Volume 11, Issue 4, April – 2026 International Journal of Innovative Science and Research Technology
ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26apr1128

Auto Project Narrator – Text to Video Converter

Venkata Lakshmi G.[1]; Susmitha N.[2]; Deva Sai Praneetha V.[3]; Anusha P.[4]; Aswitha S.[5]

[1]Assistant Professor, Department of CSE (AI&ML), GVP College of Engineering for Women, Visakhapatnam, AP, India.
[2;3;4;5]Student, Department of CSE (AI&ML), GVP College of Engineering for Women, Visakhapatnam, AP, India.

Publication Date: 2026/04/28

Abstract: The growing demand for a better way to describe how a technical project will be completed has created the need for intelligent systems that automatically create multimedia (e.g., video) explanations of a project from text only. Preparing a presentation using traditional methods typically requires manually putting together slides, creating images, narrating the presentation, and doing a video demo; all of which can take time and effort to accomplish. This project proposes an artificial intelligence-based system to create a video description of a project automatically using NLP, RAG, Stable Diffusion and neural TTS technologies. The system takes a user-provided description of a project and builds structured documentation using a combination of retrieve systems and Large Language Models (LLMs). The documentation that is generated is converted into a multi-scene storyboard that uses the content from the documentation to create each scene of a video. Each visual prompt created from the storyboard is used as input to the Stable Diffusion model in a high-performance GPU cloud computing environment to generate images for each scene. Finally, the speech output from the neural TTS synthesis process is used to produce clear and natural audio to accompany each video scene. The generation methodology's final output, an explanatory video with a project report summary, will consist of images and audio generated and synchronized together by the new TDMs and then assembled using a video-processing method. Empirical evaluation of the new system indicates that it will automate the creation of a substantial portion of the presentation of a multimedia project, thus reducing the number of hours required for manual production of completed media and increasing the clarity of presentation as well as the attractiveness of the presentation. The new framework is designed to allow for easy scalability to enable the automated generation of educational media and provide intelligent documentation of multimedia.

Keywords: *Text to Video Generation, NLP, Retrieval Augmented Generation (RAG), Stable Diffusion, Generative AI, Storyboard Generation, Multimedia Automation, Neural TTS, Automated Video Generation.*

How to Cite: Venkata Lakshmi G.; Susmitha N.; Deva Sai Praneetha V.; Anusha P.; Aswitha S. (2026) Auto Project Narrator – Text to Video Converter. *International Journal of Innovative Science and Research Technology*, 11(4), 2184-2194. https://doi.org/10.38124/ijisrt/26apr1128

I. INTRODUCTION

In recent years the progression in AI technology and multimedia devices has cut small impact on the copywriting cycle. Educational institutions, researchers and software developers use visual presentations and explanatory videos more and more to best express project ideas. On the other hand, preparing such multimedia content usually is a labour-intensive process that involves writing scripts, designing visuals, recording narration and editing video. It requires technical and time-consuming work that is difficult for people who do not have multimedia production experience.

As intelligent language models and generative AI systems become more accessible, it has proven fully feasible to automate some facets of content writing. This is possible because NLP (Natural Language Processing) techniques allows machines to understand textual descriptions and extract meaningful information. Likewise, generative models like diffusion-based models can generate high fidelity images in the presence of textual cues or prompts while neural text-to-speech systems can read written passages aloud as if they were being narrated. With the combination of these technologies, intelligent systems can be designed to autonomously convert textual information into compelling multimedia presentations.

This potential has sparked an exciting area of exploration for these technologies: automatically creating explanatory videos of a software or research product. These systems can help learners, instructors, and builders visualize project ideas in a more intuitive and aesthetically pleasing fashion. If a user enters a description of his project, the system should generate storyboard scenes automatically, draw all visual illustrations, create audio narration from a text (text-

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26apr1128

to-speech) and finally compile everything into an explanatory video.

While generative AI technologies have progressed rapidly, crafting multiple AI components into a single pipeline for the generation of automated multimedia output has yet to be fully realized. Most solutions are based on different components of the process such as image generation or speech synthesis, not an entire processing block that can convert a textual definition into a structured explanatory video.

Sensing the need for tackling these challenges, this research work proposes an automated paradigm which offers to convert textual project descriptions into explanatory videos by utilizing artificial intelligence techniques. By combining Retrieval Augmented Generation (RAG) along with large language models to generate storyboards, image synthesis through the help of Stable Diffusion, neural text-to-speech to read off the narration as well as video rendering techniques, forms a comprehensive media pipeline. By automating the entire process, the proposed system reduces the time and effort required for preparing project presentations while improving clarity, accessibility, and visual engagement. This approach contributes toward the development of intelligent tools that support automated educational content creation and advanced multimedia documentation.

II. RELATED WORK

The recent advances in AI, deep learning, and multimodal learning-focused models can automatically generate multimedia content from text descriptions. Text to video generation systems are designed to convert natural language input into videos of visual scenes with narration, which can be used for creating automatic storytelling, educational contents, and visualization of projects. Conventional video creation is time-consuming and resource-intensive, involving manual script writing, animation, voice narration, and editing. Recent research has taken up the pursuit of automating this step using diffusion models, transformer-based architectures, and multimodal AI systems.

This section reviews important research works related to text-to-video generation, diffusion models, multimedia narration systems, and automated video synthesis.

In [1], Xing et al. (2025) introduced Make-Your-Video: Text and Structural Guidance for Custom Video Production, a diffusion-based system for controlling text-to-video generation. The authors use textual prompts together with structural motion guidance, move-byid frame-wise depth maps to synthesise temporally consistent videos. The new approach modifies existing latent diffusion models with the addition of temporal modules and causal attention masking. Experimental results show that our method achieves better temporal consistency and prompt alignment than existing video generation methods.

In [2], Kesharwani et al. (2024) introduced *DigitalAvatarGen: Text-to-Digital Person Video Generator*, which converts textual input into animated digital avatar videos. The system integrates Google Text-to-Speech and SadTalker models to generate synchronized facial animation and speech. The approach provides an effective solution for automated digital presentations and AI-based communication.

In [3], Zhang et al. (2023) proposed *ControlVideo: Training-free Controllable Text-to-Video Generation*, which utilizes diffusion models combined with ControlNet guidance to generate videos directly from text prompts. The framework introduces cross-frame attention mechanisms and hierarchical sampling to improve temporal consistency. However, the system struggles with generating complex motion purely from textual input.

In [4], Khachatryan et al. (2023) developed *Text2Video-Zero*, a framework that converts pre-trained text-to-image diffusion models into video generators without requiring additional training. The method introduces latent motion dynamics and cross-frame attention to maintain temporal consistency between frames. Although effective for short videos, the model is limited in generating long sequences.

In [5], Hong et al. (2023) introduced CogVideo, a large-scale transformer-based text-to-video generation model. The system is trained on massive text-video datasets and generates videos by sequentially predicting frames from textual descriptions. CogVideo demonstrates the effectiveness of transformer architectures for capturing temporal relationships in generated videos.

In [6], Singer et al. (2023) developed Make-A-Video, a diffusion-based framework capable of generating videos from textual prompts without requiring large-scale video datasets. The model leverages pre-trained text-to-image diffusion models and extends them with temporal attention mechanisms to produce coherent video frames.

In [7], Ho et al. (2022) proposed Imagen Video, a cascaded diffusion architecture for high-definition video generation from text. The model progressively generates low-resolution frames and refines them to high-resolution outputs, producing visually realistic and temporally consistent videos.

In [8], Rombach et al. (2022) introduced Latent Diffusion Models (LDMs), which perform diffusion processes in a compressed latent space rather than pixel space. This approach significantly reduces computational cost while maintaining high-quality image generation, forming the foundation of modern generative systems such as Stable Diffusion.

In [9], Radford et al. (2021) proposed CLIP (Contrastive Language-Image Pretraining), a multimodal model that learns the relationship between textual descriptions and visual representations. CLIP enables AI systems to align text and images semantically, which is essential for guiding generative models in text-conditioned image and video synthesis.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26apr1128

In [10], Ramesh et al. (2022) introduced DALL·E 2, a deep generative model capable of producing high-quality images from textual descriptions. The system combines diffusion models and CLIP-based embeddings to generate realistic images aligned with textual prompts.

In [11], Saharia et al. (2022) developed Imagen, a text-to-image diffusion model that produces photorealistic images using large language models and diffusion processes. Imagen achieves state-of-the-art performance in text-guided image synthesis and demonstrates strong alignment between text prompts and generated images.

In [12], P. Esser et al. (2023) proposed a diffusion-based approach for structure and content-guided video synthesis. The method uses diffusion models to generate visually consistent video frames while maintaining spatial structure and semantic content. The model iteratively refines noisy representations to produce high-quality video outputs. This work demonstrates the effectiveness of diffusion models in automated visual content generation, which is related to the image generation stage used in the proposed system.

In [13], Wu et al. (2022) presented Tune-A-Video, which adapts image diffusion models for video generation using a small number of video frames. The model fine-tunes attention layers across frames to maintain temporal coherence while generating new video content.

In [14], Ho et al. (2020) introduced Denoising Diffusion Probabilistic Models (DDPM), a generative framework that synthesizes images by iteratively removing noise from random samples. Diffusion models have become a core technique in modern generative AI systems for image and video synthesis.

In [15], Bain et al. (2021) developed WebVid-10M, a large-scale dataset containing millions of video-text pairs for training text-to-video models. The dataset provides diverse multimodal data that improves the ability of models to understand complex visual scenes described in natural language.

In [16], Shen et al. (2018) proposed Tacotron 2, a neural text-to-speech synthesis system capable of generating natural human-like speech from text. Tacotron 2 has been widely used in automated narration systems for multimedia applications.

In [17], Baevski et al. (2020) introduced wav2vec, a self-supervised speech representation learning model that improves speech generation and recognition. Such models contribute to modern AI-based narration systems used in automated video generation pipelines.

In [18], Xu et al. (2020) proposed AI-based storytelling systems that automatically generate multimedia presentations from textual descriptions by integrating natural language processing, image generation, and video composition techniques.

Table 1 Related Works on Text-to-Video Generation

Category	Researcher	Method	Description	Limitation
Traditional Multimedia Systems	Xu et al. (2015)	Attention-based Neural Networks	Neural image captioning model for generating descriptions from images.	Limited ability to generate dynamic video content.
Traditional Multimedia Systems	Shen et al. (2018)	Tacotron (Neural TTS)	Deep learning speech synthesis converting text into natural speech.	Requires large datasets and GPU resources.
Machine Learning Methods	Radford et al. (2021)	CLIP Model	Aligns text and visual representations to guide image generation.	Cannot generate videos directly.
Machine Learning Methods	Bain et al. (2021)	WebVid Dataset	Large dataset with millions of video–text pairs for training video generation models.	High storage requirements and dataset bias.
Deep Learning Methods	Rombach et al. (2022)	Latent Diffusion Model	Diffusion in latent space to generate high-quality images.	Requires powerful GPU hardware.
Deep Learning Methods	Ramesh et al. (2022)	DALL·E 2	Diffusion-based model generating images from textual prompts.	Limited temporal consistency for videos.
Deep Learning Methods	Saharia et al. (2022)	Imagen	Photorealistic text-to-image generation using diffusion models.	Requires large datasets and high computation.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26apr1128

Video Generation Models	Singer et al. (2023)	Make-A-Video	Generates videos from text prompts using diffusion models with temporal layers.	Limited control over motion dynamics.
Video Generation Models	Hong et al. (2023)	CogVideo	Transformer-based text-to-video generation model.	Needs large training datasets and GPU memory.
Video Generation Models	Khachatryan et al. (2023)	Text2Video-Zero	Converts text-to-image diffusion models into video generators.	Produces only short video clips.
Video Generation Models	Zhang et al. (2023)	ControlVideo	Diffusion model with motion guidance for controllable video generation.	Limited ability to generate complex actions.
Proposed System	Proposed Work	NLP + Stable Diffusion + TTS	Generates storyboard, images, narration, and explainer video from project text.	Requires GPU for faster image generation.

III. METHODOLOGY

In this work, we present the Auto Project Narrator framework as an end-to-end pipeline that automates the process of translating textual descriptions of projects into informative videos by leveraging a wide spectrum of artificial intelligence techniques, from Natural Language Processing (NLP) at the beginning stages to Retrieval-Augmented Generation (RAG), Large Language Models (LLMs), diffusion-based approach for image generation and neural text-to-speech narration followed by automation of video production. The input project description is processed, A structured storyboard is generated, visual scenes are created using generative models and narration audio is produced, finally the three components will be assembled into an explanatory video.

➢ *The Complete Pipeline has a Total of Five High-Level Stages:*

- Text Input and Preprocessing
- Storyboard Generation
- Image Generation using Stable Diffusion
- Audio Narration Generation
- Video Composition and Rendering

These stages help in converting textual descriptions of project progress into multimedia content.

➢ *System Architecture*

The proposed system architecture includes multiple modules that each perform another part of the data processing before passing their results as input to one or more following modules. Step 1 — A user describes his project via a web interface; the input is transformed using NLP techniques. This processed text is then fed into a language model that creates an entire storyboard with multiple scenes. Stable Diffusion transforms every scene into an image, and a text-to-speech model converts narration text into audio. The generated images and audios are combined and synced together in order to create the final video.

In a more abstract sense, you can formulate the system architecture as:

The system architecture can be expressed as:

$$P = \{T, S, I, A, V\}$$

Where

- Trepresents input text
- Srepresents generated storyboard scenes
- Irepresents generated images
- Arepresents narration audio
- Vrepresents the final generated video

- *System Architecture*

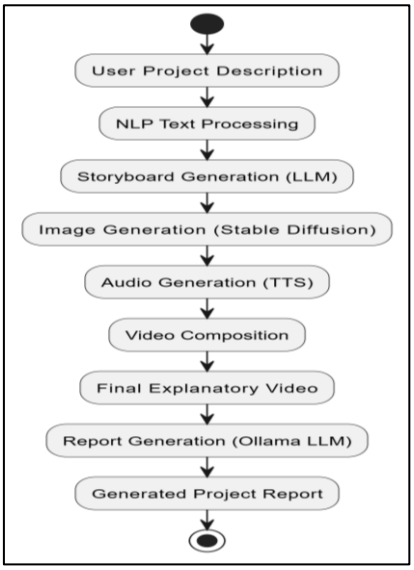

Fig 1 Proposed Auto Project Narrator System Architecture

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26apr1128

The architecture illustrates the complete pipeline from textual project description to final explanatory video generation.

➢ *Text Processing using NLP*

The system receives textual project description from user. This input text uses Natural Language Processing techniques to provide meaningful data.

The input sentence could be Look like this

$$T = \{w_1, w_2, w_3, \dots, w_n\}$$

Where,

- w_i represent tokens in the input text
- n(in the first equation) is total number of tokens

The preprocessing stage includes:

- Tokenization
- Stop-word removal
- Sentence segmentation
- Keyword extraction

A preprocessing transformation can be defined as:

$$T' = f_{NLP}(T)$$

Where f_{NLP}represents the text preprocessing function.

The processed text $T'$is then passed to the storyboard generation module.

➢ *Storyboard Generation*

Our storyboard generation module uses LLMs to convert the processed text into structured output scenes that represent parts of our project explanation.

The process of story board generation can be defined as:

$$S = g_{LLM}(T')$$

Where,

- $S = \{s_1, s_2, \dots, s_m\}$represents generated scenes
- g_{LLM} represents the language model function
- m is the number of scenes generated.
- Each scene is defined as:

$$s_i=(n_i, v_i)$$

Where,

- n_i, represents narration text
- v_i represents visual prompt.

This structure ensures that each scene contains both a textual explanation and a visual representation.

➢ *Image Generation using Stable Diffusion*

Generative model Stable Diffusion, based on latent diffusion, which can also generate high-fidelity images from text prompts. The forward diffusion process progressively adds noise to the latent image representation:

$$x_t = \sqrt{\alpha_t} x_{t-1} + \sqrt{1-\alpha_t}\epsilon$$

Where,

- x_t represents the noisy image representation
- α_t represents the noise scheduling parameter
- ϵ represents Gaussian noise.

This process of reverse diffusion removes the added noise in several steps to produce the final image.

$$x_{t-1} = p_\theta(\, x_{t-1} \mid x_t \,)$$

The generated image for each scene is expressed as:

$$I_i = f_{diffusion}(v_i)$$

Where, $v_i$is the visual prompt.

➢ *Audio Narration Generation*

Text-to-Speech (TTS) synthesis is used to convert narration text for each scene into speech.

Here's the definition of the audio generation function:

$$A_i = f_{TTS}(n_i)$$

Where,

- A_i represents generated audio narration
- n_i represents narration text.

This step makes sure that every scene will have a voiceover corresponding to the project concept.

➢ *Video Composition*

In the last step, it combines the image and narration audio into a video. The images are sequenced in accordance with the storyboard structure and narration audio.

The video generation process can be defined as:

$$V = \{(I_1, A_1), (I_2, A_2), \dots, (I_m, A_m)\}$$

The final video is produced using:

$$Video = combine(V)$$

Where combine() represents the video rendering function.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26apr1128

➢ *Report Generation*

Apart from crafting explanatory videos the system also facilitates automated project report generation using a large language model that is deployed through Ollama inference framework. A report generation module that produces the project documentation from the structured storyboard information and narration text.

Storyboard scenes are created with narration text that describes the various elements manipulated in this project. These segments of narration are then compiled and fed into the report generation model to form structured parts namely Introduction, Methodology, Implementation, Results and Conclusion.

- *Express a Report Generation in the Fashion*

$$R=f_{LLM}\ (S)$$

Where,

✓ R represents the generated project report
✓ $S=\{s_1,s_2,...,s_m\}$ denotes the storyboard scenes
✓ f_{LLM} denotes that the LLM inference performing by Ollama.

Panel Scene s_i contains narration text n_i, which serves as the main textual input for generating report content.

$$s_i=(n_i,\ v_i\ )$$

All extracted narration components from all scenes are flattened.

$$N=\{n_1,n_2,...,n_m\}$$

Based on this set of narrations, the model for report generation ultimately writes out formatted documentation:

$$R=f_{ollama}(N)$$

Where,

- f_{ollama}: the Ollama-based large language model for report synthesis.

The output of this report will then be, in essence a set of structured sections.

$$R=\{r_1,r_2,r_3,...,r_k\}$$

Where,

- r_1 represents the introduction section
- r_2 represents the methodology section
- r_3 represents the implementation section
- r_k represents the final report sections.

The video explanation is formatted to a document like docx that gives documentation of the generated content parallel with other output.

It extends the Auto Project Narrator system to provide video and textual explanations from a single project description, combining visually interactive descriptions with automatically created representations.

➢ *Algorithm 1: Pipeline for the Auto Project Narrator*

- Input: Project Description
- Output : Video Gen, Report Gen
- Receive project description from user
- Perform NLP preprocessing on text
- Generate storyboard scenes using LLM
- For each scene
- Generate image with Stable Diffusion
- Create narration audio with TTS
- Synchronize generated images and audio
- Render final explanatory video
- Generate project report using Ollama
- Return *V,R*

IV. EXPERIMENTAL SETUP AND IMPLEMENTATION

This section gives a detailed description about the hardware environment, software configuration and implementation which is used to develop and execute the proposed Auto Project Narrator system.

It does not need a large labelled dataset as is the case with traditional machine learning systems, but it operates on the project descriptions provided by the user and dynamically generates multimedia using pretrained models.

➢ *Hardware Environment*

The system was powered by a combination of local development resource and cloud-based GPU infrastructure. The generation for several part of the actual work done in Google Colab NVIDIA T4 GPU acceleration.

Table 2 Hardware Configuration

Component	Specification
Processor	Intel Core i5
RAM	16 GB
GPU	NVIDIA T4 (Google Colab)
Storage	512 GB SSD

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26apr1128

➢ *Software Environment*

The system was implemented using Python and several AI frameworks for text processing, image generation, and video rendering.

Table 3 Software Configuration

Software	Purpose
Python	Core programming language
Flask	Backend web framework
Stable Diffusion	Image generation model
Text-to-Speech (TTS)	Audio narration generation
MoviePy	Video composition

Visual Studio Code was used as the primary development environment.

➢ *Implementation Pipeline*

This system works as a pipeline, where each module handles the output of its predecessor.

- *The Implementation Workflow Includes:*

✓ Regular user enters project description via the web interface
✓ NLP module preprocesses the text
✓ LLM generates storyboard scenes
✓ Stable Diffusion generates scene images
✓ TTS module generates narration audio

Also, you can listen audio, while the video module synchronizes it with your images for final video generation.

➢ *System Modules*

Table 4 System Modules

Module	Function
NLP Processing	Extracts structured information from text
Storyboard Generator	Converts project description into scenes
Image Generator	Generates scene images using Stable Diffusion
Audio Generator	Produces narration using TTS
Video Composer	Combines images and audio

V. RESULTS AND DISCUSSION

In this section, we evaluate the Auto Project Narrator system that generates explanatory videos from project descriptions in text. Unlike traditional classification models where evaluation is straightforward by measuring the accuracy of class predicted versus ground truth, these systems are generative in nature (the system generates data samples) and therefore, the evaluation focuses on scene quality, image-relevance to narration, clarity of narration and video coherence.

The results show how the proposed pipeline successfully converts textual project descriptions into structured storyboard scenes and generates visual audio representations from these.

A. Storyboard Generation Results

Storyboard generation module: Transforms the project description into a series of scenes that illustrate the flow of the system. In each scene there is the narration text and a visual prompt for image generation.

The scenes generated are in order so the system can walk you through the project, step by step.

B. Image Generation Results

Images are produced for each storyboard scene from the Stable Diffusion model. These visuals illustrate the concepts in the project description.

The generated images show good semantic alignment with the prompts from storyboard scenes.

➢ *Below is the Example of the Storyboard that is Generated by the System:*

- *Scene 1:*

✓ Narration: The AI Skin Recommendation System project aims to develop an AI-powered skin recommendation system that can analyse user skin types, concerns, and preferences to suggest customized skin care routines and products.
✓ Visual: A developer sitting at a desk with a laptop open to a code editor, surrounded by notes and diagrams.

- *Scene 2:*

✓ Narration: The system will require data on skin types, concerns, and preferences from APIs and user input. This data will be used to train an AI model that can generate personalized skin care recommendations.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26apr1128

✓ Visual: A person holding a tablet with a graph showing different skin types and concerns, surrounded by books and research papers.

- *Scene 3:*

✓ Narration: The system architecture consists of a user interface, data processing module, AI model training module, and output generation module. The modules will be designed to work together seamlessly.

✓ Visual: A whiteboard with diagrams and flowcharts showing the system's architecture, surrounded by developers working on laptops.

➢ *Generated Scene Image*

Fig 2 Image of a Person Holding a Tablet with a Graph Showing Different Skin Types and Concerns, Surrounded by Books and Research Papers, while Another Person is Using the System on a Separate Screen
(Image Generated Using Stable Diffusion for a Story Board Style)

This is a visual at loggerhead system workflow as stated in the storyboard.

C. Narration Generation Results

The narration generation module transforms the text description of the scene into speech using a TTS engine. And for every scene, a narration is generated to accompany with.

It enhances the overall quality of the video output and makes it easier for users to clearly understand what is being explained in the project.

Table 5 Narration Output Example

Scene No	Narration Output
Scene 1	The AI Skin Recommendation System project aims to develop an AI-powered skin recommendation system that can analyse user skin types, concerns, and preferences to suggest customized skincare routines and products.
Scene 2	The system requires data on skin types, concerns, and preferences from APIs and user input. This data will be used to train an AI model capable of generating personalized skincare recommendations.
Scene 3	The system architecture consists of a user interface, data processing module, AI model training module, and output generation module that work together to generate recommendations.
Scene 4	User input includes skin type, concerns, and preferences. This information is collected through the system interface and integrated with skin-related data obtained from APIs.

D. Video Generation Results

The last stage takes the generated images and narration audio to pull together into the explanatory video.

Video generation module aligns narration audio with visual scenes to create a coherent explanation of the project.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26apr1128

➢ *Generated Video Frame*

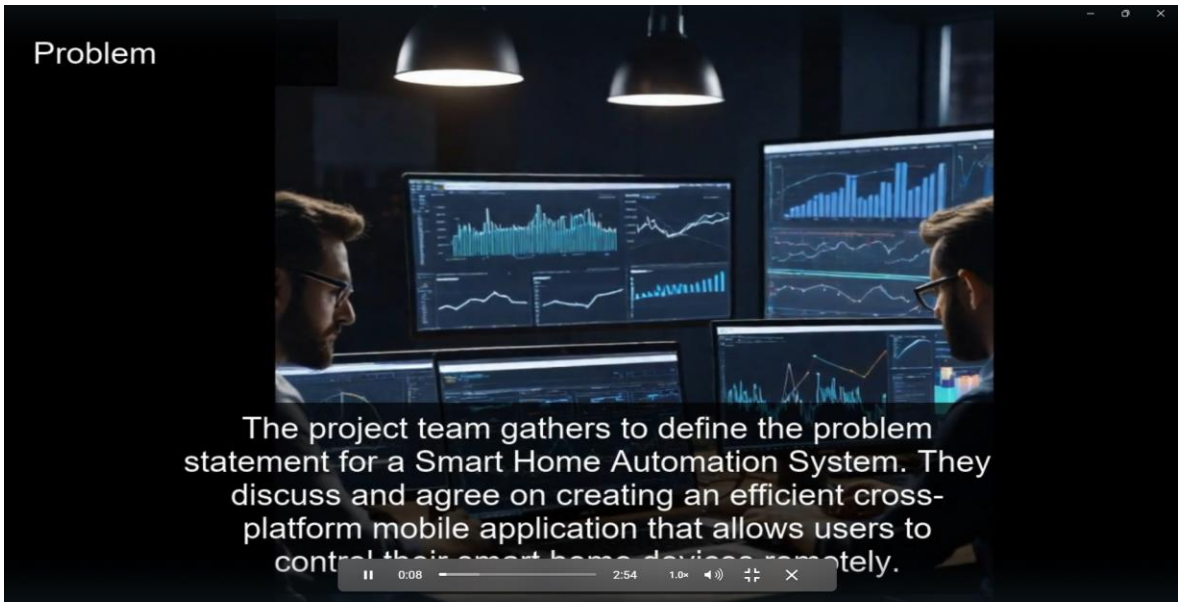

Fig 3 Frame from the Generated Explanatory Video Produced by the Proposed System.

The video provides a structured explanation of the project workflow.

E. System Evaluation

Since the proposed system focuses on generative multimedia output, the evaluation is performed using qualitative analysis based on scene generation quality, image relevance, narration clarity, and video coherence.

Table 6 System Evaluation

Evaluation Aspect	Observation
Scene Generation	Scenes logically represent project workflow
Image Relevance	Generated images match visual prompts
Narration Quality	Speech is clear and synchronized
Video Coherence	Final video explains project clearly

The evaluation results indicate that the proposed system successfully generates multimedia explanations from textual project descriptions.

F. Discussion

The experimental results show that the Auto Project Narrator system proposed in this paper effectively combines multiple 450 AI technologies to generate explanatory videos from textual inputs. The NLP module accurately extracts information-responsible clauses from project descriptions, and the LLM-based storyboard generator arranges the explanation in scenes to include in the video.

The Stable Diffusion model is capable of generating visually relevant images corresponding to the applied text scene prompts, and the text-to-speech module creates narrations that further articulate and clarify the explanations. The last three-dimensional creative confirmation module aligns the pictures and narration sound, to create an informative explanation video.

The results validate the effectiveness of our approach in automatically generating explanations for multimedia projects.

VI. CONCLUSION

This paper introduces an automated system that can generate so-called explainer videos from textual project descriptions, using various artificial intelligence technologies. We propose the Auto Project Narrator system which uses Natural Language Processing (NLP), Large Language Models (LLMs), diffusion-based image generation, text-to-speech narration and video rendering to automatically create multimedia explanations from textual input.

The stages of our methodology include text preprocessing, the creation of a storyboard, image synthesis with Stable Diffusion, narration generation with speech synthesis technology, and final video composition. Experimental results show the system can generate structured storyboard scenes, visually relevant images and synchronized narration creating an explanatory video describing the project workflow.

The proposed approach minimizes the manual effort involved in making project explanation videos and provides an automated solution to visualize technical concepts. The

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26apr1128

system incorporates various AI modules into a pipeline, allowing for text to be converted into video, audio, or other, forms of media. The results indicate that the proposed framework can effectively support educational presentations, technical demonstrations, and automated documentation generation.

FUTURE WORK

While the proposed plot generation system is able to successfully generate explanatory videos from textual project descriptions, several improvements can be implemented to allow even better performance of the pipeline.

First, future work can look into the incorporation of advanced video generation models to wire together more realistic and dynamic animations rather than simple scene images. Recently, text-to-video diffusion models have advanced, greatly enhancing the visual quality of generated videos.

Another area of improvement involves enabling different language narrations in order to provide viewers with additional translations for the explanatory videos created using this technology. With the ability to create explanatory videos in numerous languages, the technology's usability will increase more globally.

Enhancements to the language model's scene comprehension will provide the system with additional ways to provide detailed visual prompts and storyboards. This means that, through this improvement, the system will be able to generate more informative and visually appealing content.

Second, real-time optimization strategies can be used to decrease the latency of video generation and make actual interactive type videos. Use the community + feedback mechanisms to allow the system to dynamically create scenes/narration that users want.

This task in the future may also delve into multimodal transformer models allowing text and whole-image/video content to be processed together. These methods can help to increase the coherence and naturalness of automatically generated explanative videos even more.

Collectively, these improvements would increase Auto Project Narrator capacity to deliver semi-remotely created multimedia content generation.

REFERENCES

[1]. J. Xing, M. Xia, Y. Liu, Y. Zhang, Y. Zhang, Y. He, H. Liu, H. Chen, X. Cun, X. Wang, Y. Shan, and T. T. Wong, "Make-Your-Video: Customized Video Generation Using Textual and Structural Guidance," *IEEE Transactions on Visualization and Computer Graphics*, vol. 31, no. 2, pp. 1526–1541, Feb. 2025.

[2]. A. Kesharwani, N. Bagdwal, D. Patel, N. Adigoppula, and M. E. Hossain, "Text-to-Digital Person Video Generator: DigitalAvatarGen," Kennesaw State University, Capstone Project Report, Dec. 2024.

[3]. Y. Zhang, Y. Wei, D. Jiang, X. Zhang, W. Zuo, and Q. Tian, "ControlVideo: Training-free controllable text-to-video generation," *arXiv preprint arXiv:2305.13077*, 2023.

[4]. L. Khachatryan, A. Movsisyan, V. Tadevosyan, R. Henschel, Z. Wang, S. Navasardyan, and H. Shi, "Text2Video-Zero: Text-to-image diffusion models are zero-shot video generators," in *Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV)*, 2023.

[5]. W. Hong, M. Ding, W. Zheng, X. Liu, and J. Tang, "CogVideo: Large-scale pretraining for text-to-video generation via transformers," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2023.

[6]. U. Singer, A. Polyak, T. Hayes, J. Yin, D. An, S. Li, and T. Baldridge, "Make-A-Video: Text-to-video generation without text-video data," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2023.

[7]. J. Ho, W. Chan, C. Saharia, W. Chan, D. Fleet, and M. Norouzi, "Imagen Video: High-definition video generation with diffusion models," *arXiv preprint arXiv:2210.02303*, 2022.

[8]. R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 10674–10685.

[9]. A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Machine Learning (ICML)*, 2021.

[10]. A. Ramesh et al., "Hierarchical text-conditional image generation with CLIP latents," *arXiv preprint arXiv:2204.06125*, 2022.

[11]. C. Saharia et al., "Photorealistic text-to-image diffusion models with deep language understanding," in *Proc. Conf. Neural Information Processing Systems (NeurIPS)*, 2022.

[12]. P. Esser, J. Chiu, P. Atighehchian, J. Granskog, and A. Germanidis, "Structure and content-guided video synthesis with diffusion models," *arXiv preprint arXiv:2302.03011*, 20923.

[13]. J. Z. Wu et al., "Tune-A-Video: One-shot tuning of image diffusion models for text-to-video generation," *arXiv preprint arXiv:2212.11565*, 2022.

[14]. J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. Conf. Neural Information Processing Systems (NeurIPS)*, 2020.

[15]. M. Bain, A. Nagrani, G. Varol, and A. Zisserman, "Frozen in time: A joint video and image encoder for end-to-end retrieval," in *Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV)*, 2021.

[16]. A. Blattmann et al., "Align your latents: High-resolution video synthesis with latent diffusion models," *arXiv preprint arXiv:2304.08818*, 2023.

[17]. A. Baevski et al., "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. Conf. Neural Information Processing Systems (NeurIPS)*, 2020.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26apr1128

[18]. K. Xu et al., "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. Int. Conf. Machine Learning (ICML)*, 2015.
[19]. X. Gu, C. Wen, J. Song, and Y. Gao, "SEER: Language instructed video prediction with latent diffusion models," *arXiv preprint arXiv:2303.14897*, 2023.
[20]. J. Shen et al., "Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, 2018.