Explore the Application of Diffusion Model in Video Content Creation

Haiyun Xiao
Illinois Institute of Technology
Chicago, IL, USA
hxiao8@hawk.iit.edu

Yu Li
Illinois Institute of Technology
Chicago, IL, USA
yli385@hawk.iit.edu

Haichen Pang
Illinois Institute of Technology
Chicago, IL, USA
hpang3@hawk.iit.edu

Figure 1. A series of experimental images generated from Stable Diffusion version 1.4, 2023.

Abstract

As we enter the era of short videos, the content landscape has witnessed a surge in video production. However, this surge still necessitates substantial creator involvement throughout the entire content creation process. Inspired by the transformative potential of Diffusion Models, we've embarked on a quest to explore the feasibility of generating videos from text, much like the concept of text-to-image generation. In this research, we present two distinct pathways for video generation using Diffusion Models. Firstly, we leverage pre-trained generative Diffusion Models (DDMs) to produce a sequence of images based on detailed prompts derived from video text scripts. Given that videos are essentially composed of numerous frames, each of these frames can be treated as an individual image. Upon obtaining this series of images, we concatenate and semantically link them together to form a cohesive video. This research including experiments and summary has been completed and documented in this report. Our implementation is available at https://github.com/byxhy/Explore-Stable-Diffusion

ACM Reference Format:
Haiyun Xiao, Yu Li, and Haichen Pang. 2023. Explore the Application of Diffusion Model in Video Content Creation. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX).* ACM, New York, NY, USA, 5 pages. https://doi.org/XXXXXXX.XXXXXXX

Conference acronym 'XX, Oct 00–00, 2023, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

In today's fast-paced digital age, short videos have become the quintessential form of entertainment, communication, and self-expression. The current era is witnessing an unprecedented surge in the popularity of short video content, with platforms like TikTok, Instagram Reels, and YouTube Shorts at the forefront of this cultural revolution. These bite-sized videos, typically lasting just a few seconds to a minute, have transformed the way we consume and create content. They offer an instant, engaging, and easily digestible way to connect with audiences, share information, showcase creativity, and even launch careers. This popular era of short videos has not only redefined our online experiences but also opened up new avenues for individuals and businesses to make their mark in the digital landscape, making it an exciting and dynamic era in the world of media and communication.

For content creators, the prospect of reducing their efforts and streamlining the video production process is indeed an

enticing one. The aim is to enable creators to focus more on ideation and creative aspects, allowing them to generate novel concepts and content. To achieve this, a viable approach is to develop an intelligent agent or system that can assist in automating various aspects of video production. The intelligent agent can be designed to take creative ideas and concepts as input from content creators and then execute numerous tasks involved in video production. Such tasks may include generating scripts, scene design, image or video synthesis, post-production editing, and more. The agent can leverage advanced technologies like AI, machine learning, and computer vision to accomplish these tasks efficiently.

By shifting the burden of technical and labor-intensive tasks to the intelligent agent, creators can devote more of their time and energy to the creative and artistic aspects of their work. This can lead to a more efficient and streamlined content creation process, ultimately fostering greater innovation and productivity in the field of video production.

The denoising diffusion models (DDMs) have been successfully applied in various tasks such as image and video synthesis , audio generation, image customization, reinforcement learning, and recently in scientific applica-tions[2][5][7]. Leveraging the capabilities of contemporary large-scale (DDMs), we have embarked on an exploration of the feasibility of inputting video descriptions in text format and tasking the diffusion model with the generation of the corresponding video content. Our research can be divided into two main phases: In the first part of our investigation, we fed video text descriptions as prompts to the diffusion model. This prompted the model to generate a sequence of consecutive images through a text-to-image conversion process. Subsequently, these generated images were synthesized into videos. This portion of our research has been executed and documented. The second part of our research delves into the realm of the latent diffusion model. In this phase, we are striving to enable the model to directly learn the underlying principles governing video content. Our aim is to explore the potential for inputting text descriptions and having the model generate full-fledged videos directly, bypassing the intermediate step of generating individual images.

By advancing into this second phase, we seek to unlock the untapped potential of diffusion models in generating video content directly from textual descriptions, thus furthering the realm of automated content creation.

2 Background

We briefly describe essential background here.

2.1 Diffusion Models

Diffusion Models are probabilistic models designed to learn a data distribution p(x) by gradually denoising a normally distributed variable, which corresponds to learning the reverse process of a fixed Markov Chain of length T. The forward process adds stochastic noises to a data sample to learn the pattern of dataset. The reverse direction, which corresponds to the generative by denoising[2].

2.2 Conditioning Mechanisms

Similar to other types of generative models, diffusion models are in principle capable of modeling conditional distributions. This can be implemented with a conditional denoising autoencoder and paves the way to controlling the synthesis process through inputs such as text, or other image-to-image translation task[5].

2.3 CLIP

CLIP (Contrastive Language-Image Pre-Training) is a neural network trained on a variety of (image, text) pairs. It is a Transformer language model to used as the language understanding component that takes the text prompt and produces token embeddings[1]. It can be instructed in natural language to predict the most relevant text snippet, given an image, without directly optimizing for the task, similarly to the zero-shot capabilities of GPT-2 and 3[4].

3 Experiments

3.1 Experimental Setup

Task Settings. We conducted experiments to explore two distinct approaches for generating video content. The first approach involved synthesizing multiple consecutive image frames produced by the diffusion model and assembling them into a coherent video. In contrast, the second approach aimed to directly generate video content from the diffusion model. The second half of our investigation is currently ongoing and inconclusive. The experimental part of this report also focuses on the first method. The inputs for both tasks are the same, which consists of a textual prompt describing the script for the desired video content. This prompt serves as the foundational input for both methods. However, the first task includes an additional step where multiple images are processed as individual frames and then combined to produce a complete video, while the latter task bypasses this intermediate step and generates the video directly from the provided textual prompt.

Model and Datasets. We test different pre-trained DDMs on various datasets for experiments. Finally, we choose Stable Diffusion from Hugging Face as our pipeline. It provide an end-to-end inference pipeline that we can use to generate images from text. It's trained on 512x512 images from a subset of the LAION-5B database. This model uses a frozen CLIP ViT-L/14 text encoder to condition the model on text prompts. With its 860M UNet and 123M text encoder. In this work, we use Stable Diffusion version 1.4 (CompVis/stable-diffusion-v1-4) as our pre-trained model.

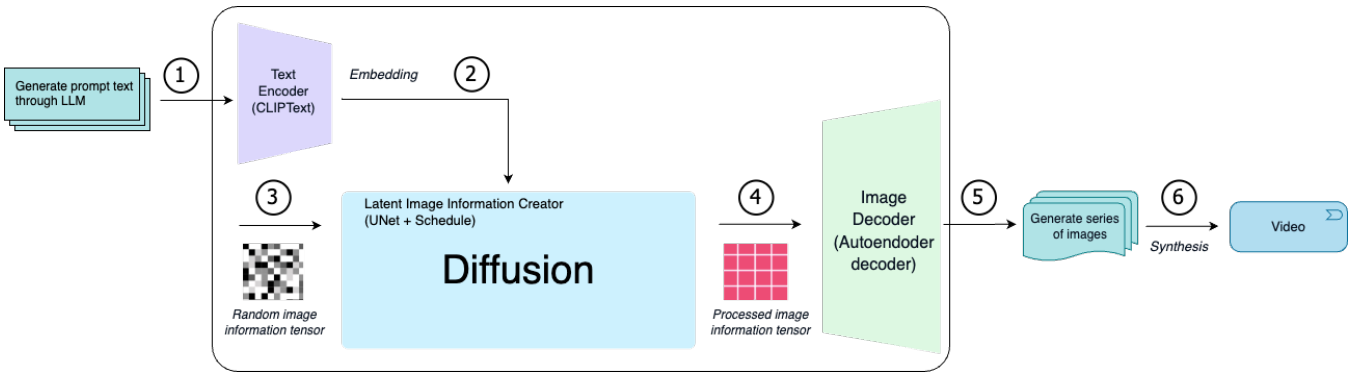

Figure 2. The pipeline of video generation

Implementations. The complete experimental process design is illustrated[1] in Figure 2. Initially, the determination of the desired video topic takes place, followed by the refinement of specific content for a particular scenario. To enhance this refinement process, we leverage the capabilities of modern Large Language Models (LLM)[3]. Creating a synopsis outlining the content of our video is the next crucial step. In this experiment, we expedite the initial topic selection by relying on ChatGPT-3.5, as shown in Figure 3. Upon obtaining detailed descriptions of the scenarios, ChatGPT can then summarize and generate the corresponding prompt words. The assistance of LLM allows us to obtain detailed text contents, which serve as prompts for the subsequent steps.

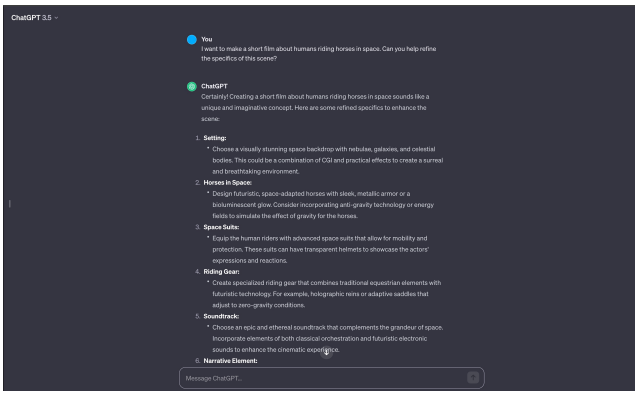

Figure 3. ChatGPT to refine the video scenes

Subsequently, we proceeded to load the pre-trained Stable Diffusion version 1.4 model[6], using the text prompt generated in the earlier step.

This specific prompt led to the generation of a series of images, which served as the foundational source frames for the subsequent video creation. The choice of this particular prompt was driven by its ability to produce more satisfactory results in line with our intended video content. We experimented with various prompt words to generate diverse series of pictures, including uplifting phrases generated by ChatGPT in previous iterations. We selected three sets that demonstrated superior effects for presentation, streamlining the subsequent video synthesis process. First, we load the pre-trained weights of all components of the model. This version can produce images with a resolution of 512x512. We're loading the weights from the half-precision branch fp16 and also tell diffusers to expect the weights in float16 precision.

Figure 4. A photograph of an astronaut riding a horse

In the first group, the prompt words are "A photograph of an astronaut riding a horse". To generate multiple images for the same prompt, we simply use a list with the same prompt repeated several times. We'll send the list to the pipeline instead of the string we used before. At last, it generates a total of 56 pictures arranged in a 7x8 grid. A selection of these pictures is presented in Figure 4.

Figure 5. Many flowers growing in the riverside park

In the second group, the prompt words are "Many flowers growing in the riverside park". At last, it also generates a total of 56 pictures arranged in a 7x8 grid. A selection of these pictures is presented in Figure 5.

In the third group, the prompt words are "A very exciting and fun speed racing game". At last, it also generates a total of 56 pictures arranged in a 7x8 grid. A selection of these pictures is presented in Figure 6.

In the final step of our process, we utilized FFmpeg to convert the sequence of images generated in the previous step into a video, employing the H.264 codec for video encoding. Additionally, we incorporated corresponding background music to enhance realism. To ensure specific video characteristics, we configured the following parameters:

1. Frame Rate: We configured the frame rate of the resulting video to 7 frames per second, thus controlling the pace at which frames appear in the video. With a total of 56 pictures, this setting yields an 8-second short video.

Figure 6. A very exciting and fun speed racing game

2. Resolution: The resolution of the output video was fixed at 512x512 pixels, determining the size and dimensions of the video frame.

3. Video Codec: We specified the video codec to be used as libx264, which is synonymous with the H.264 video codec. This choice ensured efficient video compression and quality.

4. Audio channel: We designated background music as the audio accompaniment for each video, contributing to a more realistic and vivid viewing experience.

As a result of these configurations and processing steps, we obtained a video in which each frame was sourced from the Stable Diffusion model, reflecting the content described by the prompt and generated through the model's capabilities.

3.2 Experimental Results

Throughout our experimentation with the diffusion model, we observed a notable sensitivity of the stable diffusion model to the provided prompt words. After exploring various prompts, we ultimately selected three groups: 'A photograph of an astronaut riding a horse,' 'Many flowers growing in the riverside park,' and 'A very exciting and fun speed racing game.' These groups were used as inputs for our model, generating a series of pictures. Subsequently, these pictures served as raw materials for synthesizing our videos. The final generated video effect is shown in Figure 7. The completion of the entire experimental process confirms the feasibility of pipeline we designed.

Despite generating a series of images from this prompt, we encountered challenges in achieving a smooth and coherent

video. When we got these series of images and synthesized them into an image, we found that the video was not smooth. Even if the same prompt sentence is used, the images generated will vary greatly and be irrelevant to each other. So the final effect of the video did not meet our expectations. The video did not look as natural as the video shot by a human.

Figure 7. The generated short video of speed racing game

4 Discussion and Conclusion

Following our previous experiments, we gained valuable insights into the model's sensitivity to prompt words. The art of designing effective prompts thus becomes a major project worthy of further exploration. Simultaneously, we encountered significant challenges when attempting to generate videos that were smooth and consistent with human perception. Just contemplating the direct synthesis of videos from images generated by stable diffusion may still demand considerable effort and fine-tuning. Considering these challenges, achieving naturalness and smoothness in videos captured by human videographers proves to be a formidable task.

However, we made an interesting discovery: for videos with lower requirements for smoothness and frame rate, such as those resembling comics, it is entirely feasible to generate them in this manner. This is because the audience's primary expectations for such videos lie in compelling storylines narrated through text, and they find it acceptable as long as accompanied by corresponding pictures and text.

Based on these findings, we are also considering a new idea: whether it is possible to directly generate videos using latent diffusion models. This involves the interesting concept of feeding video material directly into a model, allowing it to learn the underlying patterns and dynamics, and then generate the video directly through backward inference. Of course, this entails a series of challenges, such as how to design the network structure, how to formulate the loss, how to train, etc., but they should all be meaningful.

References

[1] J Alammar. 2022. The Illustrated Stable Diffusion. https://jalammar.github.io/illustrated-stable-diffusion/

[2] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems* 33 (2020), 6840–6851.

[3] Yiheng Liu, Tianle Han, Siyuan Ma, Jiayue Zhang, Yuanyuan Yang, Jiaming Tian, Hao He, Antong Li, Mengshen He, Zhengliang Liu, Zihao Wu, Lin Zhao, Dajiang Zhu, Xiang Li, Ning Qiang, Dingang Shen, Tianming Liu, and Bao Ge. 2023. Summary of ChatGPT-Related research and perspective towards the future of large language models. *Meta-Radiology* 1, 2 (Sept. 2023), 100017. https://doi.org/10.1016/j.metrad.2023.100017

[4] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. arXiv:2103.00020 [cs.CV]

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

[6] Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, and Thomas Wolf. 2022. Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers.

[7] Ye Zhu, Yan Yan, Yu Wu, Olga Russakovsky, and Zhiwei Deng. 2023. Boundary Guided Learning-Free Semantic Control with Diffusion Models. (2023).