SVD*iff*: Compact Parameter Space for Diffusion Fine-Tuning (Supplementary Material)

Ligong Han[1,2*] Yinxiao Li[2] Han Zhang[2] Peyman Milanfar[2] Dimitris Metaxas[1] Feng Yang[2]
[1]Rutgers University [2]Google Research

A. Implementation Details

Implementation The original DreamBooth was implemented on Imagen [12] and we conduct our experiments based on its StableDiffusion [9] implementation [16]. DreamBooth [10] and Custom Diffusion [7] are implemented in StableDiffusion with Diffusers library [15]. For LoRA [6], we use our own implementation for fair comparison, in which we also fine-tune the 1-D weight kernels, and use rank-1 for 2-D and 4-D weight kernels. This results in a slightly larger delta checkpoint (of size 5.62MB) than the official LoRA implementation [11].

Learning rate Our experiments show that the learning rate for these spectral shifts needs to be much larger (1,000 times, *e.g.* 10^{-3}) than the learning rate used for fine-tuning the full weights. For 1-D weights that are not decomposed, we use either the original learning rate of 10^{-6} to prevent overfitting or a larger learning rate to allow for a more rapid adaptation of the model, depending on the desired trade-off between stability and speed of adaptation.

B. Cross-Attention Regularization

Multi-subject We explore an extension of adding regularization to the cross-attention maps during fine-tuning. The resulting pipeline is illustrated in Fig. 12. This regularization is motivated by the visualization of the cross-attention maps of fine-tuned models. As shown in Fig. 24, the dog's special token ("sks") attends largely to the panda. Therefore, we use MSE on non-corresponding regions of the cross-attention maps to enforce separation between the two subjects. Intuitively, this loss promotes the dog's special token to focus solely on the dog and vice versa for the panda. Our initial experiment on full weight fine-tuning shows that adding this regularization greatly eliminates the stitching artifact.

Single-subject Our observations also show that the cross-attention map associated with the special token may attend to unwanted areas, even in the case of single-subject generation. For instance, as shown in Fig. 20, the attention of the special token "$[V_1]$" leaks to the background (whereas the attention of "$[V_2]$" does not). To address this issue, we explore the use of regularization on cross-attention maps to improve single-subject generation. The main idea is to limit the attention of the special token to be no more spread-out than that of the coarse class token (e.g. "dog"). To achieve this, we first obtain a binary mask M_t indicating the subject by thresholding the coarse class token's attention map. Then, we add a L2 regularization loss on the special token's attention map A_t^V, as follows:

$$\mathcal{L}_{reg} = ||A_t^V - \texttt{sg}(A_t^V \odot M_t)||_2^2, \tag{8}$$

where $\odot$ denotes elementwise multiplication and $\texttt{sg}$ is a stop gradient operator. The results of using this CA regularization are compared to the case without regularization in Fig. 19, and as expected, the regularization reduces overfitting to the background.

C. Single Image Editing

DDIM Inversion We show comparisons of with and without DDIM inversion [13] using ours ("SVD"), LoRA [6] ("LoRA"), and DreamBooth ("Full") on single-image editing in Fig. 22. If inversion is not used, DDIM sampler with $\eta = 0$ is applied. If inversion is employed, we use DDIM sampler with $\eta = 0.5$ and $\alpha = 0$, except for edits in Fig. 22-(d,f,h) where $\eta = 0.9$ and $\alpha = 0.9$. Interestingly, for the chair example (row 2) we need to inject large amount of noise to get desired edits. For other edits (a,b,e,g,i,j) DDIM inversion improves editing quality and alignment with input images for "SVD", but makes results worse for "Full" in edits (b,g,i) and for "LoRA" in edits (b,i). We can conclude that DDIM inversion improves editing quality and alignment with input images for non-structural edits when using our spectral shift parameter space. We also observe that LoRA in general tends to underfit the input image, as shown in (c,d,e,i) (without inversion).

Comparison with other methods Furthermore, we compare our method with the popular Instruct-Pix2Pix [1] in Fig. 23 (marked as "ip2p"). The comparison is *not* entirely fair as Instruct-Pix2Pix does not require fine-tuning on individual images. Nevertheless, it is worth investigating fast

*Work done during an internship at Google Research.

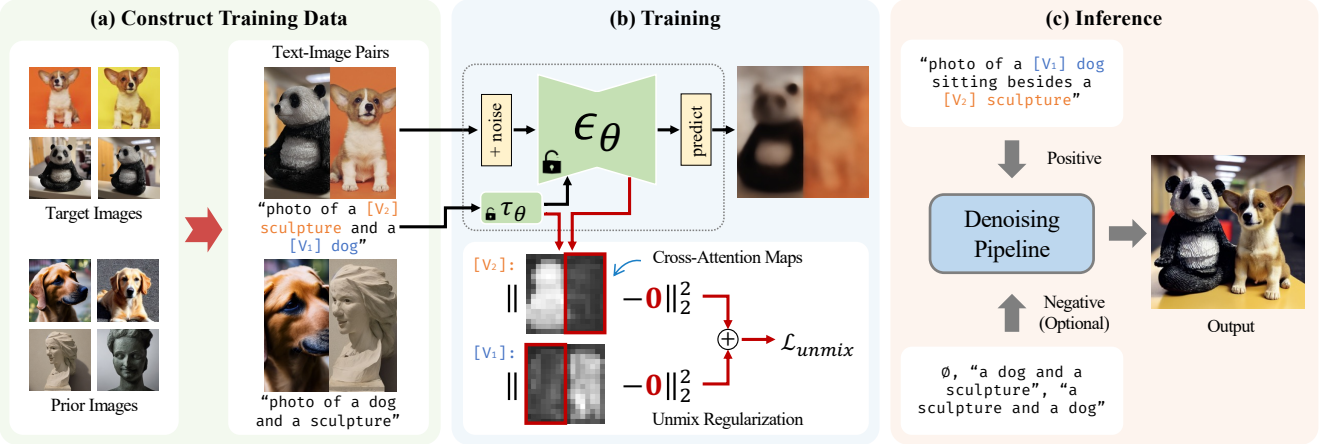


Figure 12: Cut-Mix-Unmix data-augmentation for **multi-subject generation**. The figure shows the process of Cut-Mix-Unmix data augmentation for training a model to handle multiple concepts. The method involves **(a)** manually constructing image-prompt pairs where the image is created using a CutMix-like data augmentation [17] and the corresponding prompt is written as, for example, "photo of a $[V_2]$ sculpture and a $[V_1]$ dog". The prior preservation image-prompt pairs are created in a similar manner. The objective is to train the model to separate different concepts by presenting it with explicit mixed samples. **(b)** To perform unmix regularization, we use MSE on non-corresponding regions of the cross-attention maps to enforce separation between the two subjects. The goal is to encourage that the dog's special token should not attend to the panda and vice versa. **(c)** During inference, a different prompt, such as "photo of a $[V_1]$ dog sitting besides a $[V_2]$ sculpture".

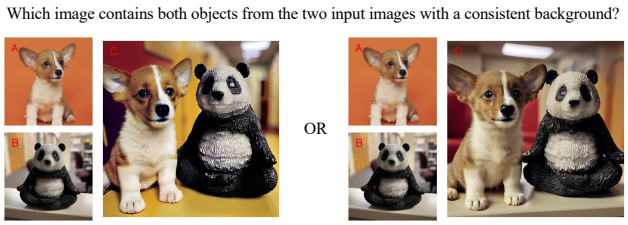


Figure 13: Example of human evaluation. Each method is represented by a collage of images with two real images on the left (labeled "A" and "B") and one synthesized image on the right (labeled "C").

personalized adaptation and avoiding per-image fine-tuning in future work.

Subject Combinations	Human Preference (SVD *vs.* Full)
Teddy + Tortoise	**53.2**% : 46.8%
Dog + Cat	**62.9**% : 37.1%
Dog + No-Face	**65.0**% : 35.0%
Dog + Panda	**62.0**% : 38.0%

Table 2: Human evaluation results comparing "SVD" and "Full" for different subject combinations, with 1000 human ratings for each combination.

D. Multi-Subject Generation

User study In Tab. 2, we present the results of human evaluation comparing our method ("SVD") and the full weight fine-tuning method ("Full"). For each of the four subject combinations, 1000 ratings were collected. Participants were shown two generated images side-by-side and were asked to choose their preferred image or indicate that it was "hard to decide" (4.1%, 1.2%, 2.1%, and 2.2% respectively). Visual examples are given in Fig. 13.

Analysis of Cut-Mix-Unmix In this section, we present additional analysis of the Cut-Mix-Unmix data augmentation technique (without unmix regularization on the cross-attention maps). Fig. 18 illustrates the results of the default "left and right" augmentations, which still generate meaningful relations such as (a) "wear", (b) "in", and (c) "ride". In the case of (a), "full" overfits to the augmentation layout. In our initial experiments, we also randomly split the left and right images and observed similar results as with a fixed 1:1 ratio. In (d), the "up and down" augmentation exhibits similar behavior to "left and right". Nevertheless, we concur that introducing a random layout (particularly with our proposed cross-attention regularization) could further mitigate overfitting. We leave this study for future work.

Negative prompt To perform negative prompting, we repurpose the prior-preservation prompts as negative prompts $\mathbf{c}^{neg}$. Recall that *Classifier-free guidance* (CFG) [5] extrap-

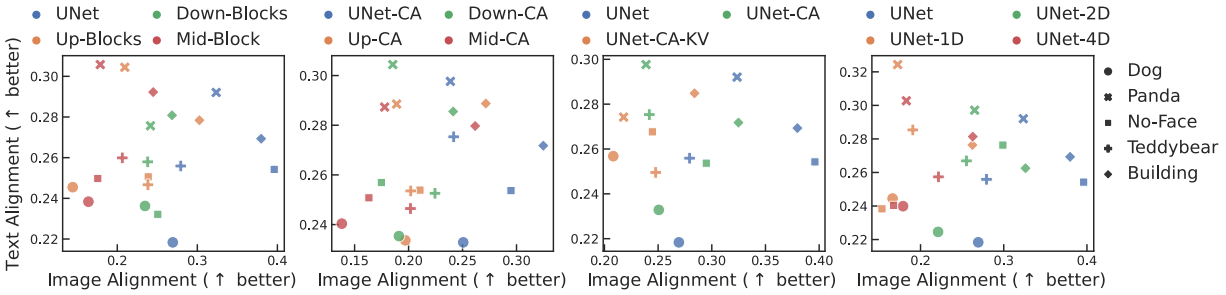
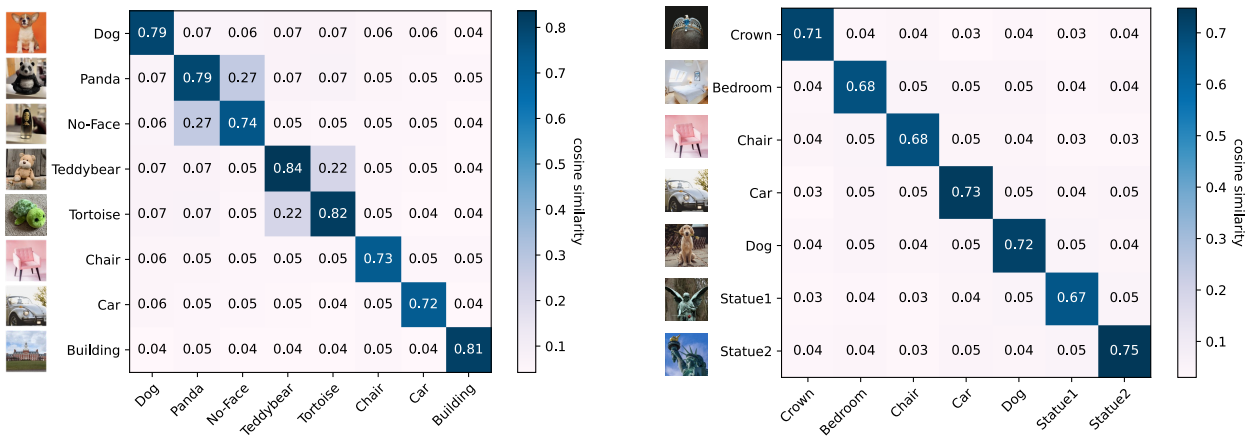


Figure 14: Text- and image-alignment scores for single-subject generation. We perform SVDiff fine-tuning on 12 subsets of UNet layers across 5 subjects.



(a) Single Subject (b) Single Image

Figure 15: Correlation of individually learned spectral shifts for different subjects/images. The cosine similarities between the spectral shifts of two subjects are averaged across all layers and plotted. The diagonal shows average similarities between two runs with different learning rates. High similarities are observed between conceptually similar subjects.

olates the conditional score by a scale factor $s > 1$,

$$\hat{\boldsymbol{\epsilon}}_{\theta,s}(\mathbf{z}_t|\{\mathbf{c},\emptyset\}) = s \cdot \hat{\boldsymbol{\epsilon}}_\theta(\mathbf{z}_t|\mathbf{c}) + (1-s) \cdot \hat{\boldsymbol{\epsilon}}_\theta(\mathbf{z}_t|\emptyset). \quad (9)$$

Similar to [14], we replace the null-conditioned score $\hat{\boldsymbol{\epsilon}}_\theta(\mathbf{z}_t|\emptyset)$ in Eq. (9) by $\tilde{\boldsymbol{\epsilon}}_{\theta,\beta}$ defined as following,

$$\tilde{\boldsymbol{\epsilon}}_{\theta,\beta}(\mathbf{z}_t|\{\mathbf{c}^{neg},\emptyset\}) = \beta \cdot \hat{\boldsymbol{\epsilon}}_\theta(\mathbf{z}_t|\emptyset) + (1-\beta) \cdot \hat{\boldsymbol{\epsilon}}_\theta(\mathbf{z}_t|\mathbf{c}^{neg}) \quad (10)$$

where $0 < \beta < 1$. This can be easily extended to including multiple negative prompts. Fig. 21 shows a few examples of using negative prompts to remove the stitching artifacts introduced by Cut-Mix-Unmix. We hypothesize that this is because the model is trained to associate the prior prompt to the stitching style so negative prompting can help removing the stitching edges. However, we observe that negative prompting may not always help.

Extensions We show a preliminary extension of Cut-Mix-Unmix to Attend-and-Excite [2]. As shown in Fig. 17, Cut-Mix-Unmix helps better disentangle respective visual features of the dog and the cat. It is also possible to extend and integrate our method to other attention-based methods [4, 14, 3].

E. Analysis on Spectral Shifts

Fine-tuning with fewer steps Here we show results of fast adaptation for single subject generation in Fig. 27. This setting is slightly different from the experiments in the main text since we limit the fine-tuning steps as 100 without prior-preservation loss [10] (for main results we fine-tune 500-1000 steps with prior-preservation loss). Thus we tune the learning rate for each method to balance between faithfullness and realism [8]. The learning rates we used are as follows:

- **SVDiff**: 1-D weights 2×10^{-3}, 2-D and 4-D weights 5×10^{-3}

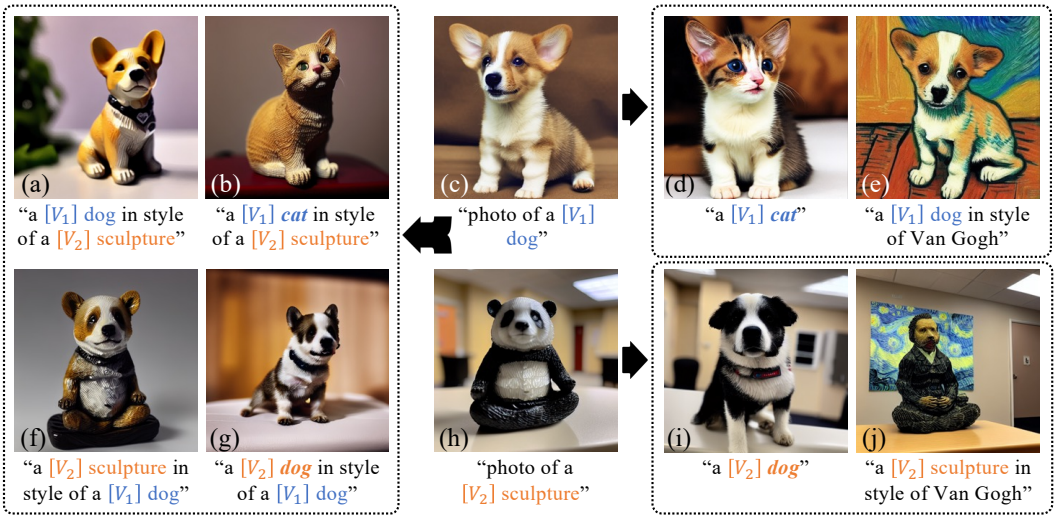


Figure 16: Simple style transfer results with SVD*iff*. Changing coarse class word: (d) and (i); Appending "in style of": (e) and (j); Combined spectral shifts: (a,b,f,g).

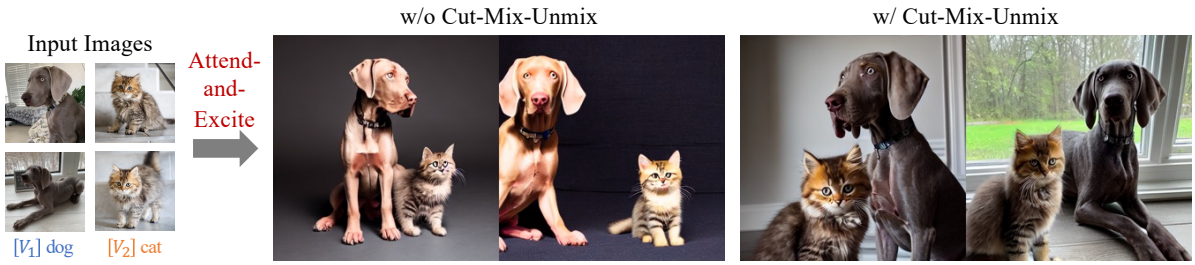


Figure 17: Results for Cut-Mix-Unmix with Attend-and-Excite [2]. Even without Cut-Mix-Unmix, Attend-and-Excite successfully separate the dog and the cat by design, despite the color of the cat is slightly leaked to the dog. Cut-Mix-Unmix helps better disentangle respective visual features of the dog and the cat.

- **LoRA** [6]: 1-D weights 2×10^{-3}, 2-D and 4-D weights 1×10^{-4}
- **DreamBooth** [10]: 1-D weights 1×10^{-3}, 2-D and 4-D weights 5×10^{-6}

In Fig. 27, the performance comparison of our method, LoRA [6] and DreamBooth [10] is shown under fast fine-tuning setting. The results indicate that all three methods perform similarly, except for the "No-Face" sculpture in (c) where LoRA shows underfitting and DreamBooth exhibits overfitting. In (e), SVDiff also shows overfitting, which could be a result of the large learning rate used.

Rank Fig. 25 shows the results of limiting the rank of the spectral shifts of 2-D and 4-D weight kernels during training. Two examples are shown for each of the three subjects, one with the training prompt (to "reconstruct" the subject) and one with an edited prompt. Results show that the model can still reconstruct the subject with rank 1, but may struggle to capture details with an edited prompt when the rank of spectral shift is low. The visual differences between reconstructed and edited samples are smaller for the Teddybear than the building and panda sculpture, potentially because the pre-trained model already understands the concept of a Teddybear.

Correlations We present the results of the correlation analysis of individually learned spectral shifts for each subject in Fig. 15. Each entry in the figure represents the average cosine similarities between the spectral shifts of two subjects, computed across all layers. The diagonal entries show the average cosine similarities between two runs with the learning rate of 1-D weights set to 10^{-3} and 10^{-6}, respectively. The results indicate that the similarity between conceptually similar subjects is relatively high, such as between the "panda" and "No-Face" sculptures or between the "Teddybear" and "Tortoise" plushies.

Scaling Fig. 26 demonstrates the effect of scaling spectral shifts (labeled as "SVD", $\Sigma_{\boldsymbol{\delta}'} = \text{diag}(\text{ReLU}(\boldsymbol{\sigma} + s\boldsymbol{\delta}))$ with scale s) and weight deltas (marked as "full", $W' = W + s\Delta W$ with scale s). Samples are generated using the same random seed. Scaling both the spectral shift and full

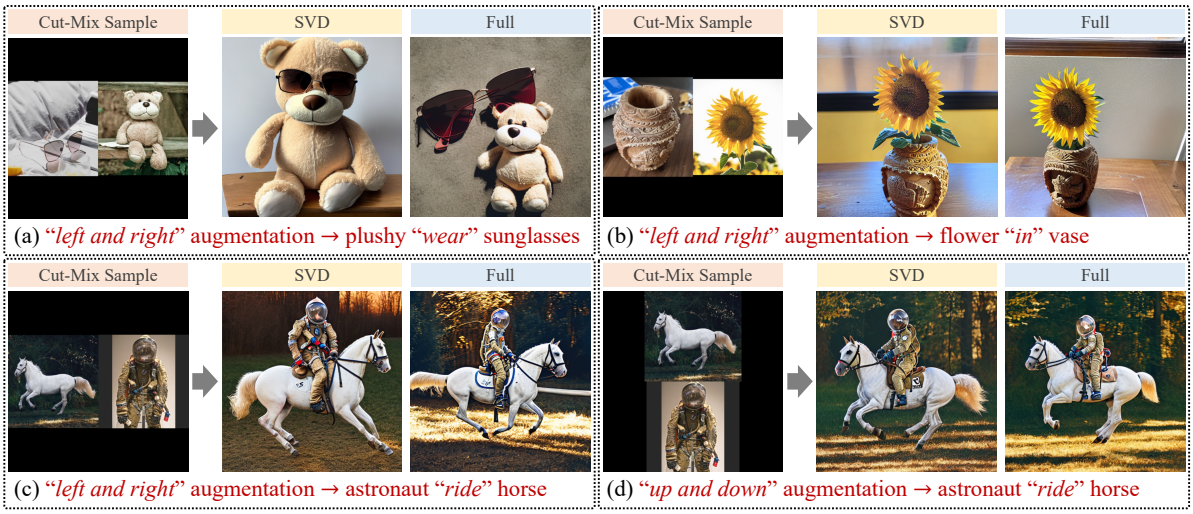


Figure 18: Additional analysis of Cut-Mix data augmentation.

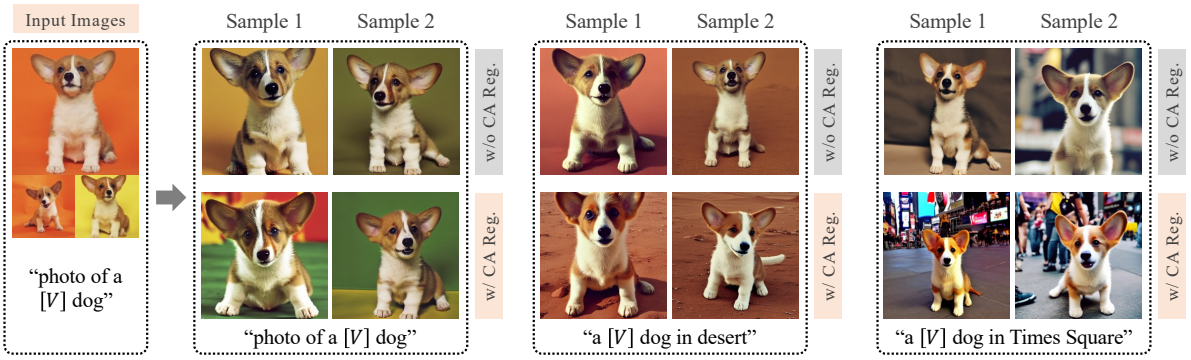


Figure 19: Comparison of Cross-Attention (CA) Regularization. The results show the effectiveness of the CA regularization in reducing overfitting to the background. The models were fine-tuned using 800 steps with the prior-preservation loss and all comparisons were generated using the same random seed.

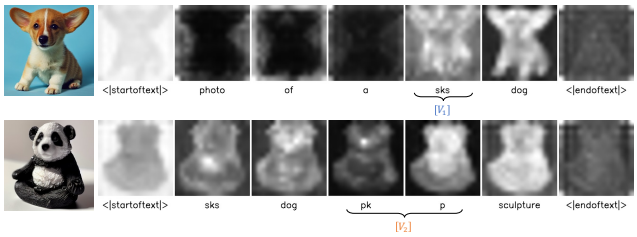


Figure 20: Analysis of cross-attention maps of the fine-tuned model. As shown, the dog's special token ("sks") also attends to background areas.

weight delta affects the presence of personalized attributes and features. The results show that scaling the weight delta also influences attribute strength. However, a scale value that is too large (*e.g.* $s = 2$) can cause deviation from the text prompt and result in dark samples.

Style transfer We demonstrate the effect of style transfer using our proposed method. We show that by using a single fine-tuned model, the personalized style can be transferred to a different class by changing the class word during inference, or by adding a prompt such as *"in style of"*. We also show that by summing two sets of spectral shifts (as discussed above), their styles can be mixed. The results show different outcomes of different style-mixing strategies, with changes to both the class and personalized style.

F. Image Attribution

Avocado plushy: https://unsplash.com/photos/8V4y-XXT3MQ.

Pink chair: https://unsplash.com/photos/1JJJIHh7-Mk.

Brown and white puppy: https://unsplash.com/photos/brFsZ7qszSY, https://

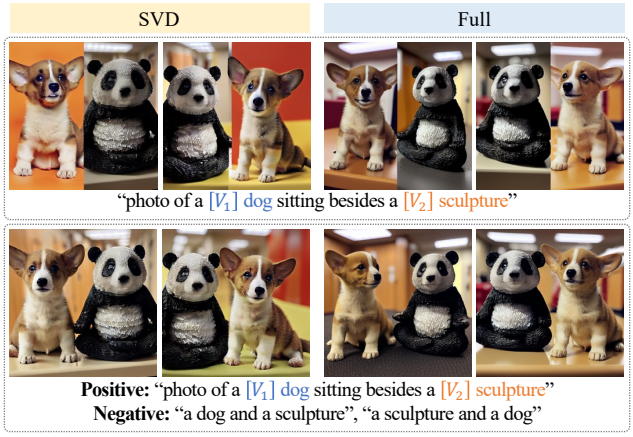


Figure 21: Using negative prompt helps to remove stitching artifact for both "SVD" and "Full".

unsplash.com/photos/eoqnr8ikwFE, https://unsplash.com/photos/LHeDYF6az38, and https://unsplash.com/photos/9M0tSjb-cpA.

Crown: https://unsplash.com/photos/8Dpi2Mb1-PM.

Bedroom: https://unsplash.com/photos/x53OUnxwynQ.

Dog with flower: https://unsplash.com/photos/Sg3XwuEpybU.

Statue-of-Liberty: https://unsplash.com/photos/s0di82cRiUQ.

Beetle car: https://unsplash.com/photos/YEPDV3T8Vi8.

Building: https://finmath.rutgers.edu/admissions/how-to-apply and luvemakphoto/ Getty Images.

Teddybear, tortoise plushy, grey dog, and cat images are taken from Custom Diffusion [7]: https://www.cs.cmu.edu/~custom-diffusion/assets/data.zip.

References

[1] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. *arXiv preprint arXiv:2211.09800*, 2022. 1, 8

[2] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *arXiv preprint arXiv:2301.13826*, 2023. 3, 4

[3] Weixi Feng, Xuehai He, Tsu-Jui Fu, Varun Jampani, Arjun Akula, Pradyumna Narayana, Sugato Basu, Xin Eric Wang, and William Yang Wang. Training-free structured diffusion guidance for compositional text-to-image synthesis. *arXiv preprint arXiv:2212.05032*, 2022. 3

[4] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 3, 8

[5] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 2

[6] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 1, 4

[7] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1, 6

[8] Chenlin Meng, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021. 3

[9] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, June 2022. 1

[10] 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. *arXiv preprint arXiv:2208.12242*, 2022. 1, 3, 4, 10

[11] Simo Ryu. Low-rank adaptation for fast text-to-image diffusion fine-tuning. https://github.com/cloneofsimo/lora. 1

[12] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022. 1

[13] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. 1, 7

[14] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. *arXiv preprint arXiv:2211.12572*, 2022. 3

[15] Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers, 2022. 1

[16] Xavierxiao. Xavierxiao/dreambooth-stable-diffusion: Implementation of dreambooth with stable diffusion. https://github.com/XavierXiao/Dreambooth-Stable-Diffusion. 1

[17] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032, 2019. 2

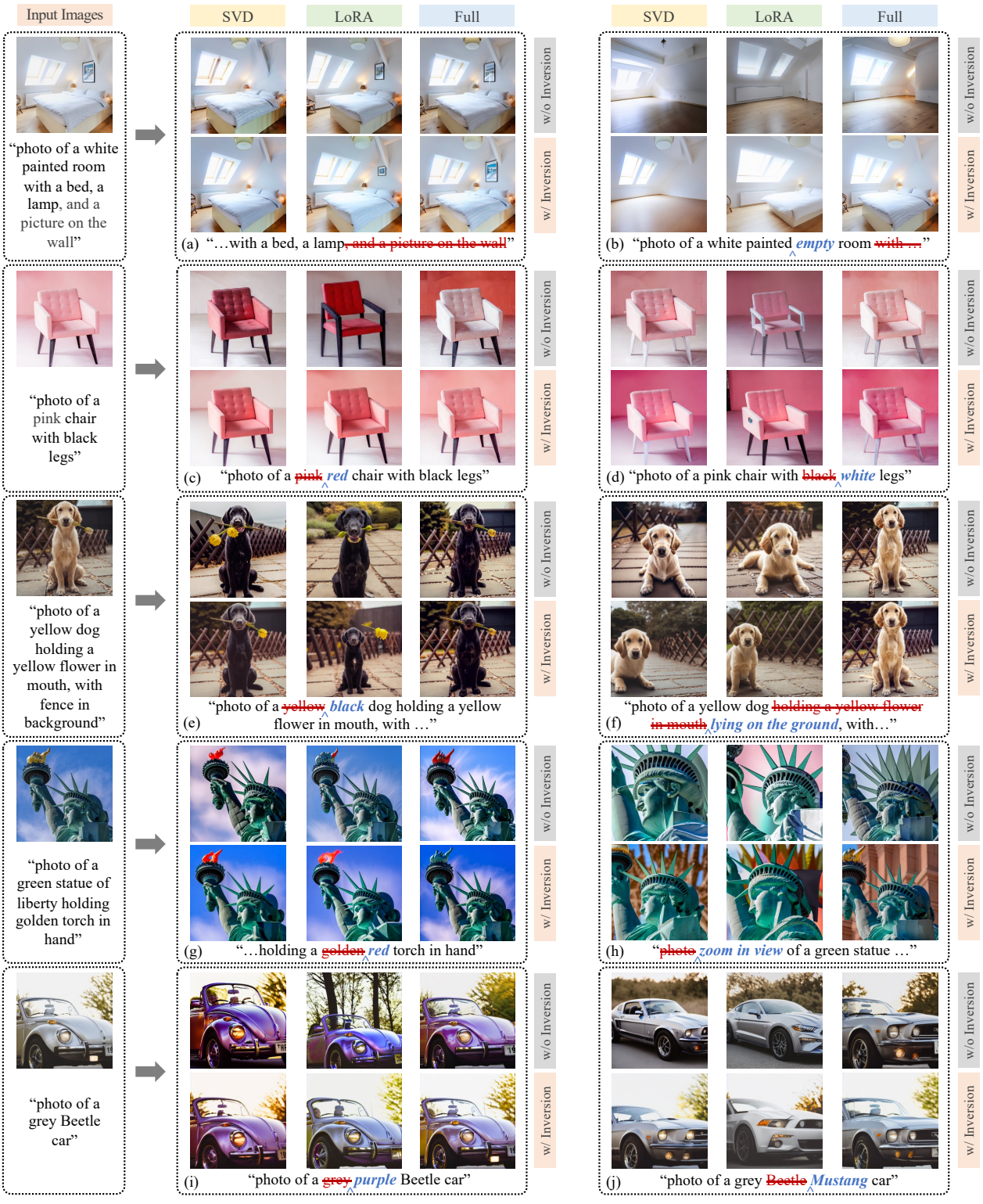


Figure 22: Results for single image editing with DDIM inversion [13]. If inversion is not used, DDIM sampler with $\eta = 0$ is applied. If inversion is employed, we use DDIM sampler with $\eta = 0.5$ and $\alpha = 0$ (α is for slerp defined in **??**), except for edits in (d,f,h) where $\eta = 0.9$ and $\alpha = 0.9$. Results show that DDIM inversion improves editing quality and alignment with input images for non-structural edits when using our spectral shift parameter space. As shown, DDIM inversion can have adverse effects on results for "Full" and "LoRA", *e.g.* (b) making the room empty, (i) changing the color to purple.

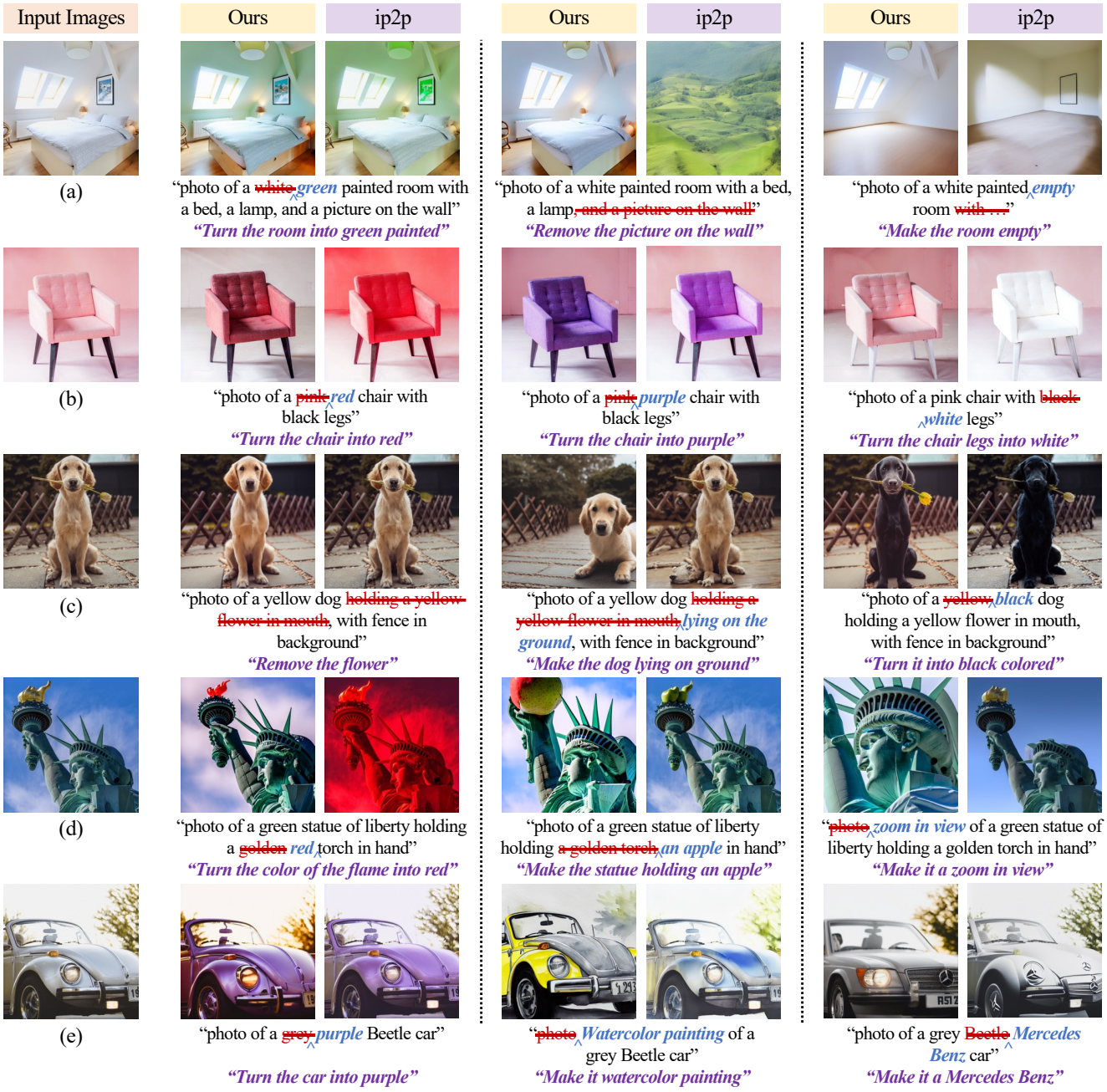


Figure 23: Comparison of our method and Instruct-Pix2Pix [1] (marked as "ip2p") on single-image editing. The instructions are displayed in bold and italicized purple text. Results show that Instruct-Pix2Pix tends to alter the overall color scheme and struggles with significant or structural edits, as seen in (a) emptying the room and (d) zoom-in view.

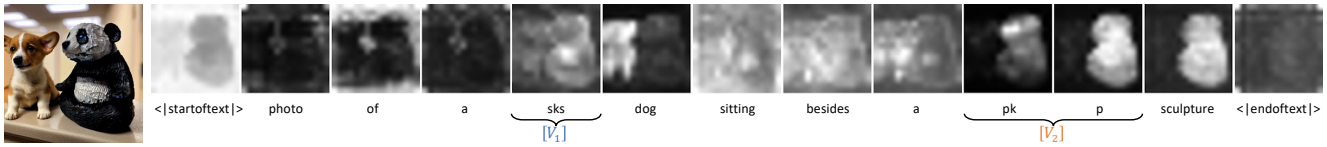


Figure 24: Analysis of cross-attention maps of the fine-tuned model without using unmix regularization. Visualization is obtained by Prompt-to-Prompt [4]. As shown, the dog's special token ("sks") attends largely to the panda.

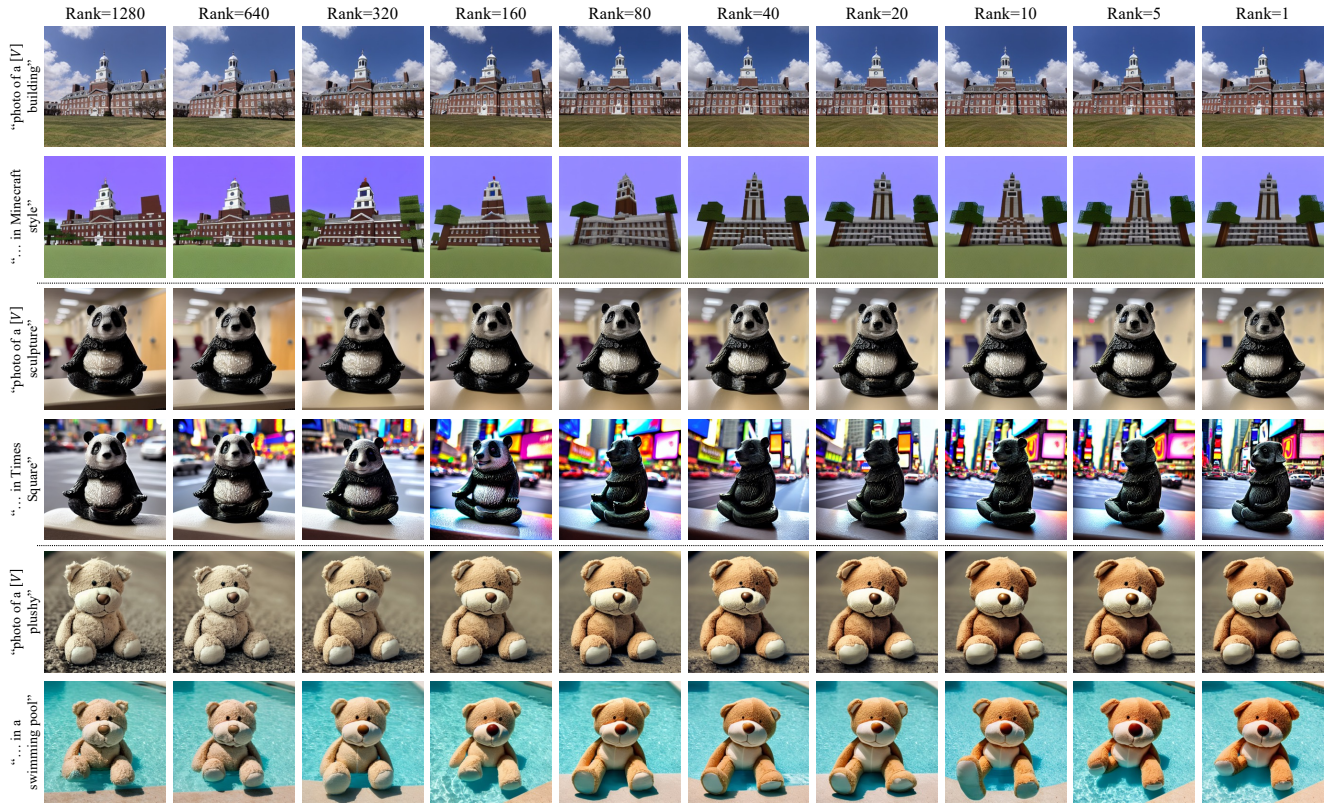


Figure 25: Effect of limiting rank of spectral shifts. The figure displays examples of the subject's reconstruction and edition with varying ranks of the spectral shifts. Results indicate that a lower rank leads to limited ability to capture details in the edited samples, with better performance observed for a subject that is easier for the model to adapt to (*i.e.* Teddybear).

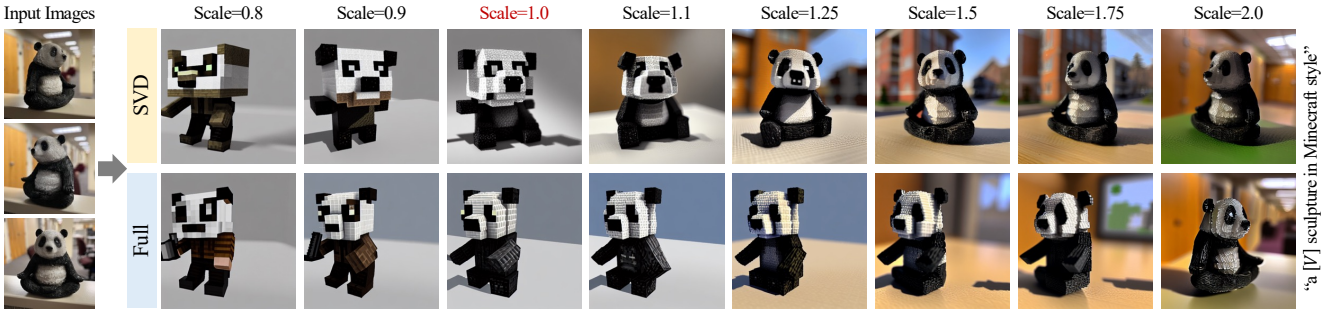


Figure 26: Effects of scaling spectral shifts ($\Sigma_{\boldsymbol{\delta}'} = \text{diag}(\text{ReLU}(\boldsymbol{\sigma} + s\boldsymbol{\delta}))$) and weight deltas ($W' = W + s\Delta W$). Note that this scale is different from the classifier-free guidance scale. Scaling both spectral shift and weight delta changes the attribute strength, with too large a scale causing deviation from the text prompt and visual artifacts.

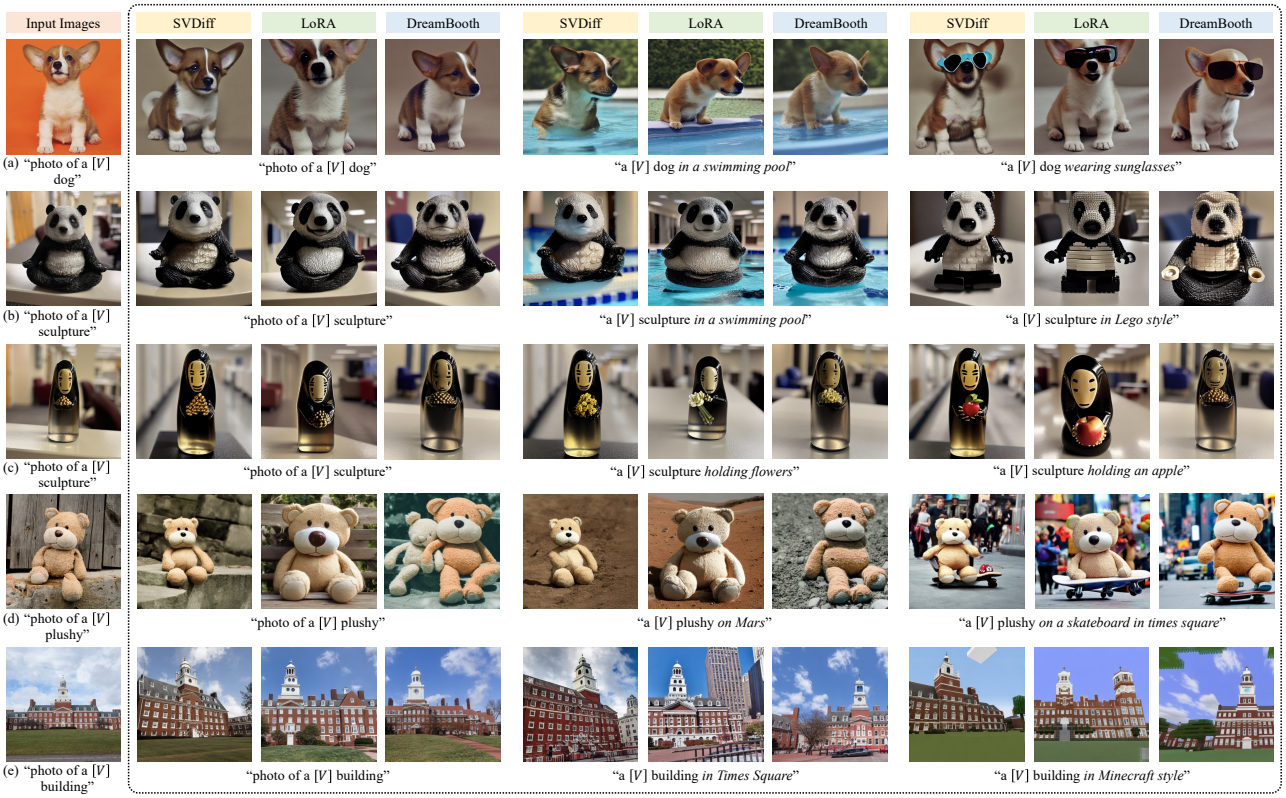


Figure 27: Single subject generation when fine-tuned with fewer steps. All models are fine-tuned for 100 steps without prior-preservation loss [10] (for main results we fine-tune 500-1000 steps with prior-preservation loss).

Figure 28: Visual samples of fine-tuning the spectral shifts of a subset of layers in the UNet.