Full Stack AI Short Video Generator

Shylaja L N[1], Anil Reddy[2], Dheeraj Jain B[3], Raman Gowda Police Patil [4] and Ravikeerthi M Y[5]
[1]*Associate Professor, Artificial Intelligence & Machine Learning, Bahubali College of Engineering Shravanabelagola,*
[2345]*Computer Science and Engineering, Bahubali College of Engineering Shravanabelagola*

***Abstract*—This project focuses on building an AI-powered Full Stack platform for creating short, attention- grabbing videos in real-time from simple text inputs. By incorporating NLP technology with a pre-trained model named Gemini, the system will intelligently break down the input data into an apt video script. A clear-cut layered structure would be used here, with separate layers for the frontend, backend, AI processing, and storage systems. Moreover, video editing operations such as trimming, transitions, or merging videos can also get accomplished with the aid of FFmpeg libraries. A prompt engineering technique can also get implemented for more accurate outputs."**

The frontend part, built using React and Next.js, is an easy- to use interface where the user can provide inputs, preview output, and edit their videos using templates, fonts, and music, among other features. The backend part of the system will be supported using Convex for enabling live database functionality as well as efficient storage functionality. The system can be scalable, efficient, and easy to use, and may have uses related to content, marketing, and educational as well as entertainment sectors. Overall, this system can provide a very efficient solution to create professional short videos very easily.

***Keywords*—Artificial Intelligence, Short Video Generation, Text-to-Video, Natural Language Processing, Full Stack Development, Gemini AI, FFmpeg, React.js, Next.js, Web-Based Multimedia Keywords— Objection detection, YOLO, Visually impaired, Real-time systems, Python, Deep learning.**

I. INTRODUCTION

The project "Full Stack AI-Based Short Video Generator" simplifies the generation of short-form video content through the intelligent use of Natural Language Processing, classical algorithms, and modern Web technologies. The consumption of short and engaging video content has increased significantly in the last few years, especially on digital platforms that promote quick visual storytelling. This increase in demand has created a deep requirement for intelligent systems which can generate high-quality video content with minimal time and maximal consistency. The proposed system is envisioned to assist users, specially content creators and marketers, in generating customized short videos from simple text prompts, without requiring technical knowledge in video editing and/or multimedia software. These systems generally use Natural Language Processing techniques at their core to analyze the input provided by the user and extract the necessary information, such as context, emotion, and intent. This extracted information is then used to create a meaningful and well-structured script for the video, meeting the user's requirements. In this way, the system eliminates manual effort at the content planning stage of script creation and ensures consistency in storytelling across different videos that are generated. This generated script is later refined using a pre-trained AI model called Gemini to pick or produce related multimedia materials such as pictures, video highlights, and audio tracks. To ensure a smooth flow of these multimedia materials, the system incorporates Dijkstra's algorithm to order these multi media components optimally. Moreover, effective management of data is realized using the B-Tree and AVL Tree data structures to enable efficient indexing, retrieval, and ordering of video components. The backend infrastructure of the system is written in the Python programming language and hosted on the Convex platform for the purpose of real-time database functionality and optimal processing of video generation functions. The processing of video generation functions like trimming videos, adding transitions to videos, and the combination of multiple videos is achieved with the FFmpeg library. The front- end of the application is created with the Next.js/React framework for developing an easy-to-use interface for the application. The system is also capable of handling various video for mats and is developed to scale based on user demands; hence, it is applicable in different fields such as making learning videos, marketing campaigns, and automation of various

video contents. Through the integration of AI concepts, Algorithmic Logic, and full-stack development within a single platform, this project offers a useful and efficient method for making current video contents. The suggested system illustrates how AI concepts can be efficiently utilized to automate the pro duction of various multimedia files and provides a basis for research and development in AI-based short video production systems.

II. LITERATURE SURVEY

In 2022, Wang et al. introduced the concept of TransPixar, an innovative method to push text to video synthesis with transparency support via RGBA video output. Unlike traditional video synthesis models based on an RGB format, TransPixar introduces an alpha channel for generating trans parent graphics such as smoke, glass, or dynamic lights. The model consists of a diffusion transformer with LoRA fine-tuned networks as well as special attention to alpha.

Experimentation revealed an improvement in motion precision as well as alignment with respect to an RGB- alpha channel, making this method valuable for video editing, animation, or visual effect algorithms. The paper lays down an excellent foundation to increase realism to video AI synthesis models.

[1] In 2022, MEVG, a multi-event video generation model based on diffusion, was proposed. It was able to generate videos from a number of text descriptions without the need for large fine-tuning. It applied the concept of last-frame aware diffusion to make sure that the visual consistency in the events is intact. Also, the model applied the use of the large language model- based prompt generator to break down the narratives. This model was able to perform remarkably well in the generation of videos from multi-event text descriptions.

[2] In a related work by Nag et al. in 2023, they examined an AI-enabled short video generation platform that combined the use of generative AI, deep learning algorithms, and large language models. The proposed solution aimed to automate content creation. The paper reviewed popular models including GANs, transformer models, and diffusion models that were applicable to text-to-video models. The researchers discussed how AI contributes to minimizing manual labor during video production. This work verifies that AI-enabled platforms can be applicable in generating short videos effectively.

[3] Wang, H., Peng, Y., Tan, K. H. (2023) proposed a scalable text-to-video generation model using a huge amount of text-less video data for pre-training. The proposed model is efficient in terms of cost, as it does not entirely rely on text video paired data sets due to a multi-stage training method and 3D U-Net temporal modeling approach. The proposed model is capable of producing videos with higher resolution, length, and temporal consistency. Nowadays, scalable model training strategies are necessary for next-generation artificial intelligence-based video generation models.

[4] Zhou et al. (2023) made an overall survey on generative AI and large language models applied to video generation, understanding, and streaming. This work discussed how LLMs improve the comprehension of videos by extracting semantic information such as objects, actions, and events. Similarly, the authors discussed the role that generative AI plays in enhancing video streaming by optimizing quality while reducing bandwidth consumption. This survey provides a far-reaching theoretical background for understanding how to integrate LLM-driven intelligence into an automated video generation platform.

[5] Weng et al. (2024) proposed ART-V, an auto- regressive text to-video generation framework based on diffusion models. Unlike the one-shot generation methods, ART-V generates video frames autoregressively while conditioning the previous generated frames, which is possible to generate videos of arbitrary length. Temporal attention mechanisms and adaptive noise scheduling were introduced into the model to improve the motion realism and content coherency. The experimental results demonstrated superior performance in motion quality and aesthetic coherency, which proved that the approach had a very good fit for high-quality video synthesis tasks.

[6] Fei et al. (2024) proposed the ”Dysen-VDM,” which is a diffusion framework for text-to-video that utilizes large language models. The proposed dynamic scene manager takes into account the information related to motion from the text

description and helps in the diffusion process through a dynamic scene graph. Experimental results showed better temporal consistency and action disorder in the output videos. This work confirms the effectiveness of incorporating LLM based reasoning into the diffusion-based video generation model.

[7] Li et al. (2024) introduced VideoGen, a reference- guided la tent diffusion method for high-definition text- to-video synthesis. The proposed method relies on a reference image produced from a text-to-image model to guide video generation. This not only improved image quality but also increased consistency in videos. Cascaded latent diffusion and temporal up-sampling were applied to produce high-definition videos. The article explains that reference-guided video generation methods play a crucial part in producing realistic AI- generated videos.

III. PROPOSED SYSTEM

The proposed study brings forward a Full Stack AI-powered system for automating the entire lifecycle of creating short format videos through the integration of different artificial intelligence technologies into one single platform. The proposed system will override the challenges in traditional workflows of video production by introducing an intelligent, scalable, and user-friendly solution that enables users to generate high quality short videos from simple text prompts with least manual intervention. The proposed solution inherently/essentially centers around an AI-driven content generation engine, which takes user provided text prompts and generates a structured video script. This engine employs Natural Language Processing to analyze the input text and extract contextual meaning, emotional tone, and intent from it. For this purpose, a pre-trained AI model generally Gemini-is used to either generate or select appropriate images, clips of video, and audio components that match the generated script. These media elements are organized in logical sequence to maintain coherence in the narrative. The application is constructed in the form of a web application with a modular approach. The frontend is built using React.js, TypeScript, and Tailwind CSS. This allows users to enter their prompts, choose the styles for their videos, preview the generated content, and control their video projects. The backend is built using Next.js. It is responsible for the handling of user authentication, prompts for AI model interaction, and controlling all the video generation tasks. The storage for all user information and final outputs is handled through a centralized storage database.

For a seamless video composition and professional-grade output, the proposed system utilizes video processing and rendering capabilities. The system uses FFmpeg for post processing functions such as cutting videos, adding transitions, and combining audio and video components. Additionally, to programmatically render videos with a precise frame- based synchronization, the system utilizes the capabilities of Remotion. To host the backend infrastructure for handling real-time database operations and efficient processing of video generation tasks, the system uses Convex.

The main features that have been integrated into the pro- posed system are:

- AI-Based Script Writing: Helps in auto-generating scripts for videos based on the text provided.
- Automated Multimedia Generation: This generates pictures, sound files, and graphical objects depending on scripts.
- Intelligent Video Sequencing: Algorithmic-based logic is used to sequence multimedia objects within optimal narrative order.
- Secure User Management: Includes authentication and personalized access features for video projects.
- Video Rendering/Storage: Facilitates high-quality video rendering, as well as secure storage of the resulting videos.

As such, with the integration of artificial intelligence, processing by algorithms, and full stack web technology, the proposed solution of the Full Stack AI Short Video Generator provides an efficient and accurate means of developing short videos automatically. The system is scalable to any future advancements that could be implemented within AI- based multimedia development.

IV. MOTIVATION OF PROPOSED SYSTEM

Currently available methods for short video production involve extensive use of human resources through manual techniques, involving considerable time, technical know-how, and

creativity. This makes authors use various platforms for script writing, speech synthesis, graphics production, and video editing separately. This not only makes the procedure complicated for authors but also inefficient. Moreover, it is difficult for authors to ensure consistency in terms of story-telling, quality, and duration when various manual methods are employed.

In response to the rising demand for short format videos across digital media platforms, it has become an urgent requirement to have systems that could automate the generation of videos in such a way that attention is also paid to ensuring that it is relevant and scalable in nature. In present systems, there is no intelligent way by which it can analyze user intentions pertaining to textual information provided for transforming it into an appropriate multimedia format seamlessly in an automated manner.

However, these challenges become even more pronounced in situations where scalable content creation is required. This may range from content creation for marketing campaigns to development on educational platforms or even social media interaction. Manual processes in video content development not only consume a lot of time but also errors associated with human intervention may result in inconsistency. It has become generate meaningful short videos from mere text inputs.

- The solution provides support for rapid prototyping and experimenting with video creation with artificial intelligence in a full-stack modular architecture that makes it easy to extend and improve.
- The combination of AI script writing, multi-media pro duction, and automatic rendering enables a streamlined and efficient means of scaled short video production.

V. METHODOLOGY

The current workflow related to the production of short videos relies fully on manual assistance and different technologies in scripting, image creation, recording voice, and video editing. The result of such a workflow may result in increased production time, variability in the quality of the output, and a full dependence on technical expertise. The current video production system lacks automation in the interpretation and compilation of the contents for video production. To solve these problems, the system uses AI/ML to fully automate the process of short video production through the combined use of Natural Language Processing, AI/ML technologies, and full- stack web development. System Architecture and Environment Configuration: The system is developed as an end-to-end web application with a modular and scalable architecture. The front-end application is developed using React with TypeScript support along with Tailwind CSS for designing an interactive and responsive user interface. The back-end application is developed using Next.js with support for hosting on Convex to enable support for real-time operations on the database as well as scalable processing. This allows seamless interaction between the UI, AI processing algorithms, and storage subsystems while also facilitating efficient processing for multiple video generation requests concurrently. AI Processing and Content Generation: The heart of the designed system belongs to the AI-powered content creation pipeline. The text inputs provided by the user are subjected to Natural Language Processing to derive context, meaning, intention, and emotions. The results of these processes are further fed into the trained AI model named Gemini to create structured video scripts. After the video script has been developed, it auto-generates multimedia files such as images and voice. Algorithmic sequence methods are employed to interlink all the multimedia files to construct a meaningful sequence.

1. Architecture Diagram

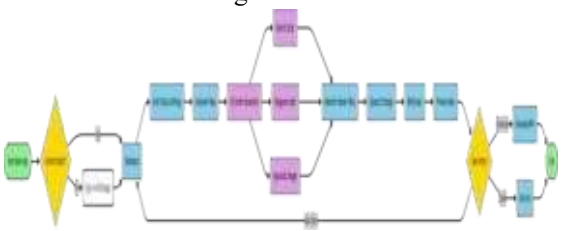

Fig.

Video Rendering and Frontend Interaction: Video composition, processing, and output are done using Remotion, together with FFmpeg, which achieves synchronization of visual elements, audio, and captions at the frame level. Also, trimming, transitions, or merging videos are done with the aid of FFmpeg, resulting in professional-looking outputs. With the frontend part of the system, one can view the videos created; in addition, one can monitor video creation progression in real-time, as well as download completed videos.

Testing, Deployment, and Scalability Factors: Thorough testing is done in each of the steps of the video production pipeline, such as the construction of the video script, creation of assets, video rendering, and storage. Its functionality in terms of speed, dependability, and error correction is also tested. Even if the proposed system is still in its prototype form with enhancement prospects, this system can still be expanded in cloud hosting for scaling up large contents in marketing campaigns, learning, or social networking sites with less adjustment of settings.

VI. SYSTEM ANALYSIS

A. Architecture Diagram

Fig 1 shows the architecture of Full Stack AI Short Video Generated, which clearly explains the functioning of different components of the system related to prompt processing, AI generation, video rendering, storage, and interaction. The designed architecture of the proposed system will focus on all possible layers of the system, which will ensure efficiency and effectiveness.

- The Web-Based Frontend Interface is the major interaction component with the users. Implemented with current internet technologies, the Frontend of the system provides the functionality of inputting the text prompt, choosing the video style, previewing the output, and controlling the video tasks of the users.
- Front-end user requests are channeled to the Backend Application Server, where the application is built using Next.js. The duties of the backend include user authentication, user input validation, control and coordination of AI services, video production workflows, as well as handling storage services.
- AI Processing Module: The heart of this entire system. This module analyzes the text inputs through Natural Language Processing and relies on the AI model, namely Gemini, to produce video scripts and manages picture generation, voice generation, and narration in such a way that it tells a cohesive story.
- Critical data like user details, prompt information, generated scripts, multimedia files, and final video outputs are all stored in a central database storage mechanism. This is ensured to be a safe storage method that provides data integrity and facilitates easy retrieval in the future.
- The Video Rendering Component is responsible for com- posing images, audio, and captions into a final short video. Technologies such as Remotion and FFmpeg are used to achieve frame-level synchronization, transitions, trimming, and merging of multimedia elements.
- The backend continuously tracks the status of video generation tasks and collects processing results from AI and rendering modules. This information is used to generate progress updates and performance insights for users.
- The final outputs, including rendered videos, metadata, and download links, are delivered to users through the frontend interface, ensuring a seamless and transparent video generation experience while maintaining secure data handling throughout the process.

B. System Workflow

The Full Stack AI Short Video Generator is designed to perform its functions systematically with an automated process involving artificial intelligence to create short videos from text inputs.
Step 1: Prompt Input The process of video creation is triggered by the user entering the type of video they require in the form of a text prompt.
Step 2: User Authentication They use the system through secured log-in sessions to access the video creation dashboard and manage their projects.
Step 3: Script Generation Based on the given prompt from the user, it uses natural language processing along with a pre- trained AI model to produce a video script.
Step 4: Multimedia Asset Creation Based on the automatically generated script, the system generates the required multimedia data like images and audio through the use of AI- based generation modules.
Step 5: Video Rendering The generated resources are organized in a manner that follows a logical order and a video is created out of it with the help of video rendering and processing tools.
Step 6: Storage Processing The final video produced, along with the related data, is safely housed within the database and storage system for later use.
Step 7: Output and Download This is The finished video can then be accessed by the user via the frontend for either previewing, downloading, or sharing of the created short video clip.

Fig. 2. Analytics Dashboard Showing Video Generation Statistics

Fig. 3. Video Generation Progress and Task Execution Report

Fig. 4. AI-Based Video Rendering and Preview Interface

Fig. 5. User Dashboard Showing Project History and Activity Logs

VII. RESULT

A. Analytics Dashboard

Fig 2 presents the analytics dashboard of the Full Stack AI Short Video Generator. It offers a real-time overview of the total number of videos generated, completed video projects, videos in processing, and overall usage statistics of the system. This will allow users and system administrators to monitor the performance of the system and track the efficiency of AI- driven workflows for generating videos.

B. Video Generation Progress Report

Fig. 3 shows the video generation progress report provided by the system. It indicates the status of each single video generation task, such as creating the script, generating images, synthesizing voices, and rendering at specific times. This will be useful for tracking the processing time and delays, if any, in a video generation pipeline.

C. AI-Based Video Rendering

Fig. 4 shows the video rendering and previewing process in the system. The user interface shows how it has compiled all these aspects: script, visual imagery, audio files, and captions to produce the short video created by the artificial intelligence algorithm.

D Dashboard and Project Activity Feed

Fig. 5 shows the overall dashboard for users with recently produced videos, project history, and activity logs. The activity log records various activities such as the generation of new videos, completion of rendering tasks, and downloads per-formed. All activities are recorded in real time. The overall dashboard helps users to organize their projects efficiently.

VIII. CONCLUSION

The Full Stack AI Short Video Generator is a complete, self- contained solution for automatically creating short-form video content using artificial intelligence. The solution is dedicated to ensuring that there is a smooth flow between the user interface and artificial intelligence processing modules, ensuring that video production is carried out in an efficient, precise, and user-friendly manner that does not require technical expertise. The solution has minimized manual work involved in video production, which in turn saves time.

The proposed system is a proof of concept for the feasibility and efficacy of AI-powered multimedia production. With the combination of technologies such as Artificial Intelligence, Natural Language Processing, React.js, TypeScript, Next.js, Convex, and FFmpeg, it is ensured to be competent, scalable, and efficient for producing a high-quality short video. The automation of script production, intelligent production of multimedia, and programmatic video output work towards increasing efficiency by eliminating human inaccuracies.

Even with the current implementation done as a functional prototype for research purposes, it provides a good foundation for upgrades. Some of these upgrades may include incorporating more video formats, enabling users for real-time modifications, incorporating mobile applications, as well as cloud implementation for mass content creation.In conclusion, the "Full Stack AI Short Video Generator" is a reference implementation for AI- based video creation systems that shows the application potential of artificial intelligence in multimedia content generation.

REFERENCES

[1] L. Wang, Y. Li, and Z. Chen, "TransPixar: Advancing text-to-video generation with transparency," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, pp. 18229–18239, 2025.

[2] G. Oh, J. Jeong, S. Kim, and W. Byeon, "MEVG: Multi-event video generation with text-to-video models," in *Proc. European Conf. Computer Vision (ECCV)*, pp. 1–14, 2024.

[3] Y. N. M. Nag, P. B. R. Purnesh, K. P. Bhat, and N. Prabhu, "AI-powered short video creation and generation," *International Journal of Advanced Research in Computer and Communication Engineering*, vol. 13, no. 11,pp. 1–7, 2024.

[4] X. Wang, S. Zhang, H. Yuan, Z. Qing, and B. Gong, "A recipe for scaling up text-to-video generation with text-free videos," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, pp. 6572–6582, 2024.

[5] P. Zhou, L. Wang, Z. Liu, Y. Hao, and P. Hui, "A survey on generative AI and large language models for video generation, understanding, and streaming," *IEEE Communications Surveys & Tutorials*, vol. 25, no. 4,pp. 1–28, 2023.

[6] W. Weng, R. Feng, Y. Wang, and Z. Zhao, "ART-V: Auto- regressive text-to-video generation with diffusion models," arXiv preprint arXiv:2311.18834, 2023.

[7] H. Fei, S. Wu, H. Zhang, and T.-S. Chua, "Dysen- VDM: Empowering dynamics-aware text-to- video diffusion with large language models," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, pp. 7641–7653, 2024.

[8] X. Li, W. Chu, Y. Wu, and Q. Zhang, "VideoGen: A reference-guided latent diffusion approach for high-definition text-to-video generation," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, pp. 1–10, 2024.

[9] J. Xu, T. Mei, T. Yao, and Y. Rui, "MSR-VTT: A large video description dataset for bridging video and language," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, pp. 5288–5296, 2016.

[10] D. Ceylan, C.-H. P. Huang, and N. J. Mitra, "Pix2Video: Video editing using image diffusion," in *Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV)*, pp. 128–138, 2023.