Customizing and Enhancing Stable Diffusion for Style LoRA and Subject-Specific Generation

Katherine Chen
Department of Computer Science
Stanford University
kathchen@stanford.edu

1 Introduction

The intersection of artificial intelligence and creative expression has led to innovative applications, with text-to-image generation being a prominent domain. This project explores the customization and enhancement of Stable Diffusion for text-to-image generation, with a focus on two distinct aspects: stylized adaptation inspired by the works of digital artist Carla Cordelia and subject-specific generation featuring a stuffed animal named Erica.

Stable Diffusion [1] stands out as a cutting-edge latent diffusion model that utilizes text-to-image capabilities to generate exceptionally high-fidelity photos based on textual descriptions. Its open-source release in 2022 has played a pivotal role in driving the generative AI revolution, showcasing the potential of scaled-up score-based models.

In this project, I was motivated to explore the adaptability of Stable Diffusion within the text-to-image generation domain. Inspired by the unique artistic blend of real painting elements and anime aesthetics showcased by Carla Cordelia, I aimed to fine-tune Stable Diffusion using Low-Rank Adaptation to capture and replicate Cordelia's distinctive art style. My second motivation centered on subject-specific generation, specifically bringing my stuffed animal, Erica, to life through techniques like Textual Inversion and DreamBooth. I thought it would be cool to generate personalized and contextually rich depictions of Erica based on textual descriptions.

In terms of results, my implementation of Stable Diffusion with LoRA successfully generated images in Cordelia's style, achieving a CLIP score of 24.980 with the AdaFactor optimizer and a learning rate of 1e-4. My model exhibited a CLIP score 29% lower than the baseline Stable Diffusion model, which scored 35.505. However, qualitative examination revealed that the output images from my model clearly resembles Cordelia's style, suggesting that CLIP score may not comprehensively capture the success of fine-tuning efforts.

For subject-specific generation, I found that DreamBooth performed better than Textual Inversion. Qualitatively, DreamBooth produced sharper and more well-defined body shapes and facial features for Erica, showcasing versatility by rendering Erica in various styles, including anime. Conversely, Textual Inversion's output suffered from deformations that made it difficult for viewers to recognize Erica. Quantitatively, DreamBooth achieved a Structural Similarity Index (SSIM) of 0.364, outperforming Textual Inversion with a SSIM of 0.346. Both methods significantly surpassed the baseline Stable Diffusion model, which attained a minimal SSIM of 0.019.

2 Related Work

The original LoRA paper "LoRA: Low-rank adaptation of large language models" [2] provides a basis for the fine-tuning approach with Stable Diffusion. The authors propose learning a low rank adaptor matrix which gets added to frozen model weights to perform fine-tuning. These adaptor matrices are parameter efficient, allow higher training throughput, and add no additional inference latency for fine-tuned language models.

"An image is worth one word: Personalizing text-to-image generation using textual inversion" [3] introduces the concept of Textual Inversion, which is one of two methods I used for subject-specific generation. The authors show that using a few input images, the concept of the images can be encoded as a new "word" in the context of the model. Using this new "word," users are able to use natural language prompts to manipulate the concept with image language models.

Building on this, "DreamBooth: Fine-tuning text-to-image diffusion models for subject-driven generation" [4] provides insights into refining text-to-image diffusion models, and I also used this for subject-specific generation. The authors fine-tune image language models by using input image and textual prompt pairs. Through fine-tuning with these image-text pairs, the authors embed the subject of the input images with as a unique identifier in the model which is then able to be manipulated later.

My project contributes a unique perspective by introducing novel datasets and scenarios for the fine-tuning of Stable Diffusion. Unlike the prior research, which primarily focused on existing datasets or general concepts, my work delves into the customization and adaptation of Stable Diffusion using specific datasets inspired by the artistic style of Carla Cordelia, as well as personal subjects like Erica the stuffed animal cat.

3 Approach

3.1 Problem Description

There were two main problems of this project. The first was to fine-tune Stable Diffusion using Low-Rank Adaptation (LoRA) to capture Cordelia's unique art style. The second was to implement Textual Inversion and DreamBooth to generate personalized images of Erica based on textual descriptions.

3.2 Dataset

I collected two datasets, corresponding to the two problems described above. The Cordelia-Style Dataset (Figure 1) contains 122 images painted by the artist, sourced from her Instagram account. The images are 890 by 490 pixels, and were used for fine-tuning Stable Diffusion with LoRA. The Subject-Specific Dataset (Figure 2) contains 10 images featuring Erica in various angles, backgrounds, and poses. These images were 3024 by 4032 pixels, and were used for subject-specific generation with Textual Inversion and DreamBooth.

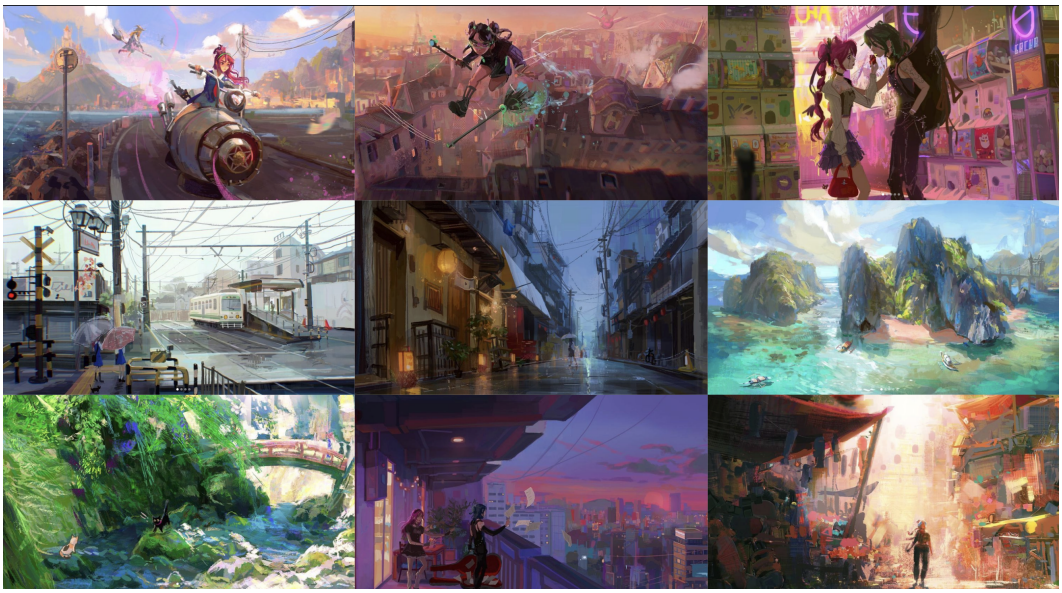

Figure 1: Example images in the Cordelia-Style Dataset.

Figure 2: Example images in the Subject-Specific Dataset.

3.3 Proposed Methods

3.3.1 Low-Rank Adaptation (LoRA)

For the first problem, I adopted a strategy of fine-tuning of a pre-trained Stable Diffusion model using LoRA, feeding it the Cordelia-Style Dataset consisting of 122 images. Stable Diffusion models, while powerful, often suffer from scalability issues and high computational requirements. LoRA was chosen as it offers a method to adapt the parameters of Stable Diffusion models with lower computational overhead, motivated by the fact that downstream fine-tunings have low intrinsic dimensions. LoRA achieves this by training the weight matrix W^{ft} to be in the form of:

$$W^{ft} = W + AB^T$$

where $W \in \mathbb{R}^{d_1 \times d_2}$ is the pretrained weights, which are kept frozen, and AB^T is the rank-r residual matrix, the only part that is fined-tuned, where $r << min(d_1, d_2)$ (Figure 3). During the fine-tuning process, the matrices $A \in \mathbb{R}^{d_1 \times r}$ and $B \in \mathbb{R}^{r \times d_2}$ are iteratively updated to minimize the original loss function of Stable Diffusion.

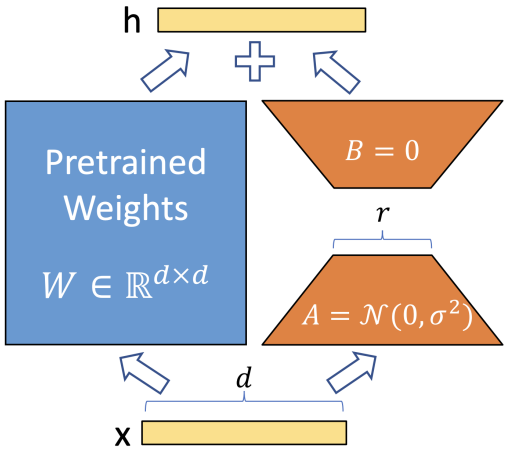

Figure 3: Reparameterization with LoRA. [2]

Implementation-wise, I followed the provided Kohya Trainer repository's starter code [5], utilizing the "Kohya LoRA Fine-Tuning" Jupyter notebook as a foundation. This notebook encapsulates the necessary procedures for loading the pre-trained Stable Diffusion model and integrating the LoRA fine-tuning. First, I uploaded the my Cordelia-Style Dataset. Next, I paired each image with textual descriptions using the provided Waifu Diffusion 1.4 Tagger. Then, I conducted the fine-tuning process with the preset LoRA network dimension of 32 and batch size of 6.

Furthermore, I experimented with various optimizers, including AdamW [6], SGDNesterov [7], and AdaFactor [8], which could be selected directly in the notebook, to identify the most suitable choice for my style-learning task at hand. Finally, I conducted hyperparameter tuning to search for an optimal learning rate.

3.3.2 Textual Inversion

Stable Diffusion models lets users create high-quality images guided by natural language prompts. But practically, users often seek finer control over the generated images, wanting to generate visuals corresponding to specific concepts or compose objects within novel scenes. Textual Inversion [3] provides a mechanism for this creative freedom. In this project, I am specifically interested in generating images of my subject, Erica, in diverse context using Textual Inversion.

During the Textual Inversion process, a text string containing a placeholder word undergoes tokenization. These tokens are then transformed into embeddings, and subsequently, a conditioning code is derived to guide the generative model. When optimizing the model, the embedding vector is refined through a reconstruction objective, aligning with the placeholder word. This approach allowed me to exert more nuanced control over the text-to-image diffusion model, giving it the ability to visualize Erica in imaginative scenes. Figure 4 presents the model architecture of Textual Inversion.

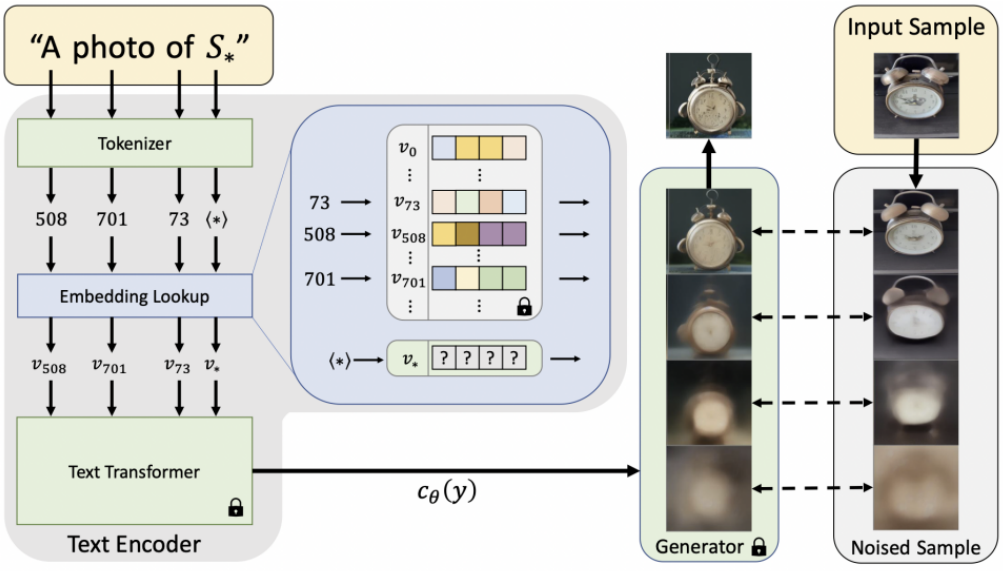

Figure 4: Textual Inversion architecture. [3]

I implemented Textual Inversion by using the Hugging Face Diffusers library, following the Jupyter notebook released by Hugging Face [9]. In this process, I uploaded my Subject-Specific Dataset consisting of the 10 images of Erica, inputted my placeholder token `<erica>`, loaded the Stable Diffusion model, added a new embedding vector in the token embeddings for `<erica>`, froze the rest of the model parameters except the newly added embedding vector, trained the model with a learning rate of 1e-04 and 2000 iterations, then evaluated the results both quantitatively and qualitatively.

3.3.3 DreamBooth

Another way that users can seek finer control over generated images of Stable Diffusion is via DreamBooth [4]. Each image is paired with a text prompt that includes a unique identifier `<erica>` and the class name, `cat-toy`, to which the subject belongs. DreamBooth simultaneously applies a class-specific prior preservation loss, which leverages the model's semantic understanding of the class, encouraging diverse image generation within the subject's class. Figure 5 shows the model architecture of DreamBooth.

In this project, I refined a Stable Diffusion model by fine-tuning it with the same set of 10 images of my subject Erica. I did this by following the DreamBooth tutorial from Hugging Face [10], which combines DreamBooth with LoRA to conserve memory usage. I used the tutorial's default learning rate, which was 1e-4, and fine-tuned the model over 1500 iterations.

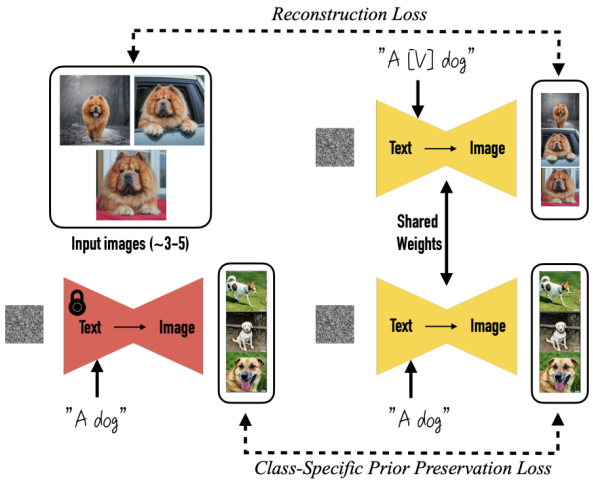

Figure 5: DreamBooth architecture. [4]

3.4 Expected Results

For both the first and second problems, the input are textual prompts. For the first problem, Cordelia-style fine-tuning, the input could be something like "A serene meadow under a starlit sky," and the expected results are generated images that match the prompt and closely mimic Cordelia's art style. For the second problem, subject-specific generation, an example input could be "Erica in a cozy reading nook with a cup of hot cocoa beside her," and the expected results are contextually rich and personalized images of Erica.

3.5 Evaluation Metrics

I compared the results of all three methods, LoRA, Textual Inversion, and DreamBooth, against my baseline of a pre-trained Stable Diffusion model without any fine-tuning. Specifically, I used the stable-diffusion-v1-4 model [11] developed by Hugging Face as my baseline.

Quantitatively, for LoRA, I reported the model's Contrastive Language-Image Pretraining (CLIP) score, which measures the semantic similarity between textual descriptions and the corresponding generated images, and compared it to that of the baseline. For Textual Inversion and DreamBooth, I reported the Structural Similarity Index (SSIM), which measures how structurally similar the generated images are to the training images in my Subject-Specific Dataset.

Qualitatively, I visually assessed the quality of images generated with three proposed methods. For LoRA, I checked if the images appear in Cordelia's art style. For subject-specific generation, I checked if my subject Erica is clearly recognizable in the different contexts that I prompted the model with. I also compared images generated using Textual Inversion versus DreamBooth to see which method performed better.

4 Results & Evaluation

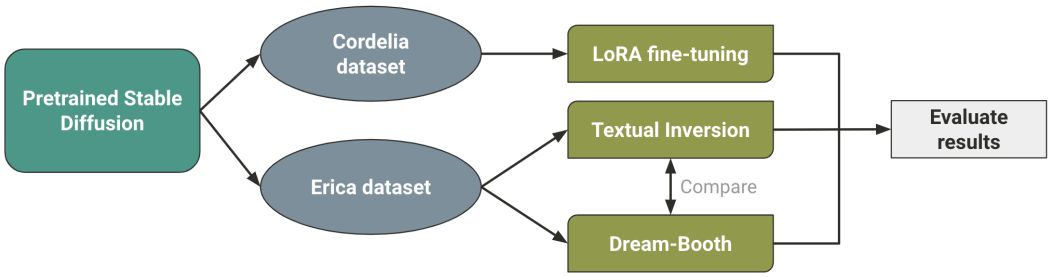

Figure 6: Experimental procedure.

4.1 Fine-Tuning Stable Diffusion With Style LoRA

Following the approaches described above, I conducted the following experiments. First, I defined my baseline model to be the stable-diffusion-v1-4 model [11] developed by Hugging Face, and measured its CLIP score, which turned out to be 35.505. Then, I ran the Kohya LoRA Fine-Tuning notebook [5], uploading the Cordelia-Style Dataset, tagging the images with captions, and subsequently training three LoRA models with different optimizers (AdamW, SGDNesterov, and AdaFactor) with network dimension 32 and learning rate 1e-4. Figure 7 shows the training loss resulting from each optimizer. From the chart, we see that AdaFactor and SGDNesterov were better than AdamW for my specific style-learning task, with AdaFactor giving a slightly lower average loss than SGDNesterov. Therefore, I selected AdaFactor as my final optimizer, then conducted hyperparameter tuning for the learning rate, with results shown in Table 1.

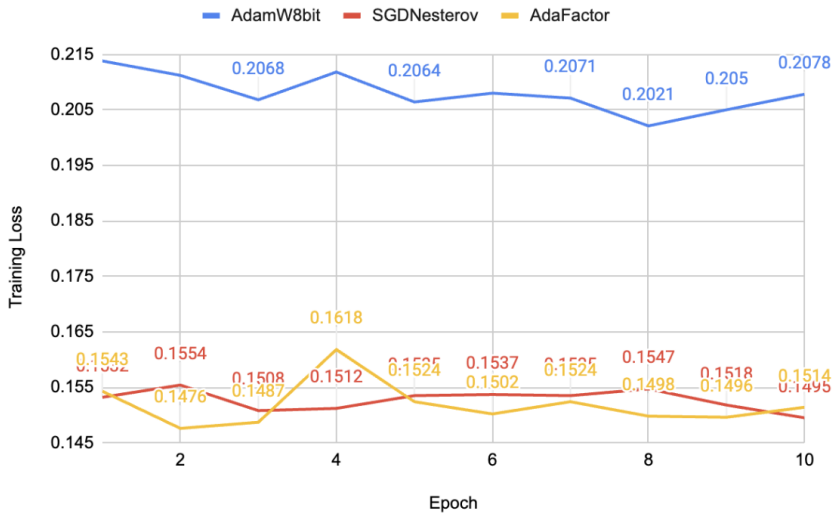

Figure 7: Training loss over 10 epochs with different optimizers.

Learning rate	5e-5	1e-4	5e-4
Avg loss	0.1596	0.1518	0.1959

Table 1: Average training loss with AdaFactor and different learning rates.

I found that the original learning rate of 1e-4 led to the lowest average training loss, so I selected it to be the learning rate of my final model. Finally, I evaluated the final LoRA model and compared it against the baseline by measuring its CLIP score (Table 2).

	CLIP score
Stable Diffusion (baseline)	**35.505**
Stable Diffusion with LoRA (AdaFactor optimizer)	24.970

Table 2: CLIP scores of the baseline versus my LoRA model.

We see that unfortunately, Stable Diffusion with LoRA did not perform better than just Stable Diffusion, achieving a CLIP score that is 29% lower than that of the baseline. A likely explanation is that since our fine-tuned model now outputs images in a special style belonging to an artist, this speciality has become an inhibitor preventing the CLIP evaluation from establishing semantic connections between text and the generated images.

With this explanation, a lower CLIP score is actually expected, and it indicates that my model did successfully learn Cordelia's art style. This conclusion is corroborated by the qualitative data shown in Figures 8, 9, and 10. The example images generated by the Stable Diffusion with LoRA model do indeed closely resemble Cordelia's art style, characterized by a unique artistic blend of real painting elements and anime aesthetics.

Figure 8: Output of the LoRA model with input "Girl with pink braids in a fantasy city scenery."

Figure 9: Output of the LoRA model with input "Cat in city alley under neon lights."

Figure 10: Output of the LoRA model with input “Train with falling sakura over a river.”

4.2 Fine-Tuning Stable Diffusion With Subject-Specific Generation

For subject-specific generation, first, I implemented Textual Inversion by using the Hugging Face Diffusers library [9], saved the model checkpoint, and made it generate a bunch of images with my placeholder token `<erica>` in various contexts. Then, I did the same thing with DreamBooth, this time following the DreamBooth with LoRA tutorial from also from Hugging Face [10]. Figure 11 shows some images generated by both Textual Inversion and DreamBooth when given specific textual descriptions containing `<erica>`.

Upon qualitative evaluation, Textual Inversion exhibits lower image quality overall, marked by noticeable deformations. An illustrative instance is observed in the image pair generated from the prompt `"<erica> in a ramen-bowl"`, where the left image depicts Erica with swollen eyes, and the right image reduces her to a featureless blue blob. In stark contrast, DreamBooth consistently produces high-quality, well-defined images. It is also interesting to note that DreamBooth showcases the ability to generate images with specific subjects in diverse styles. For instance, when prompted with `"<erica> on a floral-pattern towel"`, the output images adopt an anime style while maintaining exceptional quality and preserving the distinct features of Erica.

Method	“<erica> on a floral-pattern towel”	“<erica> in a ramen-bowl”
Textual Inversion		
Dream-Booth		

Figure 11: Generated images of Textual Inversion and DreamBooth.

In terms of quantitative assessment, I computed the Structural Similarity Index (SSIM) between the training images and the generated images for three models: baseline, Textual Inversion, and DreamBooth. The results are presented in Table 3. Notably, DreamBooth exhibited the highest SSIM value, reaching 0.364. This suggests that the generated samples from DreamBooth are structurally more similar to the images of Erica in my subject-specific dataset. This finding aligns with the qualitative evaluation above. It is also logical that the baseline model achieved a very low SSIM. This is attributed to the fact that the Stable Diffusion model, before fine-tuning, lacked recognition of the concept of Erica, given its absence from the specific-subject images during the fine-tuning phase.

	SSIM
Stable Diffusion (baseline)	0.019
Textual Inversion	0.346
DreamBooth	**0.364**

Table 3: SSIM of baseline, Textual Inversion, and DreamBooth.

5 Conclusion

In conclusion, this project delved into the customization and enhancement of Stable Diffusion for text-to-image generation, focusing on two key aspects: stylized adaptation inspired by the works of digital artist Carla Cordelia and subject-specific generation featuring a stuffed animal cat named Erica.

My attempt to fine-tune Stable Diffusion with LoRA to capture Cordelia's art style yielded promising qualitative results. Despite a 29% lower CLIP score compared to the baseline, the generated images exhibited a clear resemblance to Cordelia's distinctive style, showcasing the successful adaptation of the model.

DreamBooth outperformed Textual Inversion in both qualitative and quantitative assessments. The generated images from DreamBooth demonstrated higher quality, sharper features, and the ability to portray Erica in diverse styles, as evident in the higher Structural Similarity Index (SSIM) of 0.364 compared to Textual Inversion's 0.346.

This project contributed novel perspectives by introducing specific datasets inspired by Cordelia's artistic style and featuring personal subjects like Erica. Unlike prior research, which often focused on general concepts or existing datasets, this work explored the adaptability of Stable Diffusion in more customized and artistic contexts.

In the future, I would like to explore more nuanced evaluation metrics that better capture the qualitative aspects of artistic style and subject-specific generation, going beyond CLIP and SSIM. I would also like to explore multi-subject generation [12], adapting DreamBooth to seamlessly portray two or more subjects within a single generated image, opening up possibilities for more intricate and complex compositions.

6 Code

GitHub link: https://github.com/katchen1/StableDiffusionFinetune

References

[1] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bj¨orn Ommer. High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 10684–10695, 2022.

[2] 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.

[3] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. arXiv preprint arXiv:2208.01618, 2022.

[4] 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. 2022.

[5] https://github.com/Linaqruf/kohya-trainer/tree/main

[6] Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. arXiv preprint arXiv:1711.05101, 2019.

[7] Chaoyue Liu and Mikhail Belkin. Accelerating SGD with momentum for over-parameterized learning. arXiv preprint arXiv.1810.13395, 2018.

[8] Noam Shazeer and Mitchell Stern. Adafactor: Adaptive Learning Rates with Sublinear Memory Cost. arXiv preprint arXiv.1804.04235, 2018.

[9] https://colab.research.google.com/github/huggingface/notebooks/blob/main/diffusers/sd_textual_inversion_training.ipynb

[10] https://huggingface.co/docs/peft/task_guides/dreambooth_lora

[11] https://huggingface.co/CompVis/stable-diffusion-v1-4

[12] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, Jun-Yan Zhu. Multi-Concept Customization of Text-to-Image Diffusion. Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1931-1941, 2023.