Mixing LoRAs for interpolating art styles across history

Fiona Mariele Lau Minh Khang Van
LMU, Munich, DE
Practical "Image & Video Synthesis" WiSe 24/25
Computer Vision & Learning Group
f.lau@campus.lmu.de, k.van@campus.lmu.de

Abstract

Low-Rank Adaptation Models (LoRAs) have been introduced to save computational costs and focus on one downstream task without retraining the entire network. In the present project, we trained three LoRAs based on the CompVis/stable-diffusion-v1-4 model and images from three art epochs: Early Renaissance, Expressionism, and Pop Art from the Wikiart dataset. We perform linear interpolation to merge the art epochs into two versions. In V1, we merge two art epochs; in V2, we merge three to look at paintings that merge across history. We did a qualitative analysis of promising results, on the one hand side from images from the trained LoRAs and, on the other hand, from the merged LoRA Versions. We have identified recurring patterns regarding style and content for specific art eras and interpolation steps. Even in edge case images, we can find features from the respective art eras. In addition, we used the fine-tuned ResNet50 to classify the merged images into our three epochs. Our results showed that the classifier can classify images created by our merged LoRAs as long as only two weights are merged.

1. Introduction

The field of generative AI is changing and growing rapidly. Advances in tools such as Stable Diffusion [15] support different areas of creative and analytical work and lead to various images with diverse features in style and content. In recent years, generative models have been trained with many parameters. Low-Rank Adaptation Models (LoRAs) have been introduced to save computational costs and focus on one downstream task without retraining the entire network [7]. LoRAs can be combined in various ways including linear interpolation and concatenation [13]. One of the conceivable application areas for interpolating LoRAs for image generation is the art sector, where they can be used to create combinations of different art styles for unique compositions [5]. Furthermore, it can help students/researchers visualize the evolution of art techniques across epochs by interpolating them in a certain number of steps. We trained three LoRAs based on the CompVis/stable-diffusion-v1-4 model and images from three art epochs of the Wikiart [17] dataset in the present project. We then decided for linear interpolation to merge our LoRAs into two versions (V1 two, V2 three eras) to analyze and the results, find patterns and detect errors.

2. Background

Latent Diffusion Models (LDM) introduce an efficient approach to synthesizing high-resolution images. Unlike standard diffusion models, which apply Gaussian noise directly in pixel space, LDMs uses a variational autoencoder to compress images into a lower-dimensional latent space. This technique preserves important image information while significantly reducing computational costs. To achieve Text-to-Image synthesis, LDMs incorporate a cross-attention mechanism within the U-Net architecture. This allows the U-Net to be conditioned on textual inputs, typically via a text encoder (e.g., CLIP [14] or BERT [3]), enabling controllable image generation. [15]

Low-Rank Adaptation (LoRA) enables efficient fine-tuning of large models for downstream tasks. Retraining all layers of a model requires significant time and computational resources. Instead of updating entire weight matrices, LoRA introduces trainable low-rank layers that decompose existing layers. Typically, model weights are updated as: $W_0 \Leftarrow W_0 + \Delta W$. However, updating the entire matrix ΔW is computationally expansive. LoRA addresses this by decomposing ΔW into two lower-rank matrices, A and B, such that: $W_0 + \Delta W = W_0 + BA$. This decomposition significantly reduces the memory (VRAM) required during training, as only the smaller matrices A and B need to be optimized. Consequently, LoRA drastically reduces the number of trainable parameters, making fine-tuning feasible even on resource-constrained hardware. [7]

While this technique was first introduced for transformer based large language models, the idea still holds true for

diffusion models where the U-Net is retrained. It is also possible to fine-tune the text encoder.
This is especially needed when the prompts correspond to new words the pre-trained model is not aware of.

3. Project Setup

The project itself can be seen on GitHub [1]. For all experiments, we used the diffusion model "CompVis/stable-diffusion-v1-4" accessible through huggingface. Due to computational constraints, training is conducted on a device with a NVIDIA RTX 2060 SUPER with 8GB of VRAM where python 3.8.10 only is installed.

The implementation is designed to be reproducible and customizable. Users can adapt the configuration files of each different tasks, such as fine-tuning a diffusion model or training an image classifier, allowing them to swap models and datasets as needed. This project is inspired by two key sources: The huggingface library Diffusers library [18] and the LoRA implementation from cloneofsimo's GitHub repository[16].

4. LoRA Training Steps

Before we fine-tune a diffusion model it is necessary to precompute data which will be used in the following processes. Since we work with a fixed dataset of images it offers it self to pass the images the to the variational autoencoder and compute the latents beforehand. This process saves time later on during training and removes the computational load on the device. As already explained in Chapter 2 LoRA adds additional layers to a given model. Thus, we can retrain a pre-trained diffusion model to our desire with LoRA. It is important to freeze the pre-trained layers in detail the U-Net and the text encoder and only activate gradients on the injected LoRA layers. In order to save vram on the GPU the model itself is loaded in float16. The LoRA layers are upcasted to float32 for better training results. Therefore, we used torch autocast for mixed precision training. What's left is to specify a specific art epoch to train on. The corresponding prompt we trained on the model on is "A painting in style of **Style1**. For each art epoch we used a different number for our **Style**. The hyper-parameters for fine-tuning were:

- **Loss Function:** Mean Squared Error
- **Learning Rate:** 0.00005
- **Batch Size:** 4
- **Optimizer:** AdamW
- **Number of Epochs:** 50
- **Rank:** 16
- **Weight Decay:** 0.001

The figure 1 shows the loss curve on training and validation dataset of the art epoch "expressionism". Even though the loss is very small, we were able to create promising results which will be discussed in chapter 7.1

Figure 1. loss curve of training and validation dataset

5. Data

The Art500k image dataset was suggested to us. It was divided into several folders, the smallest of which had 7GB and only contained images from Midage, Early Renaissance, and Baroque. All other folders were too large to use the data [11]. Together with the Stable Diffusion Model, it would have been too much to process on the available hardware if we wanted a variety of epochs. We, therefore, looked for a different dataset and opted for Wikiart with 81000 images in total to need fewer GB and have a wide selection of different art styles such as Expressionism, Impressionism, Realism, or Pop Art [17]. Ultimately, we decided on the following three epochs, which have a relatively high temporal distance and are also stylistically different.

Early Renaissance (1391 images)
Expressionism (6736 images)
Pop Art (1483 images)

About the epochs:
For the Renaissance, 14th and 15th centuries, the colors gold, red, and blue were quite dominant, and nature and religion were predominant themes of the time; a famous artwork of the epoch was "Coronation-of-the-virgin" by Filippo Lippi [19] [17] - we focused on Early Renaissance images. In Expressionism, the brush strokes are dominant, the colors are bold but muddy, and the objects are often distorted [12]. The art epoch developed at the beginning of the 20th century in Europe. Emotional states are, among others, one of the themes of Expressionism [**?**]. A famous artwork of the Expressionism epoch is "The Scream", created by Edward Munch [17]. In contrast, Pop Art is a very colorful epoch that peaked around 1960. There are paintings, collages, and printings, and the epoch's themes are, among others day to day life and the consumer society [10] [8]. Andy Warhol became famous for his creative artwork "Marilyn Monroe" in this art epoch [17].

6. Interpolation of LoRAs

We developed two different versions of interpolating LoRAs. In V1, we merged two LoRAs, and in V2, we merged three LoRAs.

$$V1 = \alpha_1 \cdot wd_1 + \alpha_2 \cdot wd_2 \quad (1)$$
$$V2 = \alpha_1 \cdot wd_1 + \alpha_2 \cdot wd_2 + \alpha_3 \cdot wd_3 \quad (2)$$

In V1, we tested two subversions: In V1.1, we merged Early Renaissance with Expressionism, and in V1.2, We merged Early Renaissance and Pop Art. In all versions and subversions, we used mainly the following interpolation steps for generating images across the respective art epochs:

$$\begin{aligned} \alpha_1, \alpha_2 &: (0,1), (0.2,0.8), (0.4,0.6) \\ &\quad (0.5,0.5), (0.6,0.4), (0.8,0.2), (1,0) \\ \alpha_1, \alpha_2, \alpha_3 &: (0.25, 0.25, 0.5) \end{aligned}$$

In addition, we conducted ”Free Style Testing” and tested beyond the alpha values (e.g., a higher number of inferences in a steps or guidance scale, not in style 2 and more). In the next section, we will look closer at our image results [13].

7. Qualitative Analysis

We conducted a qualitative analysis of the image results we generated on the one-hand side by the use of the trained LoRAs and on the other-hand side from the merged LoRAs. We created comprehensive text prompts that we categorized regarding the topics: Style-based prompts, content-based prompts, and edge cases. The categories are the same across both qualitative analyses, but the prompts change for analyzing the trained LoRA to merge LoRA results.

7.1. LoRA Output Images

A comprehensive qualitative analysis of the LoRA results we provide in our Github repository in a jupyter notebook with the respective name.

7.2. Merged LoRA Output Analysis

We used the following prompts for the merged LoRA analysis to test V1.1, V1.2, and V2 respectively.

Content-based Prompts:

- ”A painting of a woman in the city in Style1 and Style2”
- ”... of a couple kissing under the tree...”
- ”... of an orange...”

Style-based Prompts:

- *Style Prompt Renaissance:* “... a girl in nature, religious, pyramidal arrangement of the figures, landscape, natural colors...”
- *Style Prompt Expressionism:* “...a girl in nature, brush strokes energetic, shining bright colors, emotional, easy, edgy, bold lines, tension, color contrasts, abstraction, immediateness...”
- *Style Prompt Pop Art:* “... of a girl in the city, bold designs, bold color, orient on comics...”

Edge Case Prompts:

- ”... aliens arriving in a multicolor sail ship...”
- ”... two horses running down a street in New York...”

7.2.1. Analysis Across Merge V1:

Comprehensive results: In general, across all merges, we can observe that with different proportions of the respective art epoch and merge steps, facial features change as well as the clothing (see figure 2). Compared to the other epochs, the Pop Art style is very intensive in merge V1. Moreover, it is more dominant regarding the content, even if we have, for example, a higher ratio of Renaissance (see figure 4). In addition, the sharpness of our output results is generally very satisfying.

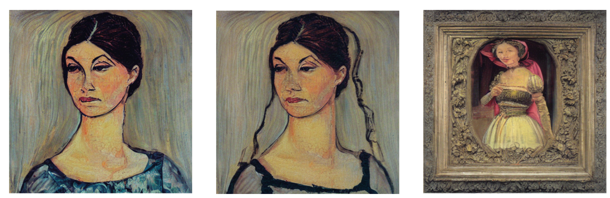


Figure 2. Left: 0.5 Renaissance - 0.5 Renaissance, Middle: 0.6 Renaissance - 0,4 Expressionism, Right: 1.0 Renaissance - 0 Expressionism

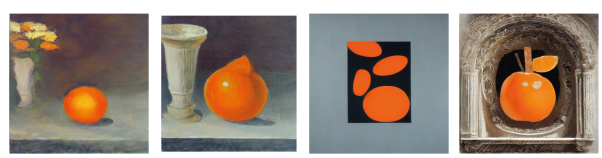


Figure 3. Left: 0.5 Renaissance - 0.5 Renaissance, Middle: 0.6 Renaissance - 0,4 Expressionism, Right: 1.0 Renaissance - 0 Expressionism

Content Based Analysis A High proportion of Renaissance leads to frames or columns appearing in multiple images of merged image results (see figure 2, 3, 5, 7). Moreover, we could observe that colors are less saturated, and texture changes to brittle (see figure 3, 4).
In contrast, with a high proportion of expressionism, typical brushstrokes are recognizable, the colors are rich but muddy, and often the objects are rather distorted and emotional (see figure 2, 5, 7).

Finally, a high proportion of Pop Art shows rich colors. Also, as mentioned, the epoch's themes are dominant (see figure 4, middle). Even with more of a proportion of Renaissance and less Pop Art, we can observe modern-looking people with modern clothes, leading to the impression that Pop Art is, despite a smaller alpha value, more dominant.

Figure 4. Left: 0.6 Renaissance - 0.4 Expressionism, Middle: 0.6 Renaissance - 0.4 Pop Art - Right: 0, 0.2 Renaissance - 0.8 Art Pop

Style Based Analysis: The style-based images all adapt to the prompts, but Pop Art was so dominant that we could still see aspects of it in the Renaissance Style Prompt.

Edge cases Analysis: Even with tested edge case prompts, we can see epoch-specific properties such as the distorted objects and bold but muddy colors of Expressionism (see Figure 5, left), bold colors and cartoonish style for pop art (see figure 5 middle), and the frame appearing, also less saturation with a high proportion of Renaissance (see figure 5, right).

Figure 5. Left: Renaissance: 0 - Expressionism: 1, Renaissance:0 - Pop Art:1, Right: Renaissance 1 - Expressionism or Pop Art: 0

Free Style In Figure 6, we can observe on the left-hand side of the three images what happens if 7-7 were used as alpha values. Furthermore, on the right-hand side, we can see the output of an image generated with tune scale 9. The middle demonstrates an image with a guidance scale of 3.

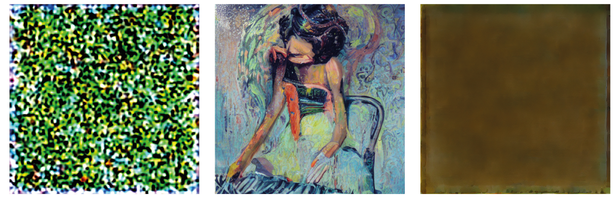


Figure 6. Left: Renaissance: 0.25 - Expressionism: 0.25 - Pop Art: 0.5, Renaissance: 0.5 - Expressionism: 0.1 - Pop Art: 0.4, Renaissance: 0.1 Expressionism: 0.5 - Pop Art: 0.4

7.2.2. Analysis V2

Regarding the qualitative analysis of merge V2, it is possible to interpret the art style with the highest ratio, but perhaps it is less straightforward than in V1.
In Figure 7, the image on the left-hand side, it can be interpreted that the hat of the woman in the image is in the style of a Renaissance; the abstract the face might belong to Expressionism, and the background and the bold and dominant colors contrast with Pop Art. In the middle and right-hand side are attributes of art epochs, with the high portion again more visible than in the image on the left.

Figure 7. Left: 0.25 R - 0.25 E - 0.5 PA. Middle 0.4 R - 0.5 E - 0.1 PA, Right: 0.1 R - 0.5 E - 0.4 - PA

8. Quantitative Analysis

While the merging of LoRAs is distinguishable for the human eye, we wanted to see if an image classifier perceives similar results when changing the alpha values. Thus, this section is divided into two parts. In the first part we explore different image classification models and how we fine-tuned them for our downstream task. The second part delves into the results of the image classifier.

8.1. Fine-Tuning an Image Classifier

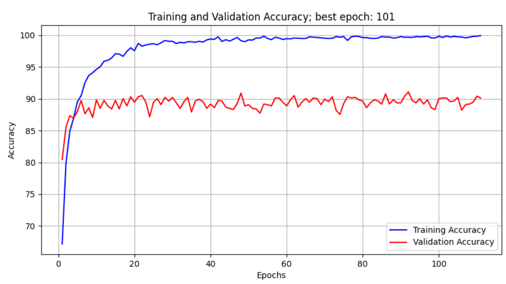
Our downstream tasks consist of classifying images into art epochs. As already mentioned in chapter 7.1 we focus on the three art epochs "Early Renaissance", "Expressionism" and "Pop Art" for simplicity. To prevent bias toward any particular epoch, we down-sampled the dataset to a common size, selecting 1483 images, which corresponds to the smallest dataset among the three epochs 5.

We also applied standard data augmentation techniques in order to avoid overfitting and enhance the model's ability to generalize.

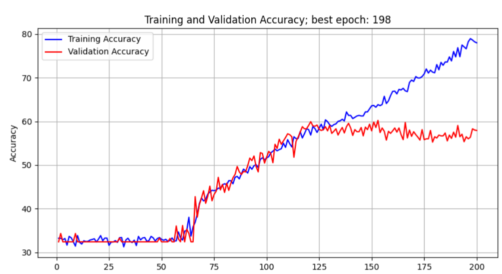
We trained two state of the art models ResNet50 [6] and the ViT "google/vit-base-patch16-224-in21k" [4] for comparative evaluation. Both models were trained with identical hyper-parameters:

- **Loss Function:** Cross Entropy Loss
- **Learning Rate:** 0.00001
- **Batch Size:** 16
- **Optimizer:** AdamW
- **Number of Epochs:** 200
- **Weight Decay:** 0.01

While both models showed promising loss curves, the accuracy measurements were disappointing. The ViT model performed barely better than random guessing, while ResNet50 was able to capture distinct features of each art epoch. Given this, we decided to proceed with ResNet50 and discard the ViT model for this task. The poor performance of the ViT is likely due to the relatively small size of our dataset (791 images). Figure 8 shows the accuracy of each model across epochs.



(a) Accuracy of the ResNet model



(b) Accuracy of the ViT

Figure 8. Accuracy plots of the image classifiers

8.2. Applying image classifier on merged images

With a promising accuracy of roughly 90%, we used the fine-tuned ResNet50 to classify images into our three art epochs. As discussed in Chapter 7, we focus on different prompts across the two versions of our LoRA merging. For V1 we used Early Renaissance as a common LoRA for comparison between two art epochs. To keep the report concise, we analyzed only one prompt for the combination of Early Renaissance and Pop Art. Figure 9 shows how the logits change when increasing the alpha value of Early Renaissance.

Although the classifier was not specifically trained to distinguish between art epochs but rather to classify individual epochs, it is still interesting to observe how the logits shift as we increase the alpha value for Early Renaissance and correspondingly decrease the influence of Pop Art. The model captures the style of the respective art epochs. However, there is a sudden jump in the logits when using an alpha value of 0.5 for both LoRAs. The model predicts Early Renaissance almost exlusively. We observed similar results across different art epochs combinations and prompts. This is likely due to the prompt itself heavily influencing the synthesis of the image. In this particular example, the prompt contains key words like "religious", a strong indicator for Early Renaissance, while Pop Art focuses on different aspects of art like vibrant colors.

When merging multiple LoRA weights - Early Renaissance, Expressionism and Pop Art - the fine-tuned classifier is confidently predicting a class. However, the prediction does not match to the true merged class. Since the model is only trained on classifying on art epoch at a time, it struggles to classify an image that blends three distinct epochs.

In conclusion the classifier is able to classify imaged created by our merged LoRAs as long only two weights are merged. Other examples can be found in the appendix A.

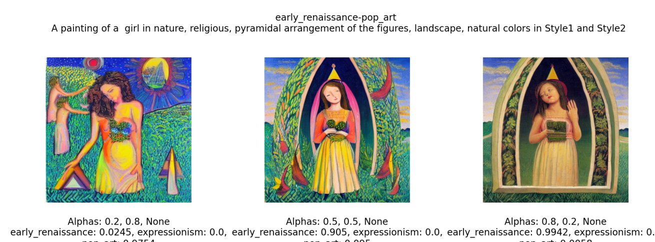


Figure 9. Alpha Values from left to right: Early Renaissance, Pop Art and Expressionism where Expressionism is None in this example

9. Limitations and Challenges

The project was successfully implemented, but we encountered several challenges along the way. Our initial finetuning of the model with LoRA resulted in no changes compared to the original model. The model failed to correctly distinguish between different artistic epochs. We identified that the learning rate scheduler was causing instability, and removing it significantly improved the results. While learning rate schedulers are commonly used in supervised learning pipelines, we found that disabling it in our case led to more promising outcomes. The final results can be seen in the Jupyter Notebook referenced in Chapter 7.1.

Another major challenge was limited access to GPUs. Training and especially inference was time consuming, as we often had to wait for available resources due to shared usage. Since our goal was to collect a diverse set of examples covering a wide range of prompts, this limitation slowed our progress.

Finally, while the webdataset library is a powerful tool for efficient data handling during training, its documentation is limited [2]. Thanks to the support of our supervisors, we were eventually able to use the ShardWriter class successfully.

10. Conclusion and Outlook

While it is already well known that fine-tuning a pre-trained model with LoRA is effective, we successfully reproduced these results. The LoRA layers were able to capture the distinct styles of each art epoch. The most interesting aspect of this project was analyzing the different effects of merging multiple LoRA weights. Our qualitative and quantitative analyses led to the same conclusion: adjusting the alpha values between different LoRA weights effectively guides the diffusion model toward a particular art epoch. Additionally, we found that the prompt itself plays a significant role in image generation. This is due to the fact that motifs from the early Renaissance are inherently different from those in Pop Art or Expressionism. Even when using a high alpha value for a later art epoch, elements of the early Renaissance remain recognizable when using keywords like ”religion” in the prompt. Despite the Art pop epoch was the most dominant.

Another aspect worth exploring is alternative techniques for merging LoRA models. Prabhakar et al. [13] experimented with different LoRA merging strategies. In addition to the standard linear merging, they found that concatenating LoRA weights produced promising results.
Lastly, our experiments were conducted using ”CompVis/stable-diffusion-v1-4” exclusively, leaving room to explore other diffusion models. One interesting candidate is Black-Forest-Labs' Flux models, which offer state-of-the-art image generation. However, training these models on consumer hardware is impractical, as they consist of 12 billion parameters [9].

For future research, one could expand the number of art epochs, focusing especially on epochs that share similar motifs—for example, a study on abstract themes or human depictions across different styles. Additionally, future work could move beyond classical painting to include sculptures or even photographs.

References

[1] Mixing Loras to interpolate art styles across history. https://github.com/blackSwan566/Mixing-Loras-to-interpolate-art-styles-across-history/tree/main. 2

[2] webdataset. https://github.com/webdataset/webdataset. 6

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. 1

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. 5

[5] Fenstermaker. How artificial intelligence sees art history, 2019. 1

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. 5

[7] 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, 2021. 1

[8] Andreas Huyssen. The cultural politics of pop: Reception and critique of us pop art in the federal republic of germany. *New German Critique*, (4):77–97, 1975. 2

[9] Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2023. 6

[10] Steven Henry Madoff, editor. *A Critical History*. University of California Press, Berkeley, 1998. 2

[11] Hui Mao, Ming Cheung, and James She. Deepart: Learning joint representations of visual arts. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1183–1191. ACM, 2017. 2

[12] Mont Marte. 8 expressionism techniques, 2023. 2

[13] Akshara Prabhakar, Yuanzhi Li, Karthik Narasimhan, Sham Kakade, Eran Malach, and Samy Jelassi. Lora soups: Merging loras for practical skill composition tasks, 2024. 1, 3, 6

[14] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 1

[15] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *CoRR*, abs/2112.10752, 2021. 1

[16] Simo Ryu. Low-rank adaptation for fast text-to-image diffusion fine-tuning. https://github.com/cloneofsimo/lora. 2

[17] Wei Ren Tan, Chee Seng Chan, Hernan Aguirre, and Kiyoshi Tanaka. Improved artgan for conditional synthesis of natural image and artwork. *IEEE Transactions on Image Processing*, 28(1):394–409, 2019. 1, 2

[18] Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, Dhruv Nair, Sayak Paul, Steven Liu, William Berman, Yiyi Xu, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers. 2

[19] Evelyn Welch. Art in renaissance italy: 1350-1500. 2001. 2

A. Appendix

The best way to understand these plots is to follow these steps:

- **Start with the Prompt:** The title text describes the base concept for all generated images.
- **Identify Art Styles:** Note the key art styles involved (e.g., early_renaissance, expressionism, pop_art).
- **Focus on One Image:** Analyze each image separately.
- **Read the "Alphas":** These values define the weighting of each style in the image.
- **Interpret the Logits:** These values indicate how much the model is adhering to each style.
- **Relate Visuals to Numbers:** Observe how the image's visual characteristics reflect the alpha values and logits.
- **Compare Images:** Compare the images to see how changes in alpha values affect the visual style.

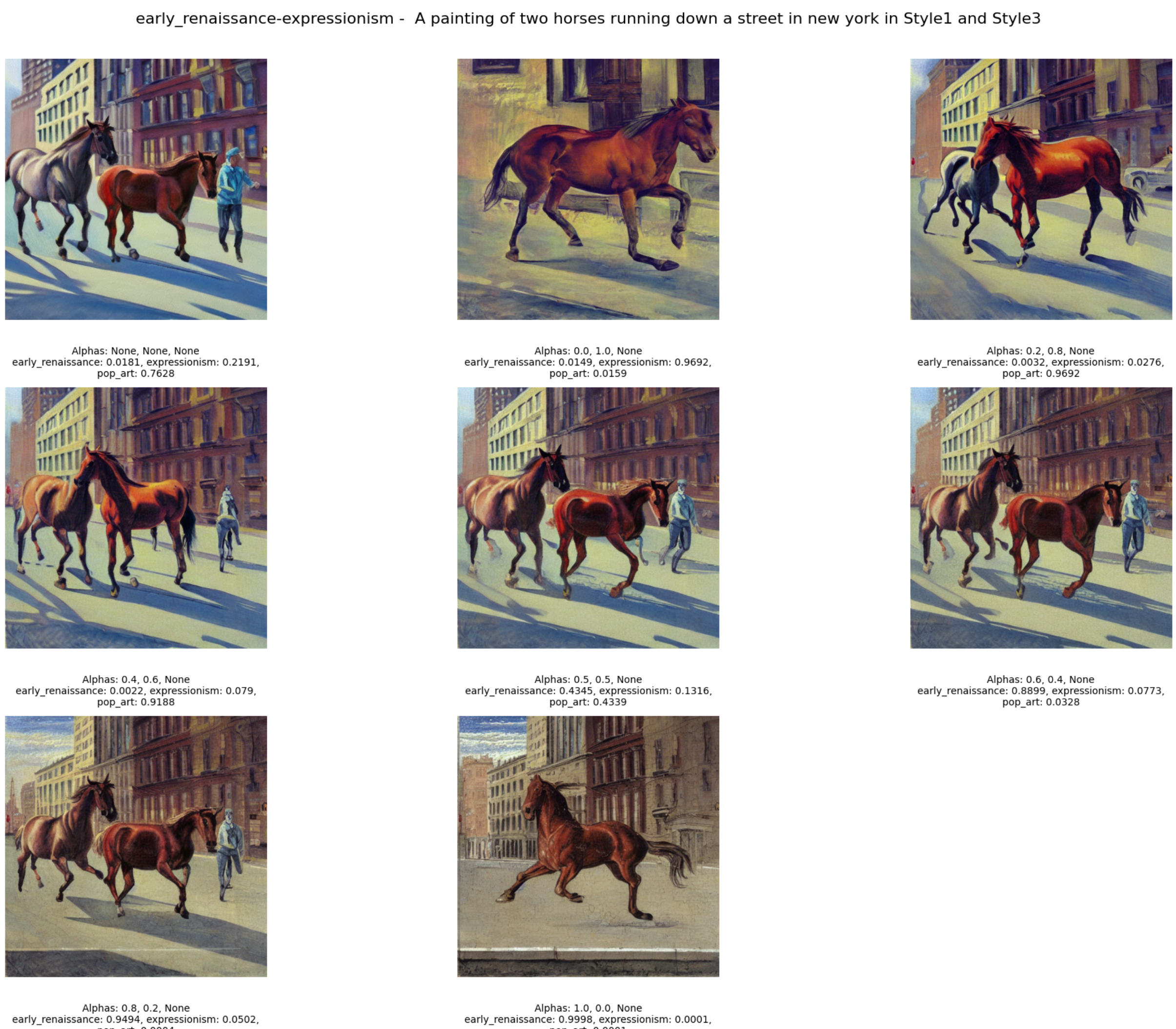
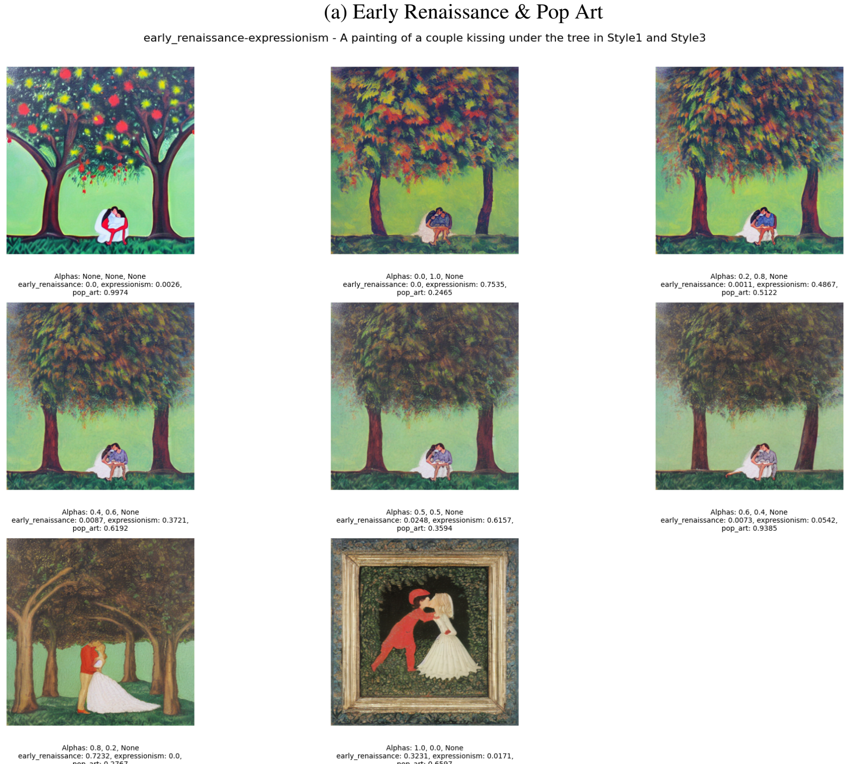


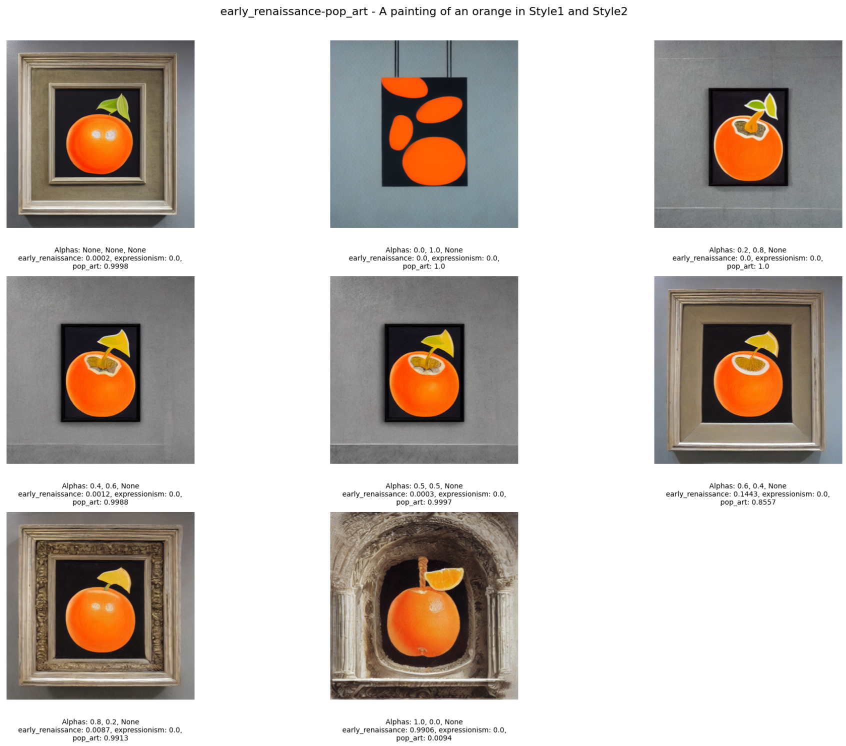
Figure 10. Illustrative Example: Style Transfer Variations

(a) Early Renaissance & Pop Art



(b) Early Renaissance & Expressionism

Figure 11. Same prompt different art epochs

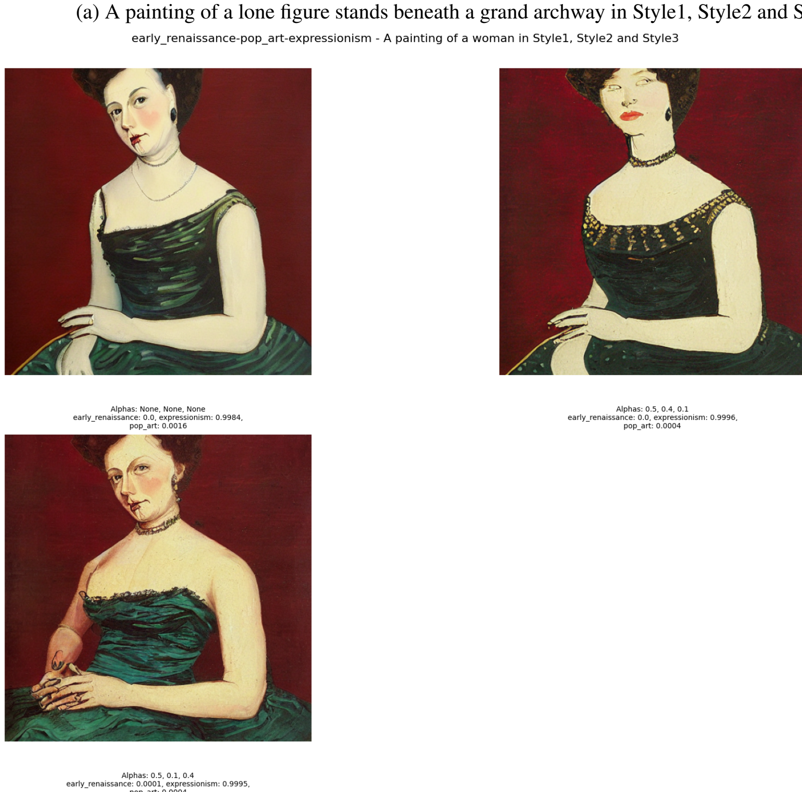


(a) Early Renaissance & Pop Art

(b) Early Renaissance & Expressionism

Figure 12. Same prompt different art epochs

(a) A painting of a lone figure stands beneath a grand archway in Style1, Style2 and Style3



(b) A painting of a woman in Style1, Style2 and Style3

Figure 13. V2 Merge with different prompts