Gen-AI Powered Text and Audio to Video Generation System

Dr.CV Madhusudhan Reddy, Dr.G K V Narasimha Reddy, Dasari Vinod, Golla Vinod Kumar, Shaik Mohammad Junaid, Polakal Dinesh Kumar.
Dept. Of Computer Science and Engineering (Artificial Intelligence), St. Johns College of Engineering and Technology, Yemmiganur, 518301, India.

***Abstract*—The need for systems that can automatically create multimedia content is getting bigger and bigger. This is because the way we communicate with each other digitally is changing fast. Normally making a video takes a lot of time and effort. You have to write a script record it edit it and make sure everything is in sync.**

This system is different. It uses something called Generative Artificial Intelligence to turn text and audio into videos. The system can understand what the text is saying because it uses Natural Language Processing. It can also create pictures. Sounds using special models. The system can even add subtitles to the videos. The system is really good, at making videos from text and audio.

Put them in the right place in the video. It can also make the video look good without needing a lot of help from people. The good thing about this system is that it can make videos quickly and the videos are quality. The people who made this system tried it out. I found that the Generative Artificial Intelligence system worked well for things like education and marketing and making media. The Generative Artificial Intelligence system is really good at making a lot of videos without needing a lot of help from people.

The Generative Artificial Intelligence system is also good at making sure the videos are of quality and that they make sense to people who watch them.

Keywords: Generative AI, Text-to-Video, Audio-, to-Video, NLP, Deep Learning, Multimedia Automation

I. INTRODUCTION

People are using multimedia communication a lot nowadays. It is a part of the digital world we live in. Websites that teach us things groups that sell us stuff and networking sites all use videos to talk to us. Multimedia communication is really important, for all these websites. They need multimedia communication to work properly.

Traditional video production is a lot of work. It involves steps like writing a script doing the narration designing the pictures editing the animation and rendering the video.

The rapid growth of digital communication platforms has greatly increased the need for automated systems that create multimedia content. Video content has become one of the most effective ways to communicate in education, marketing, social media, and entertainment. However, traditional video production involves many manual tasks, such as writing scripts, recording narration, designing scenes, editing animations, synchronizing subtitles, and rendering the final product.

These tasks require specialized skills, professional tools, and a significant amount of time, making video production resource-heavy and expensive. Recent developments in Generative Artificial Intelligence (Gen-AI) have allowed for smarter automation of content creation tasks. Deep learning models, like transformer-based architectures, enhance understanding of context in text. Generative models can create realistic images and speech. Using these technologies, this project proposes a Gen-AI Powered Text and Audio to Video Generation System, which automatically turns written or spoken input into structured and synchronized video output.

The system uses Natural Language Processing (NLP) for understanding meaning, AI-based visual generation for creating scenes, neural text-to-speech synthesis for narration, subtitle synchronization modules, and an automated engine for rendering video. By combining these parts within a modular design, the proposed system reduces the need for manual work, speeds up content production, and increases scalability. This solution is especially

beneficial for educators, digital marketers, content creators, and organizations looking for efficient multimedia automation. The project shows how generative AI can change traditional video production processes into smart, automated systems.

Recently Artificial Intelligence has gotten a lot better Generative AI. This means we can now automate tasks that used to take a lot of time. The Transformer model is really good, at understanding what the text is saying. Generative diffusion and adversarial networks can create pictures that look real. When we put these technologies together we can make systems. These systems can take text or speech. Turn it into a multimedia show that makes sense.

This research aims to develop a modular Gen-AI video automation system.

II. LITERATURE SURVEY

Video is a part of the internet these days. We use video to learn things on websites that teach us stuff. Companies use video to sell things to people in online groups. Even social media sites, like Facebook use a lot of video. Making videos the old way is really work. Video production is a lot of work. It involves steps like writing a script recording a voiceover, designing images editing animations and rendering.

Automated multimedia content generation has been a deal in research for a while now because it can really change the way we create digital content. On people made systems that used templates and rules to generate content but these systems were not very good at understanding what the content was about and could not adapt to different situations. They used the old templates and animation sequences over and over which made them bad at creating content that was dynamic and aware of its context.

Then deep learning came along. Things started to get better. Generative Adversarial Networks, which were introduced by Goodfellow and his team allowed for a way of training models where a generator and a discriminator worked against each other. This made it possible to create images from just a few ideas. People started using GANs a lot for image generation. It laid the groundwork for creating visual content that could be controlled and changed.

The Transformer architecture, which was proposed by Vaswani and his team was a change for natural language processing. It used something called self-attention, which lets models look at pieces of text and understand how they are related. Transformers have been used for text generation understanding what text means and tasks that involve types of media like creating images from text. These models made the content that was generated realistic by understanding how the words in the text were related to each other.

III. SYSTEM ARCHITECTURE

The proposed system is based on an modular structure, which includes the following modules:

1. User Interface Layer – This module accepts text or audio inputs from the user. It is like a window where the user can type or talk to the system.
2. Input Processing Module – This module is responsible for cleaning and checking the data that comes in.
3. The NLP Engine is a part of the system that breaks down text into pieces finds important words understands what things mean and figures out what is happening in a scene. The NLP Engine does all these things to help the Gen-AI video automation system work properly. The NLP Engine is really good at understanding text and speech.

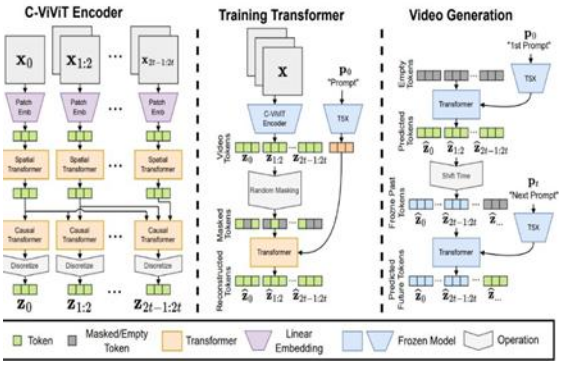

Fig no 2: System Architecture

4. The Visual Generation Module makes pictures using a computer. These pictures are related to what the NLP Engine found out. The Visual Generation Module generates images that make sense and look nice.
5. The Audio Synthesis Module creates audio that sounds like a real person talking. It uses computer models to make the audio sound natural. The Audio Synthesis Module is good at making audio that sounds like it was done by a being.
6. The Subtitle Synchronization Module makes sure the text and audio go together. It takes the audio made by the Audio Synthesis Module and the text. Makes

sure they are in sync. The Subtitle Synchronization Module does this so that the audio and text match up perfectly.

7. Video Rendering Engine – This module combines frames, audio, transitions and subtitles. It is like putting all the pieces of a puzzle to make a complete video.

8. Database Layer – This module stores. References to the generated content. The Database Layer is like a library where all the videos and information are stored.

The proposed modular structure of the Gen-AI video automation system is highly scalable and adaptable to integrations with Artificial Intelligence models. This means the Gen-AI video automation system can be easily changed and improved in the future. The Gen-AI video automation system is a tool, for making video production simpler and easier.

IV. SYSTEM FLOW CHART EXPLANATION

The workflow of this thing starts with the user putting in some words or talking into it. If the user talks into it a special program turns what they say into words on the screen. Then the language part of the program breaks down the words into things that make sense and divides the content into parts that go together. These parts are then connected to pictures that the computer makes through computer programs.

At the time the computer generates a voice that goes along with the pictures. The timing of the words on the screen is matched with the timing of the voice. Finally the program that makes the video combines the pictures, sound, movement between pictures and words, on the screen into a video file that can be played.

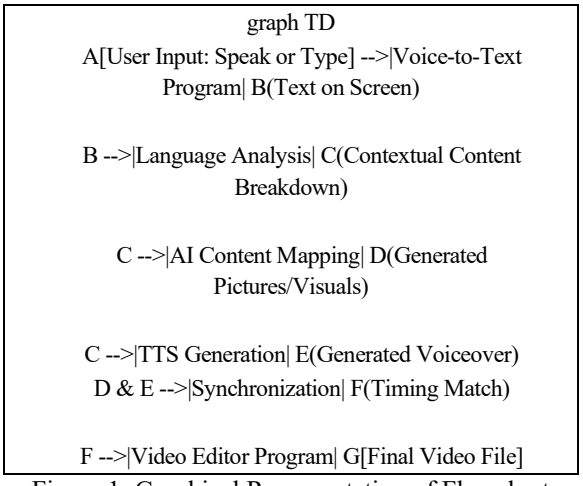

Fig no 1: Graphical Representation of Flowchart

V. METHODOLOGY

The process comprises seven major phases:

1. Input Validation and Preprocessing
2. Speech-to-Text Conversion (if necessary)
3. Tokenization and Semantic Embedding
4. Scene Segmentation and Keyword Mapping
5. Visual Frame Generation via AI models
6. Neural Text-to-Speech Synthesis
7. Video Stitching and Rendering

Mathematically, the process can be represented as:

Text Input → Tokenization → Embedding Vector → Scene

Clustering → Visual Mapping + Audio Mapping → Final Video Output.

This process ensures semantic alignment and synchronization of multimedia elements.

VI. IMPLEMENTATION

The system is built using Python and the Flask framework. We use a few libraries to get things done. NumPy is used for numbers and math. Pandas is used for working with data. OpenCV is used for images. We use machine learning libraries to run our AI models. We store information in SQLite or MySQL.

To run the system you need a computer with least 8GB of memory and 256GB of storage space. It is also an idea to have a special graphics card to make things faster. The system is, on the web so you can use a browser to get to it. You can access the system from any computer with a browser. The system uses Python and the Flask framework to make it all work. NumPy, Pandas, OpenCV and machine learning libraries are all parts of the system.

VII. ADVANTAGES

- Reduces production time and cost
- Ensures automation and scalability
- Maintains semantic synchronization
- User-friendly interface
- Applicable in education, marketing, and content automation industries

VIII. CONCLUSION

It should be noted that the "Code Quality Fixer" project effectively demonstrates the capability for improvement in good coding practices.
The new Gen-AI Powered Text and Audio to Video Generation System is a way to make multimedia production easier. This system is good because it has useful tools like NLP, generative models, speech synthesis and rendering all in one place.

The Gen-AI Powered Text and Audio to Video Generation System makes it easier to produce videos. The people who made the Gen-AI Powered Text and Audio to Video Generation System have plans for it. They want to add languages to the Gen-AI Powered Text and Audio to Video Generation System.

They also want the Gen-AI Powered Text and Audio to Video Generation System to be able to understand emotions when it is reading something loud.

The Gen-AI Powered Text and Audio to Video Generation System will be able to make videos faster and better.

They will put the Gen-AI Powered Text and Audio to Video Generation System on the cloud so people can use it from anywhere.
The Gen-AI Powered Text and Audio to Video Generation System will have options so people can make it work just the way they want.
The Gen-AI Powered Text and Audio, to Video Generation System is going to change the way people make content.

static code analysis techniques by employing AI code reviews. With the increasing complexity of completing the software development process daily, the need to ensure proper code quality is also increasing. This, in turn, significantly affects the code in terms of its performance and efficiency. Although employing manual code review techniques might be helpful when used in the proposed system, it would take longer owing to the impractical nature of the entire process in educational institutions because of the complexity faced in the process itself. Thus, an efficient method for determining code quality is proposed.

It accepts user code submissions and analyzes the code in a structured manner without executing the code in any program state. Hence, it processes the code safely while detecting syntax, structural complexity, and maintainability-related issues in the code. In addition, the quality scoring feature provides an idea of the condition of the code developed by developers and students.

Most importantly, the integration of AI-related modules will allow intelligent suggestions to be provided to programmers in future improvements of the code and will even suggest optimized versions of the code. The evaluation results show that the code effectively detects issues in code quality and allows the user to correct inefficient and incorrect code patterns in programming. For this purpose, it may help improve coding skills using better suggestions and improvements than before. It would also allow the system to be flexible to extend in the future, allowing it to become an academic and professional development system. In conclusion, Code Quality Fixer helps fill the gap between manual code reviews and automated code evaluations by creating a realistic and easy-to-use tool/platform to improve programming quality. This project helps improve software development efficiency and simultaneously assists learners who are also programmers. This project contributes to high efficiency in software development and simultaneously assists learners who are also programmers.

IX.FUTURE WORK

Although the main goal of the Code Quality Fixer tool is to perform code quality analysis in an automated way, there are opportunities to improve this tool further. The way to improve this tool further is considered to be an option because it supports different programming languages, such as C++, JavaScript, and Go, and it can be further improved in the future to support more programming languages. In other words, it supports different types of programming environments.
Other features that can be added in the future include readability and code smell.
The detection mechanisms could be used to perform some in-depth analysis on the code as well. In addition to this, the inefficient code duplication,

inefficient loops, and design patterns in code could also be used to enhance the overall quality of the evaluation process

Except for the said feature, it has also been identified that there exist some possibilities for incorporating vulnerabilities in terms of evaluation of the injection vulnerabilities and the resources. This again demonstrates that the identified scope for the tool is applicable for secure code as well.

Moreover, it would also offer a facility to expand it for adding persistent storage as well as a user, wherein the programmer would have a chance to see how his improvement is over time, along with a report of the quality of code he is writing. Another way through which this system could improve is in relation to the improvement of the reasoning ability of the AI, in order to provide more specific optimization suggestions as well as reduce the inaccuracies involved in the process itself. Also, real-time collaboration would be added to the system, which could be very helpful to an education center like schools. Moreover, the system itself would be a Software as a Service.

ACKNOWLEDGEMENT

would like to extend our heartfelt thanks to our project guide and the faculty members for their continuous support and suggestions in the development of the Code Quality Fixer Project. This research work would not have been possible without their vested interest and contributions in a deep and impactful manner. There are moments in the development of the project, where the project would not have been developed without the "enhance and proper suggestions." We extend our heartfelt thanks to the Department of Computer Science and Engineering for providing us with the necessary infrastructural facilities and academic environment to develop the project successfully. The support and assistance we received from our fellow course mates, who too have been through the intense testing period, are invaluable. Lastly, we would like to place on record our appreciations to all the authors, developers, and contributors for their assistance in the successful implementation of the static code analyzer and the inclusion of the artificial intelligence concept in the project.

decision support systems might potentially extend this utility into a clinical environment; however, this would also require a substantial safety infrastructure.

REFERENCES

[1] I. Goodfellow et al., Generative Adversarial Networks, 2014.
[2] A. Vaswani et al., Attention Is All You Need, 2017.
[3] J. Ho et al., Denoising Diffusion Probabilistic Models, 2020.
[4] Research on Neural Text-to-Speech Systems.
[5] Advances in Transformer-Based Language Models.