How to generate new versions of an original character?

An application of LoRA and DreamBooth Fine-Tuning of Stable Diffusion Models

Maëlys Boudier, Natalia Beltrán, Arianna Michelangelo

GitHub: Character-Generation-for-Comics

Under the supervision of Hannes Mueller and Jesus Cerquides Bueno

July 1, 2024

July 1, 2024

Acknowledgments

We would like to extend our deepest gratitude to the talented artist Jordane Meignaud for generously providing the character images that were crucial to the success of this project.

We also sincerely thank the esteemed comic book artist Bertrand Escaich, known as BeKa, for his invaluable collaboration. His partnership in creating a real-world comic book example using our tuned model has been instrumental in showcasing its capabilities.

We are grateful to Hannes Mueller and Jesus Cerquides Bueno for their valuable feedback during milestone presentations. Additionally, we extend our special thanks to Elliot Motte for his assistance and support throughout the process.

Copyright Statement

July 1, 2024

Abstract

This research explores the use of LoRA (Low-Rank Adaptation) and DreamBooth fine-tuning techniques on Stable Diffusion models to generate new versions of an original comic book character. By addressing the challenge of data scarcity, these techniques enable the creation of varied character poses from limited images, significantly enhancing efficiency in the comic creation process. The study demonstrates that fine-tuning pre-trained models with minimal computational resources can produce high-quality, consistent character images, reducing repetitive tasks for artists. A collaborative project with professional comic scriptwriter Bertrand Escaich and illustrator Jordane Meignaud validated the practical application, resulting in a one-page comic story featuring an original character. This research paves the way for integrating AI into comic book creation, allowing artists greater creative freedom and productivity.

Key Words: Generative AI, Comic Strip Automation, LoRA (Low-Rank Adaptation), DreamBooth, Stable Diffusion Models, Data Scarcity, Fine-Tuning Techniques, Character Generation, AI in Art, Image Generation

July 1, 2024

Contents

1 Introduction

The focus of this research is on generating new versions of an original character using LoRA (Low-Rank Adaptation) and DreamBooth fine-tuning techniques on Stable Diffusion Models. The creative process of comic strip creation is both highly iterative and time-consuming, particularly when artists need to repeatedly draw the same character in various positions. This project aims to explore how generative AI can enhance efficiency in this process, allowing artists to concentrate on their core strengths: inventing new worlds, styles, characters, and backgrounds.

A significant challenge in image generation is data scarcity. If an artist only has six images of a character they invented, it becomes difficult to computationally generate new versions of this character. Our research addresses this issue by employing fine-tuning techniques like LoRA and DreamBooth on Stable Diffusion Models. These techniques involve 'teaching' the concept (the character) to a pre-trained model to generate new images from text prompts specifying the character and a desired pose. The goal is to generate the same character in new poses that were not included in the training data, such as lifting an arm or closing eyes. This process involves fine-tuning the model to understand and replicate the character's appearance while placing it in novel poses, thereby expanding the character's visual repertoire.

Limiting the use of large models is crucial due to limited computational resources and the need to reduce CO2 emissions. Efficiently generating high-quality images with minimal data and computational power is both an environmentally and economically beneficial approach. Although our paper focuses on a single character, the techniques explored here could eventually be integrated alongside other characters and backgrounds. The ability to generate consistent and varied images from limited data can significantly enhance productivity and creativity in the comic book industry.

Image generation technology has evolved from GANs (Generative Adversarial Networks) to diffusion models, which generally perform better and may include a textual component for more guided generation. Our goal is to explore these advancements and determine what types of models can be created with scarce data. By investigating the potential of fine-tuning stable diffusion models using LoRA and DreamBooth, this research aims to provide a valuable tool for artists to streamline the creative process, reduce repetitive tasks, and focus on the more innovative aspects of comic book creation.

2 Literature Review

The intersection of comic books and artificial intelligence (AI) has seen substantial progress in recent years, particularly in colorization, style transfer, and character generation. This literature review provides an overview of the key studies in these areas, highlighting the use of Generative Adversarial Networks (GANs) and diffusion models in the comic book industry.

2.1 Colorization

The colorization of images is a significant area of research in AI applied to comic books. This process involves adding color to black-and-white images, a task that can be time-consuming and requires a high degree of artistic skill. Olivier Augereau et al. conducted a comprehensive survey of content generation techniques, including colorization, character generation, and media conversion (Augereau et al., 2018). Their work provides an overview of the advancements in AI that facilitate the coloring process, thereby enhancing the efficiency of comic book production. More recent works have started including diffusion models for this same purpose, such as a GitHub repository by Erwann Millon and a paper by Liu et al. (Liu et al., 2023; Millon, 2023).

2.2 Style Transfer

Style transfer refers to the application of a particular artistic style to an image, such as transforming a photo into a cartoon-like image. This technique has been used to generate anime images or cartoons by applying neural style transfer on photos. Chen et al. introduced AnimeGAN, a lightweight GAN for photo animation (J. Chen et al., 2020). While this technique shows promise in generating stylized images, the results still differ significantly from the detailed and polished cartoons typically used in comics or anime strips. The study underscores the challenges and potential of using neural style transfer to bridge the gap between photos and cartoon illustrations. However, *Diffusion in Style* is a novel method for tailoring Stable Diffusion to a specific target style, requiring only 50 to 200 images from the target style. This greatly enhances its practicality for scenarios where extensive image collections are unavailable (Everaert et al., 2023).

2.3 Character Generation

Character generation in the context of comic books has primarily utilized GANs. Two main variations of this approach have been explored: one that generates variations of the same character and another that generates different anime faces with specific features such as hair color and eye color inputted by the user. Marnix Verduyn's study investigates

the application of GANs for generating variations of a single character (Verduyn, 2022). This approach is instrumental in reducing the repetitive tasks faced by comic book artists, allowing them to focus on more creative aspects of their work. However, this technique was unsuccessful as Verduyn demonstrates that the images generated from the GANs remained in the uncanny valley and the generated comic character was inconsistent and had lots of weird features. Yanghua Jin et al. explored the training of GAN models specialized on an anime facial image dataset, specifically implementing DRAGAN (Jin et al., 2017). This study, while limited to face generation, demonstrates the potential of GANs to create diverse and customizable anime characters based on user inputs such as hair color and eye color.

2.4 Fine-Tuning Techniques

Our main goal for this project is character generation, for which we will study the use of fine-tuning diffusion models and textual embeddings to have some control over the generated outputs. Recent advances in training methodologies now allow for the expansion of pre-trained models beyond their original scope of knowledge. This means that individuals can take foundational models like Stable Diffusion, train them with just a few images of a specific subject, such as a new character or person, and subsequently generate new images of that subject in various settings. While developing large image-generation models from scratch requires significant computational resources and is often impractical for individuals without substantial funding, fine-tuning an existing model is far more accessible. Two innovative techniques, DreamBooth and LoRA, have emerged as noteworthy in this area (Hagström and Rydberg, 2024).

3 Data

Given that the goal of our research is to tackle the issue of data scarcity in image generation and work with small models for environmental concerns, we limited our training input to just six images. This approach aligns with our commitment to minimizing computational resources and energy consumption, which are critical in the context of sustainable AI development.

The six training images were designed by Jordane Meignaud through a small iterative process. Meignaud proposed initial sketches (see Figure 23 in the Appendix), which our team reviewed and validated. Our primary request was to ensure that the drawings featured different angles and poses, providing the model with diverse inputs and as much information as possible. The images were initially of various sizes and shapes, but all featured white backgrounds to clearly delineate the character.

During the design process, we debated color choices from an artistic standpoint, although these choices were not expected to significantly impact the model's performance beyond subjective visual enjoyment. Ultimately, the artist chose a mix of salmon pink and teal blue, creating a unique visual identity for the character.

The training data consisted of six poses of a character named Unicorn Girl: a small girl with pink hair, wearing a blue onesie adorned with a rainbow-colored unicorn horn, a tail, and she is barefoot. The input images included three from the front, one from the back, one from the side mostly front-facing, and one from the side mostly back-facing. All except one are full-body views, with the other being a portrait. These images were designed to capture different angles and poses to maximize the diversity and information available to the model.

Figure 1: Unicorn Girl Data

4 Baseline Models

To address the challenges of data scarcity in image generation and explore the efficiency of various generative AI models, we established three baselines for our experiments. These baselines include training a GAN from scratch, training a diffusion model from scratch, and using the pre-trained Stable Diffusion (version SDXL) model with prompt engineering. Each baseline provides a different perspective on the capabilities and limitations of generative models in the context of comic book illustration.

4.1 Generative Adversarial Networks

Generative Adversarial Networks (GANs) are a class of machine learning frameworks designed to generate new data samples that resemble a given training dataset. They consist of two neural networks, a Generator and a Discriminator, which are trained simultaneously through adversarial processes. The generator network takes random noise as input and generates data samples that mimic the training data, aiming to produce realistic data to fool the discriminator. In contrast, the discriminator network evaluates the authenticity of the data samples, distinguishing between real data from the training set and fake data produced by the generator. The adversarial training process continues until the generator produces data samples indistinguishable from the real data to the discriminator (see Figure 2).

First, we built a baseline GAN trained on six images, but this was unable to converge as the discriminator was too effective so the model generated random noise (see Figure 3). Traditionally, GANs are trained on datasets with 100,000+ images to achieve proper convergence (Salian, 2020). Due to the small dataset, the images generated by our model were of very poor quality. This model is the main approach explored to date in the literature regarding comic book character generation (Verduyn, 2022). However, even if this approach had worked, it would not have been practical for our purposes. One main downside is that GANs are not inherently linked to any textual components. Thus, even if we had been able to generate coherent images, they would have been uncontrollable by us. As we aim to demonstrate the use of image generation for potential efficiency gains in comic illustration, this lack of control makes GANs impractical for our specific application.

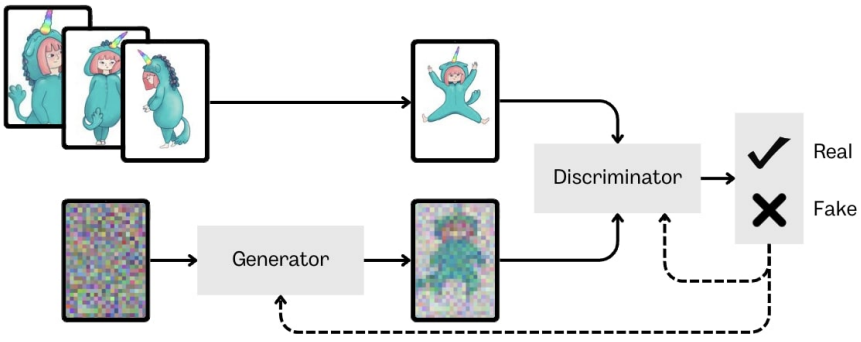


Figure 2: GAN Diagram

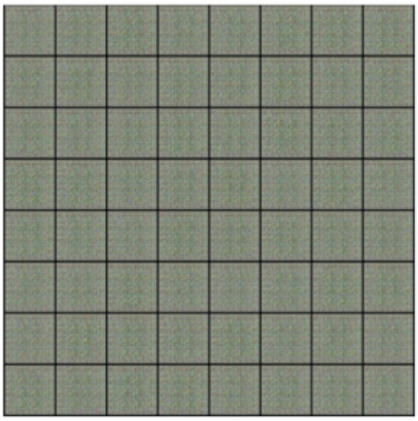


Figure 3: 8x8 Grid of GAN Baseline Results

4.2 Diffusion Models

Given the low quality of the GAN results, we decided to work on setting a diffusion baseline. Diffusion models are a type of generative model that gradually transform a simple noise distribution into a complex data distribution by learning to reverse a noising process. Essentially, these models start with a noisy version of the data and iteratively refine it to produce high-quality samples. They are used because of their ability to generate high-fidelity images, their robustness in handling complex data distributions, and their effectiveness in creating diverse outputs. Diffusion models have gained popularity for their superior performance in image synthesis and the generation of realistic and detailed images (Luo, 2022).

Interestingly, the results with the diffusion baseline were more conclusive than with GANs. We observed that the generated pixels were in the correct color palette (blue/teal), albeit requiring some imagination to see clearly (see Figure 4). However, the images were hyper-pixelized. We experimented with multiple versions, finding that downsampling our images allowed the model to learn the colors better but without any coherence, while using higher quality images resulted in less coherence. Overall, the results were not conclusive, and the generated images still lacked any inherent textual component. It is important to note that the most notable version of this model (stable diffusion) was trained extensively with billions of input data to achieve high-quality results (Baio, 2022).

Given that we will be working with fine-tuning stable diffusion models, setting the diffusion baseline by training from scratch is very important. This step ensures that we understand the foundational aspects of the model and its performance with our specific dataset. However, we are fully aware and conscious of the significant costs associated with training stable diffusion models from scratch. For instance, training a stable dif-

fusion model can cost over half a million dollars in GPU alone (Mostaque, 2022). This underscores the importance of optimizing our approach and leveraging existing pre-trained models to minimize resource expenditure. We will delve deeper into the mathematical formulas behind this model in Section 5, where we further explore the use of diffusion models.

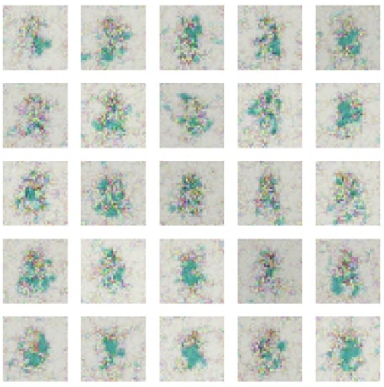


Figure 4: 5x5 Grid of Diffusion Baseline Results

4.3 Stable Diffusion Models

Our next step was to use a pre-trained Stable Diffusion Model and prompt engineering to generate a higher quality version of our character. We chose Stable Diffusion models over alternatives like DALL-E due to their efficiency on consumer hardware and open-source nature. These models, publicly available with accessible weights and architectures, are valuable for experimentation and application by researchers, developers, and the general public (Cortés, 2023).

Diffusion models excel in image and audio generation tasks, forming a fundamental part of synthesis systems such as DALL-E, Stable Diffusion, and Diffwave. Unlike GANs, diffusion models require multiple iterations of the neural network to gradually generate samples. While this process can be slower, diffusion models offer greater potential in terms of quality and diversity in generated content (M. Chen et al., 2024).

Stable Diffusion models employ a comprehensive approach, utilizing a text encoder, a latent space diffuser for denoising, and an autoencoder decoder structure to produce final images (Cortés, 2023). This architecture has been extensively trained on a diverse dataset of 512x512 images from a LAION-5B database, enhancing its adaptability and versatility across various tasks and domains. (HuggingFace, 2022).

For our project, we specifically focus on several versions of Stable Diffusion: stable-diffusion-v1.4, stable-diffusion-v1.5, and stable-diffusion-xl-base-1.0. These versions have

undergone rigorous training on billions of images and training steps, making them well-suited for the approaches we intend to employ. To establish a third and final baseline, we generated images using prompts describing our character with the pre-trained weights. The prompts included "a girl with pink hair wearing a blue unicorn onesie." While the generated images were an approximation of the desired drawing, they did not accurately replicate the artist's style. Therefore, we will focus on fine-tuning the Stable Diffusion model to maintain the desired quality level.

(a) "a photo of girl in a blue unicorn onesie, pink hair"

(b) "a cartoon of girl in a blue unicorn onesie, short pink hair and bangs"

Figure 5: SDXL Images Generated from Prompt

5 Fine-Tuning Techniques of Diffusion Models

5.1 Mathematical Foundations of Diffusion Models

In recent years, diffusion models, inspired by the physics of non-equilibrium thermodynamics (Sohl-Dickstein et al., 2015), have achieved groundbreaking performance, surpassing previous state-of-the-art models like GANs and Variational Autoencoders (VAEs) (Dhariwal and Nichol, 2021). These models have been widely applied in various domains, including image and audio generation, control and reinforcement learning, life sciences, and black box optimization.

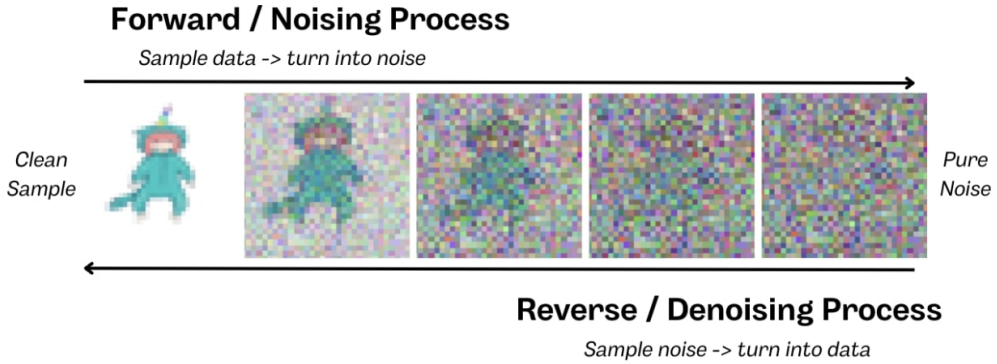


Figure 6: Diffusion Training - Forward and Backward Processes

Diffusion models operate through two key processes: the forward process and the backward process. The forward process involves gradually corrupting a clean sample from the data distribution by adding Gaussian random noise. This corruption is carried out sequentially, following a Markov chain, where each time step depends only on the previous one. As this process continues, the data distribution is progressively transformed into pure noise. In the infinite-time limit, the original data becomes indistinguishable from pure Gaussian noise.

The forward process can be mathematically described as a sequence of steps where noise is added to the data, transforming it from a clean sample x_0 to noisy sample x_T. This is represented by a Markov chain:

$$q(x_{1:T} \mid x_0) = \prod_{t=1}^{T} q(x_t \mid x_{t-1})$$

At each step t, the noise is added according to a Gaussian distribution, controlled by a variance schedule β_t, where $\beta_1 < \beta_2 < \cdots < \beta_T$ and each β_t is in the interval (0,1). The transition from $x_t - 1$ to x_t is given by:

$$q(x_t \mid x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}\, x_{t-1}, \beta_t\, I)$$

As t increases and $\beta_t \to 1$, the noise dominates, and the distribution of x_t approaches a standard Gaussian distribution (Ho et al., 2020). Thus, for a sufficiently large T, the final state is:

$$q(x_T \mid x_0) \approx \mathcal{N}(0, I)$$

In the backward process, a denoising neural network, typically a Convolutional Neural Network (CNN) named UNet, is employed. The UNet architecture is so named because of

its U-shaped structure, which allows it to capture both high-level and low-level features of the data. The denoising network is trained to reverse the forward process by sequentially removing the noise added during the forward process. This network takes noisy data and restores it to its original, clean state. Essentially, it learns to traverse the Markov chain in reverse, transforming pure noise back into meaningful data.

The reverse process is modeled as a Markov chain of conditional probabilities:

$$p_\theta(x_{0:T}) := p(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1} \mid x_t)$$

At each step t, the network predicts the mean μ_θ and optionally the variance Σ_θ of the distribution from which the next step is sampled:

$$p_\theta(x_{t-1} \mid x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t))$$

In practice, the variance is often fixed, simplifying the reverse transition to:

$$p_\theta(x_{t-1} \mid x_t) = \mathcal{N}(x_{t-1}; \alpha_t(x_t - (1 - \alpha_t)\beta_t \epsilon_\theta(x_t, t)), \beta_t I)$$

Where ϵ_θ is the neural network's prediction of the noise (Ho et al., 2020).

One notable limitation of Stable Diffusion is its potential for inconsistencies when generating a series of images depicting the same character or object with precise attributes. This variability arises because the model relies on both text and random noise, leading to high variability. This challenge is particularly evident in applications requiring detailed control and consistency across multiple images. For instance, creating a series of images featuring the same character or object with precise attributes and appearances in different poses or scenarios often proves challenging due to the inherent variability in interpreting textual descriptions (Ruiz et al., 2023).

Recently, there has been a notable shift in focus from generic image generation to subject-specific image generation, which is the focus of our research. Our aim is to generate diverse images of a specific character in various positions — a task that until recently would have been financially and computationally intensive, requiring the complete training of models from scratch. However, the increasing accessibility and popularity of publicly available models have paved the way for the development of advanced techniques that utilize large pre-existing models to explore generation of specific subjects or concepts. The techniques we focused on our paper are DreamBooth, LoRA, and the integrated approach of DreamBooth + LoRA.

5.2 DreamBooth

DreamBooth operates by retraining the entire gradient to associate a specific concept, such as a character represented by a series of images, with a unique identifier. This process begins with two primary inputs: a small set of images depicting the character and a textual prompt containing the identifier. The prompt is converted into a text embedding, where each word corresponds to a vector that encodes semantic information.

During training, the input images are subjected to different levels of noise, creating two versions: one with heavy noise and another with moderate noise. The model's objective is to transform the more noisy image at time step $t-1$ to approximate the less noisy version at time step t guided by the text embedding of the unique identifier. This iterative process allows the model to gradually enhance its ability to accurately recreate images associated with the identifier, ultimately improving its understanding and representation of the specific character.

A loss function is used to measure the disparity between the model's output and the expected less noisy image, guiding gradient updates that refine the model's parameters. Through repeated iterations, DreamBooth evolves the model's internal structure to better associate the identifier with the character concept, gradually enhancing its ability to denoise and recreate images associated with the identifier. Figure 7 shows a visual representation of the DreamBooth technique. In this case, the unique identifier would be "SKS", and the input images are of our original character.

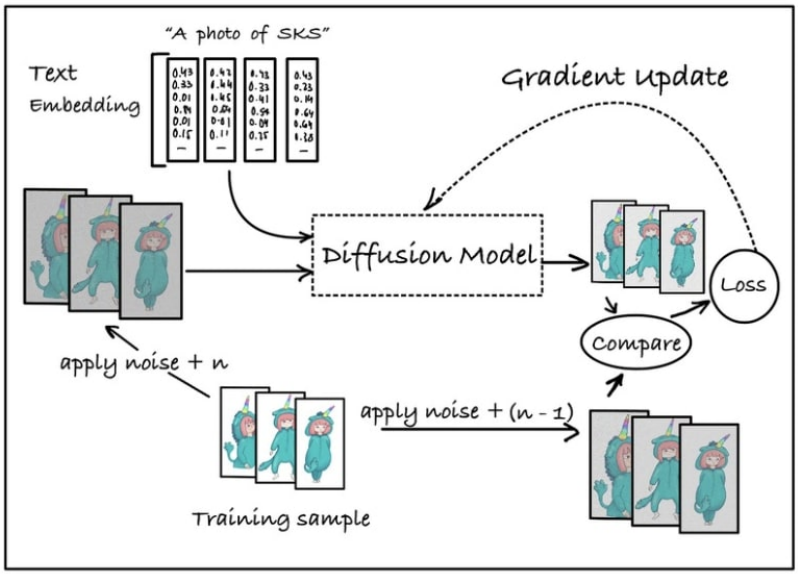


Figure 7: DreamBooth Fine Tuning Technique Process

One notable consideration of DreamBooth is its requirement to create a distinct model for each concept trained, leading to potential storage inefficiencies due to the large amounts of memory needed for each distinct concept model. Despite this drawback, Dreambooth is highly effective for training specific concepts into the Stable Diffusion model, enhancing its capacity to generate accurate outputs based on the given identifier and associated images. Our thesis will explore the application of this technique on stable-diffusion-v1.5 in generating a single character in different positions, particularly within the comic art domain. However, its versatility suggests potential applications across various other fields.

We enhanced our DreamBooth model through fine-tuning, achieving the best image results with 1000 training steps and 50 inference steps. We observed that adjusting these parameters, particularly the inference steps, had a significant impact on the final image style. Figure 8 illustrates the varying impacts of inference settings. Incorrectly set inference levels resulted in grainy, brushed-style images with inaccurate color palettes that did not match our data. In contrast, an inference of 50 produced images with a more consistent style that better aligned with our dataset.

(a) 25 Inference Steps (b) 50 Inference Steps (c) 100 Inference Steps

Figure 8: DreamBooth Inference Tuning Results

Figure 9: DreamBooth Tuning Results

Initial results from DreamBooth show promise, particularly in achieving close color matching with our dataset, as seen in Figure 9. These images represent some of the better outcomes from this approach. Although overfitting was not an issue, the model struggled to refine its understanding beyond 1500 steps and did not surpass the performance of earlier stages. When evaluating the results, some images generated class issues, as the model incorporates randomly generated class images, leading to inconsistencies with the character. Despite these inconsistencies, the method demonstrates the model's ability to grasp the concept. However, the variation in results suggests limitations when applied to real-world applications.

5.3 LoRA

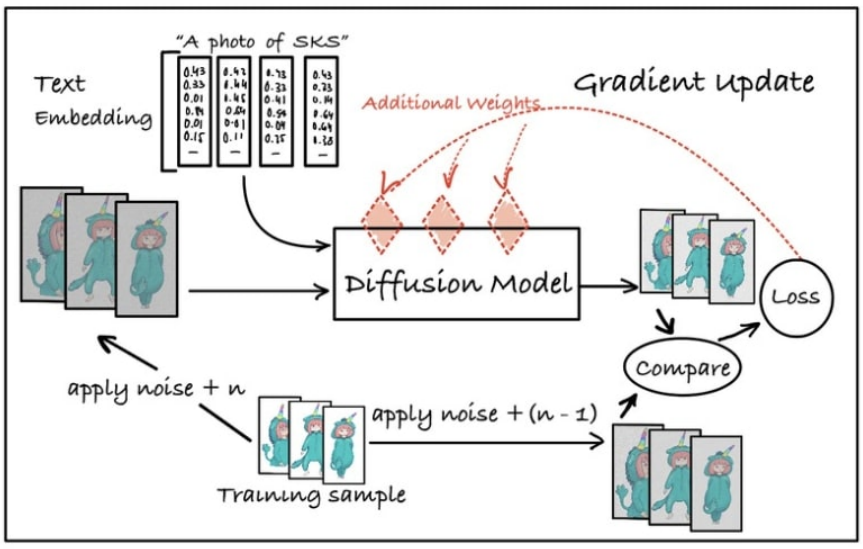


Figure 10: Low Rank Adaptation Fine Tuning Technique Process

Low-Rank Adaptation (LoRA) is a technique used to fine-tune Stable Diffusion models, specifically stable-diffusion-v1.4 (Hu et al., 2021). Similar to DreamBooth, LoRA offers a more efficient solution by creating and training only a few new layers instead of adjusting the entire model's gradient. Implemented within the cross-attention layers that connect image data with textual prompts (Yang et al., 2023), LoRA allows the diffusion model to recognize new words as unique concepts. This enhancement improves the model's effectiveness without altering its fundamental structure or requiring a complete retraining of all weights for each new concept (HuggingFace, 2023c).

Figure 11: LoRA Tuning Results

Figure 11 shows examples of images generated using the LoRA technique. During tuning, the model creates class images to artificially augment the data, aiding in character learning. However, the resulting images often resembled the class images more than the target character. While some generated images vaguely resembled the cartoon character and were conceptually closer, they still fell short in accuracy. The colors did not match the original, and there was a lack of consistency between the generated images. Consequently, these images are not suitable for practical applications.

Among the three fine-tuning methods applied, LoRA yielded the poorest results. This underperformance is largely attributed to the limited GPU power available during tuning. Tuning the model to 600 steps took roughly 6 hours, which was already longer than any other method. Attempting to tune to 900 steps took more than 12 hours, exceeding our allowed session time in Kaggle. This extended runtime is due to factors inherent in the LoRA approach. Unlike DreamBooth, which directly integrates new concepts into the model weights, LoRA adjusts specific low-rank structures within the pre-trained model.

This process involves complex parameter selection and optimization steps to minimize disruption to the model's existing knowledge, increasing computational overhead and memory management requirements, leading to prolonged processing times.

Although LoRA theoretically offers finer control over parameter adjustments, which could enhance image quality and preserve original features, it requires substantially more time and resources than the other methods. For our research, which aims to generate diverse images with minimal computational resources to handle limited data and reduce environmental impact, LoRA is not a viable option. Further tuning might improve LoRA's results, but achieving the same level of consistency as DreamBooth + LoRA would require significantly more time and resources, making additional tuning both unnecessary and counterproductive to our project's efficiency and sustainability goals.

5.4 DreamBooth + LoRA

We employ a script to fine-tune the pre-trained stable-diffusion-xl-base-1.0 (HuggingFace, 2023b) using both the DreamBooth and Low-Rank Adaptation (LoRA) techniques(Github, 2023). Research has shown that the Stable Diffusion XL (SDXL) model outperforms all previous versions of Stable Diffusion. This enhanced performance is attributed to its larger UNet backbone, the addition of two new conditioning techniques, and a separate diffusion-based refinement model (Podell et al., 2023). The goal was to train the model to generate images from specific prompts, focusing on customization to produce high-quality, personalized images for particular subjects or styles. To achieve this, we utilize pre-trained models, including a Variational Autoencoder (HuggingFace, 2023a), text encoders, and the UNet model used in Stable Diffusion. LoRA layers are integrated into the UNet and text encoder models to facilitate fine-tuning, the loss is used to update the model weights, including the additional the LoRA weights.

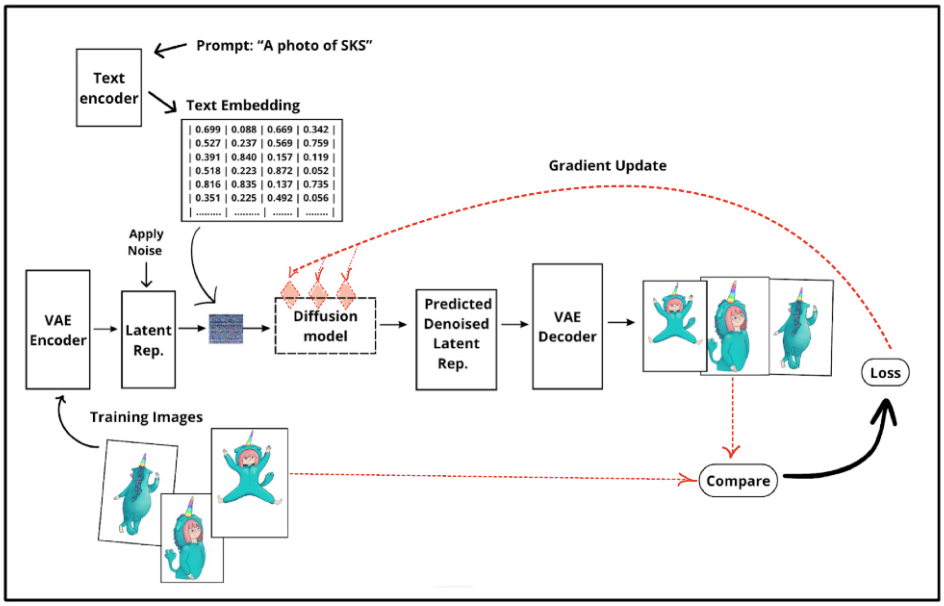


Figure 12: DreamBooth + LoRA Fine Tuning Technique Process

After experimentation, it was noticed that adding a latent tensor before the diffusion model and combining DreamBooth and LoRA techniques offered several key advantages. Firstly, it ensures the model retains general knowledge while learning new concepts through the incorporation of class images and loss functions. Secondly, the resulting file is significantly smaller and faster than the original DreamBooth. Moreover, it outperforms DreamBooth and LoRA alone in both speed and quality. This approach allows the model to accurately capture new character features and artistic styles, producing superior results.

Model Name	**Weight Size**	**Run Time**	**Visual Quality**
Dream Booth	4 GB	2 hours	low
LoRA	3.23 MB	12 hours +	low
Both	25 MB	1 hour	high

Table 1: Comparison of Models

In order to truly leverage the potential of this model, it is crucial to carefully adjust several hyper-parameters during the training and inference processes. Three key aspects of tuning that has shown to profoundly impact the model's effectiveness are the learning rate, the number of training steps, and the inference steps (Jindal, 2024). By optimizing these parameters, one can achieve a balanced model that not only learns effectively but also performs efficiently during deployment. This thesis explores the importance of fine-tuning each of these parameters and the benefits they bring to the model training and

inference processes.

In the context of fine-tuning models with DreamBooth and LoRA, the number of training steps is particularly important. These methods often involve small, specialized datasets and require a delicate balance to integrate new knowledge without overwhelming the model's existing capabilities. By carefully tuning the number of training steps, users can ensure that the model adapts well to new tasks while retaining its ability to perform across a broad range of scenarios.

The cost associated with running a large number of training steps is linked to increased computational expenses. During our model training, we estimated that approximately 6 minutes are required per 100 steps. Thus, it is essential to find a balance between the quality achieved and the computational expense. Research papers such as "Diffusion Models Beat GANs on Image Synthesis" by Dhariwal et al. typically recommend around 1000 steps to achieve this balance (Dhariwal and Nichol, 2021). In the Figure 13, we illustrate the output variation across our model checkpoints from 100 to 2000 training steps.

Figure 13: Training Steps Tuning

Adjusting inference steps can help leverage the enhanced capabilities of the fine-tuned model more effectively. However, similarly to the training steps, it is crucial to consider the computational cost associated with running additional inference steps. Through experimentation, we estimated that each increment of 10 inference steps incurred a cost of approximately 30 seconds. After testing five different prompts with increments from 10 to 100 steps, we observed that the quality did not improve beyond 50 steps.

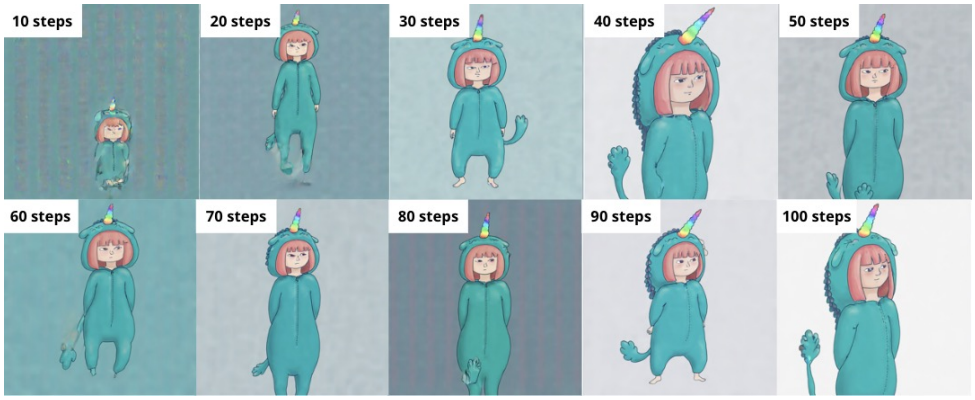


Figure 14: Inference Steps Tuning

Finally, the learning rate is one of the most critical hyperparameters in machine learning and deep learning, dictating how quickly or slowly a model learns during training. Setting the learning rate too high can cause the model to converge too rapidly to a suboptimal solution, or worse, cause the training process to become unstable and diverge. On the other hand, a learning rate that is too low can lead to excessively slow convergence, where the model takes a very long time to learn, potentially getting stuck in local minima (Pedersen et al., 2017).

Fine-tuning the learning rate is particularly important in the context of training large models with methods like LoRA and DreamBooth. These techniques rely on efficiently adapting a pre-trained model to new data or specific tasks. A well-tuned learning rate allows these adaptations to occur smoothly, ensuring that the new information is integrated effectively without disrupting the pre-existing knowledge embedded in the model (Han et al., 2023).

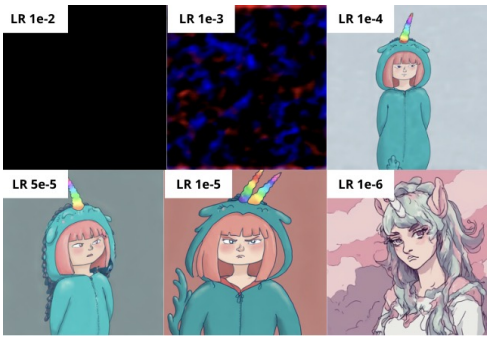


Figure 15: Learning Rate Tuning

The images above demonstrate that different tuning of the learning rate is crucial in determining the resulting image quality. Diffusion models such as Stable Diffusion XL (SDXL), utilizing both DreamBooth and Low-Rank Adaptation (LoRA), can accommodate various

styles, ranging from realistic to anime-style or a new artist's style Kabir et al., 2024).

6 Performance Metrics

6.1 Quality Ratio

The table below illustrates that an increased number of training steps does not necessarily correlate with an improved training quality ratio. The quality ratio is defined in this thesis as the visual accuracy of the model's output compared to the input data, as determined by the weights generated during training. For every 100 steps, 39 outputs created from the same 13 prompts were manually compared to the original data to identify any imperfections. Notably, the model trained for 1800 steps exhibited the highest quality ratio.

Num of steps	Quality ratio
100	0.00
200	0.03
300	0.03
400	0.00
500	0.08
600	0.30
700	0.12
800	0.23
900	0.15
1000	0.23

Num of steps	Quality Ratio
1100	0.15
1200	0.18
1300	0.10
1400	0.18
1500	0.26
1600	0.08
1700	0.12
1800	**0.35**
1900	0.21
2000	0.12

Table 2: Quality Ratio by Number of Training Steps

6.2 Training Loss

Since the Dreambooth+Lora model was our best-performing model, we decided to analyze its training loss over time across six different learning rates. This analysis aims to provide insights into how each rate affects the model's training dynamics and, ultimately, its ability to generate detailed and stylistically accurate images.

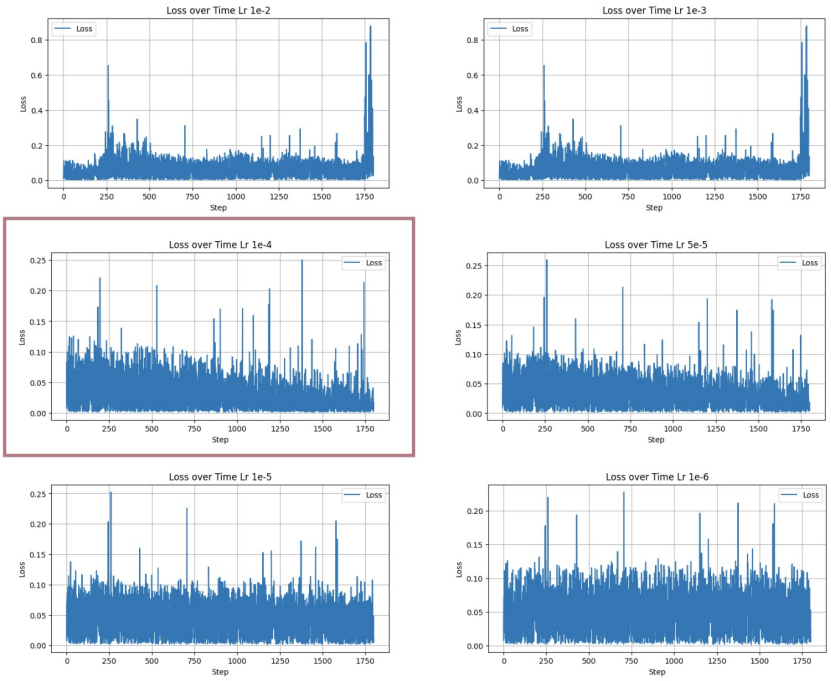


Figure 16: Loss Across Learning Rate Tuning

From Figure 16, it is evident that the learning rates have a significant impact on the loss behavior and model convergence. High learning rates, such as 1e-2 and 1e-3, result in high and volatile loss values, suggesting an overshooting of optimal weight adjustments and a lack of stable convergence. On the other hand, more moderate learning rates like 1e-4 and 5e-5 show a lower, more stable loss trajectory, indicating more effective learning. Very low rates, such as 1e-5 and 1e-6, exhibit minimal loss reduction, suggesting that the rate is too low for the model to adequately adapt and improve.

The optimal learning rate for this particular model tuning was found to be 1e-4, which demonstrated the best balance between loss minimization and visual image quality, retaining fine details and maintaining stylistic accuracy. This aligns with both the loss plots and the outputs in Figure 15, confirming that these settings produced the best results. Rates like 5e-5 and 1e-5, while still producing high-quality images, showed slight variations in detail and style, with the latter rate indicating a slow convergence that might not be efficient in practical scenarios. Understanding and monitoring the loss during training is crucial as it directly correlates with the model's ability to learn and perform tasks. Therefore, careful management of the learning rate is key to leveraging the full potential of advanced generative models like Stable Diffusion XL in producing customized, high-fidelity images.

7 Results

7.1 Dreambooth + LoRA model

From previous analyses of the DreamBooth + LoRA model, it has been established that the learning rate, number of training steps, and inference steps significantly influence the final output. Specifically, the model achieved its best results when trained with 1,800 steps, used 50 inference steps, and set a learning rate of $1e-4$. These settings produced high-quality results that closely resembled the original style and character of the artist.

These settings were implemented in code on the Kaggle platform using NVIDIA GPU P100 and the necessary packages included: Diffusers, Torch, Peft, and Transformers. Under these conditions, the training process took approximately ninety minutes, a much shorter time than previous methods. Furthermore, to maintain the uniqueness of the generated character and avoid duplicating concepts already present in the pre-trained Stable Diffusion XL model, the newly developed character was given the unique identifier of "UnicornGirl" in the training prompt.

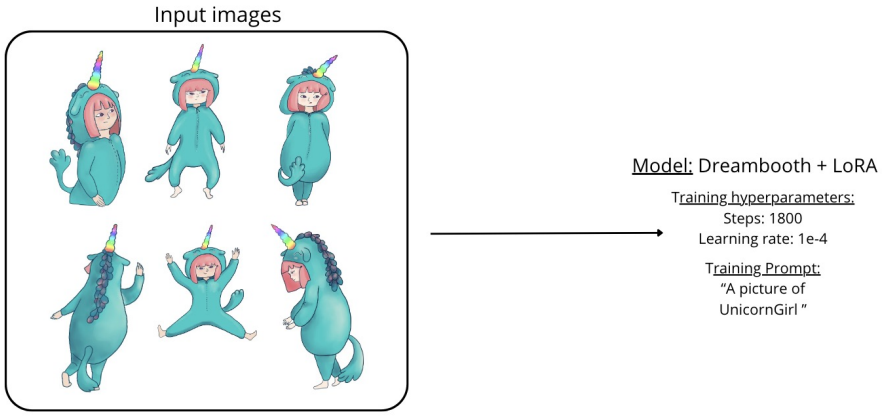


Figure 17: Input Diagram

After training, the model's weights were uploaded to a private HuggingFace repository to make them shareable. Upon downloading the weights from HuggingFace, we conducted tests using different prompts with 50 inference steps. Generating three images took approximately 150 seconds per prompt. The objective was to evaluate whether the model could learn new body positions and facial expressions while retaining concepts from its pre-training phase.

Examining the results shown below [18, 19, 20 and 21] we observed that the style and characteristics of the original character were preserved while achieving the set goals. With a quality ratio of slightly more than one-third, the model effectively showcased the chosen character in various positions, displaying different emotions, and interacting with

objects and different backgrounds. Although the backgrounds consistently matched the character's art style and color palette, their study is beyond the scope of this thesis.

(a) "A picture of UnicornGirl sad"

(b) "A picture of UnicornGirl smiling"

(c) "A picture of UnicornGirl angry"

Figure 18: Different Facial Expressions

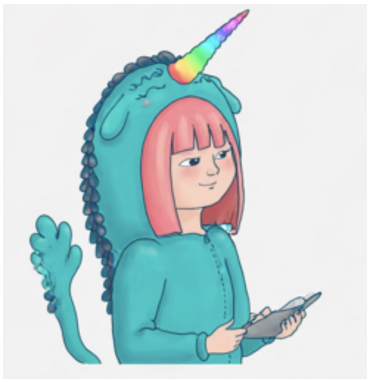
(a) "A picture of UnicornGirl praying"

(b) "A picture of UnicornGirl with hands on face"

(c) "A picture of UnicornGirl laughing"

Figure 19: Different Body Movements

(a) "A picture of UnicornGirl with ice cream"



(b) "A picture of UnicornGirl holding a book"

(c) "A picture of UnicornGirl eating a sandwich"

Figure 20: Pre-trained Concepts

(a) "A picture of UnicornGirl in front of a castle"

(b) "A picture of UnicornGirl watching the sunset"

(c) "A picture of UnicornGirl walking in the woods"

Figure 21: Backgrounds in the Same Art Style

7.2 Comic Application

Can generative AI applications contribute to the creative process of creating a comic strip? This was a key question when we embarked on this project. As generative AI continues to advance, it prompts the question of whether artificial intelligence can support creative endeavors. We acknowledge that creating a comic strip involves a significant creative component, and we do not intend for these applications to replace that aspect. Instead, our aim is to explore their potential to assist artists in generating mock-ups or aiding in various stages of the creative process.

With collaboration from esteemed French comic writer Bertrand Escaich, known as BeKa, we applied our model to a real-world scenario. We crafted a story featuring Lily, a young girl who stumbles upon a magical onesie while preparing for school, leading her on an

adventure. Our models specifically focus on generating individual characters, thus background settings or additional characters were not fine-tuned but we aimed for consistency when possible. This comic strip serves as a brief demonstration of how further research and tuning of these models could potentially support artists in their creative processes.

The story development began by showing Bertrand the images drawn by the artist and discussing some of our story ideas, emphasizing themes of adventure, school, and everyday life. We aimed to include a second character, represented twice by our own LoRA model, to demonstrate the potential for generating multiple characters. Bertrand crafted a detailed storyboard with perspective changes and specific panels relating to the story, providing detailed instructions for each panel, including dialogue and visual elements. This guidance was crucial for our prompt engineering process.

As data scientists, we iterated through several prompts for each image, spending about an afternoon to create the story by writing and adjusting the input prompts. This process illustrated the potential time savings on comic book projects. Ideally, multiple LoRA-DreamBooths would be used for each character, backgrounds, settings, and more, showcasing the significant efficiency gains possible with further development and refinement of these models.

We conclude that we were able to very closely generate the character to match the expected positions. Consistency was maintained in certain aspects, such as the background colors: pink for indoor room scenes and teal blue for outdoor forest scenes. Additionally, we successfully included two characters, although both were our original character, which presented challenges as the algorithm struggled to replicate the same character twice accurately. Exploring the effects of running multiple DreamBooth+LoRA weights within the same model could provide insights into efficiently adding new characters. Consistency was also achieved with specific elements like rounded, wooden doors with handles on the right side when closed from inside the room. In addition to evaluating visual consistency, we are also able to showcase proficiency in portraying character emotions and interactions as specified by the artist. The comic successfully showcases a range of emotions such as surprise and excitement at various scenes, effectively demonstrating the ability to interpret and depict emotional cues throughout the narrative even if not trained on that specific data.

However, the generated content was not without flaws. For instance, some doors featured windows while others did not, indicating inconsistencies in the background elements. Despite this, backgrounds were not the core focus of our project; the primary goal was character consistency. We observed that the character generated was very consistent with the style established by Meignaud and matched the storyboard framings as outlined by

Escaich (see Figure 24 in Appendix). Overall, while there are areas for improvement, the results demonstrate significant progress in using AI to streamline and enhance the comic creation process. The prompts used to generate each cell in the comic can also be found in appendix (see Table 3).

Figure 22: Comic Strip Application

8 Conclusion

This research explored the potential of generative AI, specifically using LoRA and DreamBooth fine-tuning techniques on Stable Diffusion models, to enhance the comic book creation process. Our focus was on generating new versions of an original character from limited data, addressing the challenges of data scarcity and the need for efficient, high-quality image generation with minimal computational resources.

The findings of this project demonstrate that fine-tuning pre-trained models, such as Stable Diffusion, with limited character images can produce satisfactory results. Among the methods explored, the combined DreamBooth and LoRA approach yielded the best outcomes, generating images that maintained the original character's style while accurately reflecting various poses and expressions. This method also proved to be efficient, both in terms of computational resources and environmental impact.

A key takeaway from this research is the significant potential of generative AI to streamline repetitive tasks in comic book creation, allowing artists to concentrate on their core strengths and creativity. By reducing the time and effort required for drawing consistent character images, AI tools can enhance productivity and open new possibilities for artistic expression. However, the study also highlighted challenges such as overfitting, the need for precise fine-tuning, maintaining stylistic consistency and low quality ratio. These challenges underscore the importance of continued research and development to optimize fine-tuning techniques and explore new methodologies.

The practical application of these techniques in collaboration with a professional comic book artist illustrated their real-world potential. The creation of a comic strip using generated character images showcased how AI can support artists in developing stories and visual content, suggesting substantial time savings and efficiency gains.

In conclusion, this research marks a significant step towards integrating AI into the comic book industry. By leveraging fine-tuning techniques on pre-trained models, artists can achieve high-quality, consistent character images from limited data, enhancing their creative process and productivity. Future research should focus on further refining these techniques and exploring their applications in generating complex scenes and backgrounds. This study opens the door to a future where the comic book creation process is more efficient, allowing for greater creative freedom.

References

Augereau, O., Iwata, M., & Kise, K. (2018). A survey of comics research in computer science [Submission received: 21 May 2018 / Revised: 15 June 2018 / Accepted: 20 June 2018 / Published: 26 June 2018]. *J. Imaging*, *4*(7), 87. https://doi.org/10.3390/jimaging4070087

Baio, A. (2022). Exploring 12 million of the 2.3 billion images used to train stable diffusion's image generator. https://waxy.org/2022/08/exploring-12-million-of-the-images-used-to-train-stable-diffusions-image-generator/

Chen, J., Liu, G., & Chen, X. (2020). AnimeGAN: A novel lightweight GAN for photo animation. In *Communications in computer and information science* (pp. 242–256). Springer Singapore.

Chen, M., Mei, S., Fan, J., & Wang, M. (2024). An overview of diffusion models: Applications, guided generation, statistical rates and optimization. *arXiv preprint arXiv:2404.07771*. https://arxiv.org/pdf/2404.07771.pdf

Cortés, M. L. (2023, June). *Subject-driven generation techniques for stable diffusion model* [Master's Thesis]. DTU, Department of Applied Mathematics and Computer Science. http://www.compute.dtu.dk

Dhariwal, P., & Nichol, A. (2021). Diffusion models beat gans on image synthesis. *arXiv preprint arXiv:2105.05233*. https://arxiv.org/abs/2105.05233

Everaert, M. N., Bocchio, M., Arpa, S., Achanta, R., & Susstrunk, S. (2023). Diffusion in style. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. https://openaccess.thecvf.com/content/ICCV2023/papers/Everaert_Diffusion_in_Style_ICCV_2023_paper.pdf

Github. (2023). Diffusers [Accessed: 2024-06-25].

Hagström, N., & Rydberg, A. (2024). Ai-based image generation: The impact of fine-tuning on fake image detection. *Stockholm University DiVA Portal*. https://su.diva-portal.org/smash/get/diva2:1837609/FULLTEXT01.pdf

Han, L., Li, Y., Zhang, H., Milanfar, P., Metaxas, D., & Yang, F. (2023). Svdiff: Compact parameter space for diffusion fine-tuning (supplementary material) [Accessed: 2024-06-25].

Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models [Available at: http://jonathan-ho.com/papers/denoising.pdf].

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2021). Lora: Low-rank adaptation of large language models. https://doi.org/10.48550/ARXIV.2106.09685

HuggingFace. (2022, August). Stable diffusion: A new approach to training language models [Accessed: 2024-06-26].

HuggingFace. (2023a). Madebyollin/sdxl-vae-fp16-fix [Accessed: 2024-06-25].

HuggingFace. (2023b). Stabilityai/stable-diffusion-xl-base-1.0 [Accessed: 2024-06-25].

HuggingFace. (2023c). LoRA: Leveraging low-rank adaptation for text generation. https://huggingface.co/blog/lora

Jin, Y., Zhang, J., Li, M., Tian, Y., Zhu, H., & Fang, Z. (2017). Towards the automatic anime characters creation with generative adversarial networks [Submitted on 18 Aug 2017]. *arXiv preprint arXiv:1708.05509*. https://doi.org/10.48550/arXiv.1708.05509

Jindal, S. (2024). Subject-to-subject: Controllable subject guided text-to-image generation and editing in diffusion models. *Department of Electrical and Computer Engineering, University of California, San Diego*.

Kabir, A. I., Mahomud, L., Fahad, A. A., & Ahmed, R. (2024). Empowering local image generation: Harnessing stable diffusion for machine learning and ai. *Revista Informatica Economica*. https://www.revistaie.ase.ro/content/109/03%20-%20kabir,%20mahomud,%20fadad,%20ahmed.pdf

Liu, H., Xing, J., Xie, M., Li, C., & Wong, T.-T. (2023). Improved diffusion-based image colorization via piggybacked models. *arXiv preprint arXiv:2304.11105*. https://doi.org/10.48550/arXiv.2304.11105

Luo, A. (2022). Gans vs. diffusion models: Putting ai to the test [Senior Computer Vision Engineer, Aurora Solar]. https://aurorasolar.com/blog/putting-ai-to-the-test-generative-adversarial-networks-vs-diffusion-models/

Millon, E. (2023). Color-diffusion: A pytorch implementation of a color diffusion model [Accessed: 2024-06-28]. https://github.com/ErwannMillon/Color-diffusion

Mostaque, E. (2022). Cost of construction [Archived from the original on September 6, 2022. Retrieved September 6, 2022]. https://x.com/emostaque/status/1563870674111832066

Pedersen, M., Frank, M., & Biele, G. (2017). The drift diffusion model as the choice rule in reinforcement learning. *Psychonomic Bulletin & Review*, *24*, 1234–1251. https://doi.org/10.3758/s13423-016-1199-y

Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., & Rombach, R. (2023). Sdxl: Improving latent diffusion models for high-resolution image synthesis. https://arxiv.org/pdf/2307.01952

Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., & Aberman, K. (2023). Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation [Google Research. Boston University]. *Google Research*.

Salian, I. (2020). Nvidia research achieves ai training breakthrough using limited datasets [Data augmentation technique enables AI model to emulate artwork from a small dataset from the Metropolitan Museum of Art — and opens up new potential applications in fields like healthcare]. https://blogs.nvidia.com/blog/neurips-research-limited-data-gan/

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., & Ganguli, S. (2015, August). Deep unsupervised learning using nonequilibrium thermodynamics. In F. Bach & D. Blei (Eds.), *Proceedings of the 32nd international conference on machine learning* (pp. 2256–2265, Vol. 37). PMLR. https://proceedings.mlr.press/v37/sohl-dickstein15.html

Verduyn, M. (2022). Comic art generation using gans. https://www.nix.be/assets/pdf/Masterproef_MarnixVerduyn_KUL_MAI_2022.pdf

Yang, Y., Wang, W., Peng, L., Song, C., Chen, Y., Li, H., Yang, X., Lu, Q., Cai, D., Wu, B., & Liu, W. (2023). Lora-composer: Leveraging low-rank adaptation for multi-concept customization in training-free diffusion models. *arXiv preprint arXiv:2301.01234*.

9 Appendix

9.1 Preliminary Sketch of Unicorn Girl (Jordane Meignaud)

Figure 23: Preliminary Sketch

9.2 Comic Strip Storyboard (Bertrand Escaich)

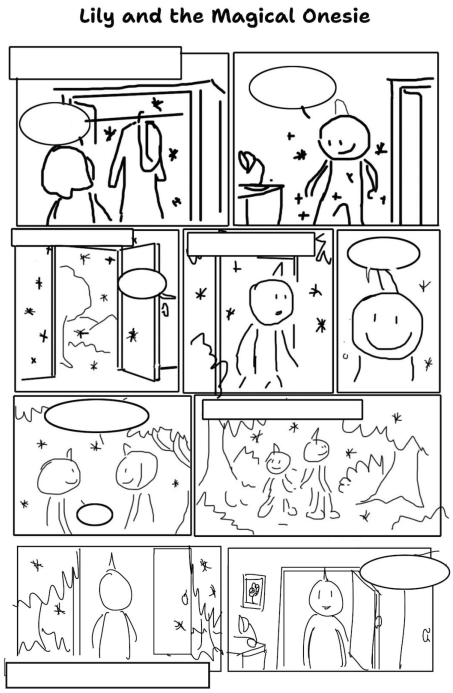


Figure 24: Comic Strip Storyboard

9.3 Comic Strip Prompts

In order to generate our comic book we utilized a variety of prompts to reach our desired images. Due to space constraints, we have opted to include the working prompts in the appendix.

Cell	Prompt
1	A portrait of UnicornGirl wearing pink looking through her closet to find her blue unicorn onesie
2	A portrait of UnicornGirl happy and excited in her bedroom next to a desk
3	A door in the bedroom open with a tree in the style of UnicornGirl
4	A portrait of UnicornGirl walking through a magical door into the woods
5	A portrait of UnicornGirl happy and excited in her bedroom
6	A picture of UnicornGirl in playing with another UnicornGirl behind a tree
7	A portrait of UnicornGirl with another UnicornGirl walking happy through the forest
8	A portrait of UnicornGirl walking through a magical door out of the forest
9	A picture of UnicornGirl next to an arched wooden door

Table 3: Prompts for UnicornGirl Images