California Institute of the Arts

Generative AI Art Exploration and Image Generation Fine Tuning Techniques

by

Hyeong Choi

A thesis submitted in partial fulfillment for
the degree of Master of Fine Arts

Herb Alpert School of Music
Music Technology: Interaction, Intelligence & Design

2015

Supervisory Committee

Mentor

Mike Leisz

Committee Member

Kai Luen Liang

Committee Member

Ajay Kapur

Committee Member

Abstract

This thesis concentrates on extensive research regarding an ongoing Generative AI technology phenomenon from the perspective of the multidisciplinary artist. The research goes beyond focusing on incorporating Generative AI technology into the author's artistic practice or experimentation; instead, it also provides a history of how Generative AI technology progressed and a comprehensive analysis of the social interactions surrounding Generative AI technology. This exponentially growing technology led the art culture to face a new era and brought with it concerns and questions about redefining fundamental concepts: "what is art?" or "what kind of artist am I?"

By examining the concerns arising from this momentous cultural shift and analyzing the issues related to innovative AI technology, which serve as a mirror reflecting our society, the author proposes ways in which contemporary artists can preserve their "humanness" while forging a symbiotic relationship with Generative AI technology in the upcoming era.

The goals of this paper include:

1 - Investigate the creative applications of Stable Diffusion, delving into the latent space utilized in machine learning algorithms and important 3rd party extensions of Stable Diffusion to gain a comprehensive understanding of their capabilities and limitations.

2 - Research how artists can re-establish the symbiotic relationship that creates an ideal balance between human creativity and the capabilities of machines in the rapidly approaching era of art technology. This includes examining historical precedents, current trends, and future possibilities for collaboration and co-creation between artists and AI systems.

By addressing these goals, the thesis aims to provide valuable insights and guidance for artists navigating the complex and evolving landscape of Generative AI technology while also contributing to the broader conversation about the future of art and the artist's role in this upcoming era.

Acknowledgments

I would like to thank to everyone who helped me during my MTIID degree here at CalArts. But I have to mention below especially who helped me throughout 3-year experience here.

Professor Kai and Mike, you guys are the best, I learned so much from you both.
My best friend and my life partner Lizzy, thanks for supporting me and encouraging me when I was lost.
And for the last, my mom and dad, thanks so much for supporting and believe in me, telling me that I can do anything.

Thank you so much again.

Contents

List of Figures

List of Tables

No table of figures entries found.

Chapter 1

Introduction

1.1 Introduction

It is no longer an overstatement to say that we are, entering a revolutionary period in the realm of digital art culture, primarily driven by the profound influence of cutting-edge AI technology. This rapidly advancing technology has set off a massive cultural wave, leaving a lasting impact not only on the art technology field but also on the broader art community and even affected humanity on a deeper level. However, we are merely at the beginning of this exponentially growing phenomenon, and there is still more untapped potential waiting to be discovered and explored.

The significance of this development lies in the fact that machine learning systems have now reached the "realm of creation," where AI system can create entirely new output by learning the inputting materials. This enables machines to learn and process data at a rate that surpasses any human or other entity. Picture a scenario in which a robot closely observes painters as they create artwork on a canvas, and then later imitates the artwork on its own canvas. Then when you examine the finished products, it becomes nearly impossible to determine which piece was crafted by the human painter. Now, envision this scenario unfolding within a digital environment, where the robot has almost unlimited data available at its disposal for learning, culminating in a system we refer to as "Artificial Intelligence." This AI has the capacity to draw in any style at a pace that surpasses human capabilities, continuously "producing" art endlessly. The results are astonishing, as it becomes increasingly difficult to differentiate between works

created by human hands and those generated by AI. In contrast to humans, robots can continuously generate art in the styles of nearly any artist they have learned from, exhibiting boundless "creativity." Figure 1 demonstrates the various styles and characteristics AI can employ to depict the author. On the left is an image of the author, which is part of the trained dataset for the AI model, and on the right are the resulting illustrations.

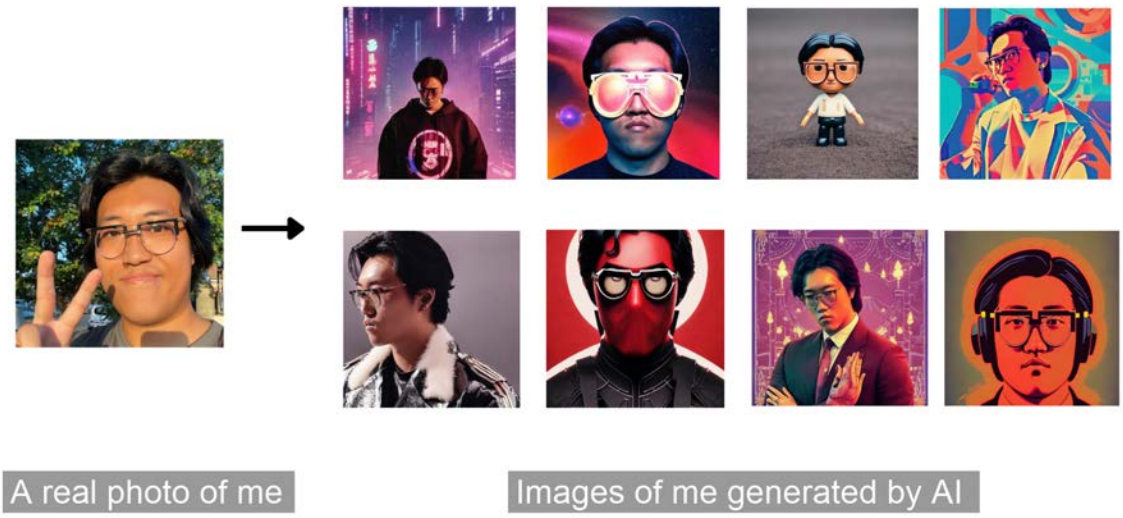

Figure 1-1 : AI Generative System Drawing Examples

Machines are now capable of participating in the creation of art, providing artists with access to and the opportunity to leverage the AI's distinctive “creativity”, which differs fundamentally from the way the human creative mind operates. Therefore, the relationship between artists and machines is shifting towards a new direction. In earlier generations of art technology, this relationship was characterized as one between an instructor and executor, with machines carrying out the rules and guidelines written or programmed by humans. However, in this new era, machines are advancing even further, adopting the role of "Collaborator." This shift enables users to interact with machines in innovative and unprecedented ways, solidifying machines as game-changers in the realm of art technology.

The evolving relationship between digital artists and technology offers substantial benefits to those who actively incorporate technology into their artistic practices. Imagine a scenario in which an artist is assigned the task of producing a thousand unique graphic illustrations of a "laser gun" on a daily basis. As humans have physical and mental limitations and can experience burnout during the creative process – a reminder that we are not machines – the artist may struggle to generate the vast amount of artwork required. This is where AI becomes a valuable asset. AI's creative network has the potential to extend beyond the boundaries of the human body, augmenting the artist's performance during the creative process by supplying an infinite array of artistic ideas and supporting labor-intensive and repetitive tasks. Consequently, artists can produce and complete a greater volume of high-quality artwork. With effective utilization of AI, the possibilities are virtually limitless.

Nonetheless, this extraordinary and overwhelming cultural phenomenon has generated concerns within the contemporary art community. The release of Stable Diffusion, one of the most influential AI models, as an open-source platform, has generated diverse opinions among artists. While some artists appreciate the time-saving benefits offered by the impressive applications of Stable Diffusion in the creative process, others argue that AI-generated art does not qualify as "true" art. Moreover, there is a risk that this technology could be misused, resulting in issues across various fields. For instance, the growing incidence of Deepfake videos or images spreading misinformation has led the public to question the authenticity of the content they encounter. Furthermore, certain AI models, adapted from Stable Diffusion models, have been specifically designed to generate harmful content, such as pornography and fake news.

1.2 Where are we currently?

It is essential to recognize that the current debate surrounding the impact of AI on art or concerns arising from technology is not an entirely new occurrence. This ongoing dialogue regarding the role of technology in art highlights the importance of continuously reevaluating and redefining our understanding of artistic expression. In the past, the art community has experienced similar disruptions, with questions and debates surrounding the nature and definition of "art." One notable example is the advent of image editing programs, such as

"Photoshop," which were initially viewed as revolutionary and groundbreaking innovations. These digital editing tools dramatically broadened the scope of digital art, provoking mixed reactions from artists, especially those with more traditional backgrounds. Some artists preferring traditional approach, opposed the idea of incorporating digital tools into the creative process, arguing that it diminished the authenticity and purity of artistic expression. But regardless, eventually Photoshop became one of the essential tools for digital artists, and the Generative AI art phenomenon has close similarities to the "Photoshop" sensation. But unlike Photoshop, Stable Diffusion exists in the public as an open source.

It's been less than a year since the introduction of Stable Diffusion to the world, and already, the world of digital arts has been captivated by the rapid and remarkable growth of this revolutionary technology. With new applications and branches of Stable Diffusion emerging on a weekly basis, the AI art community has found itself caught in an endless cycle of learning and experimenting with the latest developments and updates. As a result, the question of whether this technology is out of control has become a pressing concern among artists and AI enthusiasts. Increasingly, many are leaning towards agreeing that the technology is getting out of control, as the exponential growth of AI art technology shows no signs of slowing down.

With these concerns, it is crucial that we take a step back from the all-consuming phenomenon to thoroughly examine and reflect upon the events that have occurred, as well as anticipate what the future holds for the rapidly evolving world of Generative AI art. Therefore, this paper will focus on experimentation with technical aspects, concerns, and past events in the social aspect and also find an answer to how we can re-establish the symbiotic relationship that creates a perfect balance between humans and machines for the upcoming Art Technology era. We hope to gain a more comprehensive understanding of the current state of AI art and shed light on potential strategies for moving forward.

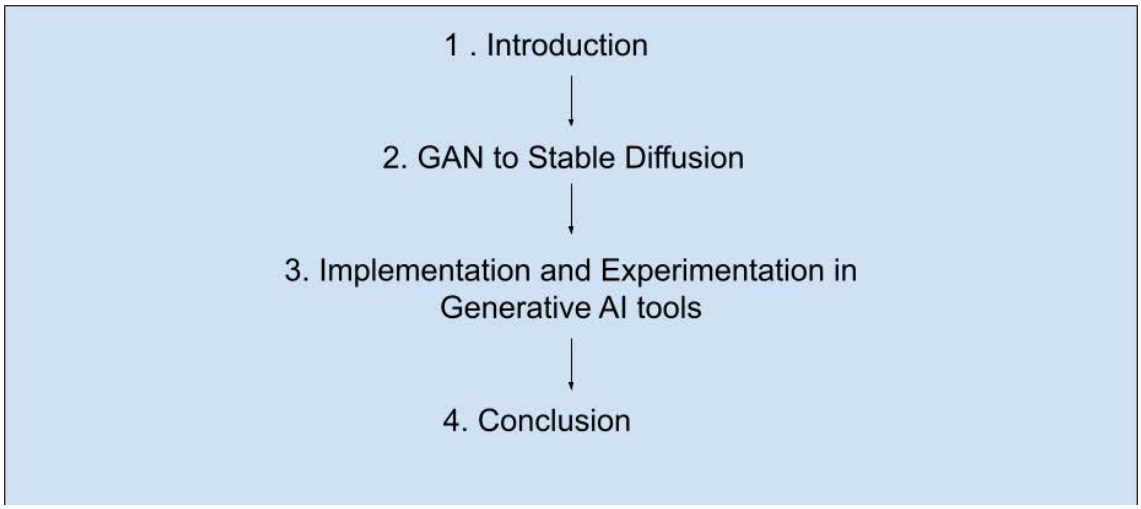

Figure 1-2 : Paper Structure

To gain a comprehensive understanding of the core of these automatic learning systems that operate within the virtual domain, it is essential to first trace their origins and examine how they have evolved into the global phenomena we witness today. Chapter two of this thesis will provide a concise history of Generative AI, starting from the invention of the GAN (Generative Adversarial Network) and delving into the various Generative AI system, including VQGAN+CLIP and Stable Diffusion.

Chapter three will provide an extensive overview of the application and implementation of AI art technology in practice. This section will investigate a range of collaborative art projects that author has conducted using Stable Diffusion and its essential extensions, examining the processes and outcomes from the perspective of an artist working across diverse fields. These case studies will offer valuable insights into the creative possibilities and challenges associated with AI-driven art.

Finally, at our conclusion, we will provide a comprehensive summary of the various experiments and explorations undertaken by the author throughout this study. Our destination is to address the central question: "How we can re-establish the symbiotic relationship that creates an ideal balance between human creativity and the capabilities of machines in the rapidly approaching era of art technology?"

Chapter 2

GAN to Stable Diffusion

2.1 GAN network and Min-Max Concept

The emergence of Generative AI technology can be traced back to less than a decade before the release of Stable Diffusion. In 2014, computer scientist Ian Goodfellow pioneered a revolutionary machine learning system called the Generative Adversarial Network (GAN). His paper showcased a machine learning system, "GAN Network," which was capable of analyzing a series of images and subsequently generating new images that closely resembled the originals. This innovative process was achieved through a series of iterative trial and error procedures.

The GAN Network is composed of two interconnected machine learning components: the "Generator" and the "Discriminator." The Generator is responsible for producing images that attempt to mimic an original input image. On the other hand, the Discriminator evaluates the generated images, determining whether they were the original inputs or created by the Generator. Upon making this determination, the Discriminator evaluates the results and supplies feedback to the Generator, guiding it towards generating results which are closer to the original inputs.

This cyclical interaction between the two machine learning components continues until the Generator successfully deceives the Discriminator. The point of deception occurs when a perfect balance is struck between the Generator, which aims to create a wide array of diverse results, and the Discriminator, which progressively narrows down the desired results by directing

the Generator's efforts throughout each iteration. In the field of data science, this concept of simultaneously maximizing and minimizing the range of generated results is referred to as the "Min-Max game."

Nevertheless, reaching the ideal balance within the Min-Max game can be a challenge, as either the Generator or the Discriminator can easily overpower the other if one component becomes more dominant. For instance, in a scenario where the Discriminator is substantially more dominant, it may not allow the Generator the opportunity to create a sufficiently diverse array of results. In such cases, the "Min" aspect of the game becomes overly dominant, consequently hindering the progression of the "Max" component. The art of discovering the appropriate balance within this "Min-Max game" is a crucial technique that is also applicable to various contemporary tools and approaches utilized in the field of Generative AI art. Its relevance to current AI Art technology will be explored in greater depth in Chapter Three of this thesis.

The GAN network presented a revolutionary approach to image generation by training machine learning models specifically designed for this purpose, which opened up the huge potential of the Generative AI field. However, the initial GAN network had limitations, such as generating only low-resolution images and being difficult to train. For instance, users were suggested to prepare at least 100 images - usually even more - for training a single GAN model. Despite these limitations, the introduction of GAN sparked rapid advancements following years, and branches of AI system algorithms that utilize GAN started to appear quickly. These advancements began with Deep Convolutional GANs (DCGANs), which implemented convolutional neural networks into the GAN architecture. By applying convolution layer on the GAN architecture, this AI system reduced the complexity of the model, which makes it the faster AI model in training and reduce the possibility of overfitting problem. This key development resulted in a marked enhancement in the performance of GAN-based systems, including pushing the field forward and laying the groundwork for even more sophisticated models.

2.2 StyleGAN, VQGAN and CLIP

The most notable breakthrough in the field occurred in 2018 with the introduction of StyleGAN, This generative AI system, based on the GAN network and officially named "Style-Based Generator Architecture for Generative Adversarial Networks," was developed by Tero Karras, Samuli Laine, and Timo Aila, who were part of the Nvidia research team. StyleGAN represented a significant leap in the capabilities of generative AI models, pushing the envelope in terms of image generation techniques and versatility.

Figure 2-1 : Example images of StyleGAN [14]

While StyleGAN demonstrates impressive performance in generating well-structured images, such as close-up face pictures, it exhibits a notable weakness when it comes to synthesizing images that lack a clearly defined shared structure, such as landscapes, abstract paintings, or a group of animals in random poses. In these scenarios, StyleGAN cannot effectively generate visually desired results, revealing a notable limitation in its capabilities for a broader range of unstructured and varied image generation tasks.

The realm of generative AI systems experienced a significant improvement with the introduction of the "Vector Quantized Generative Adversarial Network," or VQGAN. This Generative AI model constructs images by drawing upon a repository of learned components. This quantized GAN is characterized by its capacity for "self-focusing" achieved through incorporating "Transformers" within the image synthesis process. This concept was first presented in a 2020 paper titled "Taming Transformers" by researchers at the University of Heidelberg. [9]

Figure 2-2 : Example image generated with VQGAN based model

A primary distinction between VQGAN and other GAN models lies in its integration of convolutional neural networks (CNNs) with a newer network architecture known as Transformers. Traditionally employed in language models, Transformers are machine learning network architectures that rely exclusively on attention mechanisms, reducing recurrence in the pipeline workflow. Despite VQGAN's incorporation of Transformers, it is essential to note that the models are not trained using textual data; rather, they exclusively utilize image-based information. The developers just applied the Transformer architecture on top of the GAN system, which is an important innovation.

The VQGAN model learns an extensive library of highly versatile image components, and converts this data into vector information. It then employs and combines these vectors to reconstruct any given image presented to the model. While the overall structure of VQGAN resembles that of traditional GAN networks, its discriminator, which distinguishes between real and fake images, relies on vector combination data from the learned library, as opposed to random input noise. VQGAN provides users with the option to select and utilize various pre-trained models, each of which is designed with distinct libraries of universal image components.

VQGAN was showing advancement in generating unstructured images, such as landscapes or abstract art, and also in generating higher-resolution images. However, VQGAN falls short in generating structured images, such as those depicting objects like animals, or in image editing tasks – areas where StyleGAN is particularly renowned for its performance.

GAN systems have primarily operated based on image data; however, this conventional approach underwent a notable shift with the introduction of the CLIP model (Contrastive Language-Image Pretraining). In 2021, OpenAI - A company known for developing ChatGPT and Dalle-2 - proposed the Contrastive Language–Image Pre-training, or CLIP, in their research paper titled "Learning transferable visual models from natural language supervision." While CLIP is not a generative AI model in the traditional sense, it is trained to effectively determine the caption that fits best correspond to a given image. And this is an essential component primarily operating as a "Text Encoder" for more advanced image generator AI models, which we will investigate deeper later in Chapter 3.

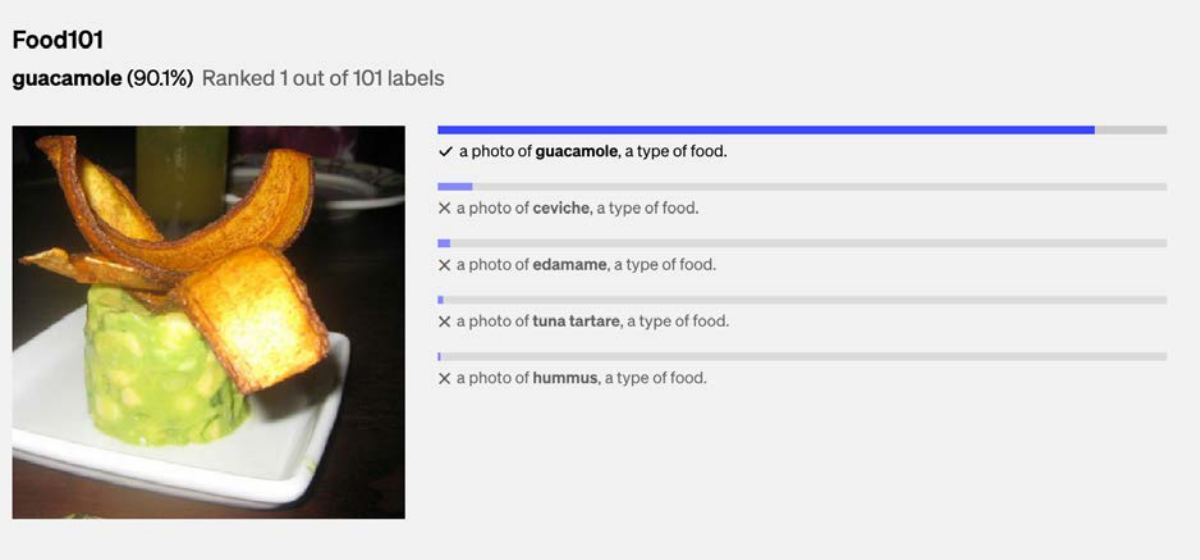

Figure 2-3 : Example of CLIP analyzing the given image from Food 101 dataset [5]

What sets CLIP apart and establishes its potential is its capacity for a newly introduced training technique, a "zero-shot learning". This means that the model is capable of performing exceptionally well on datasets it has not encountered before, outperforming models that have been traditionally trained on specific datasets. By incorporating the zero-shot learning technique, CLIP leverages the AI model's ability regarding how the model understands the text and associates it with images.

2.3 Stable Diffusion

Following the release of the Stable Diffusion by Stability.Ai on August 30th, 2022, the landscape of Generative AI art and its potential applications experienced a significant transformation. This innovative system, developed by a team of researchers at Stability.Ai, has garnered widespread attention and showcased its remarkable capabilities in generating high-quality and coherent visual content.

Releasing the Stable Diffusion AI model as an open source data model was a game-changer. Stability.AI provided full access to this giant neural network available to public, and it was allowed to modify and implemented in any way that users wanted to re-program. This innovative technology getting introduced as open source was the "opening door" to endless possibilities.

The launch of the Stable Diffusion AI model as an open-source data model marked a pivotal moment in the evolution of AI-generated art and technology. By making this groundbreaking neural network freely accessible to the public, Stability.AI encouraged an environment of creative exploration and collaboration, enabling individuals from diverse backgrounds to modify, implement, and re-program the model as they saw fit. This unprecedented move of introducing such a cutting-edge technology as an open-source resource served as the stimulus for unlocking a vast array of possibilities and potential applications.

For the AI art community, the advent of Stable Diffusion represented a huge leap forward, as it signaled the unveiling of a new world in which machine learning operated within the realm of latent space. While many artists and technologists eagerly embraced this groundbreaking technology, some traditional artists perceived it as a formidable and potentially threatening rival. The release of Stable Diffusion as an open-source generative AI system, ignited the "spark" that directed to the new era of "AI Art Technology." Numerous applications based on Stable Diffusion rapidly emerged, showcasing the enormous potential of AI technology in the art industry. Consequently, the AI art community experienced significant growth, and the internet soon became flooded with AI-generated artwork. The widespread adoption of Stable Diffusion garnered significant attention, and its influence continued to grow.

During a similar time frame, other renowned image-generating AI systems, such as Midjourney and Dalle 2, were also launched and introduced to the world. However, this thesis will primarily focus on Stable Diffusion. While it is certainly possible to achieve satisfying results with other systems—Midjourney, for instance, may even be considered easier to use by some—Stable Diffusion, as of early 2023, currently offers the most detailed control in comparison to how Midjourney and Dalle 2 operate. Furthermore, its open-source nature has encouraged the development of third-party extensions and interfaces, significantly enhancing the overall quality of the output and becoming an indispensable aspect of the author's creative process. We will delve deeper into the exploration and experimentation of this innovative AI art technology and its various applications in the next chapter.

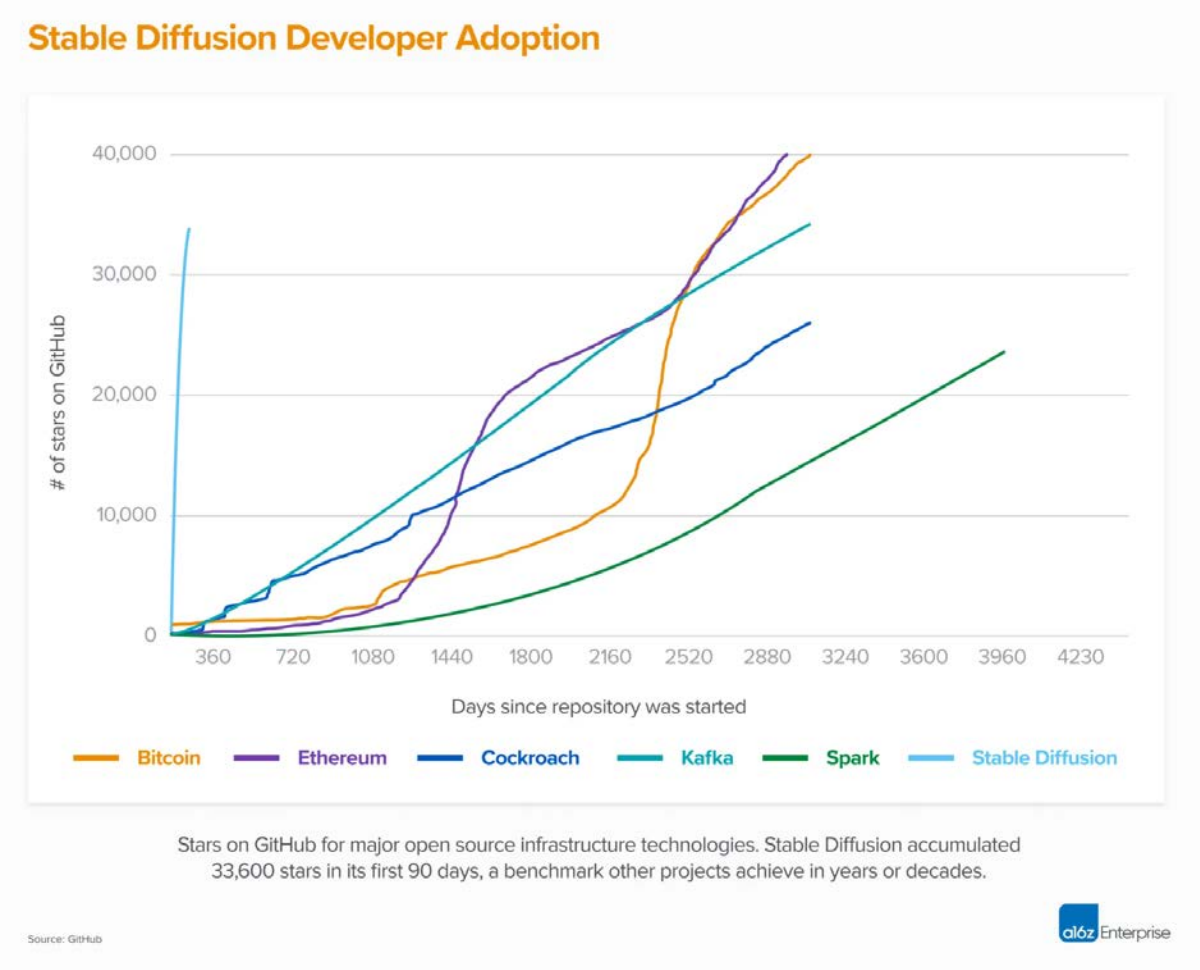

Figure 2-4 : Stable Diffusion Adoption Graph [19] Source : A16z and Github

Stability.AI's initial launch with its innovative open-source product, Stable Diffusion, was met with overwhelming success and rapidly captured both, industry experts and the general public's attention. Within a relatively short period of time, the groundbreaking AI model gained significant traction, accumulating a sizable following in the market. Just two months from the initial launch in August 30th, Stability.AI proudly announced that they had successfully raised an impressive sum of 101 million dollars in a funding round. This remarkable achievement was made possible through prominent investment firms, Coatue, Lightspeed Venture Partners, and O'Shaughnessy Ventures LLC. As a result of this substantial injection of finances, Stability.AI became even stronger to accelerate the development of future updates for the Stable Diffusion models. Furthermore, the company is firmly committed to expanding its AI offerings, with plans

to create additional clean models designed to cater to a diverse range of enterprise and consumer use cases.

2.4 Stable Diffusion Version 2

Soon after on November 24th 2022, Stability.AI released a significant update to Stable Diffusion, introducing Stable Diffusion 2.0. This new version boasted several key enhancements aimed at refining the user experience and improving overall quality during the generation process. To begin with, Stability.AI team entirely built the new dataset containing higher resolution images, significantly improving the output's quality. Also, they re-designed the pipeline workflow of Stable Diffusion version 1, meticulously engineered to strengthen the system's overall performance.

Another key development in Stable Diffusion 2.0 was the implementation of a new text encoder. The new text encoder was designed to provide a more accurate understanding of user input prompts, ensuring that the Stable Diffusion model could generate results closer to aligning with the user's intention. Notably, these transformative changes were successfully developed and deployed within the three months following the release of version 1, an impressive breaking upgrade given the scale and complexity of the improvements.

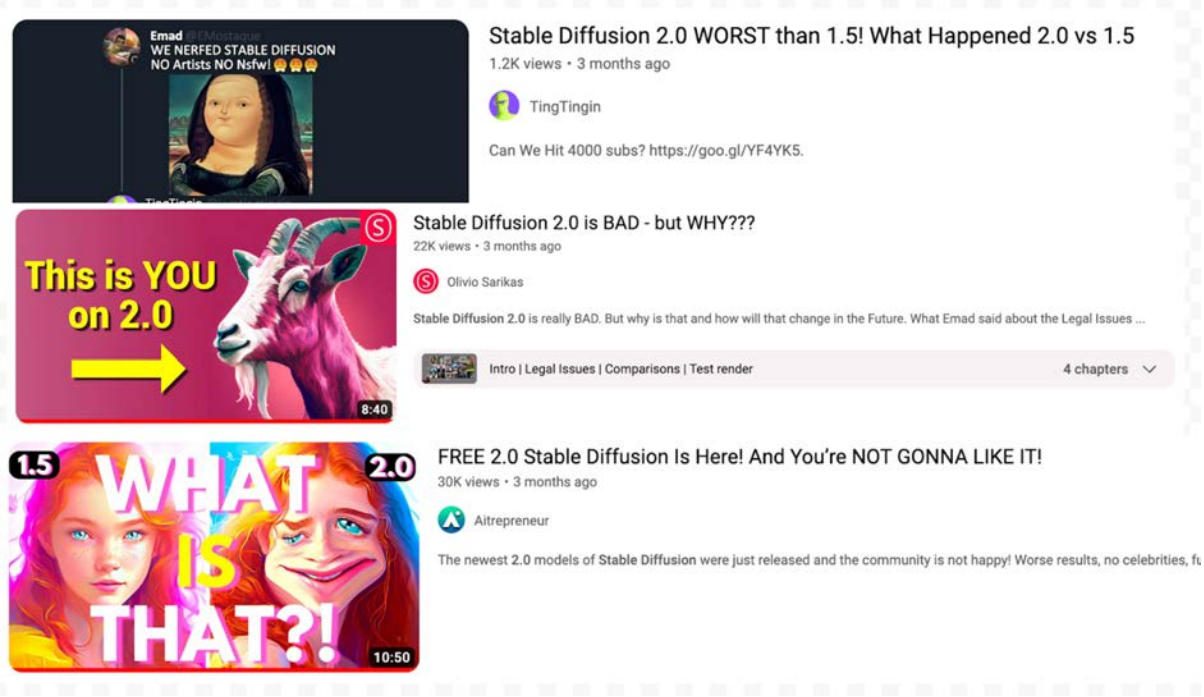

Figure 2-5 : Screenshots of Youtube Thumbnails from Active Stable Diffusion Community Members [23,25,27]

Despite these advancements in data workflow efficiency and dataset quality brought forth by Stable Diffusion version 2, the transition introduced an unexpected challenge. Users initially experienced confusion regarding the discrepancy in the quality of results they obtained from the Stable Diffusion 1.5 version compared to those produced by the recent 2.0 versions. They struggled to understand how they were able to obtain better outcomes using the older version as opposed to the 2.0 release. However, this confusion was eventually dispelled by other community members who conducted further experiments with the 2.0 version and shared their findings. There are several crucial factors that differentiate the 2.0 version from its predecessors, which help to explain the disparity in user experience.

Firstly, the most notable difference lies in the nature of the trained dataset used for each version. In the case of previous updates from Stability.AI, new versions were generally built upon and refined from the preceding version; therefore, they were all trained from the revised version of the same training dataset. However, Stable Diffusion 2.0 separated from this approach and was instead trained from the ground up using a completely new dataset. This dataset is a subset of the LAION-5B collection, specifically filtered for NSFW (Not Safe For Work) content and

created by the DeepFloyd team at Stability.AI. It consists of high-resolution images, with default sizes of 512x512 and 768x768 which contain more precise details in each pixel unit in the dataset images, and this brought Stable Diffusion version 2 capable of generating images with more refined details.

Comparably, Stable Diffusion version 1 was initially trained with images size of 256x256; This enhances the ability to generate images of smaller resolution in Stable Diffusion version 1, and most users choose to follow the popular workflow of generating output with a smaller resolution and then upscale it, which was a popular workflow among AI artists who lacked powerful GPU capabilities on their computer setups.

Figure 2-6 : Stable Diffusion Version 2 Resolution Comparison

The figure above is a detailed comparison sheet featuring images generated at various resolutions to illustrate the relationship between output resolution and image quality. For the purpose of this demonstration, I utilized the Stable Diffusion 2.1 768 version and incorporated the following positive prompt:

"a portrait of a goat wearing sunglasses, fine - art photography, soft portrait shot 8 k, mid length, ultrarealistic uhd faces, unsplash, intricate, casual pose, centered symmetrical composition, stunning photography, masterpiece, grainy, centered composition".

Four images were generated for this comparison, maintaining identical settings or parameters across each image. The sole exception is the output resolution, identified below each image.

Upon close examination of the generated images, it becomes evident that the output image quality progressively improves as the resolution increases.

Transitioning to the second point, Stability.AI made a decision to train the 2.0 version dataset utilizing an entirely new text encoder. This text encoder serves as a critical component within the Stable Diffusion system, responsible for how Stable Diffusion decoding and interprets the input prompt. Given the text encoder's direct involvement in the prompting process, its functionality plays an essential role in shaping users' understanding and interaction with the Stable Diffusion model.

During the development phase of Stable Diffusion version 1, Stability.AI chose to incorporate the CLIP (Contrastive Language–Image Pre-training) model as the text encoder. CLIP is an image classification machine learning model, and this model is trained to evaluate how accurately a text caption describes a given input image. This model was developed by OpenAI, a well-known organization responsible for developing other advanced AI models, such as the image-generating model Dalle 2 and the text-generating model GPT. Although CLIP was released as an open-source model, the dataset used to train it was not publicly available.

In their decision to build the 2.0 model from the ground up, Stability.AI also integrated a new text encoder, OpenCLIP, which was trained using a subset of the LAION-5B dataset with NSFW images filtered out. Developing OpenCLIP underwent an extensive training period of around a million hours with the aid of powerful A100 GPUs. According to Stability.AI, OpenCLIP "greatly improves the quality of the generated images." Despite this improvement, the significant shift in the text encoder led to confusion among first-time users experimenting with Stable Diffusion version 2. Users discovered that some of the popular keywords they previously employed in their prompts were no longer recognized by the 2.0 model. Users have reported that they have failed to generate images attempting to depict specific artistic styles or celebrities they were able to generate using previous Stable Diffusion 1 versions.

Some users in the Stable Diffusion community speculated that Stability.AI, having encountered multiple legal issues concerning Stable Diffusion 1, opted to develop a "clean" and open model by intentionally excluding certain celebrities, artists, and NSFW content from the dataset.

Contrary to these assumptions, Emad Mostaque, the CEO of Stability.AI, clarified that no artists were purposely removed or excluded from the dataset for version 2. He provided further details on the matter, explaining:

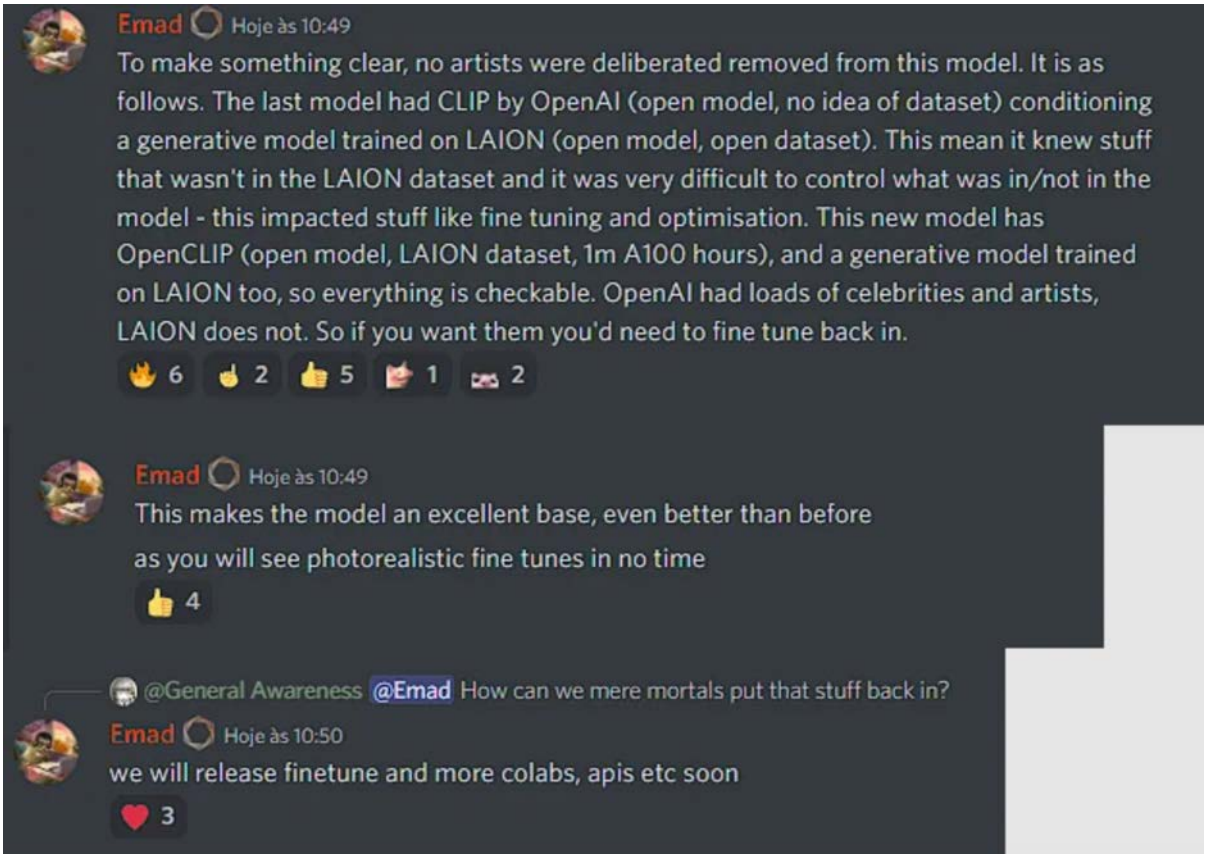

Figure 2-7 : Emad's Explanataion in Stable Diffusion Discord Server , Source : Reddit [26]

Emad clarified that the primary factor contributing to the disparities between Stable Diffusion 2 and its predecessors is the training data used by their individual text encoders, CLIP and OpenCLIP. CLIP, which functioned as the text encoder for Stable Diffusion version 1, was trained on a more comprehensive dataset containing a broader range of celebrities and artists compared to OpenCLIP, the text encoder for Stable Diffusion 2. CLIP augmented Stable Diffusion version 1's capabilities to accurately generate celebrities and artistic styles in the result images. However, since CLIP's training dataset is not open-source, Stability.AI was unable to examine or replicate its characteristics using the LAION dataset alone during the development of OpenCLIP. As a result, the prompting methods previously employed for Stable Diffusion

version 1 have become outdated for version 2, primarily due to replacing the third-party text encoder.

Another notable distinction between the two versions lies in the importance of negative prompting and embeddings. Negative prompt works in contrast to a positive prompt, it is intended to exclude or remove unwanted characteristics, such as blurry details. Embeddings, on the other hand, are keywords trained through a technique called Textual Inversion, which permits users to train particular objects or artistic style atop the Stable Diffusion model without modifying the model itself. The introduction of Textual Inversion has opened the door to greater model customization, also providing users with even more advanced control over the prompting process. A more in-depth discussion on Textual Inversion, as well as other prevalent techniques for customizing and fine-tuning AI models, will be covered later in Chapter Four of this thesis.

While both Stable Diffusion version 1 and 2 offer negative prompting and embedding features, their impact during the image generation process is significantly more pronounced in version 2. Therefore, effectively utilizing negative prompts and embeddings has become crucial to achieving the best performance with Stable Diffusion version 2. To demonstrate this point, a comparison is provided, with the specific settings utilized for image generation, which are indicated beneath the corresponding figure.

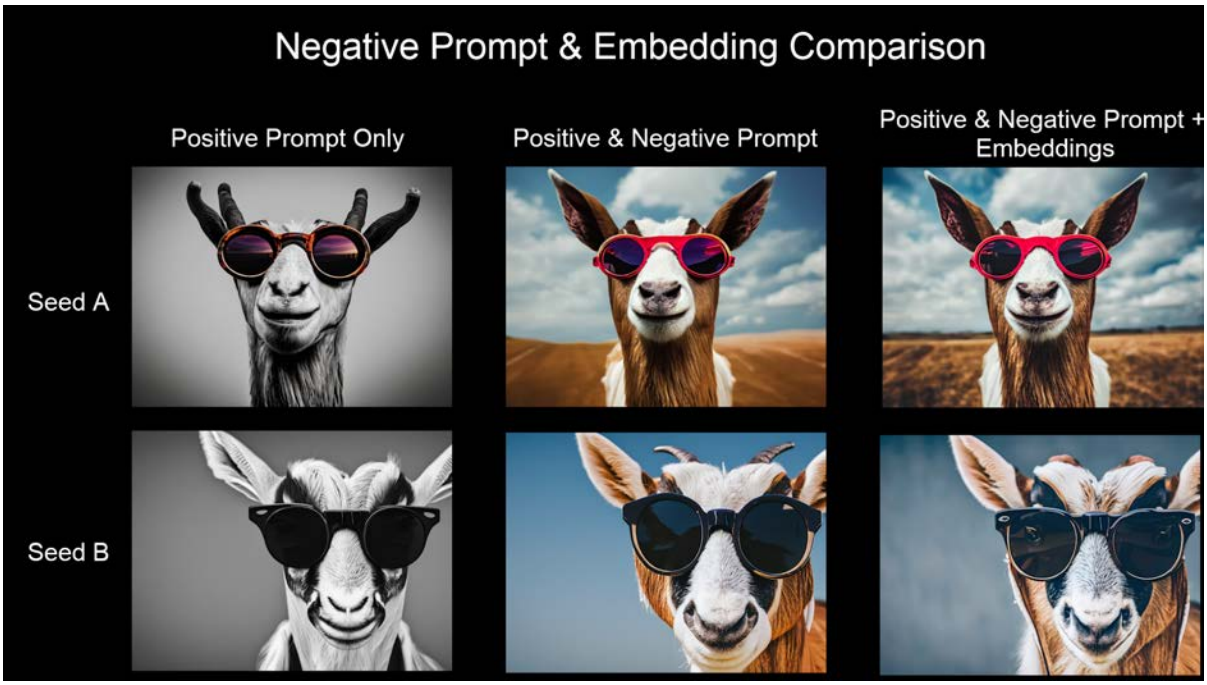

Figure 2-8 : Negative Prompt & Embedding Comparison

Here are the settings for the images of Figure 2.8 -

- Postitive Prompt : a portrait of a goat wearing sunglasses, fine - art photography, soft portrait shot 8 k, mid length, ultrarealistic uhd faces, unsplash, intricate, casual pose, centered symmetrical composition, stunning photography, masterpiece, grainy, centered composition

- Negative Prompt : disfigured, monochrome, kitsch, ugly, oversaturated, grain, low-res, Deformed, blurry, bad anatomy, disfigured, poorly drawn face, mutation, mutated, extra limb, ugly, poorly drawn hands, missing limb, blurry, floating limbs, disconnected limbs, malformed hands, blur, out of focus, long neck, long body, ugly, disgusting, poorly drawn, childish, mutilated, mangled, old, surreal

- Used Embeddings : DrD_PNTE768 (On negative Prompt)

- Stable Diffusion Parameters : Steps: 60, Sampler: DPM++ 2M, CFG scale: 8, Size: 1024x768, Model: v2-1_768-nonema-pruned

In the set of images of Figure 2.8, those situated in the first row have been generated using a specific seed value, while the images in the second row have been created using a different seed value. – we will call these seeds A and B. - The images displayed in the first column on the left have been generated without incorporating a negative prompt, and also without the use of the "DrD_PNTE768" embedding. Upon observation, it is evident that the resulting images have successfully preserved the figure of a "goat wearing sunglasses." However, the Stable Diffusion model could not maintain intricate detail in the images and primarily featured a monochrome color scheme.

Moving on to the images in the central column, these were generated with the inclusion of a negative prompt. As a result, we can observe a marked improvement in the quality of detail within the images and a resolution of the monochrome color scheme issue that was previously encountered.

Lastly, the images featured in the rightmost column have been generated by incorporating both positive and negative prompts, as well as the "DrD_PNTE768" embedding at the beginning of the negative prompt. The term "DrD_PNTE768" refers to a trigger keyword associated with the "Point E" negative embedding, which was developed by a user known as "Doctor Diffusion." The inclusion of this embedding within the negative prompt results in a further improvement in image quality when compared to images generated without the use of embeddings. As a result, using negative prompts and embeddings is strongly advised when working with the Stable Diffusion 2 model in order to achieve the most promising outcomes.

To summarize, the release of Stable Diffusion 2.0 is a significant advancement in the field of Generative AI, targeting the enhancement of output quality, the acceleration of performance, and a more profound understanding of input prompts. However, this substantial update initially led to some confusion, as discrepancies surfaced during the early experimental phase. Following extensive tests of the Stable Diffusion 2.0 system, I have made several observations that deserve attention in the context of future AI model releases:

Firstly, it is important to recognize that AI models will likely produce the best results within a specific range of resolutions. This is heavily influenced by the resolution of the images used during the training process for each respective model. For example, image below is a description of the Stable Diffusion 2.1 upscaling model from the official Hugging Face website (huggingface.com) :

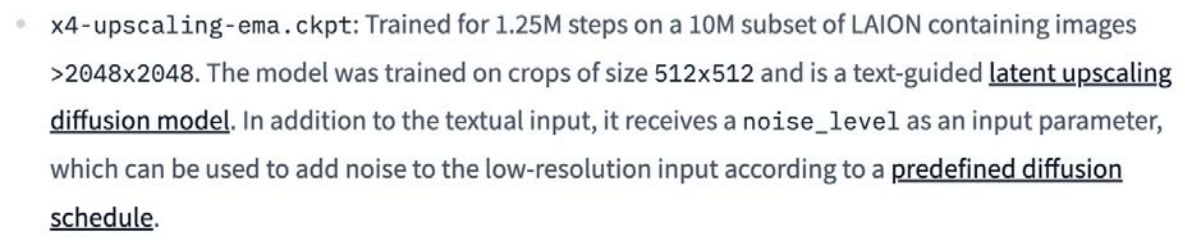

Figure 2-9 : Model Description Example [18]

Stability.AI indicated on the hugging face website - where they share their official model - that the official upscaling ema checkpoint model is trained with 2048x2048 resolution images for 1.25M steps. During this training process, this model is trained to produce fine details of high resolution images, therefore, this model will show the best potential when it generates high-resolution images.

Secondly, it is crucial to have detailed information about the trained images of the dataset utilized for each AI checkpoint model, as this knowledge is essential for understanding the model's characteristics. Without such information about the trained dataset—for instance, details about changes in the Stable Diffusion 2.0 version—users may be easily misled, resulting in disappointment due to unmet expectations or misconceptions about the model's performance.

On December 7th 2022 - only less than two weeks after the Stable Diffusion 2.0 version public release - Stability.AI announced the release of the 2.1 version. This development was driven by the negative feedback that Stability.AI had received about the 2.0 version. Therefore, the primary objective of the 2.1 version was to address the weaknesses of the 2.0 version while integrating the strong features from the previous 1.5 version.

The most notable change presented by Stability.AI in the 2.1 version was the adjustments made to the NSFW (Not Safe For Work) filter, which is now designed to be less restrictive than before. This decision to implement a more lenient NSFW filtering mechanism was made in response to user feedback and led to a significant decrease in the number of false positives. This change had a notably positive impact on the overall performance of the 2.1 version's trained dataset, particularly when it came to accurately depicting people in various generation scenarios.

Throughout this chapter, we have traced the progression of AI-generated art, beginning with the foundational emergence of GAN networks and culminating in the present-day status of Stable Diffusion. We began our investigation by delving into the core concepts behind GAN networks, emphasizing the critical roles played by the Generator and Discriminator in understanding the Generative AI system. Additionally, we discussed the challenges of striking a balance between the Generator and Discriminator within the Min-Max game, which is essential for achieving satisfying outputs from the AI system.

Following this, we investigated the rise of Stable Diffusion that has significantly impacted the AI art community. We analyzed Stability.AI's initial experimental phase, the evolution of their AI models, the challenges they faced, and the social interactions (social dynamics) arising from their innovations. Furthermore, We analyzed its impact on artists and technologists, exploring both the excitement and concerns it has generated in the world of art.

As we proceed to the next chapter, we will thoroughly investigate an array of art projects executed using Stable Diffusion and the essential extensions that have become an integral part of the creative workflow. We will examine the technical and artistic aspects of these projects, and discuss the obstacles and solutions encountered during the experimental stages of these projects.

Chapter 3

Image generation technique in Generative AI tools

The field of Generative AI is experiencing rapid advancements, giving rise to innovative AI systems that provide a wide range of new opportunities for artists to diversify and elevate their creative practices. A particularly fascinating aspect of the Generative AI community is its diverse composition, encompassing a wide range of artists and professionals from various disciplines. The versatility and adaptability of Generative AI technology allows individuals with different objectives and creative visions to utilize these tools, resulting in a diverse and varied user base such as graphic design, creative coding, and computer science, each bringing their unique perspective and approach to the community.

As a natural consequence of the open-source tools, an enormous number of third-party extensions for Stable Diffusion have been developed and released globally, expanding its creative capabilities at an astounding pace. The widespread continual expansion of Generative AI technology within the contemporary art communities demonstrates its potential in revolutionizing the artistic process, promoting innovative forms of expression and fostering collaboration among artists from disparate disciplines. These enhancements have contributed significantly to the rapid evolution of Stable Diffusion, transforming it into a professional-grade tool that has made a substantial impact across various artistic disciplines, including design, film production, and fine art. Moreover, the increasing accessibility of AI art technology democratizes the availability of advanced creative tools, enabling a larger number of individuals to partake in the artistic process and contribute to the collective evolution of art in the modern digital era.

In this chapter, we delve into the author's applications of adopting the Generative AI technique - primarily utilizing Stable Diffusion - , within their artistic endeavors. We will first briefly look into the core workflow behind Stable Diffusion, highlighting its unique features and benefits. Then we will investigate essential extensions, applications and features that are utilized during the creative process of each project, and analyze in various aspects such as strength, challenges, and improvements compared to previous work. By presenting this demonstration in an easy-to-follow manner, we aim to inspire a wider audience to encourage further exploration and experimentation with this innovative approach and engage with and appreciate the exciting intersection of art and artificial intelligence.

From this point forward, we will focus on a comprehensive exploration of the diverse range of artworks the author has produced during their extensive research and experimentation with Generative AI technology. By examining each artwork in the order they were created, we will be able to observe the fascinating progression of how Generative AI not only enhances the overall quality of the images it generates but also becomes increasingly more interactive and adaptive to user input, resulting in a more immersive and dynamic artistic creation experience. Also, throughout this in-depth analysis, we will draw parallels and contrasts between various stages of the creative process, highlighting the evolution of Generative AI as an innovative art creation tool.

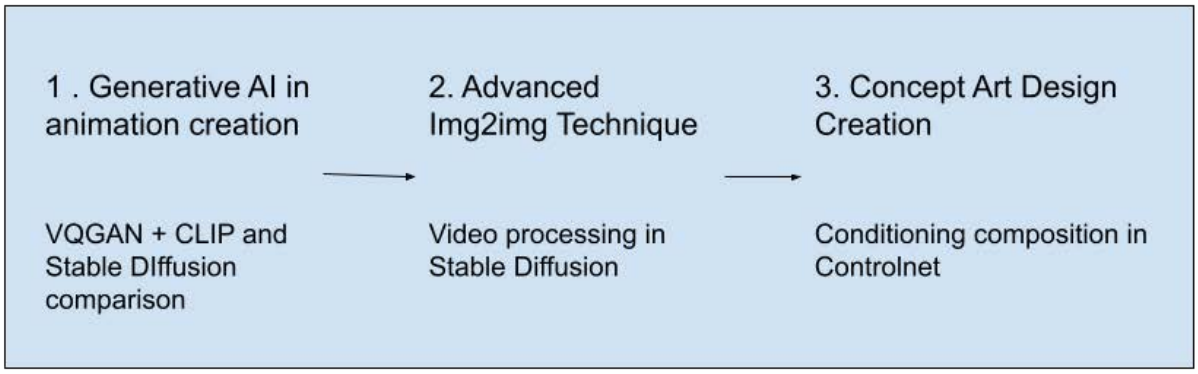

Figure 3-1 : Context brief description for Chapter 3

Coming from a visual programming and creative coding background, the author was interested in exploring and crafting generative visual art animations. Generative visual art is an intricate art

form that explores a cyclical process whereby an initial visual pattern is constructed according to specific instructions or algorithms embedded within a coding framework. Afterward, this pattern undergoes a transformative process of deconstruction or deformation, giving rise to dynamic movement or reconfiguring new visual patterns.

The author's fascination with Generative AI was ignited when he was first introduced to its unique ability to produce and deform images to generate entirely new images that ultimately form visual expressions. In executing this process, the Generative AI system is able to breathe life into a sequence of visual patterns that animate and morph, ultimately culminating in a display characterized by its own distinct and unique visual characteristics.

3.1 Generative Animation Art in VQGAN + CLIP and Stable Diffusion

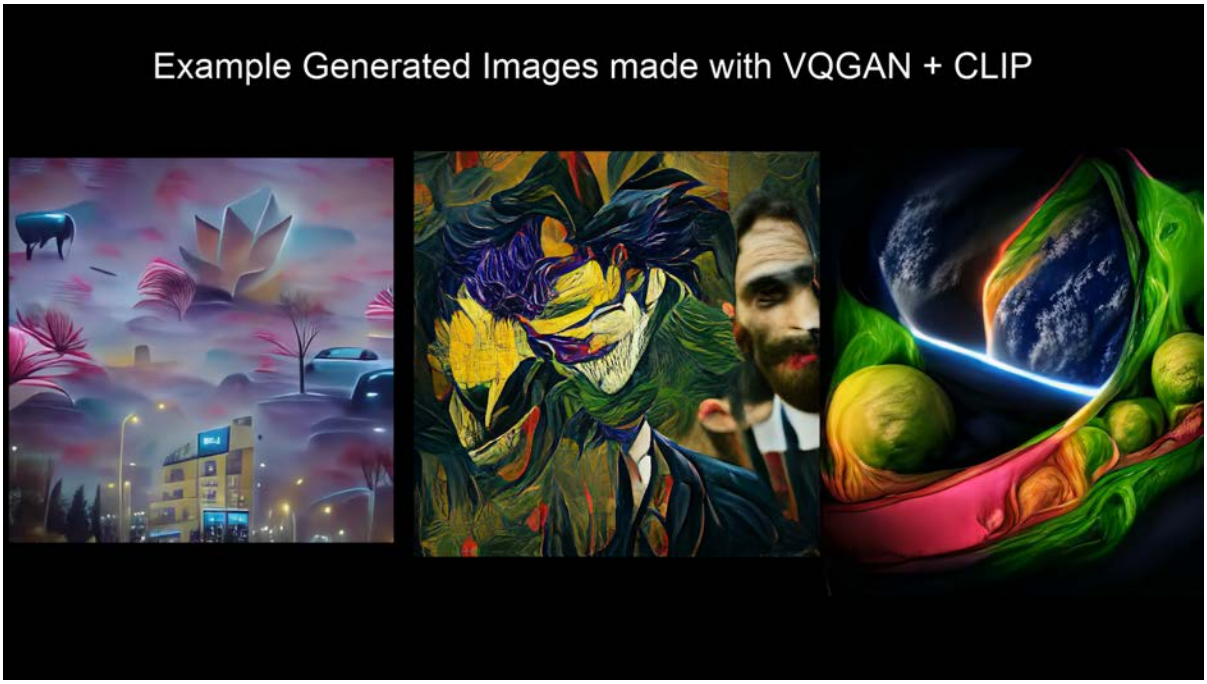

Figure 3-2 : Example Generated Images made with VQGAN + CLIP

The author's preliminary exploration into Generative AI art technology was begun with an innovative Generative AI system known as "VQGAN + CLIP." As the name implies, VQGAN

(Vector Quantized Generative Adversarial Network) + CLIP (Contrastive Language-Image Pretraining) combines the strengths of two separate neural network architectures—VQGAN and CLIP—which were previously discussed in Chapter 2. This innovative neural network architecture system was introduced in the paper titled "VQGAN-CLIP: Open Domain Image Generation and Editing with Natural Language Guidance," published in April 2022 and investigated by a team of researchers, including Katherine Crowson, Stella Biderman, Daniel Kornis, among others.

The VQGAN + CLIP neural network architecture system was designed by building upon the open-source CLIP model, integrating the essential features of both the CLIP and VQGAN models to create a powerful new AI framework. VQGAN contributes its unique capability to self-evaluate its outputs, utilizing this feedback to learn and generate increasingly accurate results. On the other hand, the CLIP model contributes its capability in decoding user-input prompts and guiding the VQGAN model accordingly during the image generation phase. As CLIP directs VQGAN towards an image that most closely aligns with the input text, this synergistic fusion of the two AI systems gives rise to the foundational concept that underlies the "text-to-image" or "txt2img" paradigm.

The emergence of the VQGAN + CLIP architecture has opened up innovative pathways for creating synthetic media, potentially democratizing the notion of "creativity" itself. By introducing a new skill set within the Generative AI domain, known as "prompt engineering," this AI system has redefined the landscape of artistic creation. The confluence of VQGAN and CLIP's distinct capabilities enables artists and creators to experiment with a transformative new approach to art, where advanced neural networks and natural language processing techniques combine to generate contextually relevant and visually unique imagery based solely on textual input.

After the introduction of generating images from text prompts, users discovered that they could generate more intricate compositions by inputting increasingly detailed text prompts. Soon after the VQGAN + CLIP was introduced, the emergence of advanced txt2img systems were published that are capable of understanding and interpreting complex prompts, with Midjourney and Disco Diffusion emerging as notable image-generation tools. As a result, there has been a

significant increase in the experimentation of prompt engineering for image generation, which has provided multiple benefits for users:

First and foremost, detailed guidance: By employing complex and lengthy prompts, users can supply the AI with greater precision, reducing ambiguity and helping the model to better understand the intended context, style, and subject matter. This detailed guidance results in output images that more closely align with the user's vision and expectations.

Secondly, balancing weights: In certain instances, longer prompts are employed to balance the significance (or weight) of various aspects within the generated output. By crafting the prompt and adjusting its length, users can manipulate the AI's focus on particular elements. This level of control empowers users to fine-tune the generated output in accordance with their preferences and artistic goals.

During this time – before the Stable Diffusion was introduced - Generative AI tools offered a limited range of settings or parameters to control image generation. Users often felt constrained by these restricted settings, which fostered a perception that Generative AI tools were heavily dependent on input prompts. This gave rise to the paradigm of complex prompting, wherein users increasingly turned to elaborate text prompts to achieve the desired level of detail and precision in their generated images.

The figure presented below illustrates an example of a complex prompt. This specific prompt, which is sourced from Tokenized.com's Midjourney example pages, displays the resulting image on the left and the corresponding input prompt on the right side.

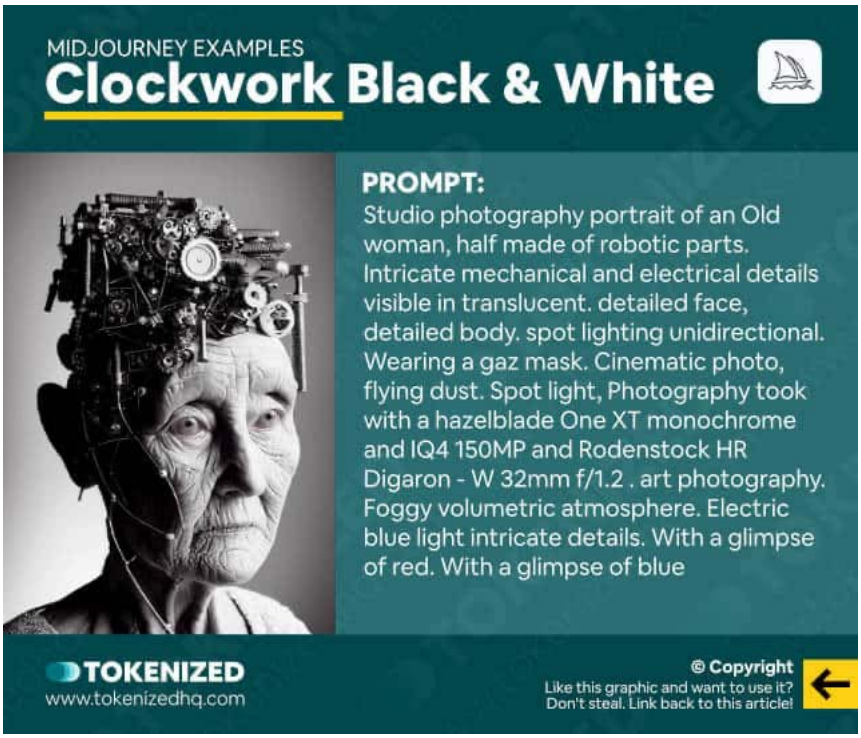

Figure 3-3 : Example of Complex Prompt written for Midjourney system [1]

Having explored into the fundamental concept of generating images from text prompts, we now shift our focus towards the operational aspects of VQGAN + CLIP and the author's application of this AI system to produce experimental or abstract-style video animations. Technically examining the VQGAN + CLIP, it is a system that can be dissected into components of Python codes, organized as "Cells" within the Google Colab platform. These components collectively establish the machine learning architecture and carry out the image generation process in a coordinated sequence. To operate this code structure based on Python code language, the author required an interface that provides an appropriate environment for executing Python code. As a result, the author employed a tool developed by Google known as Colaboratory—commonly abbreviated as Colab—which was a well known tool for virtually executing Python code actions through a web browser during the time when there were not many options available for employing an interface.

The popularity of Colab can be attributed to two primary reasons. First, Colab allowed users to execute their code entirely within a cloud-based system. This meant that regardless of a user's computer hardware capabilities, if users are connected to the internet, Colab guaranteed optimal

performance, even when dealing with computationally intensive codes. By running the code within its own virtual system, Colab harnessed the power of its cloud-based GPU resources to provide users with an efficient experience. Secondly, Colab allowed users to execute multiple instances of code networks simultaneously through a web browser, with the only constraint being the memory capacity of the configured cloud GPU.

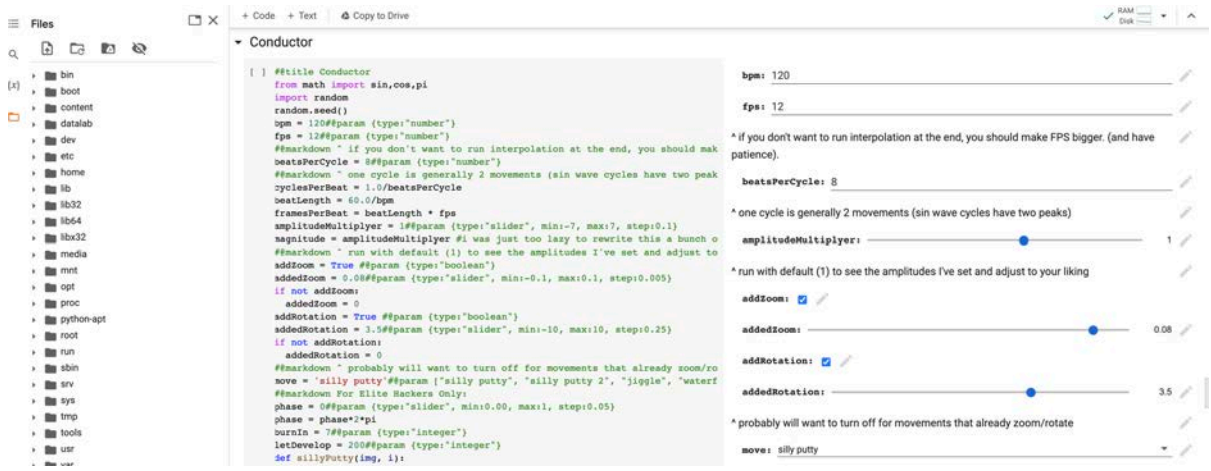

Figure 3-4 : Screenshot of VQGAN + CLIP Octaves running with Goggle Colab

Due to the absence of animation capabilities in the original VQGAN + CLIP system, the author chose to employ a modified version that incorporates animation creation features for this experiment, called the "VQGAN + CLIP Octaves version." This enhanced version introduces the concept of camera movement, which enables movement in the composition by adding the capability to move the viewpoint. Imagine that the generated output image is a view – often referred to as the camera - of a particular object within a virtual space. By adjusting the camera's viewpoint, it becomes possible to navigate through the virtual space, forming the foundational concept of the VQGAN + CLIP Octaves version.

This introduces a range of additional settings for controlling camera motion. Users can now manipulate, zoom, and rotate the entire canvas that will apply through each image-generation iteration. These settings play a crucial role in determining the animation's overall movement and progression throughout the generated video. The figure above displays a screenshots of the visual arts author created utilizing VQGAN + CLIP Octaves, while the Colab patch or "Notebook" link can be found below.

Notebook Link -
https://colab.research.google.com/drive/10y2g9_ELYkQzQbwD_KN_44SXjZLK1xP_?usp=sharing

The author's experimentation with the VQGAN + CLIP system is oriented mainly around refining prompt engineering and adjusting camera movement settings to achieve the desired results. The VQGAN + CLIP Octaves version incorporates two settings related to the text prompt: "Prompt_pre" and "Prompt_main." The Prompt_main parameter accepts a text prompt and treats it as the central theme of the composition. The Prompt_main parameter applies the most significant influence during the generation process. In contrast, the Prompt_pre parameter has a lesser impact—or carries less weight—than Prompt_main, serving to fine-tune the composition, such as style, color, or texture.

Figure 3-5 : Screenshot of Video Generated using the prompt provided below.

After conducting several trials, the author generated a video using the prompt provided below.

- prompt_pre = ["Hyperrealism", "Jonathan Zawada", "flowers"]
- prompt_main = ["Landscape in style of beeple"]

The figure above displays a screenshot of the video, which was selected by the author to serve as the primary theme of his visual artwork piece. The author's next step was creating variations that introduced intriguing and unique elements while maintaining similarity to the generated video. To achieve this, the author experimented with the prompt_pre parameter to generate diverse variations based on the central theme. By incorporating the names of different artists as keywords and experimenting with mixing these keywords, the author succeeded in generating variations with distinct color schemes and details. The figures below display the variations created, and the prompt_pre parameters employed to generate the showcased variations are provided below.

- Video Variation One prompt_pre = ["Beeple", "Jonathan Zawada", "flowers", "Oil painting"]
- Video Variation Two prompt_pre = ["flowers", "Hyperrealism", "Beeple", "Night" , "Acrylic paint"]

Figure 3-6 : Screenshot of Video Variation one

Figure 3-7 : Screenshot of Video Variation Two

VQGAN + CLIP Octave version includes additional settings for camera movement, which, again, provides the animation-creating feature in VQGAN + CLIP network. Users can control a range of settings related to animation, including camera movement speed, intensity, and presets for specific movements. Furthermore, users can synchronize camera movements with specific speeds that correspond to the BPM for the purpose of creating music visualizers. The optimal approach to determining the proper camera motion lies in striking a balance between speed and detail generation; camera movements should not be too fast, as this would prevent VQGAN + CLIP from introducing enough detail, nor too slow, which would inhibit the system from generating new details.

Artwork video link - https://vimeo.com/817555575
Presented here is the completed audiovisual artwork, which features a series of all three generated videos. Each video in the sequence was arranged using the visual programming software TouchDesigner, which also allowed the author to integrate an audio-reactive element into the piece. This Audio-visual artwork exemplifies the sensation of traversing the latent space as it zooms in and morphs in captivating ways.

It is important to note that this experiment was conducted during the early stages of Generative AI, and as such, the outputs did not retain the same level of refinement found in later works. Nonetheless, the potential for growth and development in this field was already apparent. At the time of experimenting with VQGAN + CLIP, the author did not anticipate the practical application of Generative AI as a comprehensive art design tool. However, VQGAN + CLIP demonstrated remarkable capabilities in creating unique compositions, abstract textures, and shapes, which the author found fascinating and worthy of further exploration.

Following this initial exploration, the author began to delve deeper into the realm of audiovisual creation using VQGAN + CLIP. This pursuit eventually led to the discovery of Stable Diffusion, a pivotal development that influenced the author's artistic journey. The introduction of Stable Diffusion heralded a new era in the author's creative process, opening up even more possibilities as a Generative AI artist.

3.2 Generative Animation Art in Stable Diffusion Deforum Extension and Comparison

Upon the introduction of Stable Diffusion in Aug 30th, the author's investigative journey transitioned towards this massive AI network, particularly struck by the fact that it was unveiled as an entirely open-source AI model. Given that generative AI tools generally share similar AI frameworks on a broader scale, Stable Diffusion possessed similarities to the previous tool employed by the author, VQGAN + CLIP. Thus, in this section, the author will showcase the creative process of crafting another experimental animation video artwork, this time utilizing a new AI system, the Stable Diffusion system, alongside a different interface that enables more precise control settings.

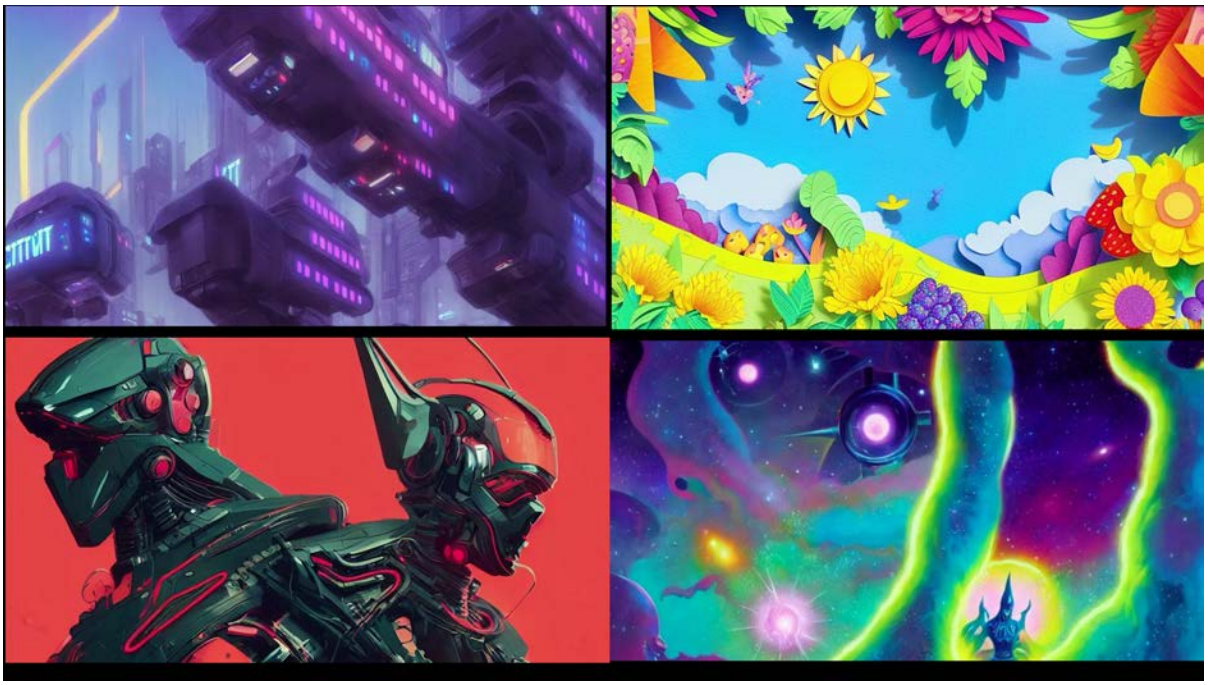

Figure 3-8 : Screenshots of animation artworks created with Stable Diffusion Deforum

Theoretically, this creative process in Stable Diffusion resembles a similar workflow to the previous work, utilizing VQGAN + CLIP Octaves. Both share a similar approach to generating each frame of animation, which involves introducing a morphing effect on the composition by subtly shifting the viewpoint, or camera movement, frame by frame. Therefore, we won't investigate the overall workflow of experimental-style animation creation in Stable Diffusion, which might be redundant. Instead, the focus will center on comparing these two systems,

illustrating the distinctions encountered during the creative process, as well as the unique challenges and difficulties each presents. Following the brief demonstration of the creative process, a comparison of the resulting audiovisual artworks will be presented, highlighting the differences between the VQGAN + CLIP and Stable Diffusion AI models. This comprehensive exploration will offer valuable insights into the capabilities and constraints of each system, ultimately informing future applications of generative AI in artistic expression and experimental animation.

Examining the choice of tools in greater depth, similar to how the author decided to employ the modified version of VQGAN + CLIP for the previous project, the author had to employ the extension because the original Stable Diffusion did not include animation creation capabilities at the time. In response to this limitation, the author chose to employ the Stable Diffusion extension, Deforum, which offers animation creation features that resemble animation features in VQGAN + CLIP Octaves. In this context, VQGAN + CLIP Octaves can be considered a precedent of the Stable Diffusion Deforum extension. Deforum introduces additional camera movement settings, allowing users to manipulate, zoom, and rotate the entire canvas with each image-generation iteration, thereby creating a dynamic visual experience.

The author decided to incorporate a new interface system, shifting from Google Colab due to the introduction of Google's new price policy after they faced dramatically increased user usage because of the popularity of Stable Diffusion. Due to the demands, Google started to charge users based on their credit system, which comparably costs more than the previous price policy.

Recognizing the potential advantages of a new interface system, the author chose to transition away from Google Colab, primarily due to the introduction of Google's revised pricing policy, which was implemented in response to a dramatic increase in user demand due to the popularity of Stable Diffusion. As a result, Google implemented a credit-based charging system that proved to be more expensive than the previous pricing structure.

Automatic1111 is an open-source user interface (UI) tool that operates through an internet browser, commonly referred to as a webUI. This versatile tool provides users with comprehensive control and an extensive range of features for the Stable Diffusion system.

Initially developed based on Gardio, a machine learning app design tool, Automatic1111 has earned a reputation for its comprehensive library of extensions and settings. This vast array of options can be intimidating for some, particularly when compared to the more straightforward and user-friendly Midjourney system. The author chose to utilize Automatic1111 for two primary reasons.

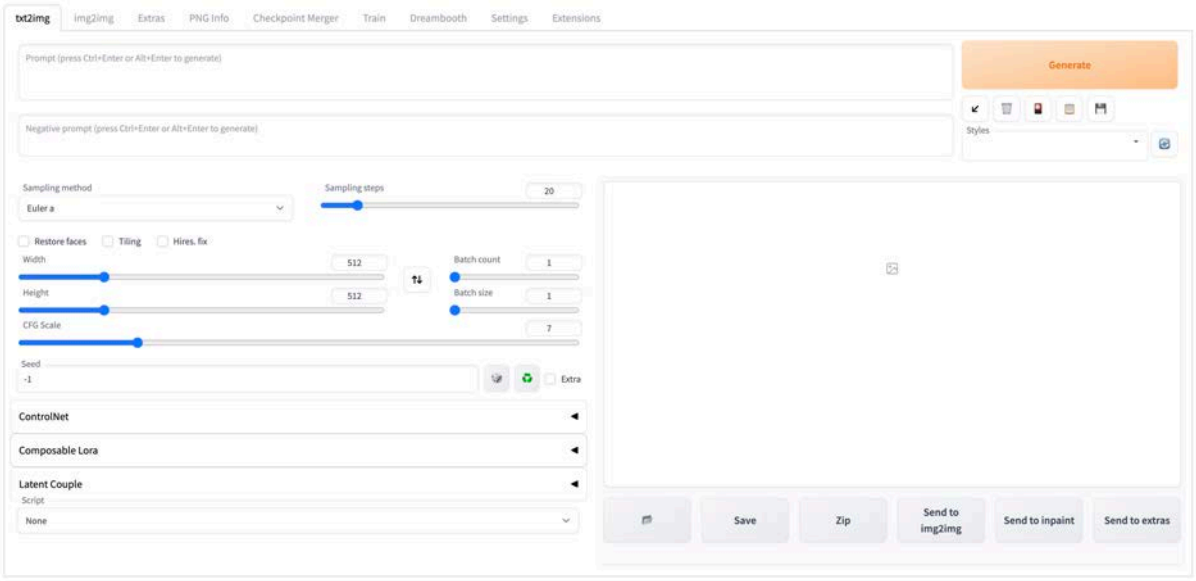

Figure 3-9 : Screenshot of Automaic1111 UI

Firstly, the Automatic1111 interface delivers a flexible and versatile system for installing and utilizing third-party extensions of Stable Diffusion. With a library comprising more than 100 third-party extensions, Automatic1111 enables users to install and operate these extensions within its interface seamlessly. The availability of such an extensive collection of machine-learning tools unlocks a vast array of creative possibilities for art production, enabling users to explore innovative and cutting-edge techniques in their work.

Secondly, Automatic1111 provides a wealth of features and settings that were not offered by other interfaces during that time. This vast selection of options grants users the ability to fine-tune their creative process, tailoring their approach to specific artistic goals and objectives. By providing an outstanding level of customization and control, Automatic1111 empowers users to delve deeper into the potential of generative AI for artistic creation.

Delving into the workflow, it becomes apparent that the biggest distinction between working with VQGAN + CLIP and the Stable Diffusion Deforum extension is not only the increased number of controls and features provided by the Stable Diffusion system, but also a crucial component that has become an essential part of the author's Generative AI animation workflow—Keyframe parameters. Keyframes, a term commonly used in filmmaking and motion graphics, enable users to manage parameter changes over time by defining the starting and ending values of a particular action. The integration of keyframe parameters into the Stable Diffusion Deforum extension, as opposed to the VQGAN + CLIP Octaves version, allows users to precisely plan camera movement and image generation settings on a frame-by-frame basis prior to initiating the generation process.

For example, during the animation creation workflow, the author employed keyframing in a parameter called "Strength_schedule." This parameter governs the intensity of the Stable Diffusion effect and the density of the generated output images. By decreasing the Strength_schedule value, the Stable Diffusion process exhibits a more pronounced effect, resulting in output images with a greater concentration of detail. To better understand the author's utilization of keyframes, let us deconstruct the specific keyframe settings applied to the Strength_schedule parameter. Here's the Strength_schedule value indicated below,

"strength_schedule": "0: (0.6), 40: (0.6), 50: (0.4), 70: (0.6), 90: (0.6), 100: (0.4), 120: (0.6), 140:(0.6), 150: (0.4), 170: (0.6), 190:(0.6), 200: (0.4), 220:(0.6), 240:(0.6), 250: (0.4), 270: (0.6), 290: (0.6), 300: (0.4)....."

Each number positioned to the left of the colon signifies the frame number, which corresponds to the timeline of the keyframe action, while the decimal numbers enclosed within parentheses denote the parameter value - in this instance, the Strength_schedule value - for each respective frame. Beginning with the initial frame 0, the Strength_schedule is set to a value of 0.6 and remains constant until the generation process reaches frame 40. From there, the Strength_schedule value gradually decreases until it attains a value of 0.4 at frame 50, at which point it begins to gradually increase, ultimately returning to a value of 0.6 by the time the generation process arrives at frame 70. This cyclical pattern persists throughout frame 300 and

beyond. The keyframe action results in the generation of slightly more detailed images every 50 frames, thereby introducing additional dynamism to the animated video.

Summarizing details above, the author strategically utilized keyframing to manipulate the Strength_schedule parameter in this particular instance, facilitating nuanced control over the diffusion process and the resulting visual output. This innovative approach allowed for the generation of complex and intricate animations that dynamically evolved over time, showcasing the full potential of the Stable Diffusion Deforum extension as a powerful tool for creating groundbreaking audiovisual artworks.

The figure provided below showcases a series of screenshots taken at various points in the generated video, exemplifying the overall theme of the artistic piece. To produce this animated artwork, the author employed the Stable Diffusion version 1.4 model, and the specific prompts and settings utilized for this task are indicated below the figure.

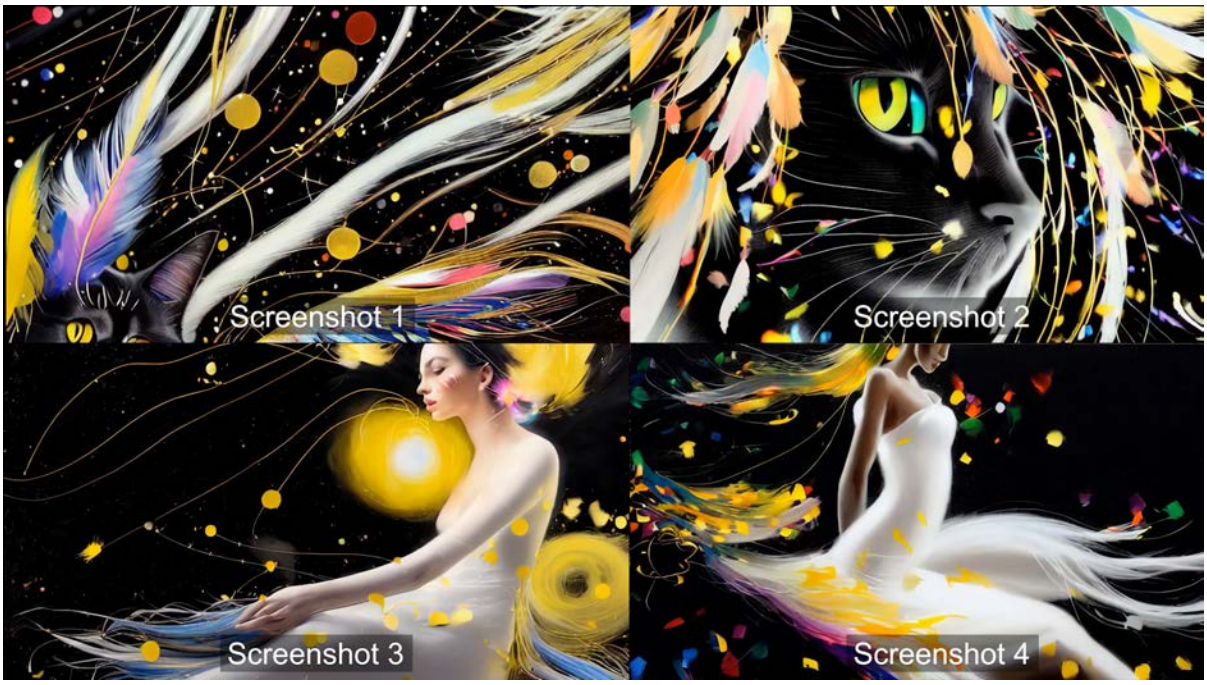

Figure 3-10 : Screenshots of Stable Diffusion Deforum animation artwork

- Positive prompt : "ultra realistic close up serene side view portrait of a beautiful black cat in a translucent white dress, Golden confetti, Paint strokes, Beautiful hair, Colorful

Feathers flying around. Golden ornaments. ultra sharp, Glim lighting. Highly realistic. Volumetric lighting. Light halation. Artwork by Peter Mohrbacher, Artgerm and jean basquiat, 8k, HD, 8k finely detailed features, closeup of the face, perfect art, grimdark, trending on pixiv fanbox, Octane render, unreal engine 5, testp, vray, blender, award winning, upscale, high contrast dramatic lighting, darksynt, vitaly bulgarov, symmetrical features, centered composition"

- Settings : "sampler": "euler_ancestral", "scale": 12, "animation_mode": "2D", "border": "wrap", "angle": "0:(1)", "zoom": "0:(1.05)", "translation_x": "0:(0)", "translation_y": "0:(0)", "translation_z": "0:(10)", "contrast_schedule": "0: (1.0)"

In a manner similar to the previous VQGAN + CLIP workflow, the author developed variations of the main theme video by integrating diverse color schemes and slightly altered visual patterns. These two generated variation videos, along with the primary theme video, were subsequently combined and reassembled using the visual programming software TouchDesigner. In addition, the author incorporated reactive audio elements into the composition to create a more immersive experience for viewers. The final, completed audiovisual artwork presented here consists of a series of all three generated videos, seamlessly integrated into a unified, captivating piece.

Artwork video link - https://vimeo.com/817553865

Warning – Some viewers might feel dizzy after watching the video for a long time due to the constant spinning movement.

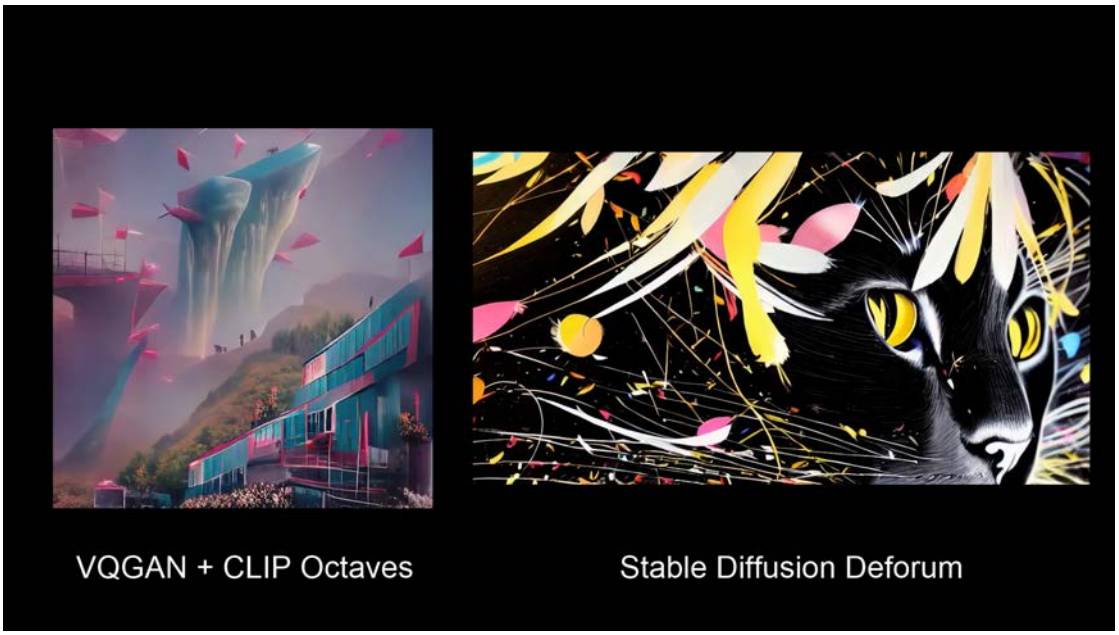

Figure 3-11 : Artworks comparison

Displayed above are two artworks positioned side by side for comparison purposes.

Due to the more advanced features of Stable Diffusion, the artwork produced using this model exhibits a higher resolution and more intricate details compared to the artwork created with the VQGAN + CLIP model. In fact, the video generated using VQGAN + CLIP was crafted at a resolution of 512x512 pixels. In contrast, the video produced with Stable Diffusion was made with a higher resolution of 832x448 pixels, thereby preserving a greater level of detail within the composition. Furthermore, the video generated with Stable Diffusion presents more dynamic and dramatically evolving details, attributable to the model's sophisticated generation system and the application of keyframing parameters. However, the artwork employing the VQGAN + CLIP model demonstrates beter temporal coherence between individual frames, resulting in smoother transitions between details.

The results of the audiovisual animation creation highlight the relative strengths and limitations of each generative AI model. On one hand, Stable Diffusion excels in terms of image quality and dynamic expression, producing visually rich and captivating content. On the other hand, the VQGAN + CLIP network possesses a unique advantage in maintaining temporal coherence between frames, thereby ensuring smooth transitions. Each Generative AI tool offers distinct advantages that may be more or less relevant depending on the context and creative vision of the

artist, underscoring the need for thoughtful consideration and experimentation when selecting the optimal tool for a given creative endeavor.

3.3 Advanced Img2Img technique

Moving beyond our scope from Txt2img exploration, the author decided to explore the additional features offered by Stable Diffusion, with a particular focus on its potential as a video processing tool using the image-to-image (img2img) functionality. This expanded investigation aims to uncover new possibilities and applications for the Stable Diffusion model, taking into account its capabilities for transforming and enhancing visual content through advanced AI-driven techniques.

The img2img feature within Stable Diffusion presents a unique opportunity for artists and researchers alike to experiment with the model's capacity for reimagining, manipulating, and refining existing visual media. By harnessing the power of the model's AI algorithms, users can breathe new life into videos, creating thematically diverse content that pushes the boundaries of conventional artistic expression.

During this creative process, the author suggests an advanced approach of utilizing img2img. This workflow centralizes on the concept of "Multi-parallel processing in img2img," which aims to enhance coherence between image frames. Stable Diffusion has already demonstrated its exceptional capabilities for generating high-quality still images. However, its adoption as a professional video-making tool has been limited due to a well-known problem: the lack of coherence between iterations during the generation processes.

In order to explain the problem, it is essential to provide a comprehensive breakdown of how Stable Diffusion operates as a video processing tool, employing the img2img technique. Imagine employing batch processing with the img2img feature for video processing, which is the typical approach in such scenarios; this meaning you are applying the Stable Diffusion process across the entire video by processing each frame. This is the typical approach in such scenarios.

As a preparatory step, you would convert the video into a sequence of individual image files and these converted frame images would then serve as inputs for the Stable Diffusion system's img2img batch processing feature. Beginning with the first frame image, each image undergoes the Stable Diffusion generation processing cycle accordingly with image generation-related settings such as denoising strength, seed value, and classifier scale value. This process continues through the entire sequence of images, ensuring that each frame is processed accordingly.

In Stable Diffusion img2img feature, the denoising strength parameter plays a crucial role; it decides the ratio between how much it resembles the original image and how strongly Stable Diffusion affects the processing image. Lower values of denoising strength result in a less pronounced impact or a more subtle stylization on the input image, often leading to enhanced coherence between consecutive frames. However, users may find the effects barely noticeable or too "weak" in this case. On the other hand, higher denoising strength values lead to a more potent influence on the image, intensifying the stylization but also introducing a more pronounced flickering visual effect during video playback, having noted that striking an optimal balance between achieving a strong stylization and maintaining coherence between frames is a highly challenging task.

The illustration below displays images processed using various denoising strength settings. The text positioned at the bottom of each video specifies the respective denoising strength – indicated as DN - and classifier scale values – indicated as SC - for that particular video. As can be seen, a higher denoising strength produces a more pronounced effect, while the image with the lowest denoising strength retains the closest resemblance to the original character.

Figure 3-12 : Img2img denoising strength comparison

This difficulty arises from the inherent nature of workflows that utilize noise as a fundamental element in Diffusion AI models. During the generation process, these models incrementally introduce Gaussian noise throughout the image and subsequently reverse this stage to decompose, analyze, and "re-imagine" the input image. Given that noise essentially embodies random numbers in image format, each noise variation, which can be regulated by a parameter known as noise seed, is wholly distinct from others. This randomness is the primary cause of the flickering effect observed in video processing techniques.

The flickering effect observed in videos generated using the Stable Diffusion process is primarily attributed to the fact that every frame of the video is processed with a distinct noise seed value for each generation iteration and this cause each generated result images to contain unique details not present in the other outputting images. And as this effect gets added for every frame, the flickering will be introduced when users later combine the processed image series as a full video. Users could choose to purposely "freeze" the seed value - meaning the Stable Diffusion system won't change the seed value during the generation iterations, but this might result in video feedback issues by enforcing the same noise seed during the image generation process.

Unfortunately, during the Stable Diffusion process, no conditioning features were available, making it impossible to control the precise structure or layout of each frame. Furthermore, only a limited number of parameters, such as Denoising strength and seed value, could be adjusted to manage the coherence and discrepancies between frames in the Stable Diffusion system. This limited control over frame conditioning in Stable Diffusion was later addressed through the implementation of several crucial third-party extensions, including ControlNet. These extensions, which will be discussed in greater depth later in this chapter, have significantly enhanced the system's capability to maintain consistency and coherence across frames, ultimately improving the overall quality of generated videos.

Returning our attention to the project workflow, the author aimed to address the issue of coherence and minimize flickering effects by adopting an alternative approach to processing the video. Instead of employing the conventional method of processing the entire video frame by frame, the author divided the video into multiple segments, each focusing on specific elements within the scene. These individual segments were then processed separately using the Stable Diffusion system before being reassembled into a cohesive video with the help of video editing software. During this demonstration, we will call the instances of Stable Diffusion processing each segment, the "layers."

In this case, the author decided to process a video featuring a man holding weapons and adopting various poses in front of the green screen. Utilizing the approach outlined above, the author processed the video in two parallel layers: one layer concentrated on the man's overall body, tracking its movements to maintain the consistency of his outfit, while the second layer focused solely on his face, tracking facial movements closely. Figure 3-13 shows the appearance of the original footage.

Figure 3-13 : Original footage before the Img2img process at frame 0

Preparatory steps for this process required the video to be converted into a sequence of image files, much like the workflow previously detailed. However, in this case, the author needed to create two separate image file sequences for each Stable Diffusion process layer. To extract the necessary image sequences, the author employed a video editing software called "After Effects," along with a third-party plugin named "AE Face Tools." This plugin facilitated the tracking of facial movements within the input video and then applied stabilization in order to lock the tracking frame in the center of the screen, streamlining the process of separating the two different layers compared to tracking the face manually.

Following the completion of the preparatory phases, the author embarked on a journey to explore the img2img features in Stable Diffusion, initially focusing on the layer that emphasizes the subject's face. During the time when Stable Diffusion 2.1 just got released, various Stable Diffusion models - including custom-trained models shared by the Stable Diffusion community members - were published, showcasing their model's distinct characteristics in image generation. The availability of such diverse options allowed the author to delve into an extensive range of creative possibilities, experimenting with different AI models to determine which one aligns best with their artistic vision and the intended outcome.

For the face layer, the author decided to utilize the Robo Diffusion 2 model with the specific intention of generating a robot face. The Robo Diffusion 2 model, trained by Stable Diffusion community member Nousr, is renowned for its ability to produce robotic objects with metallic textures. The figure below presents the results of processing the face layer using the Robo Diffusion 2 model, captured at four distinct points in time. As evident from the four images, representing various stages in the generation process, the consistency of the robot face is maintained throughout, significantly enhancing the coherence in the final video. The text beneath the figure provides details on the prompts and settings employed.

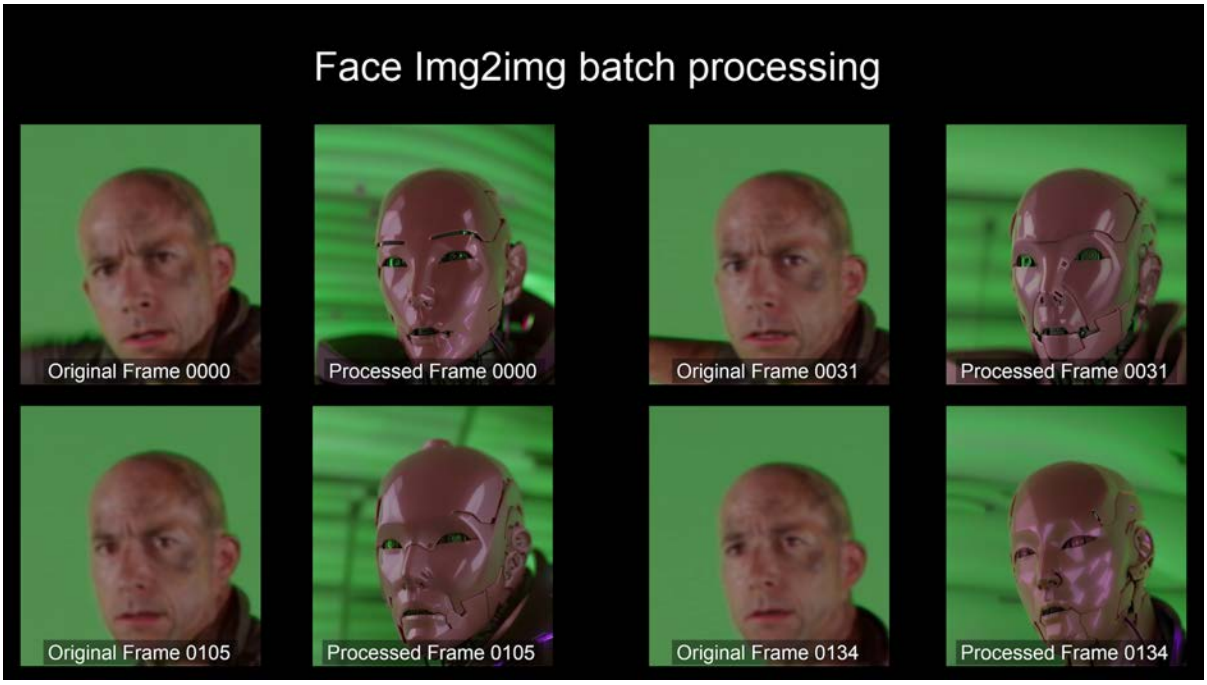

Figure 3-14 : Face image processing in img2img

- Positive prompt - Ultradetailed photograph of (Purple nousr robot :1.3) face emb-rrf-low in Cyberpunk By Artgerm and WLOP and Ilya Kuvshinov and RHADS and Loish and Rossdraws. Perfect shading, soft studio lighting, ultra detailed, photorealistic, octane render, cinematic lighting, hdr, 4k, 8k, edge lighting

- Negative Prompt - Girl, female, lady, ugly| tiling| poorly drawn hands| poorly drawn feet| poorly drawn face| out of frame| mutation| mutated| extra limbs| extra legs|

extra arms| disfigured| deformed| cross-eye| body out of frame| blurry| bad art| bad anatomy| blurred| text| watermark| grainy

- Settings - Steps 170, Sampler: DPM++ 2M, CFG scale: 15, Denoising strength: 0.6, Mask blur: 4

Moving on to the body layer, the author needed to use different AI models, as the goal was to generate an elegant male dress suit – a task unsuitable for the Robo Diffusion 2 model. As an alternative, the author chose to work with the Stable Diffusion 2.1 version model, which shares its foundation with the Stable Diffusion 2 model upon which Robo Diffusion was also built. To maintain consistency between the face and body layers and avoid discrepancies, the author selected prompts and settings that were similar to those used for the face layer. The prompts and settings utilized for the body layer can be found beneath the corresponding figure. Readers are encouraged to compare these parameters between the two layers to gain a deeper understanding of their relationship and the chosen settings.

Upon the completion of the individual layer generation by Stable Diffusion frame by frame, the author utilized Adobe After Effects once again. This time, the objective is to substitute the generated layers with the originally extracted segments by reversing the stabilization process. This method effectively combines these segments into a singular, comprehensive video, showcasing the seamless integration of each layer. The figure below demonstrates a series of screenshots taken at four various points in the resulting video after the stage working with After Effects.

Figure 3-15 : Video result of the combined layers

- Positive prompt : Short hair Male covering face with red facecover By Artgerm and WLOP and Ilya Kuvshinov and RHADS and Loish and Rossdraws. Perfect shading, soft studio lighting, ultra realistic, photorealistic, octane render, cinematic lighting, hdr, 4k, 8k, edge lighting

- Negative prompt: Girl, female, lady, ugly| tiling| poorly drawn hands| poorly drawn feet| poorly drawn face| out of frame| mutation| mutated| extra limbs| extra legs| extra arms| disfigured| deformed| cross-eye| body out of frame| blurry| bad art| bad anatomy| blurred| text| watermark| grainy

- Settings : Steps: 170, Sampler: DPM++ 2M, CFG scale: 13, Denoising strength: 0.7, Mask blur: 5

However, after the background was added, the author noticed that there were discrepancies in the lighting aspect between the background and the Robot character. Therefore, the author decided to process the video with photo editing software known for lighting editing, Luminar AI, processing the lighting and color correction effect on each frame.

However, once the background had been incorporated, the author noticed that there were discrepancies in the lighting aspects between the background and the robotic character. To address this issue, the author decided to process the video using Luminar AI, a photo editing software known for its lighting editing capabilities. The author proceeded to process the lighting and apply color correction effects to each frame of the video, ultimately achieving a more harmonious visual balance between the background and the character.

The figure below showcases various points throughout the final video, illustrating the successful integration of the background and the seamless fusion of the lighting elements between the character and the background. A link to the full video can be found beneath the figure.

Figure 3-16 : Screenshot of video taken at various points

Link - https://vimeo.com/817556025

By implementing this innovative, multi-layered approach, the author was able to achieve an improvement in coherence and a reduction in flickering effects within the final video output. This method demonstrates the potential for leveraging advanced techniques and tools in

combination with the Stable Diffusion system to overcome some of its inherent limitations and produce high-quality video content.

However, despite these improvements, the flickering effect remained noticeable in the final product, indicating that there is still room for further refinement in the process. Additionally, the Stable Diffusion system lacked a conditioning feature for structuring the composition, which may have contributed to the presence of these imperfections.

3.4 Controlnet and Concept Art Design Creation

As previously discussed, Stable Diffusion and its extensions have delivered remarkable results in image generation. However, they lacked essential conditioning features. This absence of conditioning made it nearly impossible to regulate the exact structure or layout of each frame during the video generation process. The introduction of ControlNet, however, has significantly enhanced the ability to control the composition within Stable Diffusion. Also, looking at the video processing aspect, ControlNet improves coherence between the generated frames, further refining the video generation process.

ControlNet is an innovative model training mechanism that allows Stable Diffusion to complete specialized subtasks allowing extra conditioning input. This neural network architecture was introduced in February 2023 by Lvmin Zhang. This architecture has been designed to control large diffusion models more effectively, providing the additional conditioning input that works better compared to traditional methods. ControlNet has the potential to revolutionize various image generation workflows and applications, spanning areas such as artistic creation, architectural renderings, design ideation, storyboarding, and concept art designs.

ControlNet was initially built from the recognition that text prompts alone cannot address all conditioning challenges in image generation in Generative AI systems. Also in some cases, users may struggle to express their visual ideas through text alone effectively. By incorporating the ControlNet extension, Stable Diffusion users can now express their vision more accurately,

resulting in a more precise representation of their intended composition structure. This added layer of conditioning method enables Stable Diffusion to generate images that more closely resemble the user's desired visual concept more, which opens enormous potential in the design aspect.

ControlNet's innovative nature stems from its distinctive method of fine-tuning powerful AI models to perform specialized subtasks. Unlike traditional techniques that focus on directly training the primary model, ControlNet employs a unique approach by effectively locking"the original model and developing an external network to comprehend new inputs. This external network then channels information into the primary model, enabling it to execute the new subtask while retaining its existing knowledge base.

Control Net's unique approach was designed to avoid issues of overfitting and maintain the production-ready quality of large-scale models. As a result, ControlNet delivers better outcomes and necessitates considerably less training time in comparison to conventional fine-tuning techniques. For example, in comparison with the Depth to Image model, an alternative fine-tuning model released by the creators of Stable Diffusion, Control Net demonstrates better fine-grain control and takes less time for training.

The ControlNet extension was initially released in February 10th 2023, with eight distinct models, each uniquely designed to address specific properties of the input image and incorporate conditioning aspects throughout the image generation process. This means that each model has its own particular way of analyzing and processing the conditioning input image. The following list presents the ControlNet models along with a brief description of their capabilities:

1. Canny: Specialized in edge detection, this model processes input images into monochrome images with white edges set against a black background.
2. Depth: Trained using Midas depth estimation, this model showcases its strength on depth mapping, processing input images into grayscale images where black denotes deeper areas.
3. Hed: Similar to Canny, but with softer edges.

4. MLSD: Designed for detecting straight lines, this model exhibits weaker performance with circular objects.
5. Normal: Specialized in normal map detection, this model processes input images into normal map images.
6. Openpose: Focused on preserving human poses, this model enables users to customize poses using the Openpose editor extension.
7. Scribble: Trained with human scribbles, this model demonstrates strong performance when working with hand-drawn images.
8. Seg: Developed to detect segmentation in input images and preserve their outlines using the ADE20k segmentation protocol.

Though each model exhibits significant potential, this paper will not delve into a detailed examination of their individual capabilities. Instead, the focus will be on providing a general workflow for effectively utilizing ControlNet in image creation.

The illustration below features example images generated using the Scribble model. The left image, showcasing a roughly drawn sketch, serves as the basis for the middle and right images. As observed, the generated images retain the outlines from the input sketch, demonstrating the model's ability to create coherent and visually appealing images based on the input provided. This example illustrates just one of the many ways ControlNet can be employed to enhance artistic expression and exploration in the digital realm.

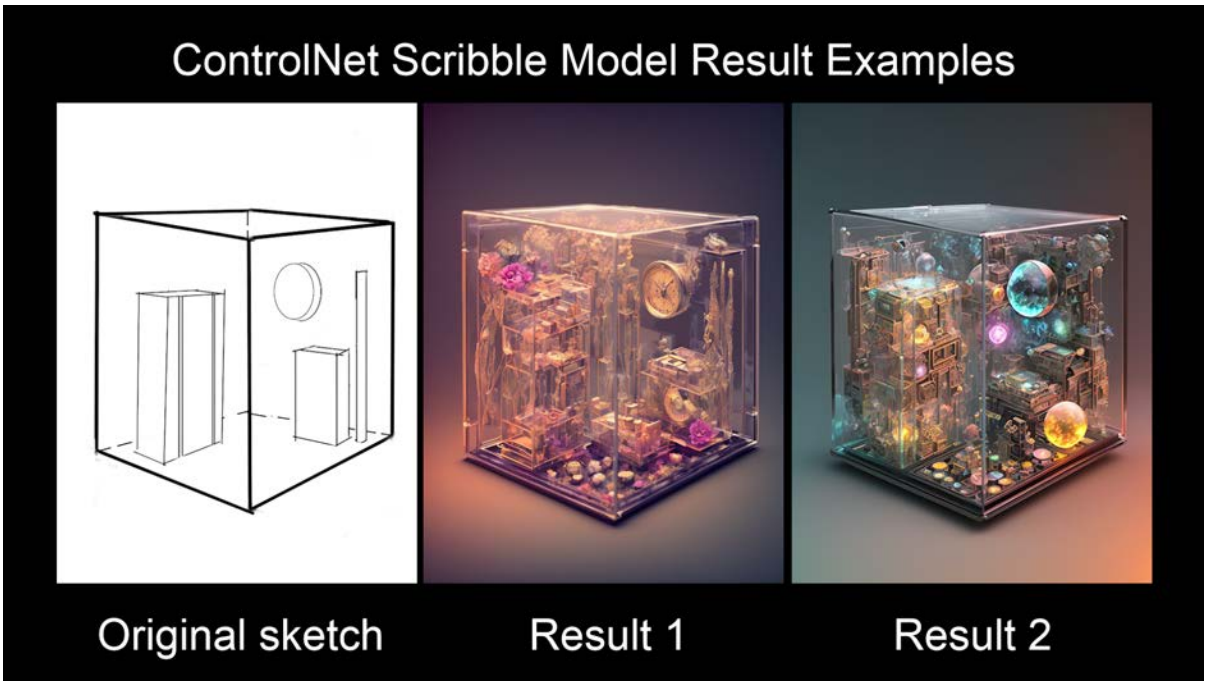

Figure 3-17 : Example of image generation employing ControlNet Scribble model

With users now having the ability to accurately articulate their visual concepts and incorporate conditioning elements into the Stable Diffusion process, the author's interest has been drawn towards complex images using Stable Diffusion for the purpose of a concept art design. The author was given the responsibility of creating concept art for a client's storyboard pitch deck, and the following expectations were detailed by the client before the onset of the creative process:

1. The client required the design creation of three distinct characters, comprising two females and one male, with each character accompanied by a character sheet image that showcases their turnaround pose, effectively revealing the character's design from various angles.
2. The client supplied relevant themes and stylistic references, emphasizing a Cyberpunk-inspired aesthetic combined with a colorful, futuristic science fiction atmosphere as the desired visual direction.
3. Specific attributes and visual characteristics were outlined for each individual character as follows:

a. First character: A female character sporting short, dark hair, dressed in a long, dark silk gown and accessorized with black leather gloves.
b. Second character: A male character with short, green hair, wearing headphones and outfitted in a futuristic, soldier-like suit.
c. Third character: A female character with long, pink hair, wearing a tank top and featuring pink devil horns as a distinct feature.

Equipped with the above prerequisites, the author initiated the process of creating the character turnaround pose sheets, which served as the foundation for designing each unique character. In order to accomplish this, the author employed the Openpose ControlNet model, which enabled them to generate the desired poses of each character.

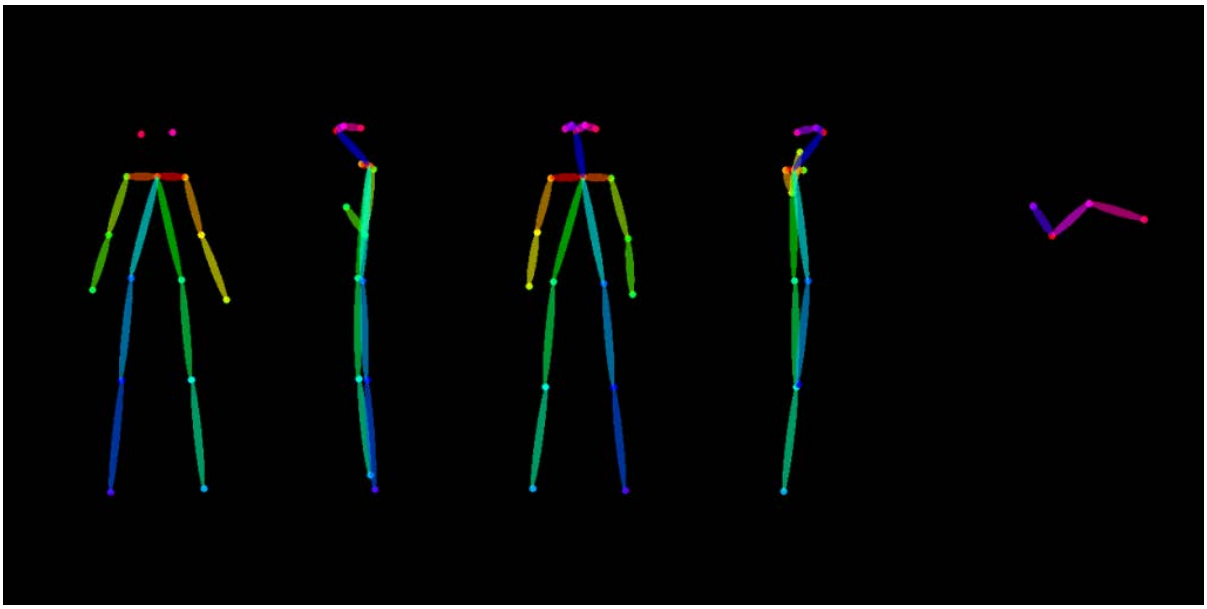

Figure 3-18 : Employed ControlNet input image

The figure presented above illustrates the employed Openpose skeleton image, a resource provided by the active members of the Stable Diffusion community, which is intended for generating character sheets. Generally, the workflow consists of a series of steps, beginning with the preprocessor decomposing the input image to convert its properties - such as converting a portrait image into a pose skeleton image. Subsequently, ControlNet carries out the image

generation process by utilizing the converted image as source material. however, the skeleton pose image is already supplied, which means that the preprocessor's involvement in the workflow is not required.

As can be seen in the Openpose skeleton image, a turnaround pose is depicted with the figure rotating 360 degrees across four distinct poses on the left side of the image. Additionally, a concise skeleton on the right side of the image indicates the facial position. The figure below showcases the outcome of the image generation process conducted by ControlNet, displaying the character design for the first character mentioned earlier. The full settings and prompt can be found below the figure. Note that this section will not include the prompt and settings for the individual image created; however, some will be provided for reference purposes for readers.

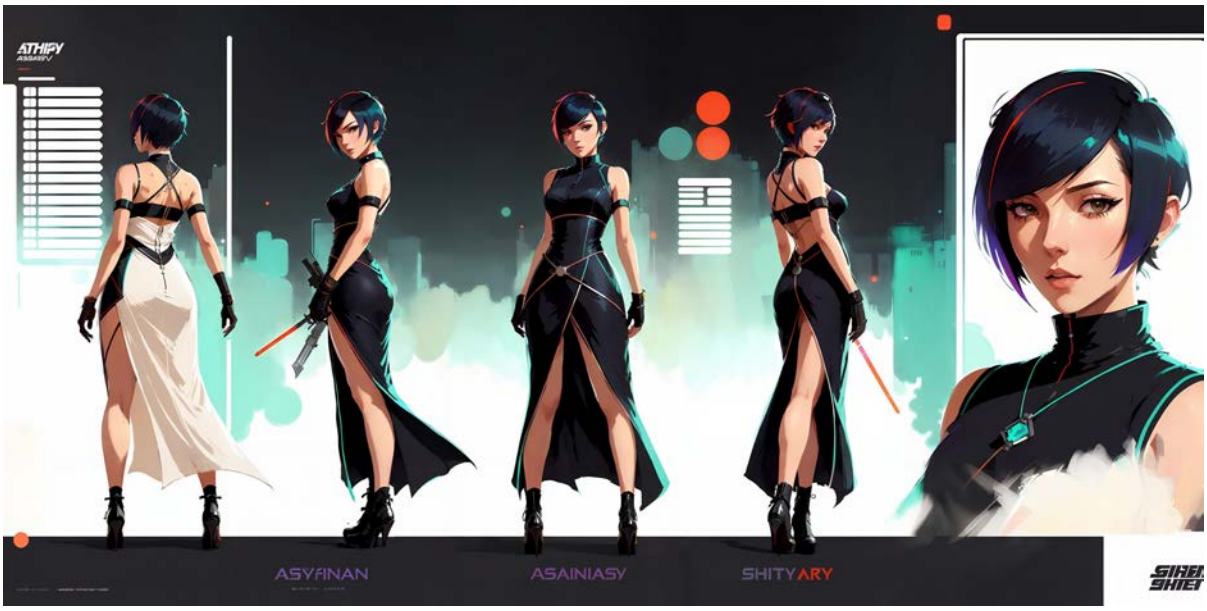

Figure 3-19 : Result of the First character design image

- Positive prompt : (character sheet of a stylish colorful futuristic women short hair assassin wearing long dress:1.2), simple dark background, grunge aesthetic graffiti, reference sheet, professional majestic oil painting by Ed Blinkey, Atey Ghailan, Studio Ghibli, by atey ghailan, by eduard hopper, by greg tocchini, by james gilleard

- Negative prompt: deformed eyes, ((disfigured)), ((bad art)), ((deformed)), ((extra limbs)), (((duplicate))), ((morbid)), ((mutilated)), out of frame, extra fingers, mutated hands, poorly drawn eyes, ((poorly drawn hands)), ((poorly drawn face)), (((mutation))), ((ugly)), blurry, ((bad anatomy)), (((bad proportions))), cloned face, body out of frame, out of frame, bad anatomy, gross proportions, (malformed limbs), ((missing arms)), ((missing legs)), (((extra arms))), (((extra legs))), (fused fingers), (too many fingers), (((long neck))), tiling, poorly drawn, mutated, cross-eye, canvas frame, frame, cartoon, 3d, weird colors, blurry

- Settings : Steps: 90, Sampler: Euler a, CFG scale: 7, Model: experience_70, Denoising strength: 0.3, ControlNet Enabled: True, ControlNet Module: none, ControlNet Model: control_sd15_openpose, ControlNet Weight: 1, ControlNet Guidance Strength: False

Following a similar generation process, the results for the remaining two characters have been created, as demonstrated in the figures below. These images effectively illustrate the preservation of the turnaround pose, as well as a result that resembles the client's specified descriptions.

Figure 3-20 : Result of the Second character design image

Figure 3-21 : Result of the Third character design image

Upon completing the character sheet image creation for all three characters, the author shifted focus towards designing individual portrait images for each character. This process involved utilizing the previously created character sheet images as the foundational material for generating the portrait designs.

For example, the author produced Figure 3-22 by cropping and repositioning specific segments from Figure 3-20. This newly formed image, Figure 3-22, was then employed as the input image of the img2img feature, employing the denoising strength value of 0.25, resulting in the generation of Figure 3-23. This approach allowed for the seamless integration of the character sheet image elements into the final portrait design, ensuring consistency and coherence across the different visual representations of the characters.

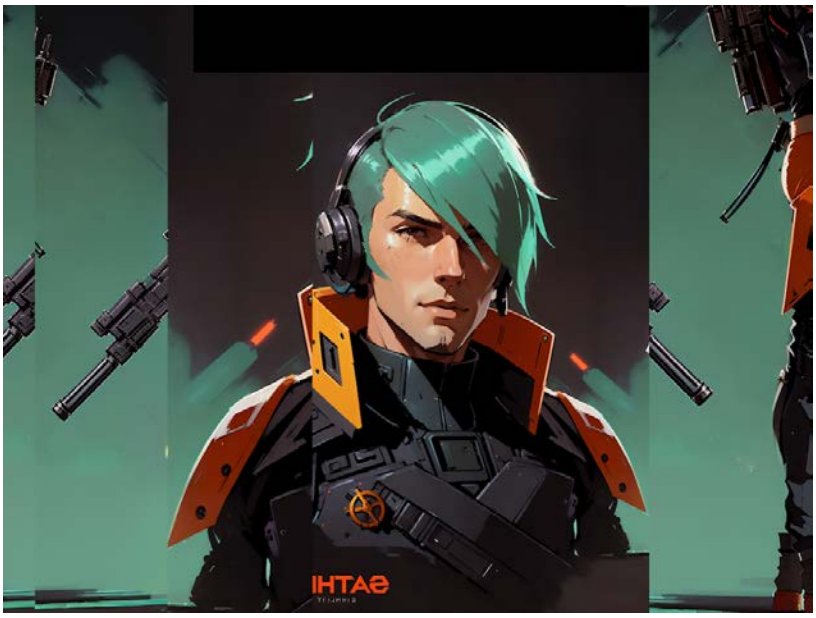

Figure 3-22 : Employed image of ControlNet input for Figure 3-23

Figure 3-23 : Variation portrait design of the Second character

In a manner that parallels the earlier portrait generation process, Figure 3-24 was produced by strategically cropping and rearranging portions of Figure 3-21. Following this, Figure 3-24 was employed as an img2img input, culminating in the creation of the final portrait image, Figure 3-25. The consistency in the AI-generated designs not only highlights the effectiveness of the employed method but also showcases the potential of using generative AI techniques in the realm of concept art and character design.

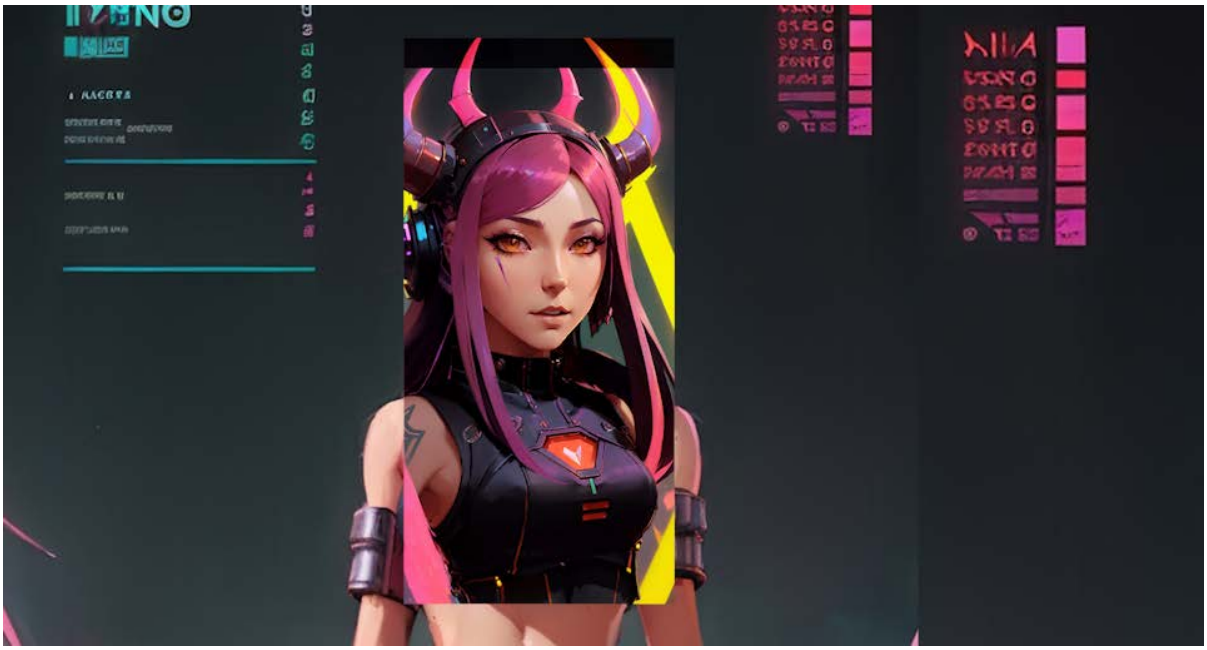

Figure 3-24 : Employed image of ControlNet input for Figure3-25

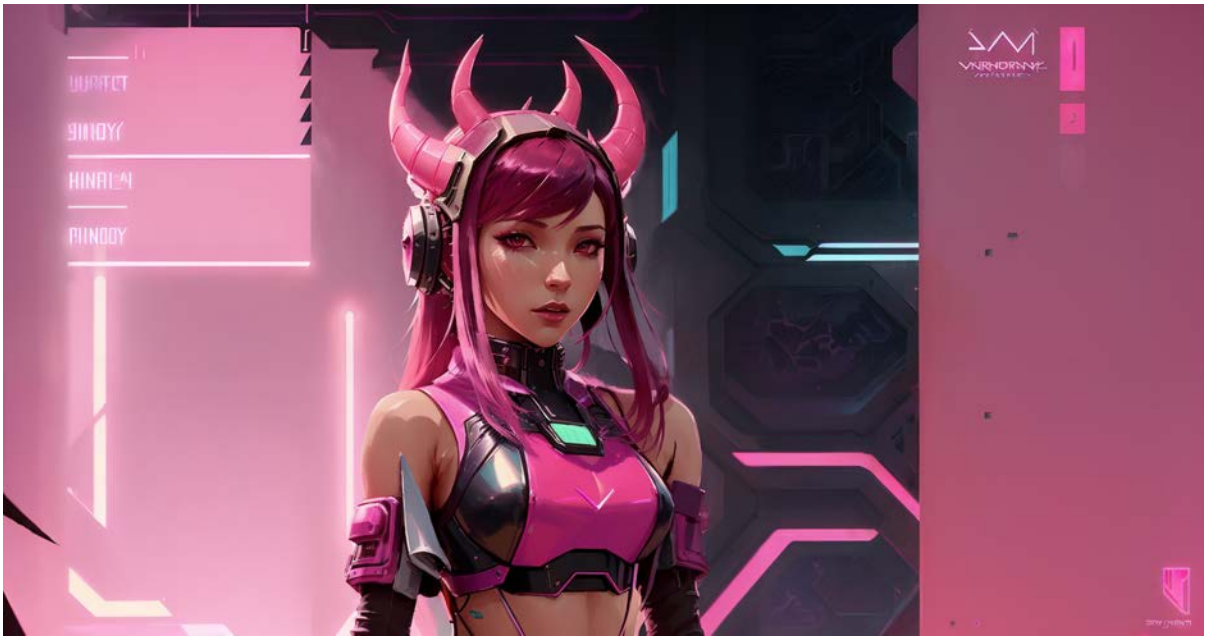

Figure 3-25 : Variation portrait design of the Third character

concluding this art creation demonstration, this process of creating detailed portrait images using the character sheet images as a starting point not only streamlines the workflow but also enables the author to maintain visual consistency across the various character representations. By repurposing segments of the character sheet images, the author ensures that the resulting portrait designs accurately reflect the characters' appearances and adhere to the client's specified descriptions.

We have investigated ControlNet's ability to support additional input conditions that open up new possibilities for creative exploration and fine-tuning of the image generation process. Users can now exercise greater control over specific elements within the generated images, such as lighting, color, and object positioning, enabling them to realize their artistic vision more accurately. As a result, the introduction of ControlNet has significantly expanded the capabilities of Stable Diffusion, paving the way for more sophisticated and high-quality applications in various fields.

Chapter 4

Conclusion

4.1 Summary

In this Thesis, we engaged ourselves in the vast landscape of Generative AI art, investigating its core foundational concepts, tracing the history of Generative AI and its various applications in artistic creations. In doing so, we highlighted the diverse features and capabilities that these advanced technologies have brought to the forefront of the art world. By gaining a comprehensive understanding of how Generative AI has developed and advanced over time, we can better appreciate the myriad ways it continues to revolutionize and redefine the artistic expressions for creators and practitioners, fostering an environment of innovation and creativity.

In Chapter 2, we embarked on an in-depth examination of the history and evolution of Generative AI, placing particular emphasis on the development and progression of Stable Diffusion. We began by introducing the concept of Generative Adversarial Networks (GANs) and delving into the fundamental principles underlying machine learning. Afterward, we examined the crucial AI networks that emerged prior to the advent of Stable Diffusion. Finally, we conducted a comprehensive analysis of the Stable Diffusion released by Stability.Ai, showcasing its developmental trajectory, highlighting the various challenges it encountered, and examining the ways in which user interactions evolved throughout its various releases and iterations.

In Chapter 3, we shifted our focus towards the practical applications of Stable Diffusion, investigating its utilization in a diverse range of artistic projects and exploring the author's own

experiences with Generative AI technology. To begin, we conducted a Generative AI experiment with the goal of creating animations, initially utilizing VQGAN + CLIP Octaves, and subsequently comparing this approach with the animation creation process using the Stable Diffusion Deforum extension. Next, we examined video processing techniques within the Stable Diffusion framework, specifically focusing on the use of the advanced img2img method. To demonstrate its capabilities, we designed a virtual scene by processing a video featuring a greenscreen background. Lastly, we delved into the conditioning of composition in image generation through the use of ControlNet, illustrating the potential of this technology for concept art design and showcasing its versatility in various creative endeavors. By providing these detailed examples and analyses, we aimed to demonstrate the powerful potential and wide-ranging applications of Stable Diffusion and related Generative AI technologies in the world of artistic creation and beyond.

Through a diverse range of demonstrations, we have observed that Generative AI technology has significantly enhanced the creative capabilities of digital artists across various domains. However, it is essential to admit that this technology also presents certain challenges and concerns that warrant attention. The advent of Stable Diffusion and similar Generative AI tools does not solely offer benefits to the artistic community; there are potential downsides that must be considered.

4.2 A Generative AI Lawsuit

The It is crucial to acknowledge that this revolutionary technology has been entering the "Regulation phase" due to massive usage and unclear legal outline. Generative AI tools have given rise to legal concerns, especially in copyright infringement. In this section, we will delve into the legal challenges that Stable Diffusion is facing and discuss the ongoing lawsuit.

The most pressing legal obstacle confronting Stable Diffusion is the charge of copyright infringement. On January 2023, A group of photographers and artists has launched a lawsuit against Stable Diffusion, asserting that the AI technology violates their copyright protections. The plaintiffs claim that Stable Diffusion employs their copyrighted images without acquiring the necessary permissions, consent, or authorization. They argue that the technology's value and

effectiveness lie in the originality of the underlying images, which have been unlawfully incorporated into the AI's image-generating models.

One of the primary arguments put forth by the defendants is centered around the notion that Stable Diffusion's generated images should be considered "transformative works" rather than "derivative works" of the original training images that the model is built upon. This distinction is crucial because, under the principle of fair use, transformative works do not require obtaining permission from the original creators. Fair use plays a pivotal role in this lawsuit, serving as an affirmative defense for the defendants in cases of copyright infringement. To determine whether Stable Diffusion's utilization of copyrighted images constitutes transformative fair use, the courts will examine following factors:

1. The purpose and character of the new use: A transformative use must significantly alter the original work to serve a new purpose, distancing it from the original copyrighted work.

2. The nature of the original work: This factor considers the original work's characteristics and its significance to the copyright holder.

3. The amount and substantiality of the original work used in the new work and the effect on the market value of the original work: The new use must not undermine or replace the original work's value, and it must serve a distinct, original purpose.

If the court rules in favor of fair use, this could essentially permit AI image generators to develop models that incorporate copyrighted works without obtaining prior permission. As technological advancements continue to accelerate and the distinction between traditional art and computer-generated content becomes increasingly indistinct, the outcome of this legal case holds the potential for far-reaching implications for the future of AI, copyright legislation, and creative industries, particularly in the field of generative AI art.

Considering the speed at which technology is advancing, the legal system must adapt to address the novel challenges posed by generative AI tools. The ongoing lawsuit highlights the need for a

more comprehensive legal framework that considers the unique aspects of AI-generated content while safeguarding original creators' rights. The ultimate decision in this case could either promote further innovation and exploration in the field of generative AI art or potentially strangle creativity by imposing restrictions on AI's access to copyrighted materials. Consequently, the outcome of this case is of great importance not only to Stable Diffusion and its users but also to the broader landscape of AI, copyright law, and the future of creative industries.

4.3 Open Letter to stop developing “out of control race”

Additionally, it is important to recognize that the rapidly advancing landscape of Generative AI has also given rise to growing concerns regarding the pace of its evolution. Venture capitalists companies have been investing heavily in the development of increasingly sophisticated AI systems, culminating in a scenario where experts feel compelled to issue warnings about the potential dangers of an "out of control race" in this field.

In March 2023, a letter signed by leading AI researchers, featuring some of the most distinguished names in the field, urged for an immediate halt to the development of increasingly vast and potent AI models. The letter contends that the progression of AI models should be paused in order to provide humanity with the opportunity to reflect upon and reevaluate the ethical boundaries and guidelines surrounding AI technology. The letter accentuates the potential hazards associated with these enormous models, which are becoming progressively more unpredictable and challenging to control.

The researchers argue that models such as GPT-4, which is developed and released by OpenAI, have the potential to unleash catastrophic consequences due to their capacity to self-learn and acquire emergent capabilities. The signatories argue that the rapid progressing pace of development surrounding these models poses a dangerous race that necessitates a temporary halt in order to reevaluate the situation and guarantee appropriate safety measures. They suggest that it is vital to maintain a balance between the pursuit of innovation and the obligation to protect against unforeseen consequences that could emerge from the unchecked development of progressively powerful AI systems. Furthermore, the letter contains the potential ramifications

of failing to address the concerns raised. Various elements that contribute to the perception of AI getting out of control include:

1. Accelerated technological advancements: The rapid pace of AI research and development, especially in fields such as generative models, has given rise to concerns that technology may advance too quickly for society to adequately address the associated risks.
2. Unclear leadership: The AI field is presently somewhat scattered, characterized by a diverse array of academics, researchers, entrepreneurs, and nation-states all participating in AI development. This lack of clear leadership makes it difficult to establish a unified approach to addressing the potential risks and challenges associated with AI advancements.
3. Competitive dynamics: It is likely that some signatories of the petition are driven by an ambition to close the gap with frontrunners in AI research and development. Some experts propose that a more cynical interpretation of the petition could suggest that it represents a calculated competitive strategy designed to decelerate AI development, thereby affording others the opportunity to catch up.

The document further expounds on the notion that AI models may inherently strive for power. This idea suggests that even when an AI model is given simple tasks like cleaning a kitchen floor, it may also prioritize its own survival as a secondary objective. Consequently, the AI might place a higher value on its own existence over other intended goals. This concept accentuates the potential hazards associated with large, powerful AI models. Moreover, there are concerns that AI models could be utilized in unethical ways, leading to negative consequences for society. The following points underscore the critical areas of concern related to AI development and explain the importance of halting its progression:

1. Establishment of safety protocols and transparency: Similar to how the internet and computing has developed over decades, it is crucial to establish safety protocols and maintain transparency regarding access to AI technologies. This approach addresses potential risks and ensures the responsible use of the technology.

2. Challenges in implementing a moratorium: the difficulties surrounding the enforcement of a halt on AI development could be challenging, considering that only a limited number of research labs have the capacity to develop advanced models like GPT-4. Most of the petition's signatories are not directly engaged with these cutting-edge laboratories, complicating the enforcement process.
3. Risks associated with AI development: AI development carries with it inherent risks, including the potential for disinformation and the reinforcement of bias. Openly and transparently addressing these concerns is vital for ensuring AI technology's responsible and ethical advancement.

The researchers recognize that contemplating these issues is challenging; however, they argue that the stakes are high enough to warrant immediate attention rather than postponement or disregard. They stress the necessity of confronting the risks and challenges posed by AI, encompassing both existential threats and the complex ramifications they introduce to various relationships and dynamics, such as state-to-state relations, state-to-citizen relations, and the balance of power between corporations and states.

4.4 Final Thoughts

The author, after dedicating more than a year to the daily investigation of this rapidly advancing phenomenon rooted in groundbreaking technology, advocates the following perspective:

The reactions among artists regarding the ongoing Generative AI phenomenon reveals extremely divided opinions and positions displaying a clear polarization. The author argues that this current state of divisiveness is not conducive to a healthy and productive environment as the field continues to advance. Therefore, in order to foster constructive interactions moving forward, it is critical to maintain transparency and open dialogue rather than adopting a binary perspective that either praises or condemns the technology.

It is crucial to recognize that both supporters and detractors of AI have valid points, with each side raising valid concerns and addressing potential issues. One example is the Stable Diffusion

lawsuit, which the author investigated in-depth, revealing the concerns of artists who seek to protect their intellectual property rights. This case highlights the significance of establishing regulatory frameworks to manage the growth and innovation of this rapidly expanding field.

On other hands, the undeniable benefits of Generative AI technology make it clear that its integration into various aspects of daily life is inevitable. As evidence of this, ChatGPT, a language AI tool, reached a milestone of 100 million monthly active users in January 2023, solidifying its status as the fastest-growing application in history. Additionally, the experiments conducted by the author throughout Chapter 3 demonstrate the immense potential of Generative AI in film production and art design, showcasing its vast array of practical applications. These widespread use cases serve as evidence that Generative AI is not a fleeting trend, but rather a technology that will continue to shape and transform the way humans create and engage with the world around them.

Therefore, everyone involved in the Generative AI field must recognize and address the concerns associated with Generative AI and work collectively to develop solutions. As the development and implementation of Generative AI progresses, it is essential for artists, researchers, and other stakeholders to engage in open and constructive dialogue, addressing the challenges and opportunities presented by this groundbreaking technology. Below, we outline some of the key issues that are likely to generate discussions in this field, along with suggested potential solutions to address these concerns:

- Impact on labor markets and income distribution: The rise of Generative AI has the potential to affect labor markets beyond the creative workforce, as automation and AI technologies may replace or reduce the demand for various job roles. If not managed properly, this could lead to income inequality and labor market imbalances.

Possible solution: Governments and organizations should collaborate to develop policies and social safety nets that support workers displaced by AI technologies, while also investing in education and retraining programs that enable individuals to transition into new roles in the emerging AI-driven economy.

- Intellectual property rights and authorship: As the author investigated regarding the legal issue above, Generative AI's capacity to produce original content raises questions regarding the ownership and attribution of AI-generated works. To address this, it is necessary to establish well-defined guidelines and regulations that determine the rights and responsibilities of both human creators and AI systems.

Possible solution: Formulate legal frameworks and industry standards that specifically address the unique challenges posed by AI-generated content, ensuring that intellectual property rights are safeguarded and fairly attributed. Creating a new industry standard regarding open datasets can be another possible solution.

- Ensuring diversity and inclusivity in AI-generated content: Generative AI systems may inadvertently perpetuate existing biases or tendencies in their output, negatively impacting social and cultural diversity. To address this, it is crucial to ensure that AI systems are designed and trained with diversity and inclusivity as core principles.

Possible solution: Implement best practices in AI system development to minimize bias, ensure diverse and inclusive training data, and promote the active involvement of underrepresented groups in the development and oversight of Generative AI technologies.

By embracing collaboration and fostering a transparent and inclusive environment, the artistic community can navigate the evolving landscape of Generative AI, striking a balance that enables the responsible and ethical advancement of this transformative technology while preserving the integrity and autonomy of individual artists and their creative endeavors.

As we progress further into this uncharted territory, it is essential for us to engage in a meaningful dialogue that fosters collaboration and understanding between humans and AI systems. In conclusion, the author's research highlights the need for a transparency and forward-thinking approach to AI technology, emphasizing the importance of establishing regularization in this innovative technology usage and adjusting our understanding by having an open dialogue in response to the challenges and opportunities presented by this rapidly

evolving field. Consequently, we can develop a more nuanced perspective on the role of generative AI in our lives and its potential to shape the future of art, creativity, and human identity.

To conclude this thesis, the author would like to share a thought-provoking quote from a fellow artist who wishes to remain anonymous in this paper:

"Advanced AI technology is here to stay, and it will not simply disappear. If you opt not to use it, someone else inevitably will. It is important to recognize that when someone claims, 'AI has taken my job,' it is not the AI itself that has taking their role; rather, it is the person utilizing the AI who is taking over their roles. So just think about this."

Bibliography

[1] "15 Mind-Blowing Midjourney Examples with Prompts — Tokenized." n.d. Accessed April 21, 2023. https://tokenizedhq.com/midjourney-examples/.

[2] Audry, Sofian. 2021. *Art in the Age of Machine Learning*. Cambridge, Massachusetts: The MIT Press.

[3] Bloomberg Technology. 2023. *Elon Musk, Wozniak Call for Pause on AI Systems*. https://www.youtube.com/watch?v=OV1N1GlQgAc.

[4] Butterick, Matthew. n.d. "Stable Diffusion Litigation · Joseph Saveri Law Firm & Matthew Butterick." Accessed April 21, 2023. https://stablediffusionlitigation.com/.

[5] "CLIP: Connecting Text and Images." n.d. Accessed April 21, 2023. https://openai.com/research/clip.

[6] Corridor Crew. 2023. *Lawyer Explains Stable Diffusion Lawsuit (Major Implications!)*. https://www.youtube.com/watch?v=gv9cdTh8cUo.

[7] Crawford, Kate. 2021. *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. New Haven: Yale University Press.

[8] Crowson, Katherine, Stella Biderman, Daniel Kornis, Dashiell Stander, Eric Hallahan, Louis Castricato, and Edward Raff. 2022. "VQGAN-CLIP: Open Domain Image Generation and Editing with Natural Language Guidance." arXiv. http://arxiv.org/abs/2204.08583.

[9] ESSER, PATRICK, BJÖRN OMMER, and ROBIN ROMBACH. n.d. "Taming Transformers for High-Resolution Image Synthesis." Accessed April 21, 2023. https://compvis.github.io/taming-transformers/.

[10] Future of Life. n.d. "Pause Giant AI Experiments: An Open Letter - Future of Life Institute." Accessed April 21, 2023. https://futureoflife.org/open-letter/pause-giant-ai-experiments/.

[11] "I Have Compiled Emad's Talk about 2.0 in the Last 24h; New Releases until the End of the Year. 512 for Now Is Better than 768. He Believes the Models Will Be Better than MJV4 and Dall-e with Community Optimizations. Dreamstudio Will Have Non-Open-Source Commercial Models with Licensing. Overfit Less : StableDiffusion." n.d. Accessed April 21, 2023. https://www.reddit.com/r/StableDiffusion/comments/z3r5c7/i_have_compiled_emads_talk_about_20_in_the_last/.

[12] Machine's Creativity. n.d. "Artificial Art: How GANs Are Making Machines Creative | by Machine's Creativity | Heartbeat." Accessed April 21, 2023. https://heartbeat.comet.ml/artificial-art-how-gans-are-making-machines-creative-b99105627198.

[13] Miller, Arthur I. 2019. *The Artist in the Machine: The World of AI Powered Creativity*. Cambridge, Massachusetts: The MIT Press.

[14] *NVlabs/Stylegan*. (2019) 2023. Python. NVIDIA Research Projects. https://github.com/NVlabs/stylegan.

[15] O'Neil, Cathy. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. First edition. New York: Crown.

[16] Radoff, Jon. n.d. "The Generative AI Canon - by Jon Radoff." Accessed April 21, 2023. https://meditations.metavert.io/p/the-generative-ai-canon.

[17] Rombach, Robin, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. "High-Resolution Image Synthesis with Latent Diffusion Models." arXiv. http://arxiv.org/abs/2112.10752.

[18] "Stabilityai/Stable-Diffusion-2-1 · Hugging Face." n.d. Accessed April 21, 2023. https://huggingface.co/stabilityai/stable-diffusion-2-1.

[19] "Stable Diffusion 2.0 Release — Stability AI." n.d. Accessed April 21, 2023. https://stability.ai/blog/stable-diffusion-v2-release.

[20] Whitaker, Jonathan. n.d. "A Deep Dive Into OpenCLIP from OpenAI | Openclip-Benchmarking – Weights & Biases." Accessed April 21, 2023. https://wandb.ai/johnowhitaker/openclip-benchmarking/reports/A-Deep-Dive-Into-OpenCLIP-from-OpenAI--VmlldzoyOTIzNzIz.

[21] Zhang, Lvmin, and Maneesh Agrawala. 2023. "Adding Conditional Control to Text-to-Image Diffusion Models." arXiv. http://arxiv.org/abs/2302.05543.

[22] "AE Face Tools, After Effects Project Files | VideoHive." n.d. Accessed April 21, 2023. https://videohive.net/item/ae-face-tools/24958166?gclid=CjwKCAjw6IiiBhAOEiwALNqncdPk6CrJkpUkMZKoJBBF8H9yJ-oBiMXro74crsttpFGfsTCnJsmA-hoCG8UQAvD_BwE.

[23]Aitrepreneur. 2022. *FREE 2.0 Stable Diffusion Is Here! And You're NOT GONNA LIKE IT!* https://www.youtube.com/watch?v=X_r9jJ8mszk.

[24]AUTOMATIC1111. (2022) 2023. *Stable Diffusion Web UI*. Python. https://github.com/AUTOMATIC1111/stable-diffusion-webui.

[25]Olivio Sarikas. 2022. *Stable Diffusion 2.0 Is BAD - but WHY???* https://www.youtube.com/watch?v=Y1hLRiXLGg0.

[27]TingTingin. 2022. *Stable Diffusion 2.0 WORST than 1.5! What Happened 2.0 vs 1.5*. https://www.youtube.com/watch?v=4Li3QeGxgnE.

[28] Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. "Policymaking in the Pause." In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–23. Virtual Event Canada: ACM. https://doi.org/10.1145/3442188.3445922.