Dynamic Narratives: Content Creation with Automated Text-to-Video Technology

Rashaz Rafeeque
Apex Institute of Technology (CSE)
Chandigarh University
Punjab, India
rashazrafeeque@gmail.com

Jeevan A J
Apex Institute of Technology (CSE)
Chandigarh University
Punjab, India
jeevanaj2003@gmail.com

Rhishitha T S
Apex Institute of Technology (CSE)
Chandigarh University
Punjab, India
rhishithats002@gmail.com

Merry K P
Apex Institute of Technology (CSE)
Chandigarh University
Punjab, India
merrypaulose@gmail.com

Abstract— Traditional video production often requires significant time, resources, and expertise, limiting the ability of individuals and organizations to leverage video's powerful communication potential. This study addresses this challenge by proposing a novel, automated text-to-video conversion system. This proposed system harnesses the power of cutting-edge machine learning to bridge the gap between text and video formats. It utilizes OpenAI's GPT-3 model for analyzing and expanding on user-provided topics, generating informative text that forms the foundation for the video. Additionally, Google's Text-to-Speech API converts the generated text into human-sounding narration, while MoviePy, a Python video editing library, merges the text, narration, and AI-generated images into engaging video clips. By democratizing video creation, this system empowers anyone to create high-quality video content, fostering a more inclusive multimedia experience.

Keywords—OpenAI, Natural Language Processing, Diffusion Pipeline, GAN, Video Summarization.

I. INTRODUCTION

The digital age has seen an increase in multimedia consumption, with video material emerging as the dominant force. Traditional video creation, on the other hand, sometimes requires a large amount of time, resources, and 1emocratizi video production expertise. This poses a hurdle for individuals and organisations looking to 1emocratiz on the power of video communication, limiting their capacity to successfully engage audiences. To solve this issue, this study presents a new automated text-to-video conversion system that intends to 1emocratizi video creation and transform content production workflows. The research explores the potential of cutting-edge machine learning algorithms to bridge the gap between text and video formats. Our proposed system leverages the power of OpenAI's Generative Pre-Trained Transformer 3 (GPT-3) model, a state-of-the-art text generation tool. By integrating with the OpenAI API, the system utilizes GPT-3's capabilities to analyze and elaborate upon user-provided topics, generating new and informative text content. This generated text serves as the foundation for subsequent video creation.

The system (Fig 1) works perfectly with Google's Text-to-Speech (gTTS) API. This integration enables the system to turn the generated text into human-sounding audio narration, providing an important element to the video presentation. The study investigates the interactions between these sophisticated APIs and MoviePy, a Python video editing toolkit. MoviePy is essential for merging produced text, voice narration, and AI-powered imagery to create cohesive and interesting video clips. The research is more than just a technological inquiry. It envisions a future in which anybody, regardless of technical ability or finances, may easily generate high-quality video content. Using automatic text-to-video conversion, you may create educational resources, commercial presentations, and even 1emocratizin video greetings. This not only improves information accessibility, but also allows content makers to engage viewers with dynamic and visually appealing narrative. By democratizing video creation, this study lays the path for a more inclusive and engaging multimedia experience for both content providers and viewers.

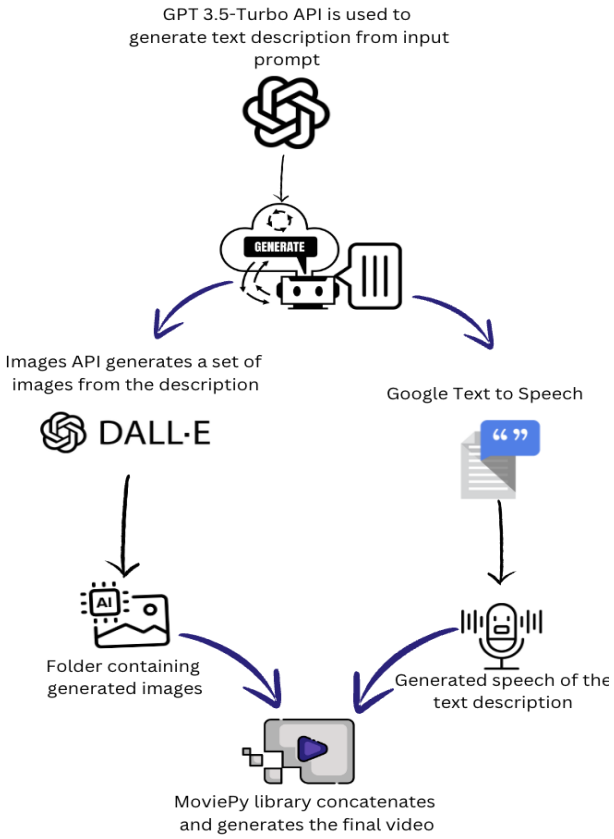

Fig 1: Data Flow Diagram

II. LITERATURE SURVEY

1. OPEN AI:

OpenAI, specifically its GPT-3 model, plays a vital role in the automated text-to-video conversion system. Its powerful text generation capabilities provide the foundation for our video content. By feeding the text input into GPT-3, we can leverage its ability to create informative and potentially creative content. This generated content serves as the initial script for our video, outlining the key points and narrative structure. This approach allows our system to move beyond simple video summarization towards generating entirely new video content based on user-provided text. The quality and creativity of the GPT-3 output directly influence the overall effectiveness of our text-to-video conversion system.[20][21]

2. STORYBOARD:

Our current system excels at direct text-to-video conversion, but research on storyboard generation offers an exciting future direction. By leveraging NLP, we could automatically create storyboards from GPT-3's output. These storyboards would visually represent key scenes and transitions, acting as a roadmap for the video. This benefits us in two ways: internally, storyboards can guide video generation, leading to a more cohesive final product. Externally, users could review and provide feedback on the storyboard before video creation, fostering a collaborative experience. Storyboards also enhance comprehension for viewers. Even non-technical users can grasp the narrative flow from the visuals, making it ideal for educational or explainer videos. Storyboards also enhance comprehension for viewers. Even non-technical users can grasp the narrative flow from the visuals, making it ideal for educational or explainer videos.[5][12][20][21]

3. NATURAL LANGUAGE PROCESSING (NLP):

Natural Language Processing (NLP) plays a critical, yet often unseen, role in our text-to-video conversion system. Imagine a skilled translator, but instead of languages, NLP bridges the gap between human text and a format our system understands. It delves into the text's intricacies, extracting the essence and transforming it into meaningful video components.

This magic unfolds in two key steps. First, NLP acts like a meticulous architect, meticulously dissecting the text. It breaks it down into logical units like sentences or paragraphs, ensuring the video narrative has a clear structure and flows naturally. Second, NLP transforms into a detective, identifying key ideas and entities mentioned within the text. Think of it like highlighting the crucial characters and plot points in a story.By grasping these nuances, the system creates videos that resonate with viewers, evoking the desired emotions and fostering a more engaging experience. Ultimately, NLP empowers our system to move beyond simple keyword searches, allowing it to capture the heart of the text input. This leads to the creation of impactful and meaningful video content that transcends a simple conversion – it becomes a true storytelling experience.[2][3][4][7]

4. GENERATIVE ADVERSARIAL NETWORK(GAN):

GAN networks are a type of deep learning model with potential applications in automated video generation. Imagine two competing AI models: one (generator) creates video frames based on the text input, and the other (discriminator) tries to distinguish these generated frames from real videos. Through this competition, the generator progressively improves its ability to create realistic and visually compelling video content. While not currently implemented in our system, exploring GANs could lead to more visually rich and dynamic video outputs in the future. By harnessing the power of GANs, we could pave the way for more immersive and interactive multimedia experiences, revolutionizing how we consume and create video content. [8][9]

5. DIFFUSION PIPELINE:

Diffusion models are a relatively new deep learning approach for image and video generation. Unlike GANs, these models begin with noise and gradually refine it into a coherent video based on the text input. Imagine adding noise to a blank canvas and progressively transforming it into a video scene guided by the textual information. This offers smoother transitions and potentially more control over the creative process compared to GANs. Exploring diffusion pipelines could be an exciting future direction for our video generation system.[6][8][17][18][19][22]

6. ENCODER-DECODER:

The encoder-decoder architecture is a fundamental concept in many deep learning tasks, including text-to-video conversion. The encoder component acts like a compressor, processing the text input (encoded) and extracting relevant information. This compressed representation is then fed to the decoder, which utilizes it to generate the corresponding video output (decoded). While our system might not explicitly use a separate encoder-decoder model, similar principles are likely applied within the chosen deep learning framework (e.g., GPT-3).[1][2][6][12]

7. VIDEO SUMMARIZATION:

Although video summarization research focuses on condensing existing videos, it holds valuable lessons for our text-to-video system. Analyzing how these methods identify key points and create concise summaries can inform our approach to video generation. By ensuring our generated videos are engaging and avoid information overload, we can keep viewers captivated and maximize content impact.[11][12]

8. TEXT-TO-SPEECH TECHNOLOGY:

Text-to-Speech (TTS) technology plays a crucial role in adding audio narration to our automated video creation process. By integrating a TTS API like gTTS, we can convert the text script generated by GPT-3 into a natural-sounding voiceover. This voiceover adds another layer of engagement and accessibility to the final video. With both audio and visual elements, the generated video becomes more impactful and resonates better with viewers.[20][21]

III. METHODOLOGY

The paper provides a novel method (Fig 2) for automatically turning written content into entertaining video presentations. The solution bridges the gap between text and video formats by utilising OpenAI's Generative Pre trained Transformer 3 (GPT-3) model and other Artificial Intelligence (AI) functionalities. In the initial stage, the system accepts a user-specified topic as input. It then uses the OpenAI API to access GPT-3's sophisticated text creation capabilities (particularly the "gpt-3.5-turbo-0125" model). This model analyses the specified topic and creates a new text document that expands on it. The second stage focuses on converting the generated text to a video format. For each paragraph, the system uses OpenAI's Image API to generate a meaningful image that visually compliments the text. Next, it uses the Google Text-to-Speech (gTTS) API to generate an audio narration of the paragraph, which includes a human-like voice reading the text. Finally, the system uses MoviePy, a Python package for video editing. MoviePy smoothly blends the created image, audio narration, and the paragraph's text overlay to create a complete video clip. This technique is repeated for each paragraph, yielding a series of discrete video segments. In the last stage, the technology combines the different video clips into a single, comprehensive video file. It accomplishes this by using MoviePy's concatenation functionality to combine the clips in a sequential manner. The final film efficiently communicates the original textual content in a visually appealing and educational manner. This completed video can then be viewed, saved, or shared as needed.

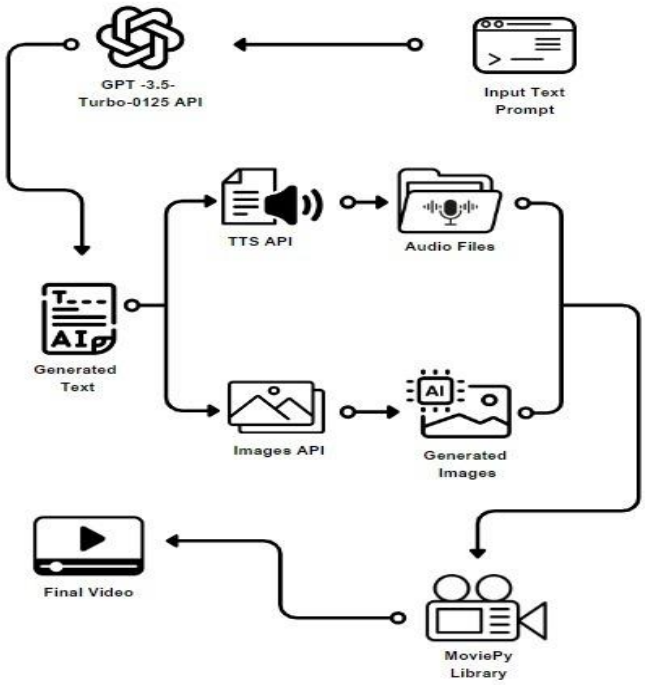

Fig 2: Methodology

The user provides a topic as input. The system leverages the OpenAI API to access the GPT-3 (gpt-3.5-turbo-0125 model) for text generation. Based on the topic, GPT-3 generates a new text document elaborating on it. The generated text is saved in a designated file. For video generation the system reads the generated text content from the stored file. Regular expressions are used to divide the text into individual paragraphs. Separate folders are created for audio, images, and videos to organize the generated media files. It processes each paragraph sequentially.

OpenAI's Image API is used to generate an image that visually represents the current paragraph. Google Text-to-Speech (gTTS) creates an audio narration for the paragraph with a human-like voice.MoviePy combines the generated image, audio narration, and a text overlay into a single video clip for each paragraph. All the individual video clips are gathered. MoviePy's concatenation functionality merges the clips sequentially into a single comprehensive video file.

IV. RESULT

This research tackles the challenge of time-consuming and expensive traditional video creation by proposing a novel automated text-to-video system. Designed to democratize video production, the system empowers anyone, regardless of technical expertise, to create high-quality content.Here's how it works: users simply provide a topic. The system then leverages the cutting-edge OpenAI's GPT-3.5-Turbo model, a powerful text generation tool, to craft a detailed and informative script based on the given topic. But GPT-3.5-Turbo doesn't stop there. It analyzes the script further, suggesting scene breaks and optimizing phrasing to ensure clear segmentation – a crucial step for creating a cohesive video narrative.

To bring the script to life visually, the system integrates with the OpenAI Images API. This powerful tool generates unique images that correspond to the content of each paragraph within the script. These AI-generated visuals serve as a springboard for storyboarding the video, providing a foundation for a visually engaging experience. Next, Google Text-to-Speech comes into play. This API transforms the segmented script into a natural-sounding, human-like voiceover narration, adding another layer of richness to the video.

Finally, MoviePy, a Python video editing toolkit, takes center stage. It seamlessly combines the generated images, the narrated audio, and text overlays for each paragraph into individual video clips. These clips are then meticulously assembled by MoviePy into a single, comprehensive video file.

The impact of this automated system is multifaceted. It empowers anyone, from educators creating engaging learning materials to businesses crafting compelling presentations, to easily produce high-quality video content. This not only improves information accessibility but also allows creators to leverage dynamic visuals to capture and retain the attention of their audience. By democratizing video creation, this research paves the way for a more inclusive and engaging multimedia experience for both content creators and viewers alike.

Prompt 1: Tsunami (in Fig 3)

Generated Text : A tsunami is a massive ocean wave caused by underwater disturbances like earthquakes, volcanic eruptions, or landslides. It can unleash devastating destruction when it reaches coastal areas.

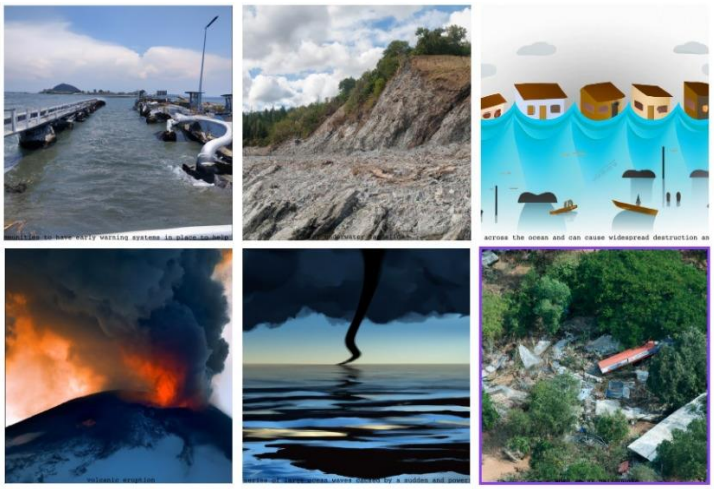

Fig 3: Tsunami

Prompt 2: World War (in Fig 4)

Generated Text : A world war is a global conflict involving multiple nations, marked by widespread military mobilization, significant casualties, and far-reaching geopolitical consequences.

Fig 4: World War

Prompt 3: Solar System (in Fig 5)

Generated Text : The solar system consists of the Sun, eight planets, moons, and various smaller celestial bodies. It provides insights into planetary formation, space exploration, and the potential for extraterrestrial life.

Fig 5: Solar System

The polarity and subjectivity of the generated texts are calculated and visualized in the form of line graph. Together, subjectivity and polarity (in Table 1) enhance the emotional resonance and engagement of the generated videos.

Prompt	Polarity	Subjectivity
Cinderella Story	0.9965	0.9841
AI	0.9042	0.4927
Winter	0.7003	0.5951
Tsunami	-0.836	0.5771
World War	-0.6369	0.9034
Solar System	0.7906	0.7500

Table 1: Polarity vs Subjectivity of Generated Text

In text-to-video generation, polarity (Fig 6) assesses the sentiment as positive, negative, or neutral, guiding decisions on visual elements and music to match the overall emotional stance of the text.

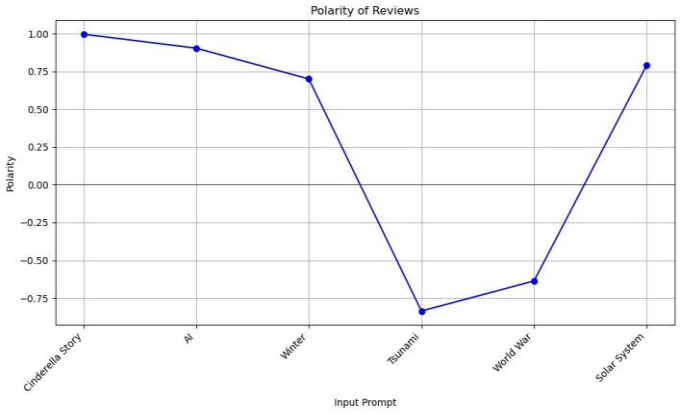

Fig 6: Polarity of Generated Text

In text-to-video generation, subjectivity refers to the expression of personal opinions and emotions, influencing the narrative style and emotional tone of the video.

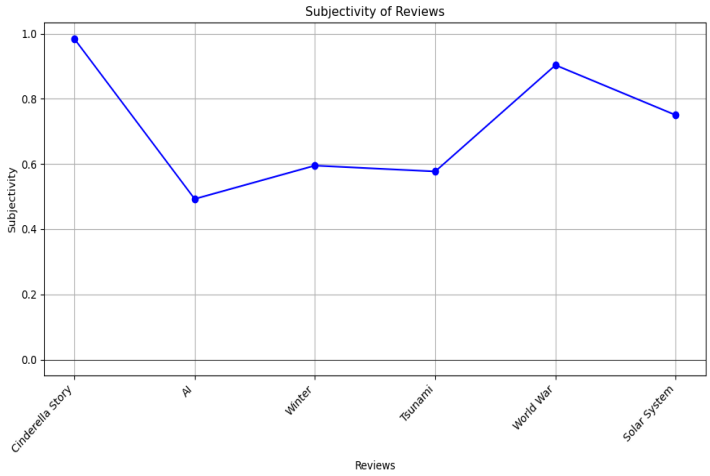

Fig 7: Subjectivity of Generated Text

V. CONCLUSION

This research presented a novel approach to democratizing video creation through an optimized text-to-video AI model. By focusing on efficiency, reliability, and user-friendliness, the system empowers a broader range of individuals and organizations to leverage the power of video communication.Developing a text-to-video model requires careful consideration of both efficiency and reliability. The project prioritized using state-of-the-art AI tools like OpenAI's GPT-3.5-Turbo and the OpenAI Images API for script generation and image creation respectively. These tools significantly streamline the video production process, allowing users to create videos in a shorter timeframe. Furthermore, the use of Google Text-to-Speech ensures consistent and high-quality narration for the generated videos.

The project acknowledges that the current version focuses primarily on efficiency and offers limited creative control for users. While GPT-3.5-Turbo provides scriptwriting assistance and the OpenAI Images API offers a foundation for storyboarding, these features could be further developed to allow finer control over creative vision. Future iterations might explore integrating user-defined parameters or artistic styles into the script generation and image creation processes. Additionally, exploring AI-powered video editing capabilities could provide users with more granular control over the final video product.By making video creation more accessible, the system fosters a more inclusive multimedia experience. Educational institutions and businesses can create engaging content without requiring extensive video editing expertise. Individuals can explore video communication for personal use or small-scale projects. This not only empowers creators but also broadens the scope of information dissemination and communication possibilities.

In conclusion, the research lays the groundwork for an effective and user-friendly text-to-video AI system. With ongoing refinement to empower user control and unlock advanced editing capabilities, this technology holds promise to revolutionize video creation for broader audiences, enriching the multimedia landscape with dynamic content and accessibility. As it continues to evolve, this innovative system stands poised to bridge the gap between text and video, ushering in a new era of creativity and engagement in multimedia production.

REFERENCE

[1] Make-A-Video: Text-to-Video Generation without Text-Video Data. Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, Yaniv Taigman

[2] VideoPoet: A Large Language Model for Zero-Shot Video Generation. Dan Kondratyuk, Lijun Yu, Xiuye Gu, José Lezama, Jonathan Huang, Rachel Hornung, Hartwig Adam, Hassan Akbari, Yair Alon, Vighnesh Birodkar, Yong Cheng, Ming-Chang Chiu, Josh Dillon, Irfan Essa.

[3] VATT: Transformers for Multimodal Self-Supervised Learning from Raw Video, Audio and Text. NeurIPS 2021 · Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, Boqing Gong.

[4] VideoCLIP: Contrastive Pre-training for Zero-shot Video-Text Understanding. EMNLP 2021 · Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, Christoph Feichtenhofer ·

[5] SparseCtrl: Adding Sparse Controls to Text-to-Video Diffusion Models. 28 Nov 2023 · Yuwei Guo, Ceyuan Yang, Anyi Rao, Maneesh Agrawala, Dahua Lin, Bo Dai.

[6] A Recipe for Scaling up Text-to-Video Generation with Text-free Videos. 25 Dec 2023 · Xiang Wang, Shiwei Zhang, Hangjie Yuan, Zhiwu Qing, Biao Gong, Yingya Zhang, Yujun Shen, Changxin Gao, Nong Sang ·

[7] Text2Video-Zero: Text-to-Image Diffusion Models are Zero-Shot Video Generators. ICCV 2023 · Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, Humphrey Shi ·

[8] Kim, Doyeon, Donggyu Joo, and Junmo Kim. "Tivgan: Text to image to video generation with step-by-step evolutionary generator." IEEE Access 8 (2020): 153113-153122.

[9] Zhao, R., Gu, Y., Wu, J.Z., Zhang, D.J., Liu, J., Wu, W., Keppo, J. and Shou, M.Z., 2023. Motiondirector: Motion customization of text-to-video diffusion models. arXiv preprint arXiv:2310.08465.

[10] Lin, Xudong, Gedas Bertasius, Jue Wang, Shih-Fu Chang, Devi Parikh, and Lorenzo Torresani. "Vx2text: End-to-end learning of video-based text generation from multimodal inputs." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7005-7015. 2021.

[11] Abohwo, J. (2023). Regis: refining generated videos via iterative stylistic redesigning.

[12] Hu, Yaosi, Chong Luo, and Zhenzhong Chen. "Make it move: controllable image-to-video generation with text descriptions." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 18219-18228. 2022.

[13] Fu, Tsu-Jui, Licheng Yu, Ning Zhang, Cheng-Yang Fu, Jong-Chyi Su, William Yang Wang, and Sean Bell. "Tell me what happened: Unifying text-guided video completion via multimodal masked video generation." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10681-10692. 2023.

[14] Kim, Taehoon, ChanHee Kang, JaeHyuk Park, Daun Jeong, ChangHee Yang, Suk-Ju Kang, and Kyeongbo Kong. "Human Motion Aware Text-to-Video Generation with Explicit Camera Control Supplementary Materials."

[15] Xin Yuan, Jinoo Baek, Keyang Xu, Omer Tov, Hongliang Fei; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops, 2024, pp. 489-496.

[16] “ModelScope Text-to-Video Technical Report”. Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, Shiwei Zhang

[17] Lee, SukChang. "Transforming Text into Video: A Proposed Methodology for Video Production Using the VQGAN-CLIP Image Generative AI Model." International Journal of Advanced Culture Technology 11, no. 3 (2023): 225-230.

[18] “Tune-A-Video: One-Shot Tuning of Image Diffusion Models for Text-to-Video Generation”. Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, Mike Zheng Shou; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 7623-7633.

[19] Wang, W., Yang, H., Tuo, Z., He, H., Zhu, J., Fu, J. and Liu, J., 2023. VideoFactory: Swap Attention in Spatiotemporal Diffusions for Text-to-Video Generation. arXiv preprint arXiv:2305.10874.

[20] https://techcommunity.microsoft.com/t5/analytics-on-azure-blog/transforming-text-to-video-harnessing-the-power-of-azure-open-ai/ba-p/3837389

[21] https://techcommunity.microsoft.com/t5/analytics-on-azure-blog/transforming-text-to-video-azure-open-ai-cognitive-services-and/ba-p/3904631