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 portion of our study has already been completed and documented in our report. Another possibility of generating video directly through DDMs is still in progress.

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, 3 pages. 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

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

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[1][2][4]. 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 [82] 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[1].

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[2].

2.3 CLIP

CLIP (Contrastive Language–Image Pre-training) builds on a large body of work on zero-shot transfer, natural language supervision, and multimodal learning. CLIP can be applied to any visual classification benchmark by simply providing the names of the visual categories to be recognized, similar to the "zero-shot" capabilities of GPT-2 and GPT-3.

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 latter part of our investigation is currently a work in progress, and definitive conclusions may not be reached until the final stages of our research. 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[4].

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.

Implementations. Choose the prompt based on the video content we want to generate. This step can be done by first determining the subject we want to shoot, and then refining the specific content of the photo. These detailed text contents are the prompt words we want. We can also leverage the capabilities of modern large language models to assist in the refinement process. We just need to come up with a synopsis of what our video will be about.

Subsequently, we proceeded to load the pre-trained Stable Diffusion version 1.4 model, using the text prompt generated in the earlier step[3].

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.

In the final step of our process, we employed FFmpeg to transform the sequence of images generated in the previous step into a video. We used the H.264 codec for this video encoding. To ensure specific video characteristics, we configured the following parameters:

1. Frame Rate: We set the frame rate of the resulting video to 10 frames per second, controlling the pace at which the frames appear in the video.

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.

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

During our experimentation with the diffusion model, we discovered that the Stable Diffusion model exhibited a high degree of sensitivity to the hint words provided. At the first stage, we experimented with various prompts, such as "a photograph of flowers are blooming in forest" "a photograph of young people dancing" and "photos of young people dancing", etc. However, the results obtained from these initial prompts did not meet our expectations. Therefore, we decided to study using the official example prompt "Photo of an astronaut riding a horse" as input to the video generation process.

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. Since the images themselves are not coherent, even if the same prompt sentence is used, the images 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.

4 Discussion and Conclusion

In the wake of our previous experiments, we gained valuable insights into the sensitivity of the model to prompt words. Therefore, the art of designing effective prompts emerges as a significant engineering challenge deserving of further exploration. Simultaneously, we encountered substantial challenges when attempting to generate coherent images. Given these challenges, achieving the level of naturalness and smoothness found in videos shot by human videographers proved to be a formidable task.

In light of these findings, we are now planning to investigate an alternative approach: directly generating videos using the latent diffusion model. This involves the intriguing idea of feeding video materials directly into the model, allowing it to learn the underlying patterns and dynamics, and subsequently generating videos through reverse inference. While this endeavor may present significant difficulties, it represents a meaningful and innovative attempt to address the complexities associated with video synthesis.

References

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

[2] 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.

[3] 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.

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