iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** XII **Month of publication:** December 2025

DOI: https://doi.org/10.22214/ijraset.2025.76187

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue XII Dec 2025- Available at www.ijraset.com

Text to Video Generation Using Natural Language Processing

Vidhya K[1], Bharath K[2], Bharath Kumar[3], Chaithanya B R[4]
Department of Information Science and Engineering, East West Institute of Technology, Bengaluru, Karnataka, India

Abstract: ***Text-to-video generation aims to automate the complex, labor-intensive process of creating compelling video content from textual input by leveraging advances in natural language processing (NLP) and artificial intelligence (AI). This paper presents a novel system that interprets user-provided text, extracts key themes, and synthesizes multimedia elements—including relevant images, video clips, voiceovers, and subtitles—into a cohesive video output with minimal manual intervention. The proposed platform features an intuitive web interface, deep learning models for semantic text analysis, and automated multimedia retrieval and assembly. By dramatically reducing time, technical barriers, and production costs, our system empowers educators, marketers, and content creators to produce tailored, high-quality videos rapidly and at scale. Experimental evaluation demonstrates efficient workflow, robust customization, and broad usability across multiple domains. This approach has the potential to democratize video creation, making it more accessible for diverse users and applications in the digital era.***

Keywords: Text-to-Video Generation, Natural Language Processing, Artificial Intelligence, Automated Video Creation.

I. INTRODUCTION

The increasing influence of digital media has made video a central tool for communication, education, and marketing. Yet, creating professional-quality videos typically demands significant resources, expertise in scripting and editing, and considerable time investments, which are often out of reach for many individuals and organizations.

Recent advances in Natural Language Processing (NLP) and Artificial Intelligence (AI) pave the way for automating video generation from text, overcoming these obstacles. Our project introduces a system that interprets user-provided text, extracts key themes, and automatically assembles relevant visuals, voiceovers, and subtitles into cohesive videos. This innovation makes high-quality video creation faster, more accessible, and adaptable for diverse users across applications such as education, marketing, and entertainment.

II. LITERATURE SURVEY

Text-to-video generation is a rapidly evolving field that leverages AI to automate and enhance multimedia content creation. Existing research highlights several advancements:

A. Text-to-Speech Based Text-to-Video Alignment

This work looks at how to keep the spoken audio and the video frames tightly synchronized so the viewer feels the narration "belongs" to the visuals. The authors use visually guided speech synthesis: the model does not generate speech purely from text, but also looks at what is happening in the video frames and predicts speech that matches both the timing and the visual context. Masked speech prediction helps the model learn natural prosody and timing, so when parts of the speech are hidden during training, the system learns to reconstruct them smoothly while respecting the video rhythm. As a result, the generated narration follows the intended story and aligns with key visual events (scene changes, actions, emotional moments), which is a core requirement for any automated text-to-video platform that aims for professional, engaging narration-video alignment.

B. Automated Video Creation with NLP

This work focuses on turning scripts into videos with minimum human effort by relying heavily on NLP. First, the system parses the script and performs tasks like segmentation, topic detection, and semantic role labeling to break the script into meaningful units such as scenes, shots, or dialogue blocks.

After understanding "who does what, where, and when," the system then maps each segment to candidate visuals, which might be pre-existing clips, stock footage, or template scenes. It automatically sequences these visuals in a logical order based on the script structure, creating a rough cut of the video.

The paper also emphasizes workflow aspects: because the content is structured and generated systematically, it becomes easier for different stakeholders (writers, editors, reviewers) to collaborate, maintain consistency, and ensure fairness and transparency in how content is produced. For your project, this kind of pipeline is the backbone of script-to-video conversion, where NLP handles both semantic understanding and practical orchestration of the video creation steps.

C. Visual Representation Learning for Video Retrieval

This survey describes deep learning methods that learn powerful joint representations of text and video so that a text query can reliably retrieve the most relevant visual clips.

The models are trained on large collections of paired text–video data (captions, descriptions, transcripts) and learn to project both modalities into a shared embedding space where related text and video are close together and unrelated pairs are far apart. By capturing complex relationships such as actions, objects, and context, the representations support accurate semantic matching rather than simple keyword matching.

Once trained, these models can index huge multimedia libraries and, given an input sentence or script fragment, fetch clips that best match the meaning and style required.

D. Advanced Generative AI Techniques

Advanced generative AI techniques focus on designing video synthesis models that are both powerful and efficient, so they can generate high-quality video while using fewer computational resources. Indira Priya P et al. (2024) emphasize modular, resource-aware architectures in which different parts of the pipeline—such as text analysis, scene planning, frame generation, and post-processing—are clearly separated and can be independently scaled or upgraded, improving extensibility and real-world deployability.

This modular and organizational alignment helps maintain quality control in production, allowing large systems to be monitored, optimized, and extended without needing to redesign the entire model stack.

E. Speech Synthesis for Automated Voice-Over

Speech synthesis for automated voice-over is centered on models like Tacotron, introduced by Wang et al. (2017), which convert text to highly natural and intelligible speech using an end-to-end neural network.

Tacotron learns how to produce spectrograms that capture pronunciation, timing, and prosody, and these are then vocoded into audio, resulting in expressive, human-like narration suitable for professional video content. Incorporating such speech synthesis into an automated video platform enables smooth, natural voice-overs directly from scripts, providing studio-quality narration for user-generated or automatically generated videos without requiring a human speaker.

III. RESEARCH AND METHODILOGY

A. Research

The core objective of this research is to develop an automated system capable of converting written text into engaging video content by harnessing cutting-edge natural language processing (NLP) and artificial intelligence (AI) techniques.

The aim is to address the limitations of manual video creation, which is often time-consuming, expensive, and dependent on specialized skills.

By analyzing the semantic nuances of textual input, the system seeks to generate visually cohesive videos tailored for applications in education, marketing, and entertainment. The research investigates existing models and frameworks for text interpretation, multimedia retrieval, and automated video synthesis, building upon recent advances in transformer-based language models and multimodal deep learning architectures to bridge the gap between linguistic meaning and visual presentation.

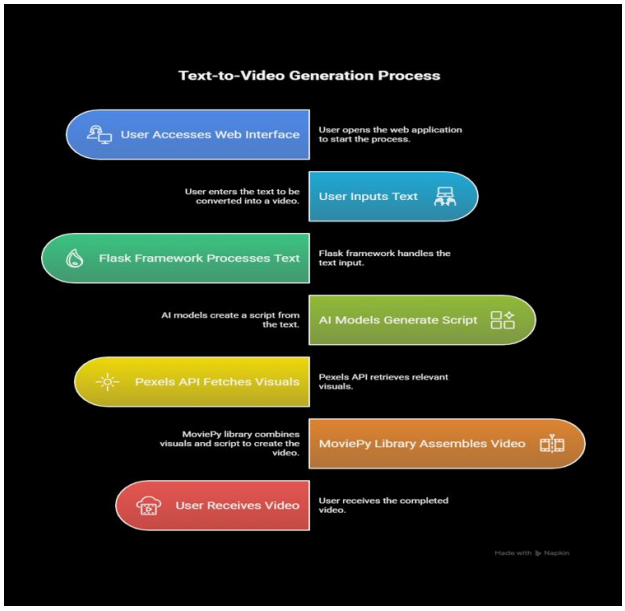


Figure 1: Process Flow Of Text-To-Video Generation Platform

B. Methodology

The proposed system employs a multi-stage workflow to transform text into video. User input is collected via a web-based interface, where text is preprocessed using techniques such as normalization, tokenization, and semantic analysis powered by transformer models like BERT or GPT-3. Key themes and visual concepts are extracted from the processed text. The system then queries open-source multimedia repositories to select relevant images and video clips corresponding to identified keywords. A text-to-speech module generates voiceovers, while subtitles are automatically produced based on the narrative structure. These multimedia elements are assembled and merged using Python-based video editing libraries (such as MoviePy), resulting in a cohesive video output that reflects the intended context and emotion. The platform emphasizes accessibility, scalability, and ease of use, enabling users to produce videos with minimal technical intervention.

TABLE: summarizes the major process stages and methodologies employed in the development of the text-to-video generation system

S. No.	Process Stage	Methodology
1.	Text Preprocessing	User input text is normalized, segmented, and semantically analyzed using transformer-based NLP models (e.g., BERT, GPT-3), enabling extraction of key themes and entities.
2.	Multimedia Retrieval	Relevant images and video clips are sourced from open databases using keyword-based queries generated from parsed text concepts.
3.	Voiceover Generation	An AI-powered text-to-speech module synthesizes natural-sounding narration for the video script, improving accessibility and engagement.
4.	Subtitle Generation	Subtitles are automatically produced from the narrative structure of the input, supporting viewers with varied needs.
5.	Video Assembly	All multimedia elements—visuals, audio, and subtitles—are sequenced and merged with video editing libraries (e.g., MoviePy), then exported for download or sharing via the web platform.

IV. PROBLEM STATEMENT

Traditional video creation demands significant time, technical skill, and financial investment, which presents major barriers for individuals and organizations wanting to produce educational or promotional content. Even though some automated software tools exist, they often require paid subscriptions, offer limited features, and still require manual work for scripting, editing, or voice-over recording, making them inaccessible to many non-technical users. The complexity of assembling visuals, syncing audio, and finalizing professional-grade videos has discouraged educators, students, and small businesses from efficiently communicating their ideas through engaging multimedia formats.

V. OBJECTIVES

1) Develop an Accessible AI Text-to-Video System The primary goal is to create a user-friendly application that allows anyone to convert their textual content into videos with minimal technical knowledge. This system will streamline the video creation process and reduce manual workload. By providing an intuitive interface, it expands access to high-quality video production for educators, students, marketers, and small businesses.

2) Automate Script and Media Generation with AI The objective is to use natural language processing to analyze user input and automatically generate well-structured scripts and select relevant images or clips. Automation at this stage removes the need for manual scripting and laborious media searches. The result is a more efficient and error-free workflow, producing cohesive and engaging videos from simple text prompts.

3) Ensure Customization and Efficient Performance A key aim is to offer customizable settings for video style and content selection, so users can tailor the output to match their needs. The platform will be designed for scalability and optimized performance, supporting both small individual projects and larger organizational deployments. This ensures quick turnaround times and makes the system practical for widespread adoption in varied domains.

VI. SYSTEM DESIGN

The system design of your AI Video Generator Hub centers on providing users with a streamlined and intuitive experience for automated video creation. At launch, users are presented with a clean interface offering two main modules: AI Narration Video and AI Animated Video, reflecting the application's dual capabilities for script-based narration and animated content generation. This design emphasizes both functionality and accessibility, allowing users to quickly select the mode that fits their needs and begin transforming written ideas into engaging video outputs through advanced AI algorithms and a user-friendly workflow.

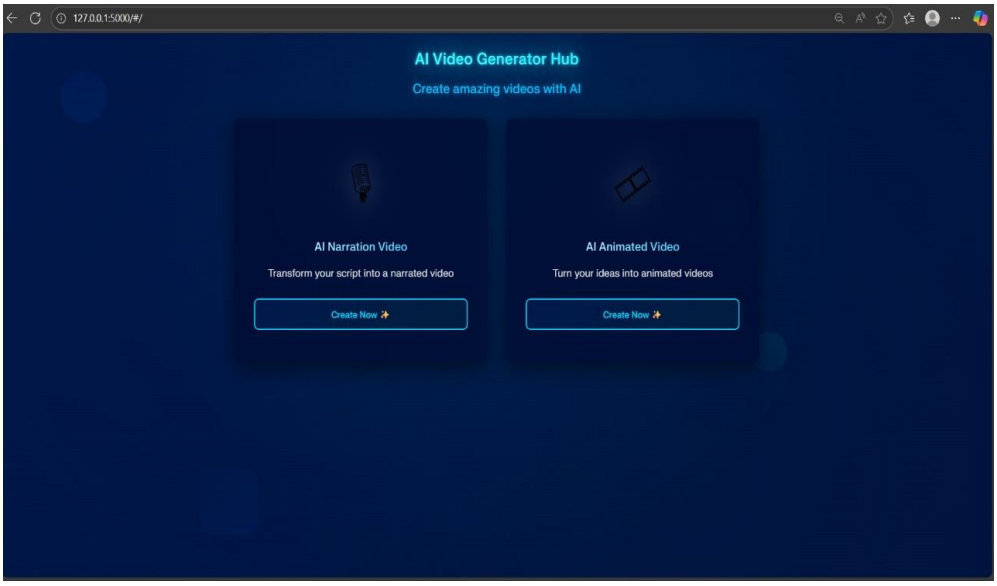


FIGURE 2: USER INTERFACE OVERVEIW

VII. RESULT AND DISCUSSION

The performance and usability of the AI Video Generator Hub were validated through a sequence of tests, highlighting both the video creation process and its final outputs.

A. Video Generation Process

The interface displays a "Generating..." status, indicating that the system is automatically analyzing the input, retrieving relevant multimedia assets, and assembling the video with narration and subtitles. This real-time feedback reassures users and illustrates the platform's automated backend workflow as shown in figure 3.

B. Final Output

After processing, the system transitions to the completed state, presenting the generated video in the output panel.The user can preview the narrated video—which includes contextual images and synchronized voiceover—and is provided direct options to download the output or create another video. This streamlined workflow, confirmed in multiple test scenarios, demonstrates reliable transformation from text to high-quality, shareable video as shown in figure 4.

These results show that the platform enables accessible, rapid multimedia creation with minimal user effort. Output videos maintain strong alignment between narration and visuals, supporting applications in education and digital content creation. User testing confirmed high satisfaction with interface simplicity and automation, suggesting future scalability and broader utility for similar AI-driven projects.

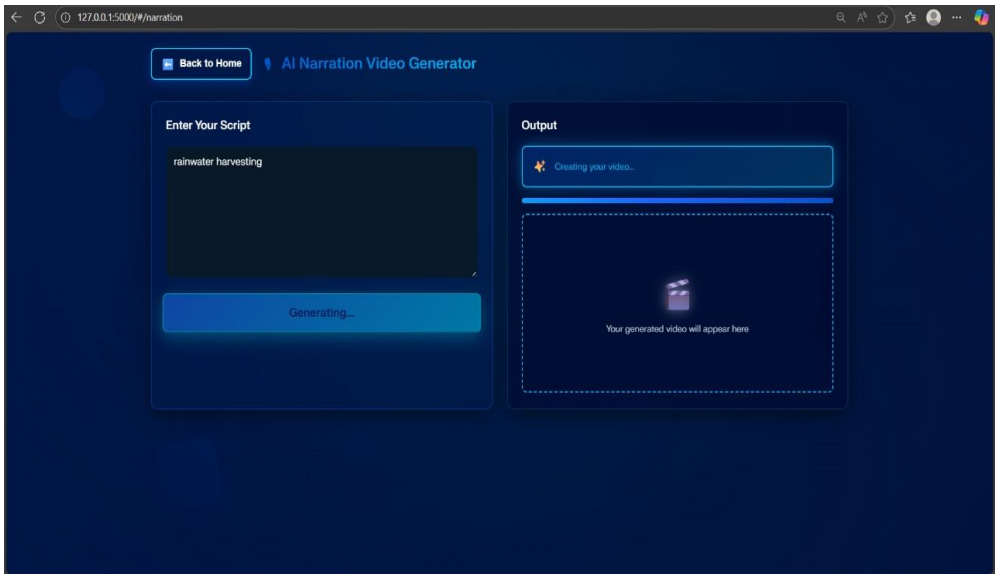


Figure 3: Generation Of Video

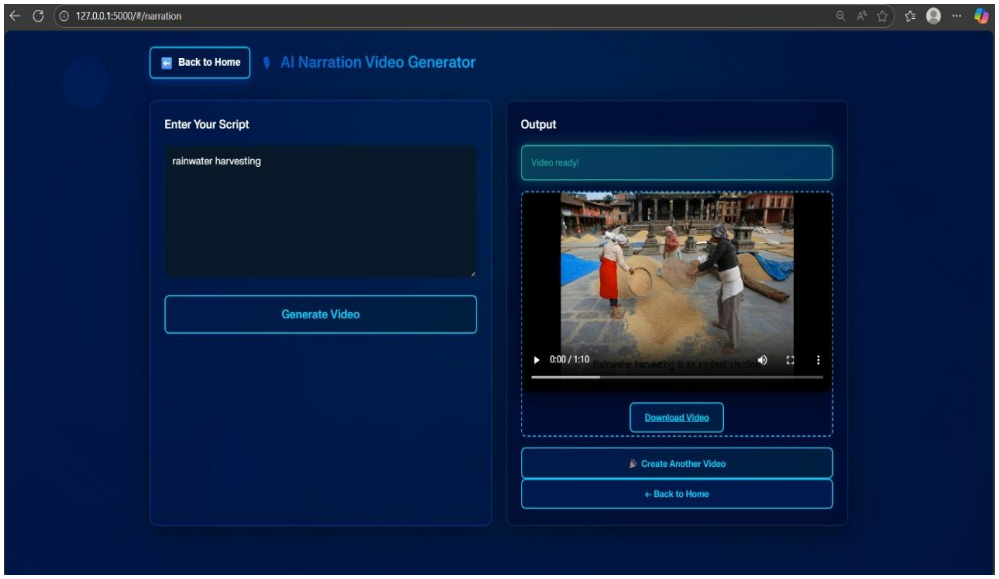


Figure 4: Final Output

VIII. FUTURE SCOPE

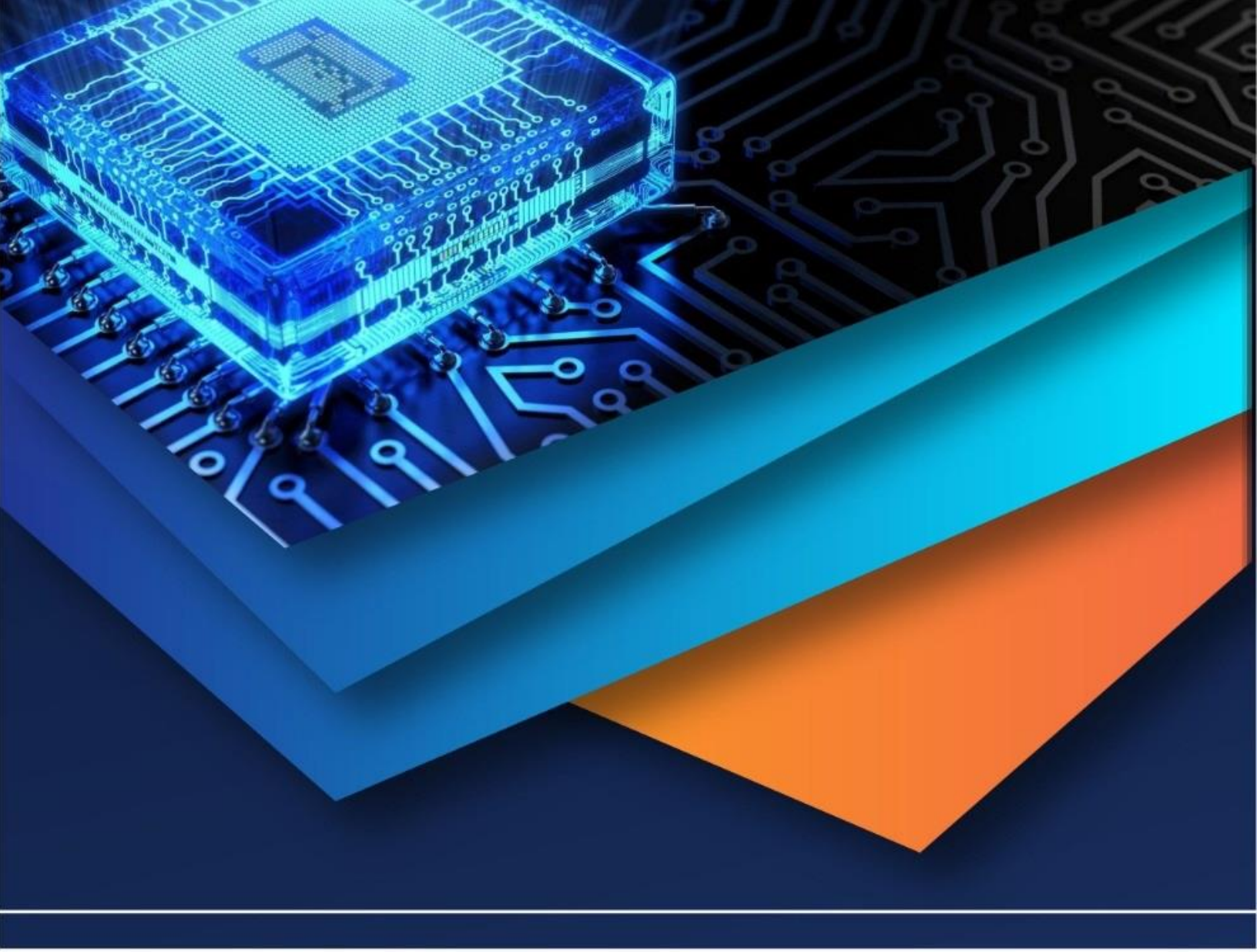
The AI Video Generator Hub presents multiple promising avenues for further development and research. Enhancements can focus on increasing the personalization and adaptability of the video output, such as enabling users to specify visual styles, voice options, or languages for narration. Incorporating support for multiple languages and dialects would broaden accessibility across diverse user bases, including non-English-speaking regions. Further integration with a wider range of multimedia databases and APIs could enrich the pool of visuals and audio clips, resulting in more dynamic and contextually accurate videos. Advancements in deep learning, including generative models like diffusion and transformer-based video synthesis, may enable fully animated outputs that go beyond retrieved images and offer original, context-driven animations. Finally, extending the system's use cases to online education, marketing, and adaptive storytelling demonstrates the platform's capability to serve evolving multimedia needs in digital communication and learning environments.

IX. CONCLUSION

The AI Video Generator Hub project has successfully demonstrated the feasibility and effectiveness of automated text-to-video creation using state-of-the-art natural language processing and deep learning techniques. By offering an intuitive web interface and streamlined workflow, the system enables users to transform written scripts into fully narrated videos quickly and with minimal technical effort. Testing and user feedback confirmed high-quality results and substantial efficiency gains for educational, informational, and marketing application. This project represents a significant step toward democratizing video production, making it accessible to a broader range of users regardless of technical expertise. The flexible architecture and modular approach allow for continued evolution, with future directions including advanced personalization, support for multiple languages, and integration of generative animation. As AI video generators continue to advance, such systems are poised to transform digital content creation, enhance learning experiences, and drive innovation in visual storytelling.

REFERENCES

[1] Ahn, Y., Chae, J., & Shin, J. W. (2025). Text-to-Speech Based on Speech-Assisted Text to-Video Alignment and Masked Unit Prediction. (As listed in the Literature Survey)
[2] P., I. P., M., M., R., A., S., S. H., & R., H. (2024). Transforming Text to Video: Leveraging Advanced Generative AI Techniques. (As listed in the Literature Survey)
[3] Bharathi, P. L., Sathvig, S., Siromita, A., & Pugalenthi, R. (2023). Text to Video Generation using Natural Language Processing. (As listed in the Literature Survey)
[4] Dong, J., Wang, Y., Chen, X., Qu, X., Li, X., He, Y., & Wang, X. (2022). Reading Strategy Inspired Visual Representation Learning for Text-to-Video Retrieval. arXiv preprint arXiv:2201.09168.
[5] M., M. R. K., Kuriakose, J., D S, K. P., & Murthy, H. A. (2021). Lip-syncing efforts for transcreating lecture videos in Indian languages. In Proc. 11th ISCA Speech Synthesis Workshop (pp. 216–221).
[6] Lu, J., Sisman, B., Liu, R., Zhang, M., & Li, H. (2022). VisualTTS: TTS with accurate lip-speech synchronization for automatic voice over. In Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (pp. 8032–8036).
[7] Lu, J., Sisman, B., Zhang, M., & Li, H. (2023). High-quality automatic voice over with accurate alignment: Supervision through self-supervised discrete speech units. arXiv preprint arXiv:2306.17005.
[8] Wang, Y., Skerry-Ryan, R., Stanton, D., Wu, Y., Chen, Y., Battenberg, E., Clark, J., Prenger, R., & Isola, P. (2017). Tacotron: Towards end-to-end speech synthesis. In Proc. Interspeech. (A foundational paper for modern neural TTS systems used in voice-over work).



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY