:

- We train both the base LoRAs corresponding to style and content using standard DreamBooth protocol on SDXL made available by diffusers python library [1]. For training, we use 1000 fine-tuning steps with batch size 1 and learning rate $5e-5$. We do not train text encoders during such fine-tuning. Further, we use rank$=64$ for obtaining both the style and content LoRAs.
- We do not use SDXL refiner model in any of our experiments, neither for training nor for inference.
- For ZipLoRA, we initialize the merger coefficient vectors with all ones. This is a natural way to initialize, since it imitates the direct merge at the beginning of the training, and gradually update the merger coefficients to minimize the alignment term along with maintaining the capability to generate both the individual concepts.
- For ZipLoRA fine-tuning, we use $\lambda = 0.01$ and learning rate 0.01 for all our experiments.
- We keep the number of diffusion inference steps fixed to 50 in all our experiments.

2 Performance of ZipLoRA on Stable Diffusion

As discussed in the main paper, LoRA fine-tuning on earlier version of Stable Diffusion (SDv1.5) fails to capture the stylization faithfully, thus the performance of ZipLoRA on SDv1.5 becomes limited by the stylization ability of the underlying style LoRA. That being said, our observations about sparsity and alignment of LoRA weights remain valid for other models of stable diffusion family, and even on SDv1.5, ZipLoRA outperforms competing methods (Direct Merge, Joint Training, Custom Diffusion, and Mix of Show) by achieving better stylizations with improved subject and style fidelity.

In this regard, we provide additional style-tuning results on SDv1.5 model in Fig. 1. One can see that the quality of the stylizations captured by SDv1.5 model is underwhelming as compared to that of SDXL.

Further, we also obtain custom stylizations by merging subject and style LoRAs using ZipLoRA on SDv1.5, and compare the results with Direct Merge, Joint Training, Custom Diffusion, and Mix of Show in Fig. 2. Note that we use SDv1.5 as a base model for these competing methods as well. For completeness, we also include the results for StyleDrop+DreamBooth. Note that StyleDrop is model-specific method that uses Muse as the base model, and since its code is not public, it is not possible to evaluate it on SDv1.5. As one can see in Fig. 2, even on SDv1.5, ZipLoRA produces superior stylization outputs and surpass all the competing methods.

We also provide quantitative evaluations on subject, style, and text alignment for our method and competing methods for SDv1.5 in Tab. 1.

Table 1: Alignment Scores for ZipLoRA on SDv1.5. While the stylization capabilities of SDv1.5 are inferior to SDXL, ZipLoRA still provides superior subject and text fidelity as compared to the existing methods when used on SDv1.5.

	ZipLoRA	Joint Training	Direct Merge	Mix of Show	Custom Diffusion	StyleDrop+ DreamBooth
	(on SDv1.5)	(on SDv1.5)	(on SDv1.5)	(on SDv1.5)	(on SDv1.5)	(on Muse)
Style-alignment ↑	**0.651**	0.579	0.581	0.618	0.574	0.646
Subject-alignment ↑	**0.413**	0.235	0.222	0.323	0.311	0.394
Text-alignment ↑	**0.283**	0.247	0.241	0.221	0.237	0.263

3 Additional Results

We provide additional results for experiments discussed in the main paper to present supporting evidence for the claims made.

Qualitative comparisons for personalized stylization on SDXL. We provide additional qualitative comparison of our method with Direct Merge,

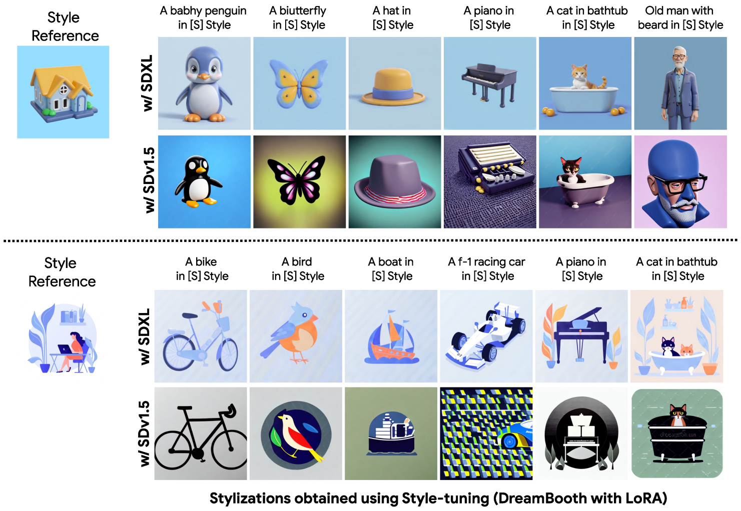

Style Reference
A babhy penguin in [S] Style
A biutterfly in [S] Style
A hat in [S] Style
A piano in [S] Style
A cat in bathtub in [S] Style
Old man with beard in [S] Style
w/ SDXL
w/ SDv1.5
Style Reference
A bike in [S] Style
A bird in [S] Style
A boat in [S] Style
A f-1 racing car in [S] Style
A piano in [S] Style
A cat in bathtub in [S] Style
w/ SDXL
w/ SDv1.5
Stylizations obtained using Style-tuning (DreamBooth with LoRA)

Fig. 1: Comparison between SDv1.5 and SDXL for Style Learning using DreamBooth with LoRA. SDXL model (top row) produces superior quality outputs when fine-tuned on a single example of a reference style (left-most column) using LoRA with a DreamBooth objective. Notice that SDv1.5 (bottom row) fails to capture the reference style consistently.

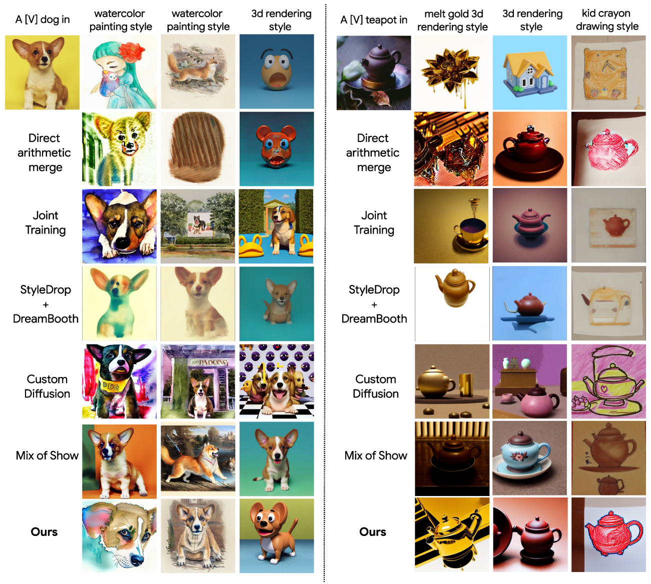


Fig. 2: Performance of ZipLoRA on SDv1.5. Even on SDv1.5, ZipLoRA outperforms Direct Merge, Joint Training, StyleDrop+DreamBooth, Custom Diffusion, and Mix of Show.

Joint Training, StyleDrop+DreamBooth, Custom Diffusion, and Mix of Show in Fig. 3. Superior results obtained using ZipLoRA further strengthens the claims of improved performance over the baselines.

Evidence that LoRA updates are sparse. In Fig. 4, we present more evidence for our claim that the LoRA updates are sparse in general, and significant chunk of low magnitude elements can be thrown away without affecting the stylization performance. As one can see, the stylization performance remains unaffected even when 80% of the elements are thrown away, while stylization degrades if this number is increased further.

Additional results of style-tuning using SDXL. We also provide additional results on Style-tuning property of SDXL model in Fig. 5. As shown, SDXL model can learn to generate stylized images in a given style through simple application of DreamBooth method without requiring any human feedback.

4 User Studies Details

We conduct user studies for a quantitative comparison of our method with existing approaches. We cast it as a binary comparison task thus conduct separate study for each pair of methods. We used Google Forms to conduct the user studies (See the user interface of our study in Fig. 6). In our study, each participant is shown a reference subject and a reference style along with outputs of two methods being compared and asked which output best depicts the reference style while preserving the reference subject fidelity. Options A and B are flipped randomly for each question. Participants for the study are selected at random from a pool of volunteers. For every study, each participant is asked 8 questions, thus each participant answers 40 questions in total across 5 studies that we conducted. 45 participants responded to our study, resulting in 360 responses for each of the five studies (1800 responses in total). As indicated in the results table in the main paper, our method is preferred over completing methods in all the five studies.

5 Societal Impact

This project empowers users to personalize both the subject and the artistic style of their images, featuring individual subjects like animals or objects, and styles like watercolor or sketch. It is important to acknowledge that, similar to other generative models and image editing techniques, this technology could be misused to create deceptive content. Addressing these potential ethical concerns remains an ongoing priority in the field of generative modeling, particularly with regards to image manipulation.

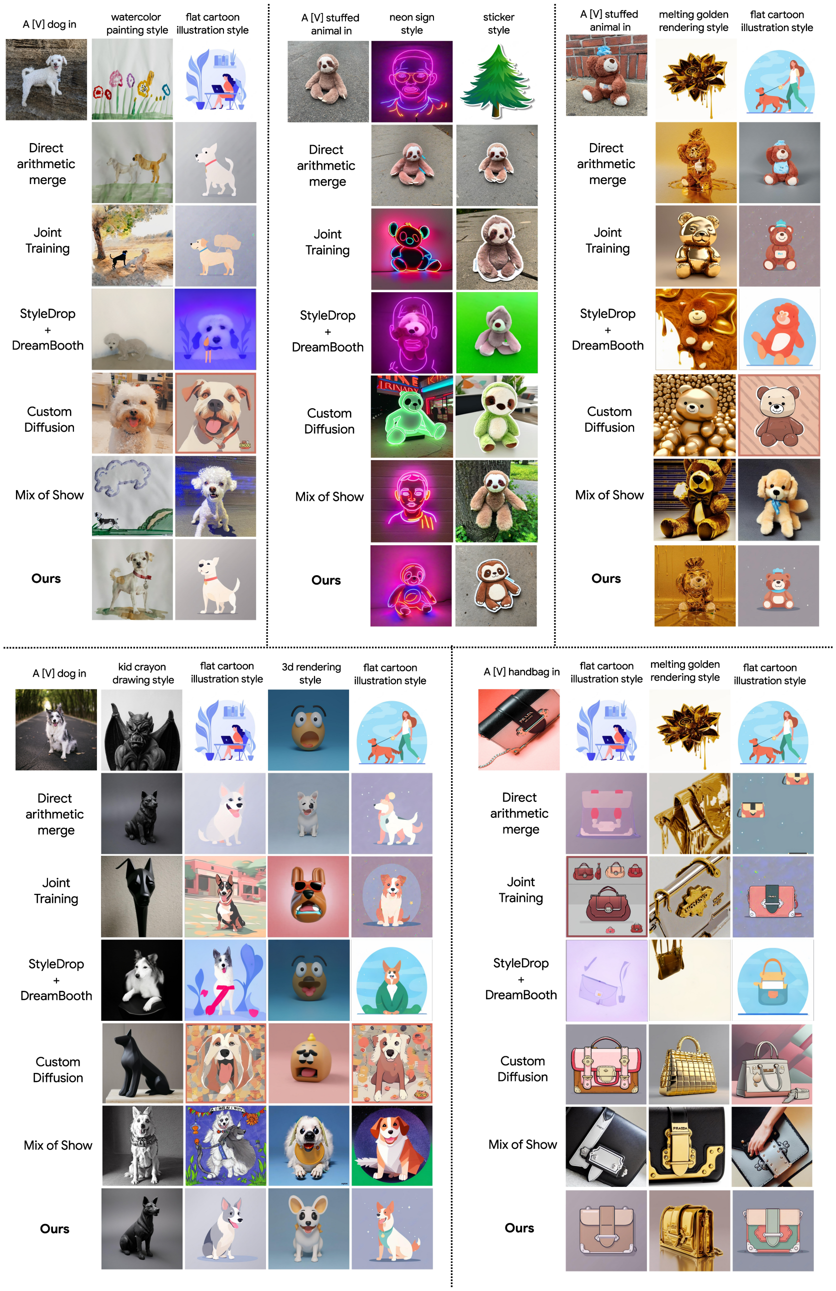


Fig. 3: Additional Qualitative Comparisons for Personalized Stylizations using ZipLoRA on SDXL. We compare samples from our method (Ours), versus Direct Arithmetic Merge, Joint Training, StyleDrop+DreamBooth, Custom Diffusion, and Mix of Show. We observe that our method achieves strong style and subject fidelity that surpasses competing methods.

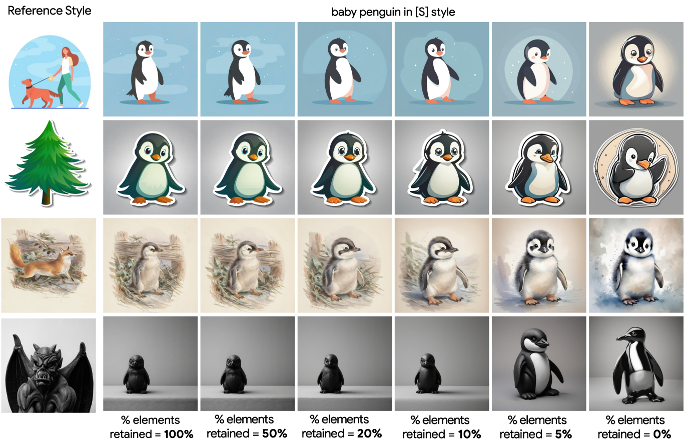

Reference Style
baby penguin in [S] style
% elements retained = 100%
% elements retained = 50%
% elements retained = 20%
% elements retained = 10%
% elements retained = 5%
% elements retained = 0%

Fig. 4: Additional Results: LoRA weight matrices are sparse. Most of the elements in ΔW have a magnitude very close to zero, and can be conveniently thrown away without affecting the generation quality of the fine-tuned model. The stylization quality is maintained even when only 20% of the elements are retained.

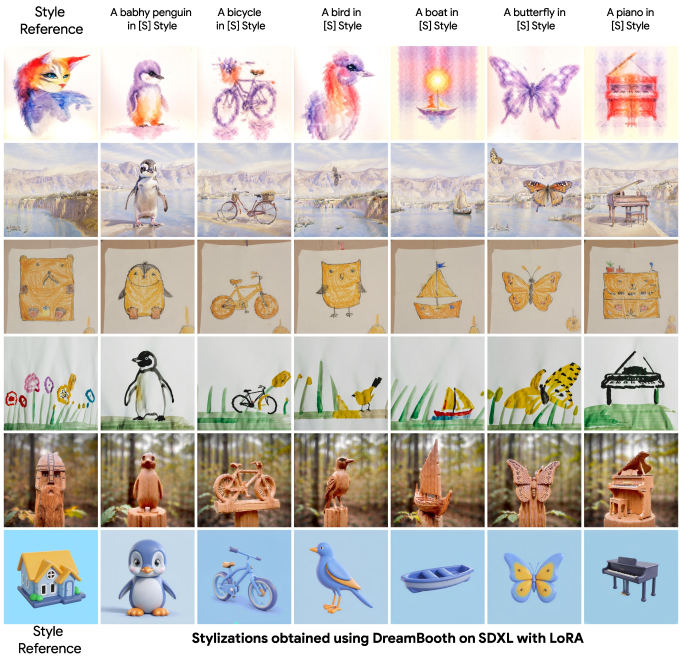
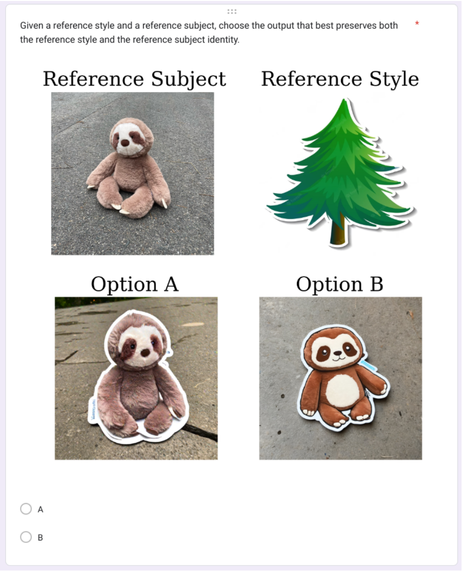


Fig. 5: Additional Results for Style Learning using DreamBooth on SDXL. SDXL model learns to produce stylized outputs when fine-tuned on a single example of a reference style (left-most column) using LoRA with a DreamBooth objective. Note that unlike StyleDrop, SDXL DreamBooth fine-tuning does not require human feedback.



Given a reference style and a reference subject, choose the output that best preserves both the reference style and the reference subject identity. *

○ A

○ B

Fig. 6: User Interface of Our User Studies. Each participant is shown a reference subject and a reference style along with outputs of two methods being compared and asked which output best depicts the reference style while preserving the reference subject fidelity.

6 Datasets and Image Attributions

We use style and content images from the datasets collected by StyleDrop [3] and DreamBooth [2] respectively. Note that these datasets do not contain any human subjects data or personally identifiable information. We provide image attributions below for each image that we used in our experiments. We refer readers to manuscripts and project websites of StyleDrop and DreamBooth for more detailed information about the usage policy and licensing of these images.

6.1 Image attributions for style references

StyleDrop project webpage provides the image attribution information here: Style Image Attribution

Specifically, the sources of the style images that we used in our experiments are as follows (linked as hyperlinks):

1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18

6.2 Image attributions for content references

DreamBooth project webpage provides the image attribution information here: Content Image Attribution

Specifically, the sources of the content images that we used in our experiments are as follows (linked as hyperlinks):

1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15

References

1. von Platen, P., Patil, S., Lozhkov, A., Cuenca, P., Lambert, N., Rasul, K., Davaadorj, M., Wolf, T.: Diffusers: State-of-the-art diffusion models. `https://github.com/huggingface/diffusers` (2022)
2. 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)
3. Sohn, K., Ruiz, N., Lee, K., Chin, D.C., Blok, I., Chang, H., Barber, J., Jiang, L., Entis, G., Li, Y., et al.: Styledrop: Text-to-image generation in any style. arXiv preprint arXiv:2306.00983 (2023)