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

Project Work Synopsis

Submitted in the partial fulfilment for the award of the degree of

BACHELOR OF ENGINEERING
IN
COMPUTER SCIENCE WITH SPECIALIZATION IN
ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

Submitted by:
21BCS6634 RASHAZ RAFEEQUE
21BCS6589 JEEVAN A.J
21BCS6272 RHISHITHA T.S

Under the Supervision of:

Prof. MERRY K P

CHANDIGARH UNIVERSITY, GHARUAN, MOHALI - 140413, PUNJAB

Feb, 2024

Abstract

Our groundbreaking synthetic video generation concept aims to elevate the art of storytelling by seamlessly translating textual explanations into visually captivating films through a harmonious fusion of language and graphics. At the heart of our system lies an integration of cutting-edge AI methodologies, with a particular emphasis on Conv3D network models, strategically employed to dynamically transmute textual content into dynamic and engaging video sequences. Through the judicious use of deep learning and multimodal processing, our solution excels in authentically conveying the substance of textual descriptions across a spectrum of scenarios.

Our research delves into the intricacies of verbal comprehension, scene construction, and temporal coherence within the realm of video production. By leveraging Conv3D networks, recurrent neural networks, and attention processes in tandem, we present a unique and robust text-to-video synthesis method that goes beyond mere translation. Our demonstrated capabilities underscore our system's proficiency in creating captivating visual narratives from textual inputs, offering a sophisticated and versatile tool for multimedia content creation. This research represents a significant stride towards closing the gap between language and visual representation in the domain of synthetic video generation.

Keywords: Conv3D, Natural Language Processing (NLP), Artificial Intelligence, Deep Learning, Generative Adversarial Network, Attention Model, Machine Learning.

Table of Contents

Title Page	i
Abstract	ii
1. Introduction	1
1.1 Problem Definition	
1.2 Project Overview	
1.3 Hardware Specification	
1.4 Software Specification	
2. Literature Survey	2-7
2.1 Literature Review Summary	
2.2 Existing System	
2.3 Problem Formulation	
2.4 Proposed System	
3. Research Objectives	7
4. Methodologies	7
5. Experimental Setup	8
6. Conclusion	9
7. Reference	10

1. INTRODUCTION

1.1 Problem Definition

In the era of multimedia content consumption, there exists a growing demand for efficient methods to transform textual information into engaging video content. Traditional methods of video creation often require significant time, resources, and expertise in video production, limiting accessibility to individuals and organizations with specialized skills and resources. As a result, there is a pressing need for automated solutions that streamline the process of text-to-video generation, enabling users to efficiently convert textual content into visually compelling video presentations. This encompasses challenges such as natural language understanding, audiovisual synchronization, content selection, visual representation, and user customization. The goal is to develop automated text-to-video generation systems that are user-friendly, scalable, and capable of producing high-quality video content tailored to diverse applications, ranging from educational videos and marketing materials to personalized multimedia presentations. Addressing these challenges will not only democratize video creation but also empower content creators to efficiently communicate ideas, engage audiences, and enhance the accessibility and effectiveness of textual information through dynamic visual storytelling.

1.2 Problem Overview

The automated text-to-video project uses cutting-edge machine learning algorithms to smoothly convert textual input into visually appealing video content, with the goal of revolutionising content creation. This project aims to create a strong model that can comprehend and analyse text in a variety of formats, including articles, scripts, and social media postings, and then translate it into dynamic visual narratives by utilising natural language processing and computer vision technology. The model aims to generate high-quality videos with captivating images, coherent storytelling, and synchronised audio elements through rigorous training and optimisation. The ultimate objective is to give consumers a simple and effective tool for producing interesting video content on a large scale, suitable for a variety of uses, such as social media posts, marketing campaigns, instructional materials, and customised video communications. This project intends to democratise content creation by automating the text-to-video process. This will enable users to convey their thoughts and stories in a visually impactful way, independent of their technological expertise or resources.

1.3 Hardware Specification

- 11th Gen Intel® i7-11800H @ 2.30GHz
- 16 GB RAM. 256GB SSD 1TB HDD

1.4 Software Specification

- GitHub
- Python
- TQDM
- Imageio
- Peft
- Torchsummary

2. LITERATURE SURVEY

2.1 Literature Review Summary

Year and Citation	Article/ Author	Technique	Source	Evaluation Parameter
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	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	Spatial-temporal U-Net with attention mechanism, video decoder.	arxiv.org	FID (Fréchet Inception Distance), visual quality, diversity, and aesthetic richness of generated videos.
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, Agrim Gupta, Meera Hahn, Anja Hauth, David Hendon, Alonso Martinez, David Minnen, David Ross.	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, Agrim Gupta, Meera Hahn, Anja Hauth, David Hendon, Alonso Martinez, David Minnen, David Ross·	Decoder-only transformer architecture	arxiv.org	Fidelity of generated videos, motion quality, performance on multi-task video creation and editing.
Abohwo, J. (2023). Regis: refining generated videos via iterative stylistic redesigning.	Abohwo, J.	Neural network integrates with existing T2V models for refinement.	arxiv.org	FID, Frechet Video Distance (FVD), visual quality, removal of artifacts and noise.

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.	Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, Boqing Gong	Video-Audio-Text Transformer (VATT)	arxiv.org	Accuracy on downstream tasks like image classification, video action recognition, audio event detection.
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 ·	Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, Christoph Feichtenhofer ·	Transformer model using contrastive learning.	arxiv.org	Accuracy on zero-shot video and text understanding tasks.
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 ·	Yuwei Guo, Ceyuan Yang, Anyi Rao, Maneesh Agrawala, Dahua Lin, Bo Dai ·	T2V diffusion model with a condition encoder	arxiv.org	Visual quality, semantic composition, controllability, applicability to various video-generation tasks.
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 ·	Xiang Wang, Shiwei Zhang, Hangjie Yuan, Zhiwu Qing, Biao Gong, Yingya Zhang, Yujun Shen, Changxin Gao, Nong Sang ·	TF-T2V framework using diffusion models.	arxiv.org	FID, FVD, visual quality, controllability, scalability.
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 ·	Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, Humphrey Shi ·	Text-to-image diffusion models like Stable Diffusion	arxiv.org	Visual quality, coherence, consistency, comparison with current text-to-video techniques
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.	Kim, Doyeon, Donggyu Joo, and Junmo Kim.	Incremental learning, Text-to-Image-to-Video GAN (TiVGAN)	IEEE	Video generation complexity, Lack of text-to-video research

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.	Lee, SukChang	VQGAN-CLIP model	KoreaScience	Modest video quality, Abstract outputs, Applicability in OTT, Cinematic, and Broadcast scenarios
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.	Lin, Xudong, Gedas Bertasius, Jue Wang, Shih-Fu Chang, Devi Parikh, and Lorenzo Torresani.	End-to-end trainable framework, Modality-specific classifiers, Generative text decoder	IEEE	Information extraction, Effective cue combination, Human-comprehensible text generation, State-of-the-art results in captioning, QA, and dialog tasks
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.	Hu, Yaosi, Chong Luo, and Zhenzhong Chen.	VQ-VAE encoder-decoder architecture, spatially aligned Motion Anchor (MA)	IEEE	Text-Image-to-Video (TI2V) generation task, Controllable and diverse video generation, Modified Double Moving MNIST, CATER-GEN datasets
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.	Fu, Tsu-Jui, Licheng Yu, Ning Zhang, Cheng-Yang Fu, Jong-Chyi Su, William Yang Wang, and Sean Bell.	Temporal-aware VQGAN, Multimodal Masking	IEEE	Text-guided video completion (TVC) task, Improved video quality, Diverse scenario evaluation (Kitchen, Flintstones, MUGEN), Comparison against previous methods (UCF-101, BAIR datasets)
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."	Kim, Taehoon, ChanHee Kang, JaeHyuk Park, Daun Jeong, ChangHee Yang, Suk-Ju Kang, and Kyeongbo Kong.	-	thecvf.org	Survey Ratings, Correct Predictions.

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. "Inflation With Diffusion: Efficient Temporal Adaptation for Text-to-Video Super-Resolution"	Xin Yuan, Jinoo Baek, Keyang Xu, Omer Tov, Hongliang Fei	Diffusion modal, Super-resolution, UNet, DDIM	thecvf.com	Peak Signal-to-Noise Ratio (PSNR), Similarity Index Measure (SSIM), Temporal coherence.
[Submitted on 12 Aug 2023] "ModelScope Text-to-Video Technical Report". Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, Shiwei Zhang	Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, Shiwei Zhang	Text-to-video synthesis, Diffusion models, Spatio-temporal blocks, VQGAN, Transformer-based text encoder, Denoising U-Net	arxiv.org	Fréchet Inception Distance (FID), Precision-Recall (Precision), Recall (Recall)
"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	Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, Mike Zheng Shou	T2I diffusion models, One-shot learning, Spatio-temporal attention mechanism, DDIM inversion.	thecvf.com	Peak Signal-to-Noise Ratio (PSNR), Similarity Index Measure (SSIM)
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.	Rui Zhao, Yuchao Gu, Jay Zhangjie Wu, David Junhao Zhang, Jiawei Liu, Weijia Wu, Jussi Keppo, Mike Zheng Shou	T2V diffusion models, Low-Rank Adaptions (LoRAs), Dual-path LoRAs architecture	arxiv.org	Fréchet Inception Distance (FID), Precision-Recall (Precision)
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.	Wenjing Wang, Huan Yang, Zixi Tuo, Huiguo He, Junchen Zhu, Jianlong Fu, Jiaying Liu	T2V generation, Spatiotemporal diffusion model, Swapped cross-attention, HD-VG-130M.	arxiv.org	Peak Signal-to-Noise Ratio (PSNR), Similarity Index Measure (SSIM), Temporal correlation.

An, Jie, Songyang Zhang, Harry Yang, Sonal Gupta, Jia-Bin Huang, Jiebo Luo, and Xi Yin. "Latent-shift: Latent diffusion with temporal shift for efficient text-to-video generation." arXiv preprint arXiv:2304.08477 (2023).	Jie An, Songyang Zhang, Harry Yang, Sonal Gupta, Jia-Bin Huang, Jiebo Luo, Xi Yin	T2I generation, Diffusion models, Autoencoder, U-Net, Latent-shift module.	arxiv.org	Fréchet Inception Distance (FID), Peak Signal-to-Noise Ratio (PSNR), Similarity Index Measure (SSIM).

2.2 Existing System

One notable existing system is a generative model based on the GPT (Generative Pre-trained Transformer) architecture, capable of generating diverse and high-quality images/videos from textual prompts. The model employs a VQ-VAE-2 (Vector Quantized Variational Autoencoder 2) to map discrete latent codes to image/videos representations. By conditioning on textual inputs, model generates images/videos that conceptually align with the provided descriptions, showcasing its ability to understand and synthesize complex visual concepts. While the model primarily focuses on image generation, its principles can be extended to generate sequences of images, forming a basis for text-to-video synthesis with the potential for diverse applications in content creation and multimedia generation.

2.3 Problem Formulation

The problem formulation centres around the development and customization of Conv3D neural network models for video classification tasks. In the realm of computer vision, understanding and interpreting video content pose unique challenges due to the temporal dependencies and spatial complexities inherent in video data. Conv3D architectures offer a powerful framework for capturing both spatial and temporal features, enabling effective analysis and classification of video sequences.

Key elements of the problem formulation include:

- Architecture selection: Finding the best Conv3D architecture (ResNet, pre-activation ResNet, etc.) based on data characteristics, computational resources, and task requirements.
- Parameter tuning: Optimizing model parameters like number of classes, sample size, and duration for performance and adaptability.
- Configuration settings: Determining optimal settings for the chosen architecture (depth, shortcut types, cardinality) for effective video processing.
- Efficiency and compatibility: Ensuring compatibility with GPUs for CUDA acceleration and designing efficient workflows for both training and inference.
- Scalability: Implementing data parallelism for efficient distribution of workload across multiple GPUs, addressing challenges of large datasets.
- Performance optimization: Seeking optimal performance and accuracy through customization, parameter tuning, and architectural refinement.

2.4 Proposed System

The project unveils a text-to-video generation system leveraging deep learning techniques. A pre-trained Conv3D network extracts visual features from video data, while a text encoder captures the essence of textual descriptions. These combined representations feed a decoder that constructs video frames sequentially, refined by an attention mechanism for coherence. An adversarial GAN framework pits the decoder against a discriminator, the latter judging real vs. generated videos. This competitive training loop ensures the decoder creates increasingly realistic videos, ultimately translating textual descriptions into visually compelling narratives. This paves the way for advancements in video editing, education, and entertainment, and future work delves into temporal constraints, attention mechanism optimization, and diverse model exploration.

3. OBJECTIVES

- To develop and train a robust machine learning model capable of accurately converting textual input into visually engaging video content.
- To optimize the model's performance through continuous refinement and iteration, enhancing its ability to generate high-quality videos from diverse text inputs.
- To incorporate natural language processing techniques to improve the model's understanding of nuanced textual cues, resulting in more contextually relevant video outputs.
- To evaluate and validate the model’s effectiveness through rigorous testing ensuring it meets or exceeds predefined performance metrics and user expectations.

4. METHODOLOGY

Our methodology involves training a Conv3D neural network model on the UCF101 dataset, consisting of videos categorized into 101 action classes, to perform video classification.

The methodology encompasses the following steps:

- We begin by importing essential libraries and modules, including PyTorch for deep learning, imageio for video processing, numpy for numerical operations, and cv2 for image manipulation. Additionally, we import custom modules for data generation and model architecture.
- We create batches of video data for model training using parameters such as window size, batch size, dataset directory, and preprocessing settings. In addition, it precomputes label mappings for videos in the dataset to maintain consistency between training sessions.
- The script's core execution flow includes specifying hyperparameters such as the number of training samples, evaluation samples, epochs, batch size, and learning rate. Video batches are made for both training and validation purposes.

- We import a pretrained Conv3D model and set up the loss function (CrossEntropyLoss), optimizer (AdamW), and learning rate scheduler (CosineAnnealingLR).
- The model architecture is composed of a predetermined build set that includes ResNet, preactivation ResNet, Wide ResNet, ResNeXt, and DenseNet. Each architecture has multiple depths and configurations to meet distinct dataset properties and processing needs.
- The model may optionally be customised to use CUDA for GPU acceleration if available. This stage guarantees efficient computing during training and inference, especially for huge datasets and sophisticated model designs.

5. EXPERIMENTAL SETUP

The experimental setup involves collecting and preprocessing a diverse textual dataset, designing and implementing the NLP-based deep learning model, training and optimizing the model, evaluating its performance using various metrics and cross-validation techniques, conducting real-time testing for accurate video creation and integrating ethical considerations to ensure responsible deployment in educational and marketing fields.

- **Dataset Preparation Batching:** Curate a diverse dataset containing paired textual descriptions and corresponding video sequences. Apply preprocessing techniques such as tokenization and stemming to textual data. Create batches of video and text data, adjusting parameters like window size and batch size for efficient training.

- **Model Configuration:** Import a pre-trained Conv3D network for visual feature extraction. Implement a text encoder and a decoder with an attention mechanism. Incorporate an adversarial GAN framework for realism.

- **Training Process & Hyperparameter Tuning:** Optimize model parameters by leveraging both visual and textual representations. Employ competitive training with the GAN framework to enhance realism. Utilize appropriate loss functions for effective learning. Specify hyperparameters (epochs, batch size, learning rate) and fine-tune iteratively for optimal performance.

- **Evaluation Metrics:** Assess the model's performance using metrics such as Mean Squared Error (MSE) for frame accuracy. Conduct qualitative analysis for visual coherence and narrative fidelity.

- **GPU Acceleration & Model Deployment:** Customize the model for GPU acceleration using CUDA. Publicly share the codebase for transparency and reproducibility.

6. CONCLUSION

In conclusion, automated text-to-video technology is a huge step forward in the content creation space, providing businesses and people with never-before-seen possibilities to easily create visually engaging stories. This novel approach has the potential to completely change the way people interact with information and communicate in the digital age by combining state-of-the-art machine learning algorithms, natural language processing methods, and computer vision capabilities.

This technology creates a multitude of opportunities by facilitating the smooth conversion of textual content into dynamic video presentations in a variety of industries and sectors. Automated text-to-video solutions enable users to deliver their messages with impact and clarity, engaging audiences and generating meaningful engagement in a variety of contexts, from marketing and advertising campaigns to instructional materials, training modules, and more.

Additionally, automated text-to-video technology is democratising content creation and levelling the playing field so that people and companies of all sizes may compete in the digital space more fairly. Content creators can use this potent weapon to unleash their creativity and tell their tales to a global audience, without being limited by the requirement for specialised skills or resources.

Further innovation and improvement are anticipated as this technology develops and reaches maturity, with improvements in speed, accuracy, and customizability imminent. Automated text-to-video solutions will keep pushing the envelope through further research and development, opening the door for a time when visual storytelling is more dynamic, impactful, and widely available than ever before.

REFERENCES

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

[4] 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.

[5] 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 ·

[6] 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.

[7] 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 ·

[8] 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 ·

[9] 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.

[10] 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.

[11] 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.

[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] “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.

[18] 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.

[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] An, Jie, Songyang Zhang, Harry Yang, Sonal Gupta, Jia-Bin Huang, Jiebo Luo, and Xi Yin. "Latent-shift: Latent diffusion with temporal shift for efficient text-to-video generation." arXiv preprint arXiv:2304.08477 (2023).