

Transforming Text into Video: A Proposed Methodology for Video Production Using the VQGAN-CLIP Image Generative AI Model

SukChang Lee¹

¹Prof., Dept. of Digital Contents, Konyang Univ., Korea
E-mail 2stonespear@gmail.com

Abstract

With the development of AI technology, there is a growing discussion about Text-to-Image Generative AI. We presented a Generative AI video production method and delineated a methodology for the production of personalized AI-generated videos with the objective of broadening the landscape of the video domain. And we meticulously examined the procedural steps involved in AI-driven video production and directly implemented a video creation approach utilizing the VQGAN-CLIP model. The outcomes produced by the VQGAN-CLIP model exhibited a relatively moderate resolution and frame rate, and predominantly manifested as abstract images. Such characteristics indicated potential applicability in OTT-based video content or the realm of visual arts. It is anticipated that AI-driven video production techniques will see heightened utilization in forthcoming endeavors.

Keywords: Text-to-Image Generative AI, Artificial Intelligence, AI, AI image generator, VQGAN-CLIP

1. Introduction

“In the early 19th century, the inception of the Daguerreotype marked the beginning of video, stemming from scientific curiosity to capture motion and humanity's innate desire to document and preserve reality[1].” “Subsequent to that initial development, as video technology has advanced at a rapid pace, video producers have engaged in extensive deliberations regarding the utilization of technology in video production[2].” “With the advancements in technology, the avenues for an artist's visual expression have broadened[3].”

With recent technological advancements, video production methods based on Generative AI have garnered significant interest. “In the swiftly evolving landscape of video production, the adoption of AI has gained prominence[4].” “Concurrently, artistic endeavors and accomplishments leveraging AI technology have matured in tandem with these groundbreaking technological advancements, establishing AI as a pivotal tool for creative expression among artists[5].” Generative AI facilitates the swift creation of pictorial images without drawing, thereby circumventing copyright concerns. Numerous individuals utilize DALL-E 2 and Midjourney platforms to produce AI images. By merely inputting desired text, these platforms can promptly generate an AI-rendered image. In light of these developments, Text-to-Image Generative AI models are garnering significant interest from both the industrial and academic sectors.

“While Generative AI technology proves efficacious for content creation, its adoption remains limited across various industries[6].” Specifically, its utilization is notably low in sectors like film, OTT, and broadcasting, where video is paramount. This can be attributed to the fact that, although Generative AI is

Manuscript received: August 3, 2023 / revised: August 22, 2023 / accepted: August 29, 2023

Corresponding Author: 2stonespear@gmail.com

Tel:+82-41-730-5332

Professor, Dept. of Digital Contents, Konyang Univ., Korea

Copyright©2023 by The International Promotion Agency of Culture Technology. This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/4.0>)

adept at producing individual images, it encounters technical constraints when tasked with video generation.

This research introduces a method for video production leveraging Text-to-image Generative AI, underscoring the potential of Generative AI technology within the video domain. Within the scope of this study, the image generation procedure and resultant video utilizing the VQGAN-CLIP model in Colab were elucidated. It is the aspiration of this research to champion the adoption of AI as a novel approach to video production, further integrating it into the video creation landscape.

2. Text-to-Image Generative AI and Video

The recent surge in interest in AI technologies has particularly spotlighted methods in Text-to-Image Generative AI. “Notable examples include OpenAI's DALL-E 2, DreamStudio, Meta's Make-A-Scene and Make-A-Video, Google's Imagen and Imagen Video, as well as Stability AI's Stable Diffusion, among others. A myriad of secondary and tertiary AI-based applications and services stemming from these technologies have become pervasive and deeply integrated into our daily routines[7].” Specifically within the domain of visual arts, generative AI represents a revolutionary intersection of human creativity and advanced technology. OpenAI's DALL-E 2, for instance, introduces an innovative capability to produce high-resolution images from mere text descriptions. Similarly, Midjourney enables users to generate images by inputting text descriptions into the command prompt of Discord platform.

However, images generated by DALL-E 2 and Midjourney have limitations when it comes to their application in video production. “Such platforms tend to produce images with a surrealistic bent, and there is an issue with the outputs being noticeably similar[8].” Furthermore, since the produced image is singular, there are inherent challenges in re-purposing it into a video sequence, which typically consists of dozens of interconnected images per second.

Beyond the aforementioned constraints, there is ongoing development in the realm of Text-to-Video technology. The company Runway has introduced Gen-2, a system capable of generating videos solely based on text descriptions. Nevertheless, its application is confined to short-duration videos, and its potential uptake in the video market appears limited due to subpar resolution and image quality.

3. Video Image Creation and Video Production Process Using VQGAN-CLIP

3.1 Subject of Research

This research employs Google Colab in conjunction with the VQGAN-CLIP model. Colab is a cloud-driven platform that allows for the execution of Python code within a web browser, obviating the need for high-performance hardware. The study undertakes video image generation, operating under the assumption of not subscribing to Colab. Within the VQGAN-CLIP model, the VQGAN component is responsible for image generation, while the CLIP component evaluates the congruence of the generated image with the given text. This procedure iteratively refines the image to more closely resemble the described content. In essence, this research employs the VQGAN-CLIP model on Google Colab to investigate images generated based on textual descriptions.

3.2 Video Production Process Using Generative AI

As illustrated in Table 1, the process of video production typically unfolds across three distinct stages: pre-production, production, and post-production. Within the AI-integrated pre-production stage, the first step involves envisaging the desired outcome of the video and subsequently identifying the appropriate VQGAN-CLIP variant that can facilitate the output. While certain models are tailored to replicate existing imagery directly, others enable the fusion of images, simulating motion via techniques like zoom transitions between videos. To achieve the desired results, it becomes essential to explore and test various models of VQGAN-CLIP for their efficiency and alignment with the intended output.

Table 1. Video production process using generative AI

Process	Contents
Pre-production	<ul style="list-style-type: none"> - Stage 1 - Objective Determination - Formulation of Video Imagery Strategy - Selection of VQGAN-CLIP Model
Production	<ul style="list-style-type: none"> - Stage 2 - Production of Video Imagery - Extraction to Image Files and Video Files - Music Composition
Post-production	<ul style="list-style-type: none"> - Stage 3 - Video Post-Processing - Final Output Extraction

During the production stage, traditional tasks include filming and music composition. However, in the context of AI-mediated video creation, filming is not necessary. This is attributed to AI's capacity to generate video imagery directly from textual prompts. In terms of music composition, the advent of AI allows for the direct creation of music tracks, obviating the need for specialized software or external commissioning.

Concluding with the post-production stage, the preliminary imagery generated through VQGAN-CLIP model forms the basis for editing. Any music synthesized during this stage can be incorporated. Video editing can span techniques from image amalgamation to speed modulation. At its core, emphasis should be placed on managing the dynamic video imagery within individual shots, as opposed to stringing together multiple shots to craft a cohesive scene.

4. Designing Video Production with the VQGAN-CLIP Model

4.1 System and Input Value Settings

Video generation using AI is conducted in Colab, leveraging the Python 3 Google Compute Engine backend. When establishing a runtime connection in Colab, concurrent active sessions can lead to computational challenges due to memory constraints. Given that the GPU RAM allocation is capped at 15GB, it is imperative to factor this in when generating videos that demand extensive computation. With a configuration of Display Rate at 5, Batches set to 5, a resolution of 640X384, and Steps fixed at 250, the preparation for modeling consumes 5.2GB of memory.

The foundational image repository can be chosen based on model classification. However, by default, the model is set to utilize the Imagenet_16384. ImageNet stands as a quintessential large-scale dataset, encompassing over a million data entries. As depicted in Figure 1, a multitude of image models are available.

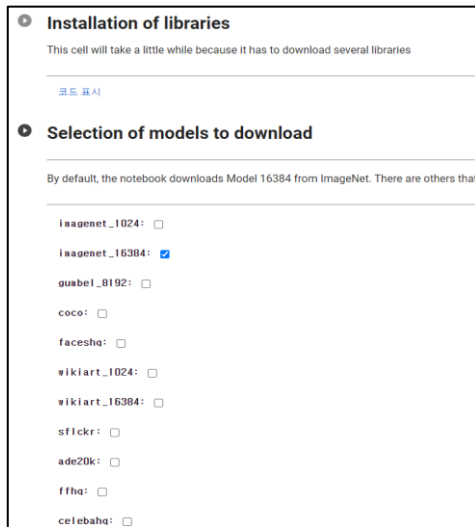


Figure 1. Selection of image models

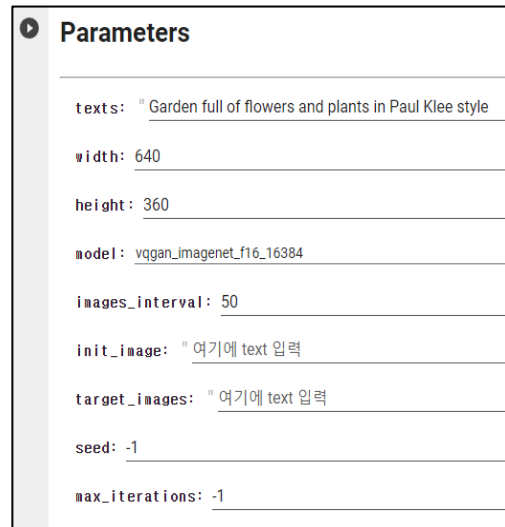


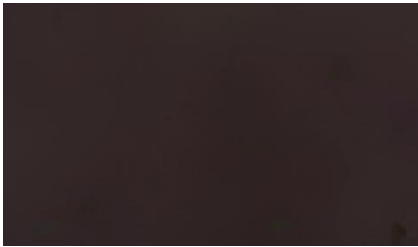

Figure 2. Set text prompt and image size



Users have the liberty to input text prompts based on their preferences. Nevertheless, it is crucial to understand that optimal results are achieved when the envisioned image is distinctly conceptualized and described with precision in VQGAN-CLIP model. Even with identical video creation commands, different executions can yield vastly different outcomes compared to prior attempts. Consequently, to realize the envisaged video, you must exhibit patience and meticulously monitor the randomly generated video. In the context of the VQGAN-CLIP model, video synthesis is conducted with reference to pre-existing images, allowing the incorporation of established artists' styles. As demonstrated in Figure 2, when an artist's name is specified within the text command, their distinctive style becomes evident in the resultant video imagery.

4.2 Video Output

After inputting the text prompt, the output is generated. The number of images showcased on the screen fluctuates based on the 'Images_Interval' value as depicted in Figure 2. By default, this is set at 50. Reducing this value diminishes the rate of transitions between images. Consequently, there is minimal discernible variance between consecutive images, necessitating the input of a suitable value. Upon the completion of video generation, the video is available for download.

Table 2. Changes in video imagery output

Process	Video Imagery	Process	Video Imagery
Stage 1		Stage 2	
Description	In the starting frame, the dark backdrop reveals nothing. However, as time progresses, variegated patterns emerge sporadically across the surface,	Description	The image undergoes transformations resonating with the content described in the text prompt.

	forming images in the video.		
Stage 3		Stage 4	
Description	The image's metamorphosis unfolds as though cellular division is taking place ubiquitously within the entire image. This culminates in the final video result.	Description	Each image in the video goes through a stabilization phase after embodying its distinct transformation. Throughout the video, there isn't a moment of complete stillness; subtle movements, akin to gentle waves, can be observed.

The resultant video imagery, as delineated in Table 2, showcases the transformation of objects within various spaces. The text prompt drew inspiration from the artistic style of Paul Klee yielding an outcome reminiscent of Klee's aesthetic. While it does not replicate the artist's work exactly, the output certainly evokes a visual artistry aligned with a similar stylistic essence.

4.3 Proposal for a Production Methodology of Video Imagery Outcomes Using Generative AI

The Generative Adversarial Network (GAN) model employs a technique to generate synthetic images that closely resemble authentic ones. To convert a user's text prompt into an image by referencing an image model, the existing image undergoes continuous modifications until the desired result is attained. This procedure manifests in the initial stages, specifically stage 1 and 2, as indicated in the aforementioned Table 2, where the video images appear to develop in a stippled or mottled manner. This progressive evolution of digital images is evident as they endeavor to emulate reference images.

Owing to the technical characteristics of GAN, they are more apt for generating abstract video images rather than photorealistic depictions. Therefore, in this research, the VQGAN-GAN model, when tasked with producing abstract results, such as images of houses and flowers, as illustrated in Figure 3, unveils a visual allure reminiscent of landscape paintings. Conversely, due to certain technical constraints, the model struggles to render highly realistic video images. For instance, attempting to generate a photorealistic portrayal of a woman results in an abstract purple representation, as demonstrated in Figure 4.



Figure 3. House image by VQGAN-CLIP



Figure 4. Girl image by VQGAN-CLIP
(Source: Reddit Reddit The Lady of Shalott)

Furthermore, the finalized video displays images emerging and transitioning against a dark backdrop. When undergoing the editing process for the final video, the evolution of the video imagery can be manipulated by adjusting the playback speed, either slowing it down or speeding it up, to achieve the desired effect.

5. Conclusion

This research assessed the current application of Text-to-Image Generative AI in the realm of video production and suggested a method for producing videos using it, aiming to diversify production approaches within the field. The principal attributes of the proposed video production technique identified through this investigation are as follows. Firstly, videos produced via Text-to-Image Generative AI lack superior resolution and frame rate. Consequently, given the relatively modest video quality of the results, it might be more suitable for OTT-based web series or individual YouTube channel content. In light of this, determining the maximum image size achievable using the VQGAN-CLIP model when extracting images is crucial. Secondly, outputs generated using the VQGAN-CLIP model tend to be abstract rather than lifelike. As images are produced in VQGAN-CLIP, they undergo creation and transformation, with the final image arising from a process of emulating tangible entities. Hence, the application of abstract video imagery offers potential for directorial strategies that showcase the temporal evolution of individual objects within video frames. Lastly, given the aforementioned attributes of Text-to-Image Generative AI, video images produced by the VQGAN-CLIP model can be aptly incorporated in cinematic or broadcast scenarios that portray fantastical timeframes and spatial settings, or alternatively, be deployed as multimedia video projects in the domain of visual arts. In light of the AI-driven video production techniques and features proposed herein, it is anticipated that the utilization of Text-to-Image Generative AI in video production will witness enhanced momentum in the forthcoming times.

Reference

- [1] Kwon, Y. Kim, C. Park, S. Shin, J. Kim, G. Philippe, J. Lee, S. Kwon, and S. Lee, "Web Server based Hologram Image Production Pipeline System Implementation," *The Journal of the Convergence on Culture Technology (JCCT)*, Vol. 7, No. 4, p. 752, 2021. DOI: 10.17703/JCCT.2021.7.4.751
- [2] J. Ahn and S. Jeong, "A Content Analysis of Visual Production Techniques Across Different Screen Sizes," *Korean Journal of Journalism and Communication Studies*, Vol. 66, No. 4, pp. 189-190, 2022. DOI: 10.20879/kjcs.2022.66.4.006
- [3] S. Lee and W. Choi, "Cinema Robotics and Digital Mise-en-Scène," *Film Studies*, No. 90, pp. 237-238, 2021. DOI: 10.17947/FS.2021.12.90.225
- [4] B. Jeon, "Artificial Intelligence and Film & Video Production: Through the Cases of the Use of Artificial Intelligence as a Film & Video Production Tool," *Preview*, Vol. 20, No. 1, pp. 134-136, 2023. DOI: 10.23120/kadmi.2023.20.1.006
- [5] H. Park, "A Case Study On Application Of Text To Image Generator AI DALL·E," *The Treatise on The Plastic Media*, Vol. 26, No. 1, p. 104, Feb 2023. DOI: 10.35280/KOTPM.2023.26.1.11
- [6] J. Son, M. Han, and S. Kim, "Artificial Intelligence-Based Video Content Generation," *Electronics and Telecommunications Trends*, Vol. 34, No. 3, pp. 34-35, Jun 2019. DOI: 10.22648/ETRI.2019.J.340304
- [7] N. Yoon, "On art and artworks accompanied with AI art generation: Focusing on the concept of 'triviality' and 'non-triviality'," *Human Contents*, No. 38, p. 38, Mar 2023. DOI: 10.18658/humancon.2023.03.37
- [8] K. Kim and H. Kim, "A case study of ChatGPT and Midjourney-Exploring the possibility of use for art and creation using AI-," *The Treatise on The Plastic Media*, Vol. 26, No. 2, pp. 3-5, May 2023. DOI: 10.35280/KOTPM.2023.26.2.1