International Journal of Innovative Research in Technology and Science
ISSN: 2321-1156 www.ijirts.org Volume 12 Issue 2, March 2024

Text-To-Image-And-Video Generator Using Machine Learning

Heena Ansari, Sukant Tekade, Chanakya Ghagre, Samiksha Kolte, Tejas Shende
Department of Information Technology
Kavikulguru Institute of Technology and Science, Maharashtra, India

Selection and peer review of this article are under the responsibility of the scientific committee of the International Conference on Current Trends in Engineering, Science, and Management (ICCSTEM-2024) at SAM Global University, Bhopal.

Abstract— The innovative idea about Text-to-Image and Video Synthesis Applications utilising state-of-the-art machine learning techniques. The primary aim is to seamlessly generate realistic visual content from textual descriptions, bridging the gap between language and multimedia. The application showcases its proficiency in generating convincing images of birds and flowers based on detailed textual input. The foundation of this project lies in leveraging deep learning algorithms, notably Generative Adversarial Networks (GANs) and recurrent neural networks, to achieve precise and high-quality image and video synthesis. The methodology involves extensive pre-training on diverse textual and visual datasets. Refining processes through user-generated data consistently improves the model's performance and adaptability. The result will be a user-friendly and efficient application, empowering content creators, designers, and storytellers to effortlessly produce captivating visual content by providing descriptive text, thereby revolutionising the creation and sharing of multimedia content.

Keywords—Generative Models, AI-driven multimedia synthesis, Machine learning, Visual Content Creation, Generative Adversarial Networks

I. INTRODUCTION

Text-to-image and video synthesis applications driven by machine learning are at the forefront of artificial intelligence and computer vision innovation. These groundbreaking applications empower visual content creation, from images to videos, solely based on textual descriptions or prompts. This transformative technology holds promise for revolutionising numerous industries, including entertainment, e-commerce, and education, and beyond the advent of machine learning models, particularly those leveraging deep learning techniques, which has propelled significant progress in generative models. These models can generate realistic and contextually relevant visual content guided by natural language descriptions. As a result, they serve as a crucial link between human language and visual representation. This document explores the fundamental aspects of text-to-image and video synthesis applications employing machine learning. It will encompass their significance, the underlying technologies driving their development, notable use cases across various industries, and potential prospects. By the conclusion of this exploration, readers will gain a comprehensive understanding of this dynamic field and its myriad applications. Text-to-image and video synthesis represent a paradigm shift in content creation, enabling streamlined and intuitive methods for generating visual content. Through the seamless integration of natural language processing and computer vision, these

applications offer unprecedented opportunities for creativity and innovation.

Moreover, generating visual content from textual descriptions opens avenues for personalised content creation, interactive storytelling, and enhanced user experiences. Text-to-image and video synthesis applications are vast and diverse, from generating product images based on textual descriptions in e-commerce to creative, dynamic educational content tailored to specific learning objectives. As research and development in this field continue to advance, we can anticipate further refinement of algorithms, improved model performance, and the emergence of novel applications that leverage the synthesis of text and visuals to enrich various aspects of our lives and industries.

II. MOTIVATION

Text-to-image and video synthesis applications can significantly enhance user experiences across various domains. These applications facilitate more intuitive and personalised interactions by allowing users to describe their ideas, desires, or preferences in natural language and transform those descriptions into visual content. This improves engagement and satisfaction in gaming, e-commerce, or educational settings. Content creation can be time-consuming and resource-intensive, requiring skilled artists, designers, or videographers. Text-to-image and video synthesis tools can automate and expedite content generation, reducing production costs and time.

III. OBJECTIVE

The objectives of text-to-image and video synthesising applications using machine learning are specific goals that drive the development and implementation of these technologies. These objectives help define the desired outcomes and guide these applications' research, development, and deployment. The primary objective is to generate visual content that is accurate and relevant to the given textual input. This includes creating images and videos that align with the provided text's context, details, and nuances. In order to create compelling and engaging visual content, these applications aim for high levels of realism and quality. These applications strive to generate content tailored to individual preferences, needs or specifications.

IV. APPLICATIONS

Our application will be smart and interpretative enough to take input from users in multiple languages, making it accessible to various communities. The application will preview the video (Any Frame from the video) so the user can specify whether the Model generates the required output, saving the user's time. Users can download the generated image and video.

V. LITERATURE SURVEY

Zhang et al.(2016) propose that Stack GAN tackle the challenge of generating high-quality images from text descriptions in computer vision. Existing methods often lack detail and realism. Stack GAN introduces Stacked Generative Adversarial Networks (GANs) to produce 256x256 photo-realistic images conditioned on text inputs. This is achieved through a two-stage process: Stage-I GAN outlines basic shapes and colours based on the text yield in glow-resolution images, while Stage-II GAN refines these outputs, enhancing details and realism. In order to enhance diversity and stabilise training, a Conditioning Augmentation technique is introduced. Experimental results on benchmark datasets demonstrate significant improvements in generating photo-realistic images. However, one limitation is Stack GAN's focus on generating

images from clear text descriptions, potentially struggling with vague or ambiguous inputs, limiting its practical applicability in real-world scenarios. Zhangjie et al.(2021) introduce Tune-A-Video for text-to-video (T2V) generation, presenting innovative solutions.

Nevertheless, the approach may face certain limitations. Despite suggesting a one-shot tuning method, the initial training of T2I diffusion models remains computationally intensive due to the extensive image data necessary. Moreover, while exceptional performance is claimed in diverse applications, the actual quality and diversity of generated videos may vary depending on factors such as the initial model's quality and the richness of the training data. The introduced components, including tailored attention and DDIM inversion, also contribute to complexity, potentially impacting implementation and usability. Success hinges on access to high-quality, diverse training data for images and text, posing data collection and preparation challenges. Further investigation is warranted to evaluate practical applicability and performance across real-world scenarios.

VI. PROPOSED APPROACH

The proposed approach involves leveraging machine learning techniques to develop an application capable of synthesising images and videos from textual descriptions. This innovative model aims to bridge the gap between text and visual media by generating corresponding images and videos based on input text. This research's development and deployment process entails several key steps. Firstly, data collection and preprocessing are crucial initial stages. This involves gathering a dataset of textual descriptions and cleaning the text data through tokenisation, punctuation removal, and lowercase conversion.

Additionally, if image data is available, it is formatted and resized to match the expected output image size. Next, text-to-image generation is facilitated by selecting an appropriate AI model, such as GANs or VQ-VAE-2, and fine-tuning it on the collected dataset. Training the model involves tuning hyperparameters, applying data augmentation techniques, and evaluating the quality of generated images using various metrics. Optional steps include image post-processing to enhance the visual appeal and video synthesis using libraries like OpenCV or FFmpeg in Python. The performance of generated images and videos is evaluated using quality metrics and subjective assessments. If necessary, a user interface can be designed to allow users to input text and visualise generated content, with options for customising video parameters.

Thorough testing and debugging ensure the reliability of the entire pipeline, followed by summarising key achievements and lessons learned. Comprehensive documentation, including code comments and a project report, is prepared to facilitate understanding and future enhancements. Potential improvements may include integrating more sophisticated AI models, enhancing text understanding through natural language processing, or optimising text-to-image generation and video synthesis performance.

System Architecture

A Data Flow Diagram (DFD) is a graphical representation of data flow through an information system shown in Figure 1. It is used to model the different processes involved in a

system and how data moves between them. A DFD consists of symbols and arrows showing the data flow.

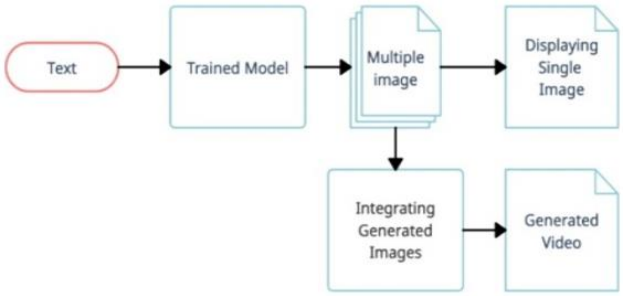

Figure 1. System Architecture

Working Text to Image and Video Generator

To begin, in the first phase of the project, you will develop an AI model that can generate multiple images based on textual input. This involves using a deep learning model, such as a GAN (Generative Adversarial Network) or a text-to-image synthesis model, pre-trained on a large dataset. The model will take textual descriptions as input and produce corresponding images. To implement this, you must preprocess and tokenise the textual input, feed it into the model, and then post-process the generated images to ensure they are coherent and visually appealing. You may fine-tune the model on a dataset containing text-image pairs to make it domain-specific and improve image quality.

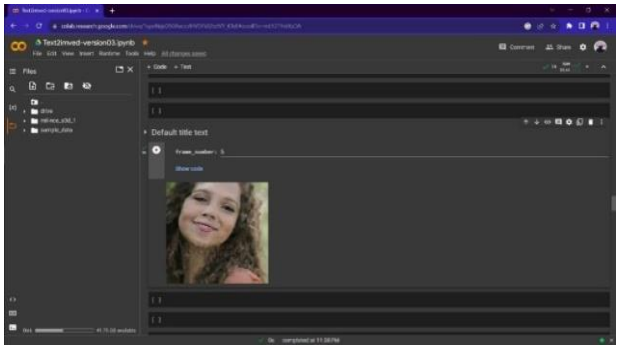

Figure 2. Generated frame for video

Data Collection: In the research second phase, which involves generating images based on textual descriptions, the focus shifts to providing a user-friendly interface for interacting with the generated images. This phase aims to create a seamless and engaging user experience by integrating a user interface component and a web or desktop application. The primary objectives are to allow users to input text descriptions and receive corresponding images in real time while ensuring the images are visually pleasing and offering options for user interaction and image management. The central goal of this phase is to create a user-friendly interface that accommodates users with various levels of technical expertise. It should be intuitive, easy to navigate, and visually appealing. The choice between a web or desktop application depends on accessibility and user base. Users need a convenient way to input textual descriptions, which will serve as the basis for image generation. This can be realised through input fields or text boxes, where users can describe the images they want to see. Effective error handling and guidance should be included to ensure accurate input. The system should provide immediate feedback by generating images based on the textual descriptions in real time. Users should see the corresponding images materialise without significant delay as they input their descriptions. This requires efficient communication with the image generation system. The displayed images should be aesthetically presented to enhance the user experience. Proper rendering, image quality, and layout should be prioritised to ensure users find the visuals engaging and satisfying. This could involve techniques such as image resizing, background rendering, and layout design. Depending on the project's goals, users should have options for interacting with the generated images. This could include zooming, panning,

rotating, or applying filters to images. The extent of interactivity will depend on the application's purpose, whether entertainment, educational, or practical. Users should be able to save or manage the generated images as needed. This involves incorporating features like download buttons, image galleries, or user accounts to save images to the cloud. The system should be designed to support image management efficiently. The interface design should align with the specific goals of the application. For instance, if the application is for creative art generation, it might focus on enhancing the user's creative process with features like customisation and editing.

In contrast, if the purpose is informational, features like image metadata and sharing options could be integrated. The interface must be optimised for performance and responsiveness. Users should not experience significant lag or delays, even during peak usage. Proper load balancing, caching, and resource management ensure a smooth user experience. The interface should provide users feedback about their requests' status and inform them about potential errors or limitations. Clear and user-friendly error messages help users understand issues and take appropriate actions.

In summary, the second phase of integrating a user interface for generating images from text descriptions is a critical step in delivering a user-centric and effective application. By focusing on intuitive design, real-time image generation, visual appeal, interactivity, and image management, the project can meet the needs of its users and achieve its specific goals, whether they are entertainment, education, or practical utility. It's essential to consider the application's purpose and tailor the interface accordingly, all while maintaining performance and responsiveness to ensure a positive user experience.

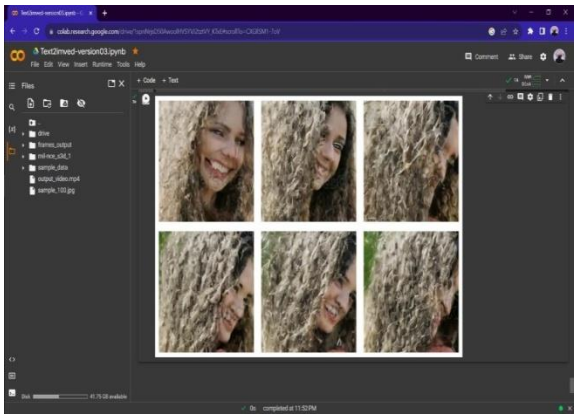

Figure 3 Multiple generated images

The third phase involves integrating the generated images into a video. To implement this, you must develop a module that takes the generated images and arranges them into a sequence, creating a video. You can use video processing libraries or frameworks to automate this task. Options for customising the video include adding transitions or audio that can be considered to enhance the final video's user experience and overall quality.

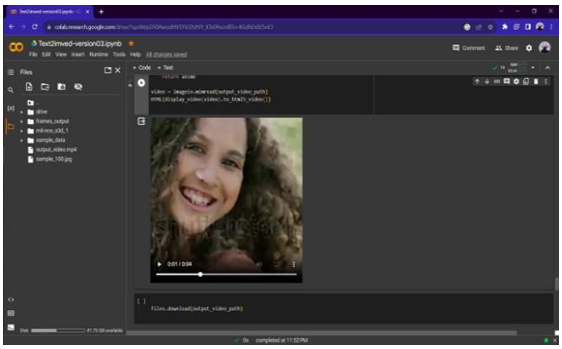

Figure 4. Generated video

Finally, the generated video is displayed to the user in the fourth phase. Similar to the second phase, this can be done through a user-friendly interface. Users should be able to view, control, and interact with the generated video. The interface should provide options for playback,

pause, volume control, and potentially even sharing or saving the video. Consider the user experience and ensure the video is displayed in a format compatible with various devices and browsers. Phase 1 of the project focuses on Image Generation from Text, leveraging a pre-trained AI model such as GPT-3 or specialised image generation models. Initially, data collection involves assembling a dataset comprising textual descriptions paired with corresponding images for potential model training. Data preprocessing is conducted to ensure alignment and cleanliness between textual descriptions and images. The subsequent step involves selecting an appropriate AI model tailored for image generation, possibly fine-tuning it on the collected dataset. The implementation then proceeds with developing a script or application capable of receiving textual input from users and generating images based on this input using the chosen AI model. The final stage within this phase entails evaluating the quality of the generated images through metrics or human assessment, enabling iterative improvements on both the model and the dataset to enhance image quality further and moving to Phase 2, in which Image Display to User, the project transitions into presenting the generated images to users in a user-friendly manner. This phase begins with designing a user interface, be it a web app, mobile app, or desktop app, facilitating user input of text and displaying generated images. Integration follows, incorporating the image generation module developed in Phase 1 into the interface. Subsequently, generated images are displayed to users in real-time or upon request, with interactive features such as zooming, saving, or requesting new images enhancing user engagement. Phase 3 shifts the project's focus towards Video Creation, where generated images are combined into a cohesive video. The sequence of generated images is organised, so they should appear in the video. Video generation utilises tools like FFmpeg, enabling the creation of videos with specified parameters such as frame rate and resolution. Optional video editing steps, including transitions, effects, or audio additions, may be undertaken to enhance the video's visual appeal. Quality control ensures that the resulting video meets the desired standards. Finally, Phase 4, Video Display to User, presents the generated video to users within the application interface. This phase integrates the video display functionality into the application, offering user controls for playback, pausing, and adjusting the video. Accessibility considerations ensure the video can be viewed across various devices and screen sizes. A feedback mechanism may also be implemented to gather user comments or ratings on the generated video, facilitating continuous improvement based on user input.

V. CONCLUSION AND FUTURE SCOPE

In conclusion, the advancements in Text-to-Image and Text-to-Video conversion using machine learning (ML) have propelled these technologies into significant domains such as content creation, entertainment, and e-commerce. The current state of these technologies showcases remarkable progress, notably in improved quality, diverse outputs, conditional generation, and efficiency. ML models, particularly generative adversarial networks (GANs), have revolutionised generated content quality, producing realistic, high-resolution images and videos. Furthermore, ongoing research endeavours aim to enhance the realism of generated content, enable interactive generation processes, develop multimodal models, facilitate few-shot learning, and address ethical concerns regarding potential misuse. Looking

ahead, future research and development in these areas hold immense promise. Envisioned directions include enhancing realism with improved details and context awareness, enabling interactive refinement of generated content, integrating multimodal capabilities, facilitating few-shot learning for niche domains, and addressing ethical considerations surrounding AI-generated content. Additionally, ensuring accessibility and inclusivity for users with disabilities remains crucial to further advancement in these technologies. As the field continues to evolve, navigating these developments with a keen awareness of both technological potential and ethical implications is imperative, fostering a responsible and inclusive approach to innovation.

VI. REFERENCES

[1]. Hadi Kazemi, Mehdi Iranmanesh, Ali Dabouei, Sobhan Soleymani, Nasser Nasrabadi. (2018). Facial Attributes Guided Deep Sketch-to-Photo Synthesis. West Virginia University.

[2]. Wengling X Jin, H. Zhang, and Zhang. (2018). SketchyGAN: Towards Diverse and Realistic Sketch to Image Synthesis.

[3]. Manish Bhattarai, Diane Oyen, Juan Castorena, Liping Yang, and Brendt Wohlberg. (2018). Diagram Image Retrieval Using Sketch-Based Deep Learning and Transfer Learning. University of New Mexico, Albuquerque, NM, USA.

[4]. Xianming Liu, Weihong Deng. (2019). Portrait Image Synthesis from Sketch via Multi-Adversarial Networks.

[5]. N. Wang, D. Tao, X. Gao, X. Li, and J. Li. (2013). Transductive Face Sketch-Photo Synthesis. IEEE Transactions on Neural Networks and Learning Systems, 24(9), 1364–1376.

[6]. D. Ulyanov, V. Lebedev, A. Vedaldi, and V. S. Lempitsky. (2016). Texture Networks: Feedforward Synthesis of Textures and Stylised Images. In ICML, pp. 1349–1357.

[7]. R. Socher, M. Ganjoo, C. D. Manning. (2013). Zero-Shot Learning through Cross-Modal Transfer. In Advances in Neural Information Processing Systems, pp. 935–943.

[8]. X. Tang, H. Jin, H. Lu, and S. Ma. (2005). A Nonlinear Approach for Face Sketch Synthesis and Recognition. In Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, vol. 1, pp. 1005–1010. IEEE.

[9]. N. Wang, D. Tao, X. Gao, X. Li, and J. Li. (2014). A Comprehensive Survey to Face Hallucination. International Journal of Computer Vision, 106(1), 9–30.

[10]. Denton, E. L., Chintala, S., Fergus, R., et al. (2015). Deep Generative Image Models Using a Laplacian Pyramid of Adversarial Networks. In NIPS.

[11]. Donahue, J., Hendricks, L. A., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., and Darrell, T. (2015). Long-Term Recurrent Convolutional Networks for Visual Recognition and Description. In CVPR.

[12]. Dosovitskiy, A., Tobias Springenberg, J., and Brox, T. (2015). Learning to Generate Chairs with Convolutional Neural Networks. In CVPR.

[13]. Farhadi, A., Endres, I., Hoiem, D., and Forsyth, D. (2009). Describing Objects by Their Attributes. In CVPR.

[14]. Fu, Y., Hospedales, T. M., Xiang, T., Fu, Z., and Gong, S. (2014). Transductive Multi-

View Embedding for Zero-Shot Recognition and Annotation. In ECCV.

[15]. Gauthier, J. (2015). Conditional Generative Adversarial Nets for Convolutional Face Generation. Technical Report.

[16]. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative Adversarial Nets. In NIPS.

[17]. Gregor, K., Danihelka, I., Graves, A., Rezende, D., and Wierstra, D. (2015). DRAW: A Recurrent Neural Network for Image Generation. In ICML.