ANIMATION USING ARTIFICIAL INTELLIGENCE

CSE3013 – Artificial Intelligence

PROJECT BASED COMPONENT REPORT

by
THIRUMANI SRINIVASAN DEEPIKA- 20BCE2047
VINAY VIKRANTH- 20BCE2059
Y. NAGA SAI SABARISH- 20BCE2370
ABHINAV KRISHNA MITTAL- 20BDS0326

School of Computer Science and Engineering

VIT
Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

NOVEMBER 2023

DECLARATION

I hereby declare that the report entitled "**Animation using AI"** submitted by me, for the CSE3013 Artificial Intelligence (EPJ) to Vellore Institute of Technology is a record of bonafide work carried out by me under the supervision of **Dr. L. Mohana Sundari.**

I further declare that the work reported in this report has not been submitted and will not be submitted, either in part or in full, for any other courses in this instituteor any other institute or university.

Place: Vellore

Date: 23-11-23

Signature of the Candidate

Acknowledgement

We extend our sincere gratitude to Dr. Mohana Sundari, our esteemed faculty for the Artificial Intelligence course, for her invaluable guidance and unwavering support throughout the duration of this project. Her expertise and insightful feedback have been instrumental in shaping our understanding and enhancing the quality of our work. We would also like to express our heartfelt thanks to our fellow team members, whose dedication and collaborative efforts have made this project a reality. Each member has played a crucial role in the successful completion of our project, "Animation using AI." Their commitment, enthusiasm, and diverse skill sets have contributed to the project's overall success. This experience has not only enhanced our understanding of Artificial Intelligence but has also fostered a spirit of collaboration and teamwork that will undoubtedly benefit us in our future endeavors.

Abstract

The exponential advancement of artificial intelligence (AI) has resulted in significant changes across all areas. This endeavor aims to foster the production of content and promote creativity via the integration of artificial intelligence, computer vision, and animation methodologies. The main aim of this work is to develop a novel system that effectively transforms real-world video sequences into engaging animated cartoons. The project presents an innovative methodology that use diffusion processes to deliberately inject controlled noise into video frames, then followed by refinement techniques with the objective of reducing undesired fluttering. The key aspect of this methodology is the deliberate introduction of noise through diffusion, resulting in the infusion of dynamic and aesthetic attributes into video frames. The careful manipulation of noise settings enables an optimal balance between artistic innovation and logical consistency, so imparting a unique aesthetic quality to the animations. The frames are subjected to a denoising procedure after being enhanced with noise, resulting in visually appealing animations that are free from noticeable flicker. The completion of this project results in the development of a full pipeline that encompasses diffusion-based animation synthesis, frame refining controlled by noise, and seamless integration of visual effects. This endeavor delves into unexplored areas in animation creation by embracing the complex interaction between diffusion, denoising, and visual effects

Table of Contents

.

List of Figures

List of Tables

List of Abbreviations

AI	Artificial Intelligence
VFX	Visual Effects
DDPM	Denoising Diffusion Probabilistic Model
DDIM	Denoising Diffusion Implicit Model
RNN	Recurrent Neural Network
GAN	Generative Adversarial Network
GPU	Graphical Processing Unit
TPU	Tensor Processing Unit

1. INTRODUCTION

1.1 Objective

Recent advances in deep learning have enabled new AI techniques for automatically converting real-world video into cartoon animations. Key innovations that make this possible include:

- Image-to-image translation models - These models can learn to translate between two visual domains, like real-world video frames and cartoon images. They are trained on dataset pairs to learn a mapping from one domain to another.
- Video prediction models - These models can learn to predict future video frames, by analyzing motion and content in previous frames. This allows an AI system to take a real-world video clip and predict how it would look as a cartoon over time.
- Motion stylization - AI techniques can also learn to abstract and stylize real-world motion from video into caricatured cartoon motion. This involves analyzing movement, exaggerating it, and rendering it with typical cartoon textures and outlines.
- Facial animation - Specialized AI methods are being developed to detect, track and recreate facial expressions and mouth movements from real-world video in a cartoon rendering. This helps automate lip sync and facial acting.

The end result of combining these kinds of AI video, image, and motion analysis techniques is an automated system that can convert raw real-world video into cartoon animation styled output. It saves significant human effort compared to manual cartoon creation. These techniques are still evolving but show promise to change animation production.

1.2 Motivation

Diffusion-Based Animation Synthesis: Develop a robust framework for utilizing diffusion-based techniques to convert video frames into animations. This involves the strategic introduction of controlled noise through diffusion processes, imparting frames with dynamic and artistic qualities.

Flicker Elimination Through Denoising: Implement denoising algorithms to eliminate unwanted flicker from frames that have undergone noise-induced diffusion. This step ensures the final animations exhibit smooth visual transitions, enhancing their overall quality.

Integration of Visual Effects (VFX): Integrate visual effects seamlessly with diffusion-induced frames to augment narrative impact and visual engagement. By weaving VFX elements into the animations, the project seeks to enhance realism and creative appeal.

Enhanced Realism and Artistic Flair: Fine-tune the noise parameters of the diffusion process to strike a balance between creativity and coherence. This aims to infuse animations with a unique artistic touch while maintaining a cohesive visual flow.

Comprehensive Pipeline Development: Create an end-to-end pipeline encompassing diffusion-based animation synthesis, noise-controlled denoising, and seamless VFX integration. This pipeline will serve as a practical tool for converting video content into captivating animations.

Pushing Boundaries of Creative Expression: By leveraging the interplay of diffusion, denoising, and VFX, explore uncharted territories in animation production. The project aspires to redefine the boundaries of creative expression by enabling the generation of animations that possess heightened realism and a distinct artistic flair.

Table 2.1 Literature Survey

Ref. No.	Paper title	Journal name and year of publication	Workdone	Technique(s) used	Disadvantages / Gaps identified
1.	**On the use of Stable Diffusion for creating realistic faces: from generation to detection**	*2023 11th International Workshop on Biometrics and Forensics (IWBF)*, Barcelona, Spain, 2023	Lorenzo Papa and colleagues (2023) propose a critical analysis of the overall pipeline, i.e., from creating realistic human faces with Stable Diffusion v1.5 to recognizing fake ones. The objective is to identify the text prompts that drive the image generation process. The spread of these technologies paves the way to previously unimaginable creative uses while raising the possibility of malicious applications.	They first propose an analysis of the prompts that allow the generation of extremely realistic faces with a human-in-the-loop approach. The chief objective is to identify the text prompts that drive the image generation process to obtain realistic photos that resemble everyday portraits captured with any camera. It involves critical analysis of the overall pipeline, i.e., from creating realistic human faces with Stable Diffusion v1.5 to recognizing fake ones.	The paper shows that Stable Diffusion is a promising technique for generating and detecting faces. However, there are still some challenges that need to be addressed, such as the ability to generate faces with different identities and expressions, and the ability to detect faces in images that are very difficult to see
2.	**DreamPose: Fashion Image-to-Video Synthesis via Stable Diffusion**	Cornell Universtitycs arXiv, 2023	DreamPose is a diffusion-based method for generating animated fashion videos from still images. The authors transform a pretrained text-to-image model (Stable Diffusion) into a pose-and image guided video synthesis model, using a novel finetuning strategy, a set of architectural changes to support the added conditioning signals, and techniques to encourage temporal consistency	The proposed method aims to produce photorealistic animated videos from a single image and a pose sequence. They use a pretrained Stable Diffusion model on a collection of fashion videos. This involves adapting the architecture of Stable Diffusion to accept additional conditioning signals, and to output temporally consistent content that can be viewed as a video. The model is based on VAE	While their method produces realistic results on most plain and simple-patterned fabrics, some of the results present minor flickering behavior on large and complex patterns. The temporal consistency on such patterns, ideally without subject-specific finetuning is another drawback faced by the author. Similar to other diffusion models, their finetuning and inference times are

				(Variational Auto Encoder)	slow compared to GAN or VAE methods.
3.	**Point-to-Point Video Generation**	Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019	The authors introduce point-to-point video generation (p2p generation) that controls the generation process with two control points—the targeted start- and end-frames. The paper explores the problem that given a pair of control points and the generation length T, the authors aim to generate a sequence x1:T with the specified length such that their start- and end-frames {x1, xT} are consistent with the control points.	While the targeted start-frame is already fed as an initial frame, the paper discussed a technique which adopts a straightforward strategy to incorporate the control points into the model at every timestep by feeding features encoded from the targeted end-frame hT to the model.	The model cannot deal with all types of videos and video formats. The authors further mention that the model cannot handle high-resolution videos. Modeling all the details such as small objects or noisy background in high-res videos is still an open problem for the existing video generation/prediction methods.
4.	**TiVGAN: Text to Image to Video Generation With Step-by-Step Evolutionary Generator**	IEEE Access, vol. 8, pp. 153113-153122, 2020	TiVGAN evolves frame-by-frame and finally produces a full-length video. In the first phase, the focus is on creating a high-quality single video frame while learning the relationship between the text and an image. As the steps proceed, the model is trained gradually on a greater number of consecutive frames. This step-by-step learning process helps stabilize the training and enables the creation of high-resolution video based on conditional text descriptions	The paper uses a technique the authors have named as Text-to-Image-to-Video Generative Adversarial Network (TiVGAN). The authors of the paper propose a novel network (GAN) that generates a video corresponding to a given description. The learning framework of the generative adversarial network is established on the basic concept that connected frames of a video have substantial continuity.	The paper briefly mentions the challenges of generating video from generated images as every frame in video is made of multiple still frames. The continuity of these frames is often not perfect hence introducing some difficulties in maintaining continuity.

5.	**Video-to-Video Synthesis**	Cornell Universitycs arXiv, 2018	The paper addresses the problem of synthesizing videos from static images. The authors propose a method that can generate dynamic and coherent videos given an input image as well as a semantic label map that describes the desired scene. The technique enables a wide range of video synthesis applications, including changing weather conditions, object manipulation, and artistic video generation.	The foundation of the approach is based on conditional generative adversarial networks (cGANs), which enable the generation of images conditioned on specific input information. In this case, the input information consists of the source image obtained by the splitting of videos into their constituent frames and a target semantic label map.	Training generative models for video synthesis is computationally intensive and requires a significant amount of data. This can make the training process time-consuming and resource-intensive. While the approach is capable of generating coherent videos, scenes with highly complex and interactive dynamics might still pose challenges in terms of generating realistic and accurate videos.

3. TECHNICAL SPECIFICATIONS

I. Automatic1111 (stable-diffusion-webui)- to create a repeating pattern like a wallpaper

II. Python 3.10.6- To run the codes

III. Git 1.5 stability ai model - Stable Diffusion is a latent text-to-image diffusion model capable of generating photo-realistic images given any text input.

IV. DreamBooth - method to personalize text-to-image models like Stable Diffusion given just a few (3-5) images of a subject.

V. Pictures of Human - to train the model for the specific person as input

VI. Pictures of animation style- to train the model for the specific animation style as output

VII. Anaconda- to train the model using dream booth

Diffusion Models:

•Diffusion models are a type of generative model that can create realistic images and videos from noise through a denoising process.

•They are trained to add noise to data, and then reverse that process to clean up the noise and generate high quality outputs.

•Different diffusion model architectures like DDPM and DDIM can be used for image and video generation.

•Fine-tuning pretrained diffusion models on custom datasets can enable generating animations mimicking a particular art style.

Noise Injection and Denoising:

•Deliberately injecting noise into video frames and then denoising can impart artistic style and fluid motion.

•Varying noise levels and diffusion steps allows controlling the trade-off between detail/consistency and abstract artistic effect.

•Advanced denoising techniques like using discriminators or denoising diffusion probabilistic models remove unwanted artifacts while retaining desired aesthetic attributes.

Conditional Generation:

•Providing class labels or text descriptions allows conditioning the diffusion model to generate specific kinds of outputs.

•The text prompts guide the model to recognize and generate the target concepts like characters or scenes.

Visual Effects:

•VFX techniques like compositing, particle effects, 3D assets etc. can enhance the visual appeal and realism of generated animations.

•Matching the rendering of VFX elements to the diffusion animation style results in a cohesive final output.

Training Methodology:

•Using a large dataset of videos in the target style allows training robust diffusion models.

•Fine-tuning on smaller custom datasets allows adapting to new styles.

•Staged training on increasing sequence lengths stabilizes video generation.

4. DESIGN

•In the realm of image transformation, cutting-edge advancement has emerged in the form of machine-learning diffusion. This remarkable process empowers computers to craft images by harnessing the power of noise. To facilitate this innovation, we're utilizing the Automatic1111 stable-diffusion-webui, a potent tool that facilitates the creation of repetitive patterns within a chosen art style, using our test model as a foundation.

•Furthermore, our exploration extends to the 1.5 stability AI model. This model stands as a pinnacle of achievement in generating images that transcend realism, all while responding to textual input with astounding accuracy. This two-way interaction between text and images ushers in a new era of creative expression and personalization.

•For a more refined touch, we've incorporated Dreambooth, a method that breathes life into text-to-image models, including the steadfast stable diffusion. Dreambooth's capabilities shine brightest when trained on images that encapsulate the same art style found in our test model. This harmonious synergy between the Dreambooth model and the stable-diffusion-webui ensures that each personalized creation resonates with the underlying aesthetic, captivating audiences with its uniqueness and allure.

5. PROPOSED SYSTEM

1. Human Style Recognition:

1. Dataset: For recognizing "human style," you would require a dataset related to human behavior and styles. Depending on your specific task, this could include datasets with information on human gestures, speech patterns, or writing styles. You might also need labelled data to train the model, indicating different aspects of human style.

2. Model: The choice of model would depend on the nature of your data and the recognition task. For example, if you're working with speech patterns, you could use deep learning models like recurrent neural networks (RNNs) or Transformers. If it's text-based, you might use natural language processing models.

3. Training: You'd train your model using this dataset to capture patterns and characteristics of human behavior and style.

2. Art Style Reference (Dragon Ball Z):

1.Dataset: To reference the art style from "Dragon Ball Z," you would need a dataset of images or artwork from the anime. This dataset can be used to guide the art style generation.

2.Model: For generating art in the style of "Dragon Ball Z," you could use image generation techniques, such as Generative Adversarial Networks (GANs) or deep neural networks. Here we are using the dream Booth model.

3.Training: The model would be trained on the "Dragon Ball Z" art dataset to learn the specific style elements present in the anime's artwork.

Training in Google Colab:

Google Colab is a popular platform for training machine learning models using Python. You can use it for both your human style recognition model and the art style generation model. It provides free access to GPU and sometimes TPU resources for faster training.

Stable Diffusion:

Stable Diffusion is a training technique for deep generative models, which is used to improve the stability and quality of the generated content. You can implement this technique during the training process of your DreamBooth model, assuming that the model architecture supports it.

The key is curating stylistic datasets and providing descriptive text prompts during training so that the model learns the visual patterns you want it to mimic. Fine-tuning on Google Colab with Python and Stable Diffusion allows you to customize the model for new styles.

6. RESULTS AND DISCUSSION

Video outputs

https://drive.google.com/drive/folders/1FF2wNaQg1Jy_B3hTQl4UCWu81UVt42Bw?usp=sharing

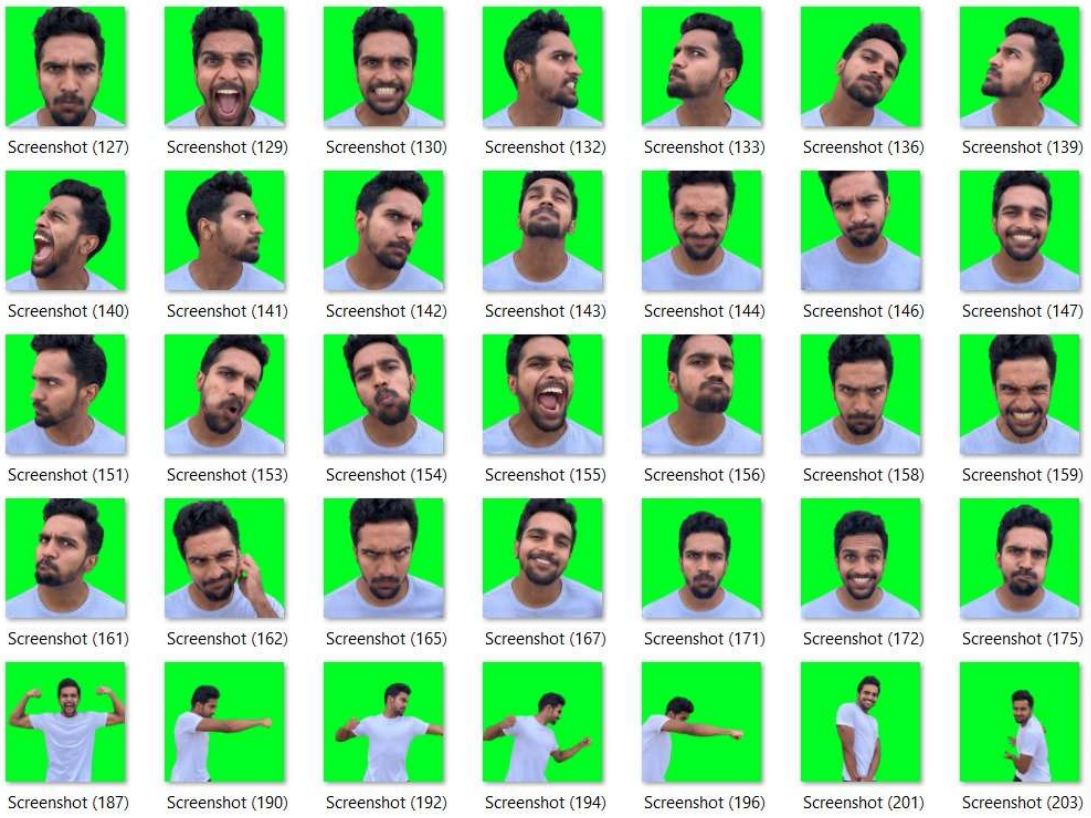

Fig. 6.1

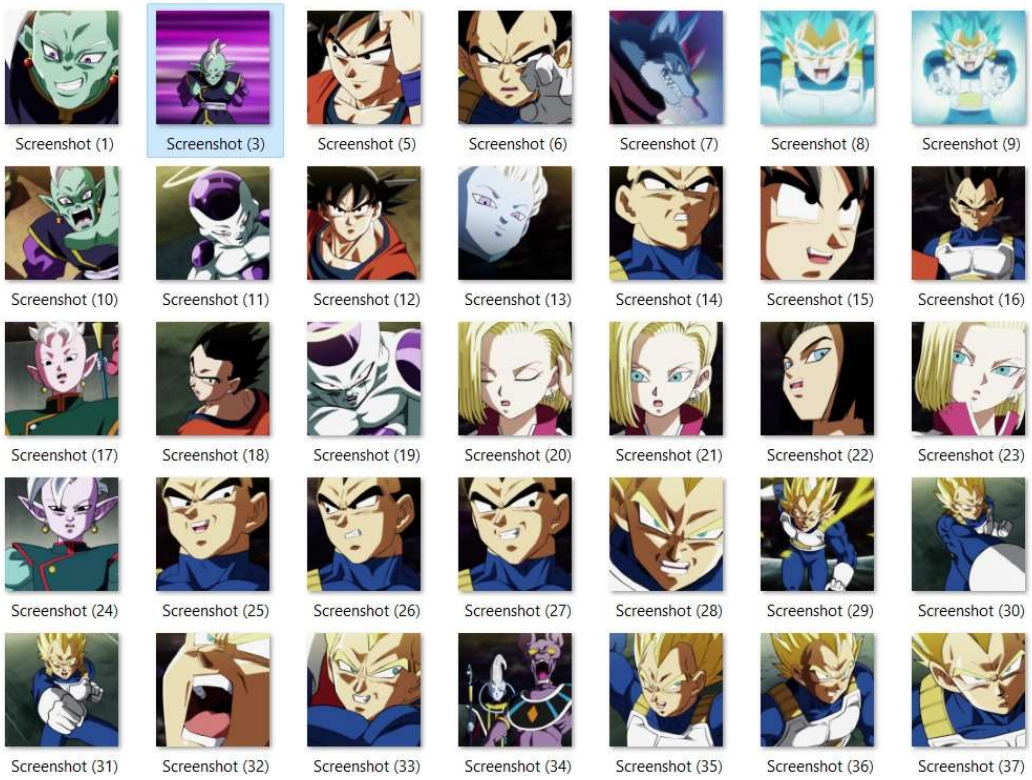
Screenshot (1)
Screenshot (3)
Screenshot (5)
Screenshot (6)
Screenshot (7)
Screenshot (8)
Screenshot (9)
Screenshot (10)
Screenshot (11)
Screenshot (12)
Screenshot (13)
Screenshot (14)
Screenshot (15)
Screenshot (16)
Screenshot (17)
Screenshot (18)
Screenshot (19)
Screenshot (20)
Screenshot (21)
Screenshot (22)
Screenshot (23)
Screenshot (24)
Screenshot (25)
Screenshot (26)
Screenshot (27)
Screenshot (28)
Screenshot (29)
Screenshot (30)
Screenshot (31)
Screenshot (32)
Screenshot (33)
Screenshot (34)
Screenshot (35)
Screenshot (36)
Screenshot (37)

Fig. 6.2

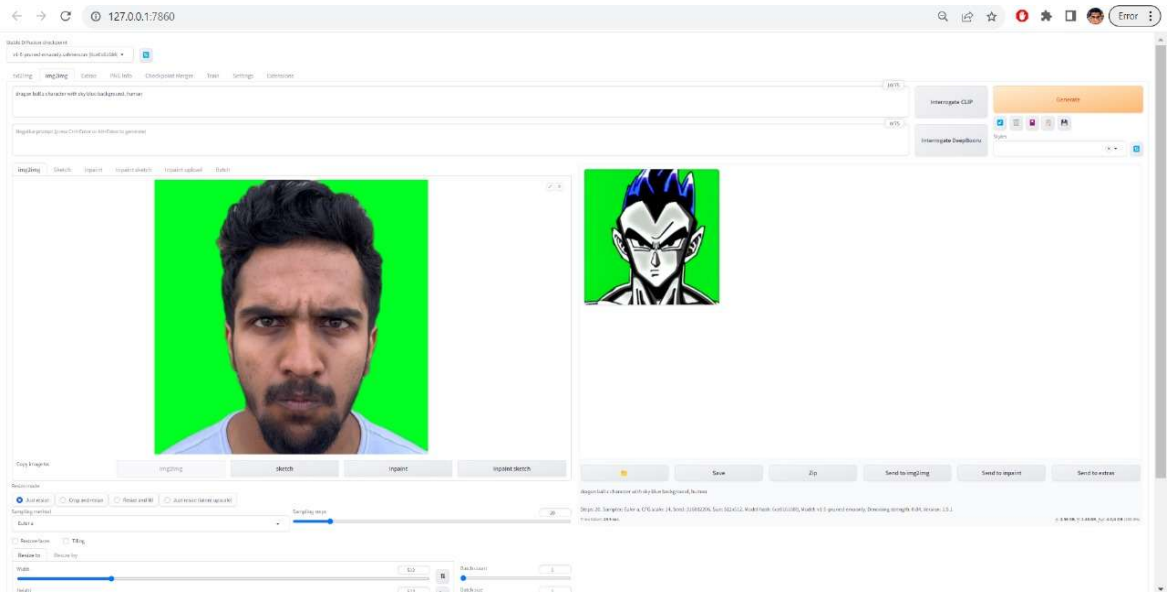
127.0.0.1:7860
Error

Fig. 6.3

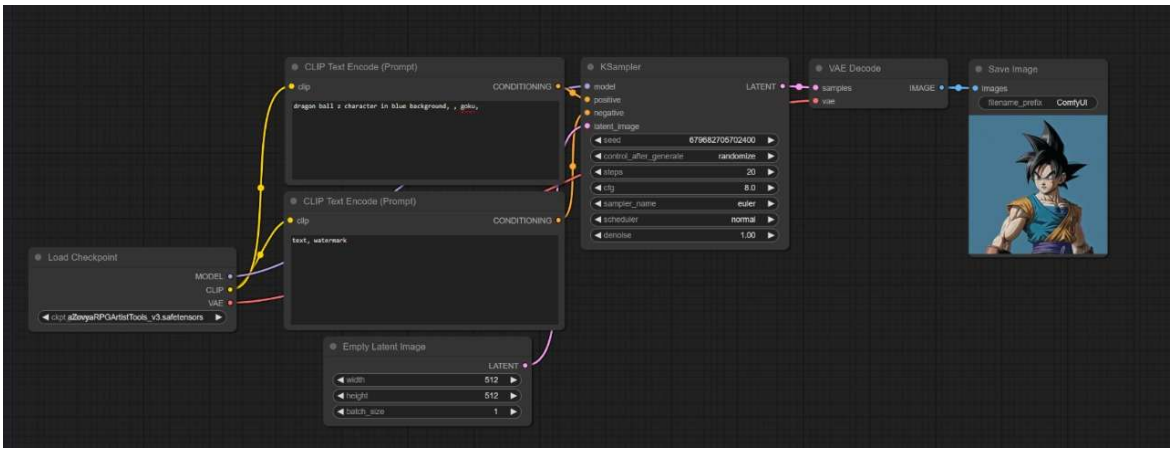
CLIP Text Encode (Prompt)
clip
CONDITIONING
dragon ball z character in blue background, , goku,
CLIP Text Encode (Prompt)
clip
CONDITIONING
text, watermark
KSampler
model
positive
negative
latent_image
LATENT
seed 679682706702400
control_after_generate randomize
steps 20
cfg 8.0
sampler_name euler
scheduler normal
denoise 1.00
VAE Decode
samples
vae
IMAGE
Save Image
images
filename_prefix ComfyUI
Load Checkpoint
MODEL
CLIP
VAE
ckpt_name aZovyaRPGArtistTools_v3.safetensors
Empty Latent Image
LATENT
width 512
height 512
batch_size 1

Fig. 6.4

CONCLUSION

In conclusion, the stable diffusion method employed in this project opens up new avenues for creative expression in the realm of animation. Stable Diffusion is an artificial intelligence (AI) model that creates images. Stable Diffusion uses a latent diffusion model (LDM). It starts with random noise and then it goes through many steps to remove noise from the picture until it matches the text prompt or the given style. The trained model demonstrated its ability to seamlessly transform live-action footage into visually striking anime sequences by using a stable diffusion model to replicate the live-action video sequence frame by frame in a specific art style, highlighting the adaptability and versatility of the approach. The future work could focus on refining the model's ability to capture specific anime styles, enhancing the user interface for accessibility, and exploring ways to optimize computational efficiency.

REFERENCES

[1]. L. Papa, L. Faiella, L. Corvitto, L. Maiano and I. Amerini, "On the use of Stable Diffusion for creating realistic faces: from generation to detection," 2023 11th International Workshop on Biometrics and Forensics (IWBF), Barcelona, Spain, 2023, pp. 1-6, doi: 10.1109/IWBF57495.2023.10156981.

[2] Johanna Karras, Aleksander Holynski, Ting-Chun Wang, Ira Kemelmacher-Shlizerman, DreamPose: Fashion Image-to-Video Synthesis via Stable Diffusion

[3] Wang, Tsun-Hsuan & Cheng, Yen-Chi & Lin, Chieh & Chen, Hwann-Tzong & Sun, Min. (2019). Point-to-Point Video Generation.

[4]. D. Kim, D. Joo and J. Kim, "TiVGAN: Text to Image to Video Generation With Step-by-Step Evolutionary Generator," in IEEE Access, vol. 8, pp. 153113-153122, 2020, doi: 10.1109/ACCESS.2020.3017881.

[5] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, Bryan Catanzaro, Video-to-Video Synthesis, doi: 10.48550/arXiv.1808.06601