International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 11 Issue: 06 | Jun 2024 www.irjet.net p-ISSN: 2395-0072

Text2Video: AI-driven Video Synthesis from Text Prompts

Shankar Tejasvi[1], Merin Meleet[1]

[1]Department of Information Science and Engineering,
RV College of Engineering
Bengaluru, Karnataka, India

Abstract - *The emerging discipline of text-to-video synthesis combines computer vision and natural language understanding to create coherent, realistic videos that are based on written descriptions. The research is an endeavour to provide a bridge between the fields of computer vision and natural language processing by using a robust text-to-video production system. The system's main goal is to convert text prompts into visually appealing videos using pre-trained models and style transfer techniques, providing a fresh approach to content development. The method demonstrates flexibility and effectiveness by including well-known libraries like PyTorch, PyTorch Lightning, and OpenCV. The work emphasises the potential of style transfer in boosting the creative quality of visual outputs by emphasising its capability to make videos with distinct styles through rigorous experimentation. The outcomes illustrate how language clues and artistic aesthetics can be successfully combined, as well as the system's ramifications for media production, entertainment, and communication. This study adds to the rapidly changing field of text-to-video synthesis and exemplifies the fascinating opportunities that result from the fusion of artificial intelligence and the production of multimedia content.*

Key Words: Text to Video, Pre- Trained Models, Style Transfer, Multimedia Content Creation, Natural Language Processing

1.INTRODUCTION

Natural language processing (NLP) and computer vision have recently come together to revolutionise the way that multimedia material is produced. A fascinating area of this confluence is text-to-video creation, which includes creating visual stories out of written prompts. Due to its potential applications in a variety of industries, including entertainment, education, advertising, and communication, this developing topic has attracted significant attention. Text-to-video generation offers a cutting-edge method of information sharing by enabling the transformation of written descriptions into compelling visual content.

The complexity of text-to-video production is explored in this work, with a focus on using pre-trained models and style transfer methods. The work's goal is to make it easier to convert textual cues into dynamic video sequences by utilising the strength of well-known frameworks like PyTorch, PyTorch Lightning, and OpenCV. Contextual information from the input text is extracted in this procedure, and the information is then converted into visual components.

The main goal of the work is to investigate how linguistic and visual clues might be combined to produce movies that accurately convey textual material while also displaying stylistic details. A key component of this system, style transfer enables the adoption of current visual styles onto the produced videos, producing visually stunning results that exemplify creative aesthetics. The system aims to demonstrate the effectiveness of its methodology in video production with a variety of styles, so showcasing the possibilities for innovation and customization.

This work contributes to the changing environment of content creation as artificial intelligence and multimedia continue to converge by providing insights into the opportunities made possible by the interaction between language and visual. The research highlights the game-changing possibilities of AI-driven multimedia synthesis by showcasing the capabilities of text-to-video production combined with style transfer.

2. LITERATURE REVIEW

The method for zero-shot picture categorization that is suggested in this study makes use of human gaze as auxiliary data. A paradigm for data collecting that involves a discriminating task is suggested in order to increase the information content of the gaze data. The paper also proposes three gaze embedding algorithms that exploit spatial layout, location, duration, sequential ordering, and user's concentration characteristics to extract discriminative descriptors from gaze data. The technique is implemented on the CUB-VW dataset, and several experiments are conducted to evaluate its effectiveness. The results show that human gaze discriminates between classes better than mouse-click data and expert-annotated characteristics. The authors acknowledge that although their approach is generalizable to other areas, finer-grained datasets would benefit from utilising different data collection methodologies. Overall, the suggested strategy provides a more precise and organic way to identify class membership in zero-shot learning contexts. [1]

The study introduces Consistent Generative Query Networks (CGQN), a novel model that might successfully construct upcoming frames in a video sequence without requiring consecutive input and output frames. In order to effectively and simultaneously sample frames that are temporally consistent at all time, the model first generates a latent representation from any set of frames. The CGQN consumes input frames and samples output frames entirely in parallel, enforcing consistency of the sampled frames. This is accomplished by training on several correlated targets, sampling a global latent, and using a deterministic rendering network. In contrast to earlier video prediction models, this. Additionally, the study offers strong experimental evidence in the form of stochastic 3D reconstruction and jumpy video forecasts to back up the methodology. [2]

Using the technique presented in this study, motion blur is produced from a pair of unexposed pictures using a neural network architecture with a differentiable "line prediction" layer. The scientists developed a synthetic dataset of motion-blurred images using frame interpolation techniques, and then they evaluated their model on a genuine test dataset. The approach is more suitable for teaching data synthesis online using deep learning and faster than frame interpolation. [3]

TiVGAN, or Text-to-Image-to-Video Generative Adversarial Network, is a network that produces films from text synopses. A single image is first produced using the framework's incremental evolutionary generator, which then gradually turns that image into a video clip with the appropriate length. The generator stabilizes the training process while conditioning on the input text using a number of techniques, including a two-stage training procedure, a progressive growth strategy, and a feature matching loss. The network is trained using a combination of adversarial loss, feature matching loss, and perceptual loss. The precise organizational structure of the network is provided in the additional material. [4]

This study reviews in-depth the most recent video Generative Adversarial Networks (GANs) models. The paper begins by recapping earlier reviews of GANs, identifying gaps in the research, and providing an overview of the main advancements made by GANs models and their variations. After dividing video GAN models into unconditional and conditional models, the research reviews each category. The conclusion of the work is the discussion of probable future directions for video GANs research. Overall, this work is a valuable resource for anyone interested in the most recent developments in video GAN models and their potential uses in a range of industries. [5]

Creating a domain-specific ontology and preparing a corpus of physics questions and answers are all steps in the process of putting PhysNLU into practice. Multiple-choice physics questions are available from PhysNLU to test one's command of natural language. It generates suggestions while evaluating the explanation's coherence with the use of automatic metrics and human annotations. A platform for crowdsourcing expert annotations is built into the product. A helpful tool for assessing how well NLP models understand physics and generate logical explanations is PhysNLU. It enhances NLU evaluation in the physics domain and increases the robustness of the NLP systems in this particular subject. [6]

This paper presents a novel method for producing videos that makes use of diffusion probabilistic models. Based on the concepts of diffusion processes, the authors propose a method that iteratively alters a noise distribution to resemble the target distribution. They emphasise how ineffective the techniques now employed for making movies are at capturing complex temporal dynamics while providing a full study of those techniques. The suggested diffusion probabilistic model effectively models video data by utilising the temporal links between frames. The authors demonstrate the effectiveness of their method on numerous video datasets and compare it to state-of-the-art methods. The results show that the diffusion model generates videos with better temporal coherence and realism that are of the highest quality. The paper presents a potent and successful strategy based on diffusion probabilistic modeling, advancing the field of video creation. [7]

The paper introduces a novel video production technique based on video diffusion models. The authors propose extending diffusion probabilistic models to handle sequential data, such as movies. They discuss the technical details of their strategy, which entails iteratively applying a diffusion process to create each frame of a video. The authors demonstrate the effectiveness of their approach by training video diffusion models on various video datasets and displaying the generated films. They also compare their strategy to other cutting-edge methods for producing videos, indicating that it performs better in terms of capturing complex temporal dynamics and visual quality. The study advances the field of video generation by providing a fresh and effective way utilising video diffusion models. [8]

CogVideo, a comprehensive pretraining technique for transformer-based text-to-video generation, is introduced in the paper. The authors propose a novel architecture that comprises a two-stage pretraining procedure to capture both textual and visual data. In two stages, a language model is refined using a huge video dataset after being pretrained on a sizable text corpus in the first stage. Additionally, they create a brand-new pretraining objective known as Cross-modal Generative Matching (CGM) in order to align the text and video representations. The performance of CogVideo on several text-to-video production jobs is evaluated by the authors in comparison to other methods. The field of text-to-video generation develops with the advent of a strong and effective pretraining technique using transformers. [9]

The invention of a way to generate text from video without text from video data. The authors suggest a novel framework called Make-a-Video that combines a text-to-image model with a video prediction model. The text-to-image model initially builds a static image representation based on the input text, and then the video prediction model gradually creates future frames by using motion information. By training these models in a self-supervised manner, the authors demonstrate that it is possible to create films from text descriptions without the need for coupled text-video data. The proposed approach is tested on several datasets, confirming its effectiveness in creating engaging and diverse videos. The study provides a contribution to the field of text-to-video creation by outlining a promising technique that does away with the need for text-video training pairings. [10]

This study introduces Imagen Video, a technique for producing high-definition videos that makes use of diffusion models. The authors propose an extension of the diffusion probabilistic model designed specifically for the production of high-quality films. They create a hierarchical diffusion approach to capture both geographic and temporal dependencies inside the video frames. After being trained with a variety of noise levels, the model creates each frame by iteratively converting the noise. The authors demonstrate Imagen Video's effectiveness using various video datasets, emphasising its ability to create high-resolution films with improved visual clarity and coherence. High-definition videos are created through an efficient procedure. [11]

The article provides a brief introduction to DiffusionDB, a massive prompt gallery dataset designed for creating and testing text-to-image generative models. The writers address the issue of limited diversity and quality in existing prompt databases by gathering a significant collection of varied and aesthetically beautiful questions. Each of the several text prompts in DiffusionDB has a high-quality image drawn from an open-access library. A vast variety of visual concepts, objects, situations, and styles were explicitly included in the dataset when it was designed. In order to assess the diversity and coverage of the prompt galleries, the authors also offer evaluation indicators. They develop and evaluate cutting-edge text-to-image generative models to demonstrate DiffusionDB's utility. The paper's large-scale prompt gallery dataset is a valuable tool that enables more thorough training and evaluation of text-to-image generative models. [12]

In this study, Text2Video-Zero, a novel technique for creating zero-shot films using text-to-image diffusion models, is introduced. The authors present a method that employs pre-trained text-to-image models to create video sequences directly from written descriptions without the need for video-specific training. By customising the diffusion process on textual inputs, the model may generate a variety of intriguing and well-coordinated video frames. They demonstrate Text2Video-Zero's effectiveness and its ability to create convincing and artistically beautiful films using a variety of video datasets. With the use of trained text-to-image models and only textual descriptions, the method offers a zero-shot approach that makes it possible to create videos. [13]

The paper presents a novel approach for text-to-image synthesis using generative adversarial networks (GANs). The authors propose a model that incorporates a text encoder, an image generator, and a discriminator network. The text encoder transfers textual descriptions to an embedding space that the picture generator uses to synthesise images. The discriminator network distinguishes between real and fake images and offers feedback to aid in the creation of images. By comparing their model to benchmark datasets, the authors show how effectively it can generate visually coherent and semantically meaningful images from text inputs. By doing this, the gap between written explanations and the development of realistic images is effectively closed. [14]

This work introduces Promptify, a method for text-to-image generation that combines interactive prompt exploration with massive language models. The authors propose a novel framework that combines human input with important language models to generate high-quality images from textual prompts. With Promptify, users may iteratively tweak the prompt and get rapid visual feedback from the model. This interactive research allows users to precisely control the production of the desired image by altering the prompt phrasing. The success of Promptify is demonstrated by the authors via user studies and comparisons with alternative strategies, emphasising improved image quality and user satisfaction. [15]

3. METHODOLOGY OF PROPOSED SYSTEM

The proposed text-to-video synthesis system employs a systematic methodology that seamlessly integrates textual descriptions with visual content to produce coherent and realistic videos.

The videos produced are available in 2 classes:

- Generic Videos
- Enhanced Videos

The following flow diagram represents the steps involved in the proposed system:

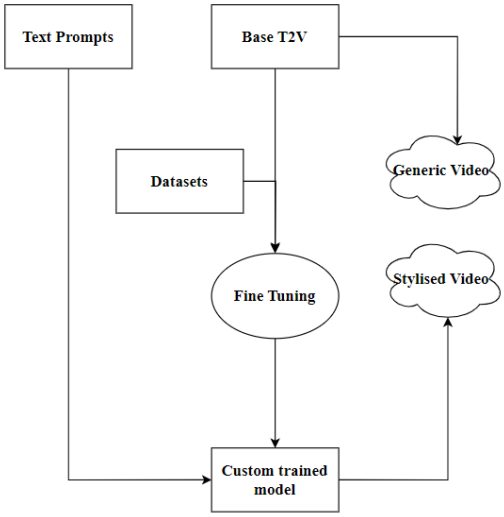

Fig- 1. Flow Diagram

The process of video generation from given text prompt involves the following 6-stage process:

1. Environment Setup:
   - Ensure CUDA availability in the notebook settings for GPU acceleration.
   - Install Miniconda, granting execution permissions and setting the installation path to /usr/local.
   - Replace the system's Python version with version 3.8 using update-alternatives.
   - Install necessary packages using apt-get and python3-pip.

2. Project Setup and Dependencies:
   - Clone the VideoCrafter project repository and navigate to the relevant directory.
   - Set the PYTHONPATH to include the project directory.
   - Install required PyTorch and related packages with specified versions for compatibility.
   - Install additional libraries such as PyTorch Lightning, OmegaConf, and OpenCV using python3-pip.
   - Install packages like AV and MoviePy for multimedia processing.

3. Model Acquisition and Configuration:
   - Clone the VideoLORA model repository and move the models to the appropriate directory.
   - Define the available VideoLORA styles and their corresponding paths.

4. Text-to-Video Generation (Base Model):
   - Set the desired text prompt and output directory.
   - Specify the path to the base model checkpoint and its configuration.
   - Execute sample_text2video.py with provided parameters for base text-to-video generation.
   - Display the resulting video using HTML to showcase the generated content.

5. Text-to-Video Generation with Style Transfer (VideoLORA):
   - Select a VideoLORA style from the available options based on the chosen LORA_PATH.
   - Set the prompt, output directory, and style parameters.
   - Execute sample_text2video.py with additional parameters for VideoLORA integration.
   - Display the style-transferred video using HTML to visualize the synthesized content.

6. Result Visualization:
   - Extract the latest video file generated in the output directory.
   - Convert the video into a data URL for display using base64 encoding.
   - Display the video animation in the notebook using HTML and the data URL.

3.1 PyTorch:

In this system, PyTorch, a flexible and popular deep learning framework, is essential. The creation of sophisticated neural networks for text-to-video synthesis is made possible by its dynamic computation graph and automatic differentiation. PyTorch excels at managing a variety of model topologies and provides a high level of flexibility that is essential for adjusting to the complexities of producing videos from text input. Additionally, the speed of computation is substantially increased by its GPU acceleration capabilities, making it a crucial element for effectively processing huge amounts of data during training and inference. The integration of cutting-edge deep learning algorithms in this system is further streamlined by PyTorch's large ecosystem of pre-built modules and community support.

3.2 PyTorch Lightning:

By removing low-level training loop details and streamlining distributed training, PyTorch Lightning boosts efficiency. This package automates data parallelism, checkpointing, and GPU allocation, enabling the system to scale up smoothly to take advantage of the hardware resources. PyTorch Lightning improves code readability and maintainability by the use of standardised procedures, promoting teamwork. Its built-in support for cutting-edge features like gradient accumulation and mixed-precision training optimises training effectiveness while using less memory. The system can concentrate on model architecture and experimental design thanks to PyTorch Lightning, expediting the

investigation of cutting-edge methods for text-to-video synthesis.

3.3 OpenCV:

To manage video data, the system relies on OpenCV, the Open Source Computer Vision Library. Its extensive collection of features enables effective image processing, frame extraction, and video file I/O. With tools to improve video quality and coherence, OpenCV's capabilities also include feature detection, object tracking, and video stabilisation. Additionally, the fact that it supports a variety of picture formats guarantees compatibility with multiple multimedia sources, which is essential when working with heterogeneous data in the text-to-video pipeline. The system uses OpenCV to provide reliable video manipulation, which is essential for turning written descriptions into aesthetically appealing and cohesive video outputs.

3.4 AV:

A multimedia library called AV gives the system the ability to manage sophisticated audiovisual data. This library makes it possible to seamlessly combine audio and video elements, assuring synchronisation and improving the overall usability of the output films. AV is excellent at managing codecs, handling multimedia metadata, and parsing and decoding video files. The system's adaptability is further increased by its support for several video formats and streaming, which accepts various data inputs and output formats. The system incorporates AV to ensure that the final movies appropriately reflect the intended storylines and styles developed from the text prompts while also maintaining the integrity of multimedia content.

3.5 MoviePy:

The system uses MoviePy, a video editing library, as its creative toolbox for enhancing video outputs. The system can smoothly incorporate transitions, apply visual effects, and assemble films thanks to its simple-to-use API. The use of text and picture overlays with MoviePy's capabilities enables the inclusion of extra information or branding into the videos. It makes it easier to create polished, professional-level videos because to its ability to concatenate videos, cut segments, and alter video characteristics. A wide range of multimedia players and platforms are supported by MoviePy's integration with several video file formats, ensuring compatibility with the output videos with a variety of viewers. The system uses MoviePy to add an artistic layer to the generated videos, improving their visual appeal and narrative quality.

4. RESULTS AND DISCUSSION

The system's results shed important light on the efficacy and potential of the suggested text-to-video generating methodology. The produced videos show an impressive conversion of written information into lively visual tales. The system's ability to grasp textual nuances and visually communicate them is demonstrated by the underlying text-to-video model's effective transformation of prompts into cohesive video sequences. By including style transfer through VideoLORA, the created videos are further improved and are infused with distinctive artistic styles that match the selected prompts. The outcomes highlight how well the system combined complex visual components with language clues to produce videos that are resonant in both substance and style.

The selection of VideoLORA styles has a big impact on the personality and atmosphere of the generated videos in the context of style transfer. The successful integration of artistic styles highlights the opportunity for individualised and flexible video content production. The system's reliance on effective libraries like PyTorch and PyTorch Lightning further simplifies the generation process and enables quick testing and improvement. The system's demonstrated capabilities, supported by reliable and repeatable findings, set the groundwork for subsequent developments in multimedia synthesis, stimulating further research and development at the nexus of computer vision and natural language processing.

The following diagrams illustrate the various image frames of videos generated using the various model classes:

Text Prompt: An astronaut riding horse in outer space

Fig- 2. An astronaut riding horse in outer space- Genearlised Video

Fig- 3. An astronaut riding horse in outer space – COCOStyle

Fig- 4. An astronaut riding horse in outer space – MakotoShinkai

5. CONCLUSION

By utilising the synergy between natural language processing and computer vision, this system acts as a creative and dynamic investigation of text-to-video production. The system successfully connects textual prompts and visually appealing video outputs by integrating pre-trained models, style transfer, and a comprehensive set of libraries like PyTorch, PyTorch Lightning, OpenCV, AV, MoviePy, and OmegaConf. The combination of language clues with visual aesthetics reveals the potential for the creation of creative content across a range of fields, including communication, education, and entertainment. The system's methodology's application and adaptability are highlighted by the methodical way in which it was implemented, which was founded on effective code execution and best practises. This system adds to the changing landscape of content creation by highlighting the capabilities of AI-driven multimedia synthesis, paving the door for interesting developments at the nexus of artificial intelligence and multimedia technologies.

6. LIMITATIONS AND CONCLUSION

Despite the established text-to-video synthesis system's excellent achievements, there are several drawbacks to its current design that must be acknowledged. The model's dependence on its training data is a noteworthy restriction; deviations from the training corpus may lead to errors. Additionally, real-time applications can be hampered by the high computing demands of training and producing movies. Future versions could concentrate on improving model generalisation and increasing computing effectiveness to address these limitations.

There are many opportunities for growth and improvement in the future. The system's realism and congruence with human perception might be strengthened by the incorporation of user feedback through human evaluations. Investigating methods to incorporate finer-grained control over video qualities, like style and mood, could produce a variety of results. The dataset's flexibility to different settings might be improved by adding a variety of textual stimuli. Additionally, improvements in transfer learning and multimodal pre-training may open up new possibilities for text-to-video synthesis. While admirable, this study system only touches the surface of a broad field, leaving plenty of potential for creativity and inquiry at the dynamic confluence of text and visual information.

REFERENCES

[1] Karessli, N., Akata, Z., Schiele, B., & Bulling, A. (2017). Zero-Shot Image Classification using Human Gaze as Auxiliary Information. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 4525-4534. doi:10.1109/CVPR.2017.679

[2] Kumar, A., Eslami, S. M. A., Rezende, D., Garnelo, M., Viola, F., Lockhart, E., & Shanahan, M. (2019). Consistent generative query networks for future frame prediction in videos. arXiv preprint arXiv:1807.02033.

[3] Brooks, T., & Barron, J. T. (2019). Generating motion blur from unblurred photos using neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 6840-6848. doi:10.1109/cvpr.2019.06840

[4] Kim, D., Joo, D., & Kim, J. (2020). TiVGAN: Text-to-image-to-video generative adversarial network. IEEE Access, 8, 153113-153122. doi:10.1109/access.2020.2986494

[5] Aldausari, N., Sowmya, A., Marcus, N., & Mohammadi, G. (2022). Review of video generative adversarial networks (GANs) models. ACM Computing Surveys, 55(2), Article 30. doi:10.1145/3487891

[6] Meadows, J., Zhou, Z., & Freitas, A. (2022). PhysNLU: A tool for evaluating natural language understanding in physics. In Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), 4904-4912. doi:10.18653/lrec-2022-4904

[7] Yang, R., Srivastava, P., & Mandt, S. (2022). Video creation using diffusion probabilistic models. arXiv preprint arXiv:2203.09481.

[8] Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., & Fleet, D. J. (2022). Video generation using video diffusion models. arXiv preprint arXiv:2204.03409.

[9] Hong, W., Ding, M., Zheng, W., Liu, X., & Tang, J. (2022). CogVideo: Large-scale pretraining for transformer-based text-to-video generation. arXiv preprint arXiv:2205.15868.

[10] Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., Zhang, S., Hu, Q., Yang, H., Ashual, O., Gafni, O., Parikh, D., Gupta, S., & Taigman, Y. (2022). Make-a-Video:Text-to-Video Generation without Text-Video Data. arXiv preprint arXiv:2209.14792

[11] Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., Kingma, D. P., Poole, B., Norouzi, M., Fleet, D. J., & Salimans, T. (2022). Imagen Video: High-Definition Video Generation using Diffusion Models. arXiv preprint arXiv:2210.02303.

[12] Wang, Z. J., Montoyo, E., Munechika, D., Yang, H., Hoover, B., & Chau, D. H. (2022). DiffusionDB: A sizable prompt gallery dataset for text-to-image generative models. arXiv preprint arXiv:2210.11890.

[13] Khachatryan, L., Movsisyan, A., Tadevosyan, V., Henschel, R., Wang, Z., Navasardyan, S., & Shi, H. (2023). Text2Video-Zero: Zero-Shot Video Generation using Text-to-Image Diffusion Models. arXiv preprint arXiv:2303.13439.

[14] Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., & Lee, H. (2016). Text-to-image synthesis using generative adversarial networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 4794-4803. doi:10.1109/cvpr.2016.296

[15] Brade, S., Wang, B., Sousa, M., Oore, S., & Grossman, T. (2023). Promptify: Interactive prompt exploration for text-to-image generation. arXiv preprint arXiv:2304.09337.