NextGen Dynamic Video Generator using AI

Anitha R
U.G. Student, Dept. of Computer Science and Engineering
Saveetha Engineering College(Autonomous)
Chennai,India
anitharamesh.cse.sec@gmail.com

Kishore N
U.G. Student, Dept. of Computer Science and Engineering
Saveetha Engineering College (Autonomous)
Chennai, India
kishore.cse.sec@gmail.com

Dr. M. Vijay Anand
Dept. of Computer Science and Engineering
Saveetha Engineering College(Autonomous)
Chennai, India
vijayanand@saveetha.ac.in

2023 9th International Conference on Smart Structures and Systems (ICSSS) | 979-8-3503-8420-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICSSS58085.2023.10407412

***Abstract*—In an era marked by remarkable technological advancements, the way we create and share information has undergone a profound transformation. This paradigm shift is epitomized by the NextGen Dynamic Video Generator using AI, a cutting-edge tool that seamlessly integrates artificial intelligence with content creation. What sets this tool apart is the customizability, which allows the users to fine-tune every aspect of the tutorials. Users could tune creative elements like humor, the depth of explanation, character appearance, and voice to tailor tutorials to precise specifications. The core functionality of the tool revolves around script generation, the Cohere's language, laying the groundwork for the tutorial's content. Furthermore, seamless integration with Edge TTS ensures that the generated scripts are delivered with utmost clarity and engagement, enhancing the overall learning experience. Character animation is powered by SadTalker, adding a dynamic and captivating dimension to these tutorials. This animated character serves as a relatable guide, facilitating a deeper connection between the content and the audience. The tool also seamlessly integrates relevant and eye-catching images from Google, which are incorporated into the presentation slides. The workflow is a well-orchestrated process involving script generation, audio dialogue creation, image retrieval, video generation, and the seamless fusion of character animations and slides. The resulting video tutorials are not only comprehensive but also engaging and ready to be shared as valuable educational resources.**

Keywords— Cohere language model, Edge TTS, SadTalker, script generation, audio dialogue creation, humor, image retrieval, next-gen, character animation, google, slides.

I. Introduction

Video generation has evolved significantly and ushered in a new era marked by remarkable technological advancements. The combination of cutting-edge technology and innovative concepts has changed the way we create, edit and consume video content. This documentary takes a journey through this transformative landscape and seeks to elucidate the critical concepts that have redefined video generation. Artificial intelligence has become a cornerstone in video generation, revolutionizing the industry. AI's role in enhancing video quality, personalizing content, and automating production processes will be a central point of discussion. The paper will delve into deep learning techniques such as recurrent neural networks (RNNs), Generative Adversarial Networks(GAN) and their contributions to video synthesis. GANs have emerged as a core technology in video generation, enabling the creation of highly realistic content. The fundamentals of GANs and how important they are to the creation of video content could be looked into deeply. Recurrent Neural Networks (RNNs) known as Neural network architectures, are made to analyze data sequences, which makes them appropriate for a variety of tasks including speech recognition, natural language processing, and time series data processing. A primary principle of RNNs is their ability to maintain and use the memory of previous data points in a sequence to influence the processing of the current data point. This video generator consists of modules like script generation, voice and image generation, character animation and finally integration of all the modules. A key feature of an RNN is its reconnection, which allows information to flow from one time step to another. This coupling ensures that the network can remember information from previous time steps and use it to make predictions or decisions at the current time step. Script generation, image generation, text-to-speech (TTS), and character animation are key components that work together to facilitate video generation. Script generation is the foundation of every video. It involves creating a structured outline of the content, including an introduction, main points, and a conclusion. This script serves as the narrative framework for the video. Image generation complements the scenario by providing visual aids. Slides contain text, images, tables, and graphs that reinforce the content of the script. They make the video visually engaging and improve understanding. TTS technology converts the script into spoken word. This audio narration complements the image in the video and makes it accessible to a wider audience, including those with visual impairments. Character animation brings the video to life by adding animated characters to act as guides or actors. These characters increase audience engagement and help convey content in a compelling way.

II. Related Works

The field of video generation has witnessed remarkable progress in recent years, fueled by breakthroughs in machine learning, computer vision, and artificial intelligence. As the demand for persuasive, engaging and informative videos continues to grow in various fields including entertainment, education and communication, research in this area is becoming increasingly important. This literature survey

979-8-3503-8420-8/23/$31.00 ©2023 IEEE

delves into the diverse landscape of video generation and examines key works, methodologies, and challenges that have shaped this dynamic field.

[1] CVGI introduces a framework for generating videos with detailed action control via textual instructions. It focuses on hand-object interactions and divides the task into control signal estimation and action generation using Generative Adversarial Networks (GAN). CVGI effectively combines motion estimation and action generation, demonstrating its success in generating realistic videos with user-controlled actions. The success of the CVGI framework provides insight into the importance of a clear division of tasks, the use of GANs, the interpretation of textual instructions, motion estimation, and the prioritization of user control and realism. By incorporating these principles, high-quality and user-driven video content could be generated.

[2] VideoFusion presents a new approach to video generation by decomposing the diffusion process. It tackles high-dimensional data spaces by decomposing single-frame noise into shared and residual noise components, resulting in high-quality video generation. This approach outperforms GAN-based and diffusion-based methods and offers a significant improvement in video quality. This VideoFusion approach would emphasize the benefits of decomposing complex processes, managing large-scale data, separating noise components, and continuously striving for quality improvement of video content.

[3] TI2V proposes a new video generation task that generates videos from still images and text descriptions with a focus on controllable appearance and motion. A Motion Anchor-based video generator (MAGE) is introduced, which uses motion anchors and three-dimensional axial transformers to provide controllability and variety. TI2V datasets and experiments confirm the effectiveness of MAGE. The TI2V approach emphasizes the importance of solving specific video generation tasks, prioritizing controllability, using advanced techniques such as motion anchors and axial transformers, and validating system efficiency through large data sets.

[4] The simulator deals with speech-driven facial animation and emphasizes speech style specific to facial identity and idiosyncrasies. Transformer uses style agnostic to optimize for an identity-specific speaking style based on a short reference video. This approach improves lip sync and preserves the actor's speech style, increasing the realism of the expressions generated. To implement a video generator project, it is essential to prioritize speech-driven animation for synchronization and realism. Use Transformer's adaptable models to optimize for different speech styles and maintain actors' identity and speech style with reference videos. Focus on achieving accurate lip sync and expressive facial animations, increasing overall video quality and engaging users in the generated content.

[5] CharacterGAN is a generative model designed for character animation with limited training samples. Handles (dis)occlusion using layering, adaptive scaling and mask connectivity. This model can efficiently create realistic animations for different characters and scales with larger datasets. Implementing a video generation system based on the principles demonstrated by CharacterGAN offers efficient content creation, realism, scalability, and diverse applications. It enables cost-effective video production with limited data, resulting in visually appealing and realistic animations and adapting to different character styles and scenarios in gaming, entertainment, e-learning and virtual assistants.

[6]Emotion Disentangling for Speech-Driven Facial Animation work presents an emotion disentangling (EDE) encoder for speech-driven facial animation. EDE separates emotion and content in speech and enhances emotional facial expressions using an emotion-driven feature fusion decoder. This approach uses a large 3D dataset of emotional speaking faces for training. When implementing a video generator project, the EDE coder approach offers insight into emotion disentanglement, content separation, and emotion-driven feature fusion. Consider using this concept to increase the emotional expressiveness of generated videos by separating emotion and content in speech, and train on large emotion-rich datasets for better realism and variety.

[7] Audio Representation Learning for Talking Head Generation aims to make the talking head generation robust to audio changes, including background noise and emotional tone. It uses separate sound representations for better accuracy of mouth movement in the presence of noise and emotional changes. In the development of a video generation project, the importance of system robustness to audio variations, including background noise and emotional tones, is highlighted by the findings from "Teaching Audio Representation for Talking Head Generation". The use of different audio representations is recommended to increase lip sync accuracy even when faced with audio distortion, thus contributing to the production of more consistent and lifelike video content.

[8] Vx2Text for Multimodal Text Generation, Vx2Text is a framework for generating text from multimodal inputs, including video, text, speech, and audio. It uses transformer networks to convert each modality into language embeddings, eliminating the need for ad-hoc cross-fusion modules. This approach outperforms the state of the art in various video-based text generation tasks. The knowledge from "Vx2Text for Multimodal Text Generation" is valuable for the video generator project. Favoring a multimodal approach that integrates video, text, speech, and audio inputs and uses transformer networks to convert input from modality to language. Eliminate the need for ad-hoc cross-fusion modules, simplifying the process and potentially outperforming existing methods in video-based text generation tasks.

[9] Emotion-Enhanced Speech-Driven 3D Facial Animation focuses on generating 3D facial expressions that correspond to speech content and emotions. It introduces an emotion extraction (EDE) encoder that separates emotion from speech content by cross-reconstructing speech signals with different emotion labels. An emotion-driven feature fusion decoder is

used to create a 3D talking face with enhanced emotions. Separate identity, emotional and content inserts enable the generation of controllable personal and emotional styles. This approach uses a large 3D emotional speaking face dataset (3D-ETF) for training, and experiments demonstrate its superiority in generating various facial movements.

[10] Robust Sound-Controlled Talking Head Generation addresses the challenge of creating controlled talking head generation robust to audio changes such as background noise and changes in emotional tone. It presents an audio representation learning framework that decomposes audio sequences into various factors, including phonetic content, emotional tone, and background noise. By conditioning the generated mouth movement on the representation of disconnected content, the model achieves significantly more accurate results in the presence of noise and emotional variations. This framework is shown to be compatible with existing state-of-the-art approaches and improves the robustness of talking head generation.

[11] Generating high-resolution images with label editing addresses the challenge of generating high-resolution photorealistic images using generative adversarial networks (GANs). It represents a variant of GAN with label adjustment, resulting in coherent 128x128 images. The study introduces new methods for image quality assessment and demonstrates that high-resolution samples carry more class-specific information and are more recognizable than low-resolution samples. These high-resolution samples outperform the low-resolution samples across 1000 ImageNet classes and show diversity similar to real ImageNet data. In the video generator project, the findings from "High-resolution Image Generation with Label Editing" highlight the importance of using GANs with label editing to create high-resolution photorealistic content. This approach, supported by effective image quality assessment methods, emphasizes the superiority of high-resolution samples in conveying class-specific information and their recognizability. By applying these principles, your video generator can create more compelling, diverse, and informative content in a variety of scenarios and classes.

III. Methodology

This video generator consists of modules like script generation, voice and image generation, character animation and finally integration of all the modules. Script generation provides a well-structured story. Image generation offers visual support and data representation. TTS converts the script into spoken words for audio narration. Character animation adds a dynamic and engaging dimension to the video. The architecture diagram is shown as Figure 1.

The first task would be script generation. Script generation is a crucial step in video production, shaping the story, dialogues and overall content of the video. A well-crafted script is the backbone of any video project, whether for educational purposes, entertainment, marketing or information dissemination. In the digital age, the demand for engaging and informative video content continues to grow, and artificial intelligence plays a key role in meeting this demand. Artificial intelligence (AI) technologies have revolutionized scripting. These technologies use natural language understanding and generation capabilities to streamline and improve the scripting process. AI-driven script generation offers several benefits, such as improving efficiency, reducing production time, and ensuring content consistency. Cohere AI is at the forefront of understanding and creating artificial intelligence-based natural language.

The platform seamlessly integrates advanced AI technologies with content creation, making it a valuable resource for script generation. Cohere AI works by leveraging the power of AI-driven language models. It begins with a well-defined prompt or description of the content to be included in the script. This can be a summary of the video's topic, key messages or an overview of the desired content. Once prompted, the Cohere AI language model processes the information and generates a script based on the input. It considers factors such as tone, style, and complexity to align the script with the intended audience and purpose of the video. This human oversight is critical to maintaining the script's coherence, accuracy, and relevance to the video's goals. Cohere AI offers a degree of customizability that allows users to fine-tune the generated script. This includes adjusting the tone, style and length to better suit the specific requirements of the video. Cohere AI's script generation capabilities offer significant benefits in terms of efficiency and content quality. It is especially valuable for a wide range of applications, including the creation of educational videos, marketing campaigns e-learning modules and more.

Applications of Natural language processing (NLP) using large language models (LLM) provide a modular and extensible way to build LLM strings and other tools that can be used to perform a wide variety of tasks such as text generation, language translation, and question answering. LangChain works by providing a set of standardized interfaces and external integrations for LLM, search, chains, agents, memory, and callbacks. These interfaces allow developers to easily combine different components to create complex NLP applications. Then the process continues to generate speech. TTS (text-to-speech) technology translates written text into spoken words. It is widely used in various applications and services to enable devices and software to communicate with users using natural speech similar to human speech.

Text-to-Speech (TTS) technology involves several stages of converting written text into spoken language. Initially, a TTS system analyzes the input text and breaks it down into smaller linguistic units, such as words, phrases or phonemes, which are the basic building blocks of speech sounds. This segmentation helps the system understand text structure and pronunciation. The text is then processed to take into account elements such as punctuation and formatting, ensuring that the spoken output remains coherent and understandable. Then the system performs a phonetic transcription and converts the text into sequences of phonemes. These phonemes serve as the basis for the construction of speech sounds and enable natural-sounding speech synthesis. Users are often given the option to choose from a range of voices or speech profiles in TTS systems, allowing the voice to be tailored to suit a particular preference or communication context.

The voice synthesis phase combines the phonetic information with the selected voice, deploying algorithms and models to generate the final speech output. Different systems can use recorded segments of human speech or use parametric synthesis to produce spoken word. After voice synthesis comes character animation. Character animation in the context of AI-driven applications involves creating and synchronizing a character's movements, facial expressions, and speech with content or dialogue.

Character animation usually starts with creating a 3D model of the character. This model includes skeletal structure and mesh that defines the shape of the figure. Motion sensing technology can be used for realistic movement. This involves recording the movements of human actors or objects and applying that data to the character's skeleton, allowing it to accurately replicate the movements. SadTalker, a deep learning model, enables the generation of realistic facial animations from a single image and an accompanying audio recording. It has its roots in Audio-Driven Facial Animation (ADFA), which uses machine learning to recognize the complex connection between audio input and corresponding facial expressions. An image of the character intended for animation is created to start the process.

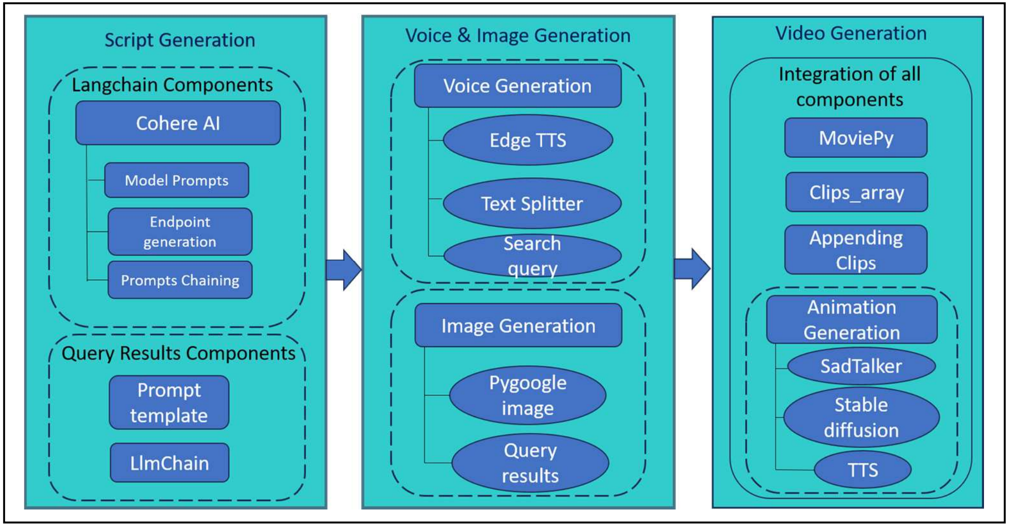


Fig. 1. Architecture Diagram

SadTalker uses this audio data to derive a series of facial motion coefficients. These coefficients encapsulate the range of facial movements associated with the character's speech. The final step involves using facial motion coefficients to create an impressive animation of the character's face. SadTalker can be used to generate realistic talking heads for video conferencing, making video calls more engaging and interactive. SadTalker generates high-quality animations of talking faces with realistic facial expressions and lip sync. Image generation involves scraping images to generate slideshow frames that are used in the video.

Generating images with Google Images and Python, 'pygoogleimage' involves creating a presentation by automatically loading images from Google Images, processing them with Python, and then integrating them into the presentation slides. Specific keywords are required to load images from google and integrate those to generate images. Integration of all modules would be the final step where all modules from script generation to character animation are integrated as one working web application and made available to the users to enhance the enchanting user experience with customizability of feature

IV. Proposed Work

Against the backdrop of remarkable technological advances, the way we create and disseminate information has undergone a profound transformation. The proposed work harnesses the power of advanced artificial intelligence technologies to facilitate the development of engaging tutorials, complete with character animations and information-rich imagery. What sets this tool apart is its high degree of customization, allowing users to fine-tune every element of their tutorials. Users can adjust creative components such as humor, depth of explanation, character appearance, and voice, and tailor tutorials to exact specifications. This video generator consists of modules like script generation, voice and image generation, character animation and finally integration of all the modules which is presented as a flow diagram in the Figure 2.

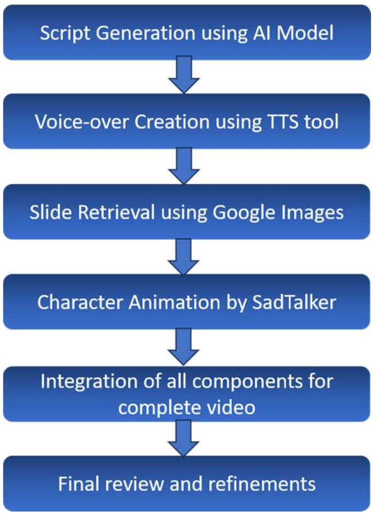

Script Generation using AI Model
Voice-over Creation using TTS tool
Slide Retrieval using Google Images
Character Animation by SadTalker
Integration of all components for complete video
Final review and refinements

Fig. 2. Flow Diagram

A. *Script Generation*

Text generation for video scripts with Cohere AI works through a sophisticated natural language processing system. Cohere AI uses advanced machine learning and language models to create detailed and understandable scripts. The process behind the model typically begins by gathering large text datasets from a variety of sources. These datasets are then preprocessed to remove noise and structure the text appropriately. Using state-of-the-art neural network architectures such as transformer models, Cohere AI trains its models on this pre-processed data. The training process involves optimizing model parameters to predict accurate results or understand the context and semantics of the text. Figure 3 depicts the flow diagram of the various processes of script generation.

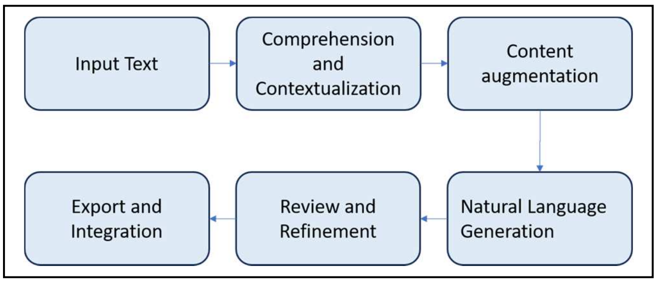


Fig. 3. Script Generation

- Input Text: Users provide input in the form of text instructions or descriptions of the content they want to include in their video script. These inputs serve as the basis for the script generation process.
- Comprehension and contextualization: Cohere AI's language model is designed to understand input text, taking into account context, desired tone, and any specific requirements. Uses contextual information to create coherent and relevant text.
- Content Augmentation: Cohere AI can augment the initial input by adding descriptive and informative elements to create a complete script. This extension ensures that the script is comprehensive and engaging.
- Natural language generation: An AI model uses its knowledge of natural language to generate human-like text. It structures sentences and paragraphs and ensures the flow and readability of the script.
- Review and refinement: The generated script is subjected to review and refinement. Users have the flexibility to modify and fine-tune the script to match their specific needs and preferences. This review process allows users to add a personal touch and further customize the content.
- Export and integration: Once the script is to the user's satisfaction, it can be easily exported and integrated into the video production workflow. The generated storyboard serves as the basis for the narration of the video, whether it is used for tutorials, promotional content or any other type of video.

Cohere AI text generation for video scripts streamlines the content creation process, making it more efficient and accessible for a wide range of applications, from educational content to marketing materials. The ability to understand context, expand on ideas, and create natural-sounding text ensures that the resulting videos are engaging and informative.

B. *Voice Generation*

Voice generation with Edge Text-to-Speech (TTS) is a critical element in the video generation process. Edge TTS is a technology that enables the conversion of written text into natural-sounding audio, increasing the overall quality and engagement of video content. Edge Text-to-Speech (TTS), also known as on-device TTS, is a system that converts written text into natural-sounding spoken language. Text Analysis, Language Modeling and Phonetic Transcription are the main elements of Edge TTS depicted in Figure 4. Unlike cloud-based TTS systems that require an internet connection to function, Edge TTS works directly on the device, making it suitable for scenarios where internet connectivity may be limited or privacy concerns are relevant. Text input: Users provide written text that they want to convert to speech. This text may be in different languages and may contain specific instructions for formatting or pronunciation.

Text Analysis
Edge TTS analyzes the input text to extract meaningful linguistic information
Language Modeling
Language models ensure contextual speech synthesis with appropriate grammar and intonation
Phonetic Transcription
Phonetic transcription helps in mapping text to phonemes for accurate pronunciation

Fig. 4. Text-to-speech elements

The various steps that are involved in converting the text generated from Cohere AI into speech are dealt below.

- On-device processing: Text input is processed locally on the user's device, such as a smartphone, computer, or embedded system. Edge TTS systems are designed to run on the device's hardware, which includes the device's CPU and memory.
- Voice Choice: Edge TTS usually offers a choice of voices, each with their own unique characteristics. Users can choose the voice that best suits the context and audience for their speech synthesis.
- Natural language processing: The TTS system uses advanced natural language processing (NLP) techniques to analyze the text you enter. This includes tasks such as text normalization, tokenization and linguistic analysis to understand the structure and meaning of the text.
- Phonetic transcription: The system converts the processed text into a phonetic expression. This representation divides text into phonemes, which are the smallest units of sound in a language. This step is critical to generating accurate and natural-sounding speech.
- Voice Synthesis: The phonetic representation along with the selected voice is used to generate speech. The Edge TTS system uses complex algorithms and models to produce speech that is natural and understandable. Depending on the quality of the system, the generated speech can be remarkably realistic.

- Prosody and intonation: To make synthesized speech sound natural, Edge TTS includes prosody and intonation. These aspects of speech include variations in pitch, speed, rhythm, and stress to convey meaning and emotion. Correct prosody and intonation increase the expressiveness of speech.
- Audio Rendering: The synthesized speech is then converted into a sound wave. This waveform represents the actual speech sound. Signal processing techniques are used to ensure high quality sound without distortion or artifacts.
- Output: Generated speech can be output through the device's speakers, headphones, or other audio output options. Users can also choose to save synthesized speech as audio files for later use.

C. Image Generation

Image generation is an important aspect of developing excellent presentations for a variety of objectives, including corporate meetings, educational lectures, and marketing materials. These images can be incorporated into the presentation that is used in video and used for explaining the concept in a much clearer manner. Visuals, particularly images, are important in delivering information and increasing audience engagement. Google Images is an excellent source to find a wide assortment of images for utilization in presentations. The PyGoogleImage library serves as a tool to facilitate the seamless integration of Google Images into the image generation process.

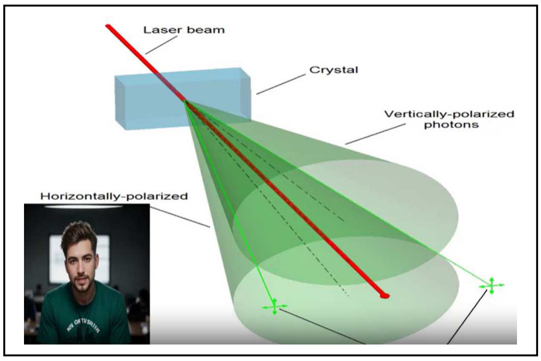


Fig. 5. Sample Image of Video

The methodology for generating images using Google images and pyGoogleImage library consists of several key steps:

- Definition of content: The content that needs to be presented in the slides are defined initially. This content can include textual information, graphics and, most importantly, images.
- Image Search: PyGoogleImage library is used to search Google Images for relevant images. Using the keywords and criteria to filter results, the matching content is inserted inside the slides.
- Image Download: Selected images from search results are downloaded using PyGoogleImage and saved to local storage for later use in slides.
- Creating Slides: Creation of slides is done using a library python-pptx. The slide layout, fonts, colors and any other elements such as headers and footers are defined.
- Insert images: Incorporating downloaded images into slides in appropriate places as in Figure 5. It is made sure that the images complement the textual content and contribute to the overall message.
- Addition of text and content: The necessary text and content are written and added to the slides. This may include descriptions, bullet points or captions that provide context for the images to convey the intended message.
- Review and Finalize: Careful review and refinement of generated images are ensured of accuracy, consistency, and overall quality. Necessary adjustments are made.
- Export: The generated presentation is exported or saved in the desired format, such as .pptx or PDF, for further process.

D. Character Animation

In the context of video generators, character animation can transform static or monotonous content into dynamic and engaging presentations. It provides a way to breathe life into the videos, making them more memorable and impactful. Whether used for marketing, education, entertainment or any other purpose, character animation is a powerful tool in the video generation process.

SadTalker is based on the principles of Audio-Driven Facial Animation (ADFA), which uses machine learning to decipher the complex relationship between audio stimuli and facial expressions. This technology makes it possible to create expressive, realistic character animations for a wide range of applications, including video games, movies and virtual reality experiences. SadTalker is the ultimate character animation solution. It bridges the gap between audio input and facial expressions, giving characters a more alive and engaging presence. By processing audio recordings and using facial landmarks, SadTalker generates facial motion coefficients, allowing dynamic synchronization of characters' facial animations with speech.

SadTalker represents a remarkable technological advance in character animation. It simplifies the process of synchronizing facial expressions with audio input, providing a versatile tool for various industries. Whether it's for creating dynamic characters for video games, movies, or virtual reality applications, SadTalker enables creators to enhance the expressiveness and realism of characters.

- Facial Landmarks: SadTalker extracts basic facial landmarks from the provided image and identifies key features such as eyes, nose and mouth. These landmarks serve as reference points for generating facial animations. In the sample character animation in Figure 6 the lip reference points are altered according to the audio generated.
- Audio Input: The audio generated by Edge TTS is used for the character's speech to start the animation process. The sound provided has to be clear and understandable, as it serves as the basis for generating facial movement coefficients.
- Facial Motion Coefficients: SadTalker processes the audio recording and creates facial motion coefficients. These coefficients represent the complex facial movements that match the

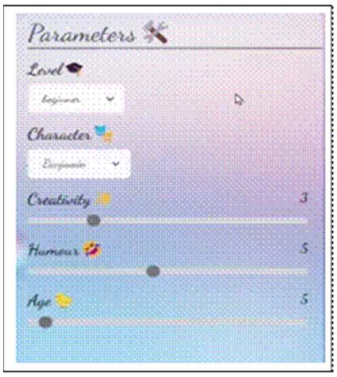
character's speech. The technology essentially deciphers the nuances between voice inflections and facial expressions.

- Character Animation: Once the facial motion coefficients are generated, users have the flexibility to animate their character. The basic principle behind the SadTalker is Stable Diffusion. Sample animation of character with the fusion of lip movements is shown in Figure 6. These coefficients serve as the guiding principles for the character's facial animation, ensuring that facial expressions are in sync with the audio.

Fig. 6. Sample Character Animation

E. *Customizability*

The Customizable Video Generator project represents a significant leap forward in content creation. This innovative tool is designed to cater to a diverse spectrum of users and allow them to customize video content according to their specific requirements. A key feature of this project is the unprecedented level of customization it offers, allowing users to fine-tune various aspects of their videos, including difficulty level, character age, humorous content and creativity depicted in Figure 7.



Parameters 🛠

Level 🎓

Beginner ⌄

Character 🎭

Benjamin ⌄

Creativity 3

Humour 5

Age 5

Fig. 7. Customizable Parameters

Customizable parameters are explained below:

- Difficulty level customization: Users can adjust the difficulty level of the content to match the knowledge and expertise of their target audience. This feature ensures that content remains accessible and engaging.
- Choosing the characters & Age of Characters: The video generator allows users to choose the age of the characters in the video, allowing them to connect with audiences of different age groups.
- Curating content with humor: Humor is a powerful tool in video content. Users have the ability to control the level of humor in their videos to be lighthearted or serious as needed.
- Creativity Tuning: By allowing users to customize the level of creativity in the content, the video generator enables creative expression. This function is extremely helpful for artistic and inventive projects.

Customizability in the video generator enhances the viewing experience by allowing viewers to personalize content based on their unique preferences, needs and demographics. This not only makes the content more engaging, but also ensures that it is relevant and applicable to a wider and more diverse audience.

V. Conclusion

The video generator redefines content creation with its customizability, linguistic foundation, audio enhancement and character animation. It allows creators to cater to different preferences and educational needs and create engaging and informative videos in a variety of fields, from education to marketing. As we move forward in this era of technological advancement, video generator is a testament to our ability to leverage AI and technology for more engaging and personalized content. This revolutionary solution stands out from other options with a wide range of customizable capabilities, making it an exceptional replacement for increasing efficiency and productivity. It paves the way for a more interactive, personalized and engaging future in video generation.

VI. Future Enhancement

In the future of video generation, the possibilities for improvement are vast and dynamic. Project development can benefit from adaptive learning mechanisms that allow the system to evolve and adapt to changing user preferences and content requirements over time. Multilingual support will open the door to a global audience and enable the generation of video content in different languages. The integration of informative video summaries simplifies the accessibility of extensive content and offers concise and engaging insights. Surround sound, enhancing the audiovisual experience, transports the user to an immersive and high-quality sound environment. Live streaming, Augmented Reality (AR) and Virtual Reality (VR) generation enable interactive and immersive content to be created in real-time for live events, immersive experiences and augmented reality applications. Together, these advances promise to reshape the landscape of video generation, enriching the content creation process and expanding its reach in a diverse and interconnected digital world.

References

[1] Ali Koksal, Kenan Emir Ak, Ying Sun, Deepu Rajan and Joo-Hwee Lim, “Controllable Video Generation with Text-based instructions”, IEEE transactions on multimedia, 2023.

[2] Zhengxiong Luo, Dayou Chen, Yingya Zhang, Yan Huang, Liang Wang, Yujun Shen, Deli Zhao, Jingren Zhou and Tieniu Tan, “Decomposed Diffusion Models for High-Quality Video Generation”, CVF Conference arxiv: 2303.08320, 2023.

[3] Yaosi Hu, Chong Luo and Zhenzhong Chen, “Controllable Image-to-Video Generation With Text Descriptions”, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022.

[4] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li and Li Cheng, “Generating Diverse and Natural 3D Human Motions From Text”, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022.

[5] Balamurugan Thambiraja, Ikhsanul Habibie, Sadegh Aliakbarian, Darren Cosker, Christian Theobalt, Justus Thies, "Personalized Speech-driven 3D Facial Animation",Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 20621-20631, 2023.

[6] Tobias Hinz, Matthew Fisher, Oliver Wang, Eli Shechtman, Stefan Wermter, "Few-Shot Keypoint Character Animation and Reposing",Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp. 1988-1997, 2022.

[7] Ziqiao Peng, Haoyu Wu, Zhenbo Song, Hao Xu, Xiangyu Zhu, Jun He, Hongyan Liu, Zhaoxin Fan, "Speech-Driven Emotional Disentanglement for 3D Face Animation", Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 20687-20697, 2023.

[8] Gaurav Mittal, Baoyuan Wang, "Animating Face using Disentangled Audio Representations", Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp. 3290-3298, 2020.

[9] Xudong Lin, Gedas Bertasius, Jue Wang, Shih-Fu Chang, Devi Parikh, Lorenzo Torresani, " End-to-End Learning of Video-Based Text Generation From Multimodal Inputs",Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7005-7015, 2021.

[10] W. Hong, M. Ding, W. Zheng, X. Liu, and J. Tang, "Cogvideo: Large-scale pretraining for text-to-video generation via transformers," arXiv preprint arXiv:2205.15868, 2022.

[11] R. Villegas, M. Babaeizadeh, P.-J. Kindermans, H. Moraldo, H. Zhang, M. T. Saffar, S. Castro, J. Kunze, and D. Erhan, "Phenaki: Variable length video generation from open domain textual description," arXiv preprint arXiv:2210.02399, 2022.

[12] U. Singer, A. Polyak, T. Hayes, X. Yin, J. An, S. Zhang, Q. Hu, H. Yang, O. Ashual, O. Gafni et al., "Make-a-video: Text-to-video generation without text-video data", arXiv preprint arXiv:2209.14792, 2022.

[13] T. Hoppe, A. Mehrjou, S. Bauer, D. Nielsen, and A. Dittadi, "Diffusion models for video prediction and infilling," arXiv preprint arXiv:2206.07696, 2022.

[14] C. Nash, J. Carreira, J. Walker, I. Barr, A. Jaegle, M. Malinowski, and P. Battaglia, "Transframer: Arbitrary frame prediction with generative models", arXiv preprint arXiv:2203.09494, 2022.

[15] W. Yan, Y. Zhang, P. Abbeel, and A. Srinivas, "Videogpt: Video generation using vq-vae and transformers", arXiv preprint arXiv:2104.10157,2021.

[16] C. Wu, J. Liang, L. Ji, F. Yang, Y. Fang, D. Jiang, and N. Duan, Nuwa, "Visual synthesis pre-training for neural visual world creation", arXiv preprint arXiv:2111.12417, 2021.

[17] W. Menapace, S. Lathuiliere, S. Tulyakov, A. Siarohin, and E. Ricci, "Playable video generation," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10 061–10 070, 2021.

[18] J. Huang, J. Liao, and S. Kwong, "Semantic example guided image-to-image translation," IEEE Transactions on Multimedia, vol. 23, pp.1654–1665, 2021.

[19] J. Xie, X. Chen, T. Zhang, Y. Zhang, S.-P. Lu, P. Cesar, and Y. Yang, "Multimodal-based and aesthetic-guided narrative video summariza tion," IEEE Transactions on Multimedia, pp. 1–15, 2022.

[20] Chuan Guo, Xinxin Zuo, Sen Wang, Xinshuang Liu, Shihao Zou, Minglun Gong, and Li Cheng, "Action2video: Generating videos of human 3d actions", International Journal of Computer Vision, pages 1–31, 2022.

[21] Mathis Petrovich, Michael J Black, and Gul Varol, "Action-conditioned 3d human motion synthesis with transformer", vae. 2021.

[22] Zhenyi Wang, Ping Yu, Yang Zhao, Ruiyi Zhang, Yufan Zhou, Junsong Yuan, and Changyou Chen. Learning diverse stochastic human-action generators by learning smooth latent transitions. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, pages 12281–12288, 2020.

[23] Ping Yu, Yang Zhao, Chunyuan Li, Junsong Yuan, and Changyou Chen, "Structure-aware human-action generation", In European Conference on Computer Vision, pages 18–34. Springer, 2020.

[24] Yu Tian, Jian Ren, Menglei Chai, Kyle Olszewski, Xi Peng, Dimitris N Metaxas, and Sergey Tulyakov, "A good image generator is what you need for high-resolution video synthesis", arXiv preprint arXiv:2104.15069, 2021.

[25] Ruozi Huang, Huang Hu, Wei Wu, Kei Sawada, Mi Zhang, and Daxin Jiang, "Dance revolution: Long-term dance generation with music via curriculum learning", In International Conference on Learning Representations (ICLR), 2021.

[26] Anindita Ghosh, Noshaba Cheema, Cennet Oguz, Christian Theobalt, and Philipp Slusallek, "Synthesis of compositional animations from textual descriptions", arXiv preprint arXiv:2103.14675, 2021.

[27] Saeed Ghorbani, Kimia Mahdaviani, Anne Thaler, Konrad Kording, Douglas James Cook, Gunnar Blohm, and Nikolaus F Troje, "Movi: A large multipurpose motion and video dataset", arXiv preprint arXiv:2003.01888, 2020.

[28] Uttaran Bhattacharya, Nicholas Rewkowski, Abhishek Banerjee, Pooja Guhan, Aniket Bera, and Dinesh Manocha, "Text2gestures: A transformer-based network for generating emotive body gestures for virtual agents", in IEEE Virtual Reality and 3D User Interfaces (VR), pages 1–10. IEEE, 2021.