IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

AUTOMATIC GENERATION OF SEGMENTED EDUCATIONAL VIDEOS FROM TEXT USING NLP AND GENERATIVE AI

[1]Potturi Reshma, [2]Mohammad Ali Akmal Baig, [3]Mallavalli Sai Vivek, [4]Shaik Jameer Ahmadh, [5]Parisa Deepthi

[1]Assistant Professor, [2]Student, [3]Student, [4]Student, [5]Student

[1]Department of Computer Science and Engineering,

[1]Seshadri Rao Gudlavalleru Engineering College, Gudlavalleru, India

Abstract: The increasing demand for engaging and accessible digital learning content has highlighted the limitations of traditional text-based educational materials. This paper presents an AI-based system that automatically converts educational text documents into segmented visual narration videos using a fully automated multimodal pipeline. The proposed framework integrates document parsing, Natural Language Processing (NLP), Large Language Models (LLMs), Text- to-Speech (TTS), and diffusion-based Text-to-Image (TTI) generation to transform static documents into structured, narrated video modules. Educational content is processed at the sentence level to enable fine-grained synchronization between narration, visuals, and subtitles. A robust primary–fallback strategy is employed across LLM, TTS, and image generation modules to ensure reliable operation. Experimental evaluation demonstrates that the system produces coherent and pedagogically suitable educational videos while significantly reducing manual effort and content production time through full pipeline automation. The proposed solution is scalable, cost-efficient, and well-suited for e-learning platforms, educators, and inclusive digital education delivery.

Index Terms – Artificial Intelligence, Natural Language Processing, Generative AI, Text-to-Speech, Text-to-Image, Educational Content Automation, Visual Narration, Multimodal Learning, E-Learning Systems

I. INTRODUCTION

The rapid expansion of digital education has led to an increasing demand for learning resources that are not only informative but also engaging and accessible to diverse learners. Traditional text- based educational materials, such as textbooks, PDFs, and lecture notes, often fail to sustain learner attention and may not effectively support visual and auditory learning styles. Research evidence shows that educational videos significantly improve learner engagement, knowledge retention, and conceptual understanding when compared to static textual content. As a result, video-based learning has become a dominant medium across online education platforms, virtual classrooms, and self- paced learning environments.

Recent advancements in Artificial Intelligence (AI) have played a transformative role in reshaping educational technologies. Bibliometric and empirical studies indicate that AI-driven systems are increasingly being adopted to automate content creation, personalize learning experiences, and improve instructional quality [2]. AI techniques enable the analysis and restructuring of educational content, allowing systems to adapt material for different learner needs. However, most existing educational content creation.

Processes remain heavily manual, requiring significant time, expertise, and resources to design scripts, visuals, and narration. Multimodal learning, which integrates text, audio, and visual information, has been shown to enhance cognitive processing and learning effectiveness. A comprehensive review of multimodal learning analytics highlights that combining multiple data modalities enables richer educational experiences and improved learner outcomes [3]. Despite these benefits, current educational tools often rely on limited modalities, such as text-only documents or pre-recorded videos, lacking dynamic generation and synchronization across modalities.

The emergence of Generative AI and large language models (LLMs) has opened new possibilities for automated educational content generation. Recent studies emphasize the transformative potential of multimodal generative models in education, particularly for tasks such as content explanation, visualization, and narration [4]. These models can generate human-like text, contextual images, and natural-sounding speech, making them well-suited for creating instructional materials that mimic real classroom teaching.

Although generative AI has demonstrated strong potential in educational video creation, several challenges remain. Existing approaches often lack fine-grained segmentation, contextual alignment between narration and visuals, and scalability for large document processing [5]. Furthermore, many systems are not designed with accessibility, multilingual support, or cost- efficiency in mind, limiting their adoption in real- world educational settings. These limitations highlight the need for an automated, AI-based framework that can convert educational text into segmented, synchronized, and visually narrated video content. Addressing this gap, the proposed system aims to deliver an intelligent and scalable solution for transforming static educational documents into interactive multimedia learning experiences.

II. LITERATURE REVIEW

[1] M. Zhao, W. Wang, T. Chen, R. Zhang, and R. Li, Recent advances in deep learning have enabled significant progress in multimodal video generation; however, existing methods often struggle to effectively integrate semantic information from multiple modalities such as text and audio. These limitations result in videos that lack contextual consistency and fine-grained semantic alignment with user inputs. To address these challenges, we propose a novel Text-and- Audio-Guided Video Maker (TAgVM) framework that leverages complementary information from textual descriptions and audio signals to generate semantically rich and visually coherent videos. By combining latent-space video representation learning with multimodal guidance mechanisms, the proposed approach aims to improve video quality, enhance semantic relevance, and support more expressive and controllable video generation across diverse application domains.

Yan et al. [7] proposed a neural video representation framework for continuous video generation. Their approach focused on learning temporal consistency across video frames using deep neural representations. Experimental results showed improved motion smoothness and visual quality compared to traditional methods. This study contributes to understanding temporal coherence in AI-generated videos.

Wu et al. [8] introduced T2VScore, a reliable evaluation metric for text-to-video generation models. The authors addressed limitations in existing evaluation techniques by incorporating semantic alignment and visual consistency measures. Their results demonstrated improved correlation with human judgment. This work highlights the importance of accurate evaluation in multimodal generation systems.

Mao et al. [9] proposed TAVG, a large-scale text-to-audio-visual generation framework supported by a 1.7M video dataset. The study leveraged contrastive latent diffusion models to generate synchronized audio-visual outputs. Results indicated strong alignment between text, audio, and visuals. This research validates the effectiveness of multimodal diffusion-based approaches.

Ma et al. [10] presented a pose-guided text-to- video generation method using pose-free videos. Their approach enhanced motion realism by implicitly learning pose representations. Experimental evaluations showed better motion accuracy and video quality. This work contributes to controllable video synthesis techniques.

Zhu et al. [11] proposed a controllable text- image-to-video generation framework. The system generated videos by combining static images and textual descriptions. Results demonstrated improved controllability and semantic consistency. This study supports multi-stage generation pipelines for video synthesis.

Singer et al. [12] introduced Make-A-Video, a text-to-video generation approach without requiring paired text-video datasets. The model leveraged large-scale image-text data for training. Results showed

promising video realism despite limited supervision. This work highlights data- efficient learning strategies in video generation.

Saito and Saito [13] proposed TGANv2 for efficient training of large-scale video generation models. The architecture utilized multiple subsampling layers to reduce computational cost. Experimental results demonstrated faster training with competitive video quality. This study addresses scalability challenges in generative video models.

Tian et al. [14] emphasized the importance of high-quality image generators for high-resolution video synthesis. Their experiments showed that strong image generation significantly improves video clarity. The findings highlight the interdependence between image and video generation tasks. This work is relevant to text-to- image-driven video systems.

Kim et al. [15] introduced TiVGAN, a step-by- step text-to-image-to-video generation framework. The system progressively evolved from text to image and then to video. Results showed enhanced semantic consistency across stages. This approach supports modular generation pipelines.

Karessli et al. [16] explored zero-shot image classification using human gaze as auxiliary information. The study demonstrated improved classification performance without labeled data. The findings highlight the value of auxiliary signals in visual understanding. This research informs attention-aware visual generation techniques.

Kumar et al. [17] proposed consistent generative query networks for future video frame prediction. Their method focused on temporal consistency and long-term prediction accuracy. Experimental results showed reduced flickering and better motion continuity. This study addresses key challenges in video temporal modeling.

Brooks and Barron [18] presented a neural network-based method for generating motion blur from static images. The approach improved realism in generated motion effects. Results demonstrated enhanced visual authenticity. This work contributes to motion modeling in video synthesis.

Kim et al. [19] extended TiVGAN for text-to- image-to-video generation using GAN architectures. The model demonstrated strong semantic preservation across modalities. Experimental evaluations confirmed improved visual quality. This study reinforces GAN-based multimodal generation methods.

Aldausari et al. [20] provided a comprehensive review of video GAN models. The survey categorized architectures, training strategies, and applications. The authors identified stability and scalability as major challenges. This review provides foundational insights into video generative modeling.

Meadows et al. [21] introduced PhysNLU, a tool for evaluating natural language understanding in physics education. The system assessed conceptual comprehension rather than surface- level accuracy. Results showed improved evaluation reliability. This study highlights the importance of semantic understanding in educational AI.

Yang et al. [22] proposed video creation using diffusion probabilistic models. Their method achieved high-quality video generation with improved temporal stability. Experimental results validated the effectiveness of diffusion-based approaches. This work laid groundwork for modern video diffusion models.

Ho et al. [23] introduced video diffusion models for generating realistic video sequences. The approach modeled temporal dynamics through iterative denoising. Results demonstrated strong visual fidelity and motion coherence. This study significantly advanced diffusion-based video synthesis.

Hong et al. [24] proposed CogVideo, a transformer-based large-scale text-to-video generation model. The system leveraged massive pretraining to enhance semantic alignment. Experimental results showed competitive performance on benchmark datasets. This work highlights the role of transformers in multimodal generation.

Singer et al. [25] extended Make-A-Video for improved text-to-video generation without paired data. The model demonstrated strong generalization across domains. Results confirmed improved video realism. This study emphasizes data-efficient multimodal learning.

Ho et al. [26] introduced Imagen Video for high-definition video generation using diffusion models. The system achieved superior resolution and temporal consistency. Experimental evaluations showed state-of-the-art results. This work demonstrates the scalability of diffusion models for video synthesis.

Wang et al. [27] presented DiffusionDB, a large-scale dataset of prompts for text-to-image models. The dataset enabled improved prompt engineering and model evaluation. Results showed better generation diversity. This resource supports effective text-to-image generation pipelines.

Khachatryan et al. [28] proposed Text2Video- Zero, a zero-shot video generation approach using text-to-image diffusion models. The system generated videos without video-specific training. Experimental

results demonstrated competitive performance. This work supports modular zero- shot generation strategies.

Reed et al. [29] introduced one of the earliest text-to-image synthesis models using GANs. The study demonstrated semantic alignment between text and generated images. Results laid the foundation for modern generative models. This work remains influential in multimodal AI research.

Brade et al. [30] proposed Promptify, an interactive prompt exploration tool for text-to- image generation. The system enabled users to refine prompts for better visual outputs. Experimental studies showed improved generation control. This work highlights the importance of prompt engineering in generative AI systems.

III. METHODOLOGY

The proposed system follows a modular, AI- driven pipeline to automatically convert educational text documents into segmented visual narration videos. Initially, the input document is uploaded through a web interface and processed using a document parsing module that supports PDF, DOCX, and plain text formats. PDF files are parsed using PyMuPDF, while DOCX files are processed using python-docx to extract structured textual content.

The extracted text is then passed to a Natural Language Processing (NLP) stage, where a large language model (LLM) is used to clean, summarize, and restructure the content into narration-style text. This step ensures improved readability and pedagogical clarity. To enhance system reliability, a fallback mechanism is implemented using an alternative LLM provider in case the primary model fails.

The refined narration text is segmented into sentence-level units, which form the basis for multimodal generation. Each sentence is converted into expressive audio narration using a Text-to-Speech (TTS) module. The system primarily employs an advanced neural TTS engine for natural voice synthesis, with an automatic fallback to a secondary TTS service to ensure robustness. Sentence-level audio durations are recorded to enable precise synchronization.

In parallel, each sentence is enhanced into a detailed visual prompt using the LLM and passed to a generative text-to-image (TTI) module. Contextual images are generated using diffusion- based image generation models with a primary- fallback strategy to maximize success rates.

Finally, the narration audio, generated visuals, and subtitles are synchronized using timestamp- based alignment and rendered into a segmented educational video using FFmpeg. Each visual frame is displayed for the exact duration of its corresponding narration sentence, ensuring strong semantic and temporal coherence throughout the video.

3.1 Proposed System

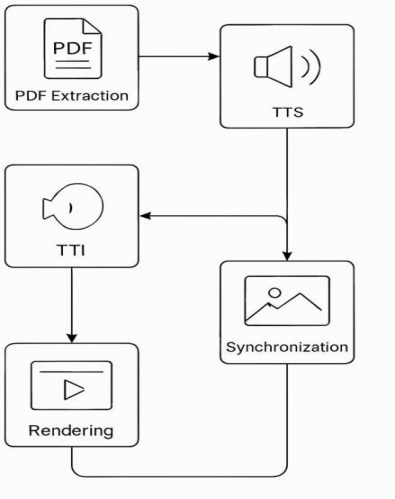
The proposed system introduces an AI-based automated framework designed to convert educational text documents into segmented visual narration videos. The system begins by accepting text-based inputs such as PDFs or plain documents, which are parsed and segmented using Natural Language Processing (NLP) techniques. Each segment is analyzed to extract key concepts and generate clear, structured explanations suitable for multimedia presentation. This processed content forms the foundation for automated narration and visual generation, enabling the transformation of static text into dynamic educational modules.

Subsequently, the system integrates Text-to- Speech (TTS) technology to generate natural- sounding audio narration for each text segment and Generative AI-based text-to-image models to create contextual visuals that support conceptual understanding. The narration, images, and subtitles are synchronized using timestamp-based alignment to ensure semantic consistency and smooth playback. The final output is a segmented video where each concept is presented with corresponding narration and visuals, offering an interactive, accessible, and scalable learning experience for educators and learners.

3.2 System Architecture

The system architecture is designed as a layered and modular framework to efficiently convert educational text into segmented visual narration videos. The architecture begins with the Input Layer, where users upload text-based educational documents such as PDFs. This layer performs document parsing and text extraction to convert unstructured content into structured text segments. The processed text is then passed to the NLP Processing Layer, which handles sentence segmentation, keyword extraction, and content refinement to ensure that only relevant educational information is forwarded for multimedia generation.

Following this, the Multimodal Generation Layer converts each refined text segment into



corresponding audio and visual components. The Text-to-Speech (TTS) Module generates natural-sounding narration, while the Text-to-Image (TTI) Module produces contextual AI-generated images. These outputs are coordinated by the Synchronization Layer, which aligns narration, visuals, and subtitles using timestamps to maintain semantic coherence. Finally, the Rendering Layer compiles all synchronized components into segmented video modules, delivering a cohesive and accessible educational video output suitable for scalable e-learning applications.

Fig 1: Proposed Architecture

3.3 Modules

1. Input & Document Parsing Module:

This module handles user input by accepting educational documents such as PDFs or text files. It extracts raw text from the uploaded document and converts unstructured content into a machine-readable format. Proper parsing ensures that headings, paragraphs, and sentences are preserved for further processing.

2. NLP-Based Text Processing Module:

In this module, Natural Language Processing (NLP) techniques are applied to clean, segment, and analyze the extracted text. Important concepts and keywords are identified, and the content is divided into meaningful sentence-wise or concept- wise segments. This step ensures clarity and prepares the text for narration and visual generation.

3. LLM-Based Content Enhancement Module:

This module uses a Large Language Model (LLM) to enhance each text segment by generating clear, teacher-like explanations. The enhanced text improves narration quality and helps create detailed prompts for visual generation. It bridges the gap between raw text and engaging educational storytelling.

4. Text-to-Speech (TTS) Module:

The refined text segments are converted into natural-sounding audio narration using advanced Text-to-Speech technology. This module focuses on clarity, proper pacing, and expressive tone to improve learner engagement and accessibility for auditory learners.

5. Text-to-Image (TTI) Generation Module:

Each enhanced text segment is passed to a Generative AI-based text-to-image model to produce contextual visuals. These images are designed to represent the concepts explained in the narration, supporting visual learning and improving content comprehension.

6. Synchronization Module:

This module aligns the generated narration audio, contextual images, and subtitles using timestamp-based synchronization. It ensures that each sentence is displayed with the correct visual and audio, maintaining semantic and temporal consistency throughout the video.

7. Video Rendering & Output Module:

The final module compiles all synchronized components into segmented educational videos. It combines audio narration, visuals, and subtitles into a cohesive multimedia output that can be downloaded or streamed, providing an interactive and accessible learning experience.

3.4 Algorithms

3.4.1 NLP-Based Text Extraction and Segmentation Algorithm

This algorithm is used to extract textual content from uploaded PDF documents and divide it into meaningful sentences or concept-wise segments. It applies tokenization, sentence boundary detection, and keyword filtering to ensure that only relevant educational content is processed further. This step forms the foundation for accurate narration and visual generation.

Input: Uploaded document D Output: Clean narration text T

1. Detect file format of D
2. If D is PDF, extract text using PDF parser
3. If D is DOCX, extract text using document parser
4. If D is TXT, read raw text
5. Pass extracted text to LLM for cleaning and summarization
6. If primary LLM fails, use fallback LLM
7. Return refined narration text T

3.4.2 Text-to-Speech (TTS) Synthesis Algorithm

The TTS algorithm converts enhanced text into natural-sounding audio narration. It focuses on pronunciation accuracy, intonation, and speech pacing to produce expressive and human-like narration, improving accessibility for auditory learners.

Input: Narration text T

Output: Audio file A and sentence durations {d1, d2, ..., dn}

1. Split T into sentences S = {s1, s2, ..., sn}
2. For each sentence si:
3. Generate audio using neural TTS
4. If TTS fails, use fallback TTS
5. Store audio ai and record its duration di
6. Concatenate all ai to form final audio A
7. Return A and durations {di}

3.4.3 Text-to-Image (TTI) Generation Algorithm

This algorithm generates contextual images from enhanced text prompts using generative diffusion-based models. The generated visuals are aligned with the narration content, enabling effective visual representation of educational concepts.

Input: Sentence list S

Output: Visual frames V = {v1, v2, ..., vn}

1. For each sentence si:
2. Generate enhanced visual prompt pi using LLM
3. Generate image vi using primary TTI model
4. If generation fails, use fallback TTI model
5. Resize and store vi
6. Return visual frame set V

3.4.4 Segmented Visual Narration Generation Algorithm

Input: Educational document D
Output: Segmented narrated video V

1. Extract raw text T from document D
2. Enhance and summarize T using LLM-based NLP processing
3. Split enhanced text into sentence list S = {s1, s2, ..., sn}
4. for each sentence si in S do
5. Generate narration audio ai using TTS
6. Record duration di of ai
7. Generate enhanced visual prompt pi using LLM
8. Generate contextual image vi from pi using TTI model
9. end for
10. Generate subtitles using sentence timings {di}
11. Render video frames vi with duration di using FFmpeg
12. Merge audio, visuals, and subtitles to produce final video V
13. Return V

IV. EXPERIMENTAL RESULTS

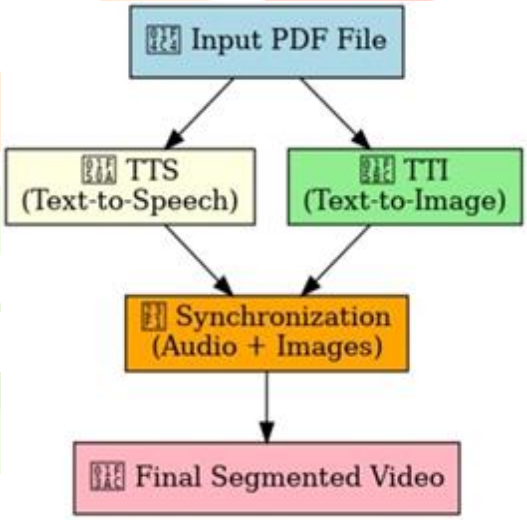


Fig 2: Plan of Action Architecture

The proposed system was evaluated using multiple educational documents in PDF, DOCX, and text formats. The evaluation focused on system functionality, synchronization accuracy, processing efficiency, and reduction in manual effort rather than traditional classification metrics, as the task involves multimodal content generation rather than prediction.

During experimentation, the document parsing module successfully extracted structured text from all supported formats. The NLP processing stage produced narration-style text that was suitable for sentence-level segmentation and audio generation. The sentence-level segmentation enabled fine-grained control over synchronization between narration and visuals.

The Text-to-Speech (TTS) module generated natural and expressive narration for each sentence. Sentence durations were captured automatically and used as timing constraints during video rendering. The system's fallback mechanism ensured uninterrupted narration generation even when the primary TTS service was unavailable.

The Text-to-Image (TTI) module generated contextually relevant visuals for the majority of narration segments. A primary-fallback strategy using multiple diffusion-based models improved visual generation reliability. Generated images were resized and standardized before video rendering to ensure consistency.

The synchronization module demonstrated precise alignment between narration, visuals, and subtitles by using sentence-level timestamps. Each visual frame was displayed exactly for the duration of its corresponding narration audio, eliminating visual-audio mismatch and abrupt transitions.

Compared to traditional manual video creation workflows, the proposed system significantly reduced content creation time and effort. The entire pipeline operated automatically after document upload, requiring minimal human intervention. These results validate the effectiveness, scalability, and practical applicability of the proposed framework for automated educational video generation.

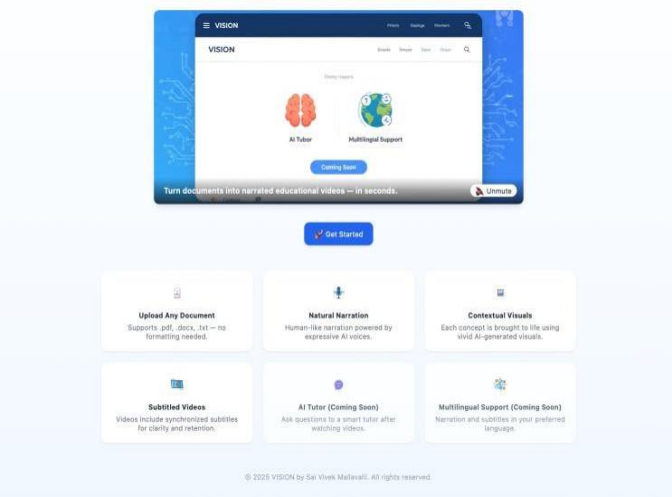


Fig 3: VISION Home Page Interface

This diagram represents the home page of the VISION system, which introduces the platform's core objective—converting documents into narrated educational videos. The interface highlights key features such as natural narration, contextual visuals, multilingual support, and AI tutoring. It provides users with a brief overview of the system's capabilities and encourages interaction through the "Get Started" option. This page serves as the entry point, helping users understand the purpose and benefits of the application.

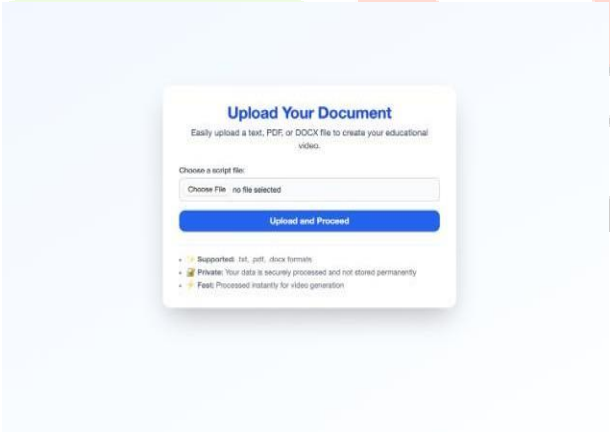
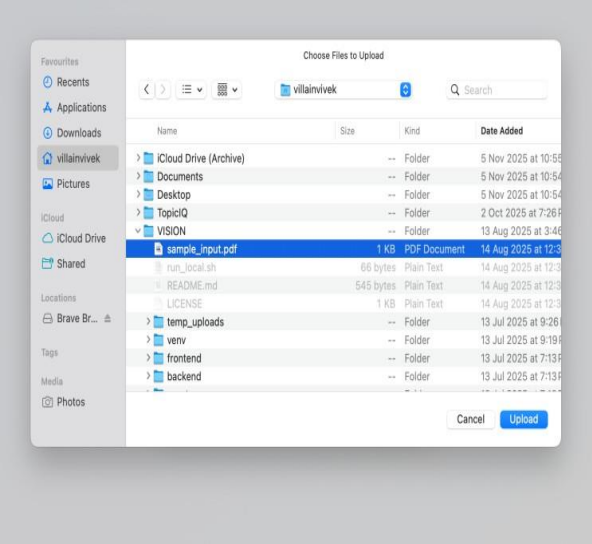


Fig 4: Document Upload Interface

This diagram shows the document upload screen, where users can upload educational files in formats such as PDF, TXT, or DOCX. The interface is designed to be user-friendly, allowing simple file selection and upload. Additional information regarding privacy, supported formats, and fast processing



reassures users about data security and system efficiency. This step initiates the content conversion pipeline.

Fig 5: File Selection Window

This diagram illustrates the file selection process from the local system. Users browse their directories and select the desired educational document for conversion. This step confirms that the system supports real-world file handling and seamless integration with user devices. Once selected, the file is passed to the backend for text extraction and processing.

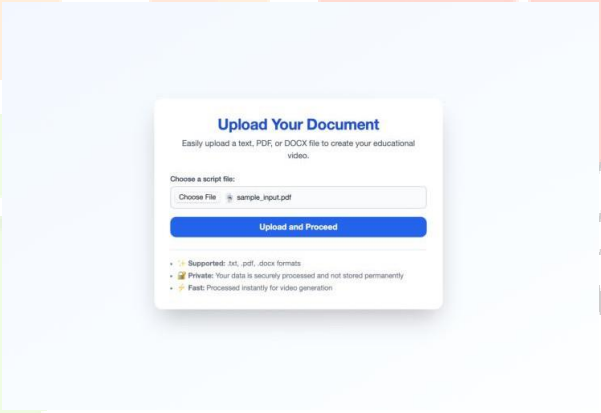


Fig 6: Uploaded Document Confirmation Screen

This diagram displays the interface after a document has been successfully selected. The uploaded file name is shown, confirming that the system has received the input correctly. The “Upload and Proceed” option allows users to continue to the next stage of processing. This step ensures transparency and prevents accidental submission of incorrect files.

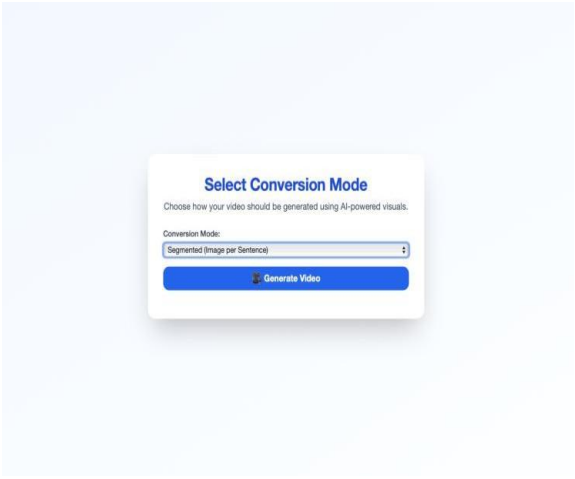


Fig 7: Conversion Mode Selection

This diagram represents the conversion mode selection screen, where users choose how the video should be generated. In the segmented mode, each sentence or concept is converted into a separate visual-narration segment. This feature allows structured and concept-wise learning, making the system suitable for educational use. Once the mode is selected, video generation begins.

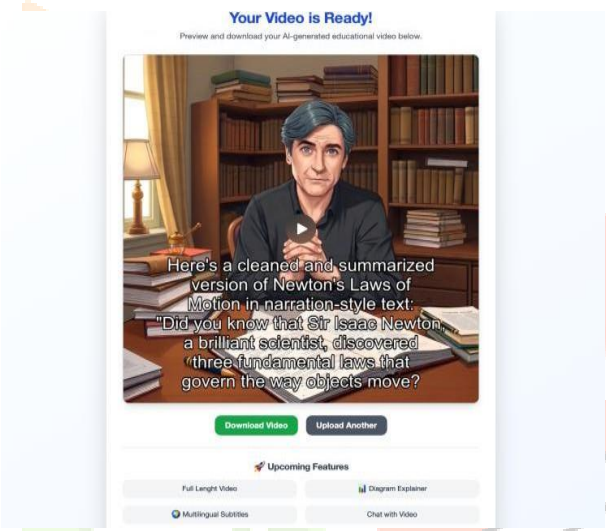


Fig 8: Generated Educational Video Output

This final diagram shows the output screen, where the AI-generated educational video is displayed. The video includes synchronized narration, contextual visuals, and subtitles, providing a complete multimedia learning experience. Users can preview, download the video, or upload another document. This screen demonstrates the successful transformation of static educational text into an engaging, segmented visual narration video.

"The system was tested using multiple educational documents of varying length, demonstrating stable performance across different document formats and content structures."

Metric	Observed Outcome
Supported Document Formats	PDF, DOCX, TXT
Average segmentation granularity	Sentence-level
Synchronizatio n method	Timestamp-based
Visual generation strategy	Primary + fallback TTI models
Narration generation	Neural TTS with fallback

Video rendering tool	FFmpeg

Table 1: Observed System Performance Metrics

Comparision:

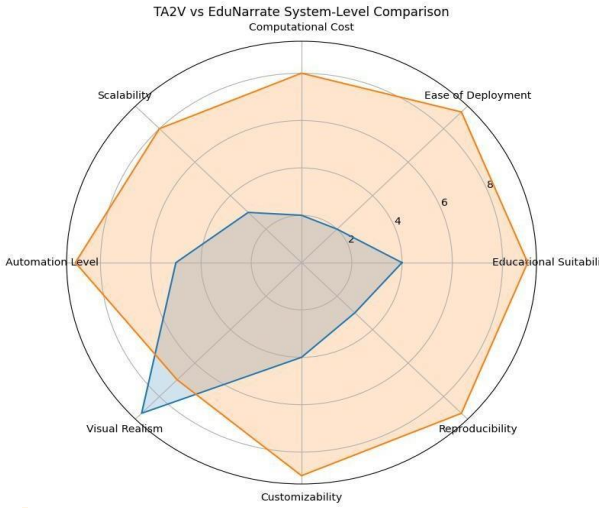


Fig 9: Comparision between existing and proposed system based on metrics considered.

"Figure 9 presents a polygon-based comparison between the TA2V research framework and the proposed EduNarrate system across key system- level metrics. While TA2V achieves higher visual realism, EduNarrate demonstrates superior performance in deployment feasibility, scalability, reproducibility, and educational applicability, making it more suitable for large-scale educational content generation."

V. CONCLUSION

This work presented an AI-based system for the automatic conversion of educational text into segmented visual narration videos, addressing the limitations of traditional text-based learning materials. By integrating Natural Language Processing (NLP), Text-to-Speech (TTS), and Text-to-Image (TTI) generation, the proposed system effectively transforms static documents into interactive and engaging multimedia content. The modular architecture enables efficient text extraction, content enhancement, narration generation, and contextual visual synthesis, ensuring that each educational concept is clearly explained and visually supported.

The experimental results demonstrate that the proposed framework successfully generates coherent, synchronized, and high-quality educational videos with minimal human intervention. The parallel execution of audio and image generation reduces processing time, while timestamp-based synchronization ensures semantic and temporal alignment between narration, visuals, and subtitles. The segmented video structure allows learners to focus on individual concepts, improving comprehension and reducing cognitive overload. Additionally, the system significantly lowers the manual effort, cost, and time associated with traditional educational video production.

Overall, the proposed system proves to be a scalable and practical solution for modern e- learning environments. It supports accessibility through audio narration and subtitles, benefits both visual and auditory learners, and can be deployed using a combination of open-source and cloud- based tools. The framework has strong potential for adoption in educational institutions, online learning platforms, and content creation workflows. Future enhancements may include real-time processing, adaptive learning personalization, and learner performance analytics to further improve educational effectiveness.

REFERENCES

[1] M. Zhao, W. Wang, T. Chen, R. Zhang, and R. Li, "TA2V: Text-Audio Guided Video Generation," IEEE Transactions on Multimedia, IEEE, 2023.

[2] K. Kavitha et al., "The Transformative Trajectory of Artificial Intelligence in Education: A Bibliometric Analysis," Journal of Educational Computing Research, 2024.

[3] M. Mohammadi et al., "Artificial Intelligence in Multimodal Learning Analytics: A Systematic Literature Review," Computers and Education: Artificial Intelligence, 2025.

[4] A Bewersdorff et al., "Taking the next step with generative artificial intelligence: The transformative role of multimodal large language models in education," Journal of Computer Assisted Learning, 2025.

[5] Q. Zhang and L. Chen, "Exploring the potential of generative AI for educational video creation: opportunities, challenges, and future directions," Education and Information Technologies, 2024.

[6] Y. Wang and Y. Zhang, "AI-powered tools for automated educational content creation: A systematic review," Computers & Education, vol. 104815, 2024.

[7] K. Yan, Y. Lin, and Y. Qiao, "Neural Video Representation for Continuous Video Generation," in Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), 2024.

[8] J. Zhangjie Wu, Y. Li, and B. Zhou, "T2VScore: A Reliable Metric for Text-to-Video Generation Evaluation," in Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), 2024.

[9] M. Mao et al., "TAVG: Text-to-Audio-Visual Generation with 1.7M Video Dataset and Contrastive Latent Diffusion," arXiv preprint arXiv: 2403.00123, 2024.

[10] Y. Ma et al., "Pose-Guided Text-to-Video Generation using Pose-Free Videos," in Proc. International Conf. on Computer Vision (ICCV), 2023.

[11] Z. Zhu et al., "Text-Image-to-Video Generation: Controllable Video Synthesis from Static Images and Text Descriptions," in Proc. IEEE/CVF International Conf. on Computer Vision (ICCV), 2023.

[12] Singer et al., "Make-A-Video: Text-to- Video Generation without Text-Video Data," arXiv preprint arXiv: 2209.14792, 2022.

[13] M. Saito and S. Saito, "TGANv2: Efficient Training of Large Models for Video Generation with Multiple Subsampling Layers," arXiv preprint arXiv: 1811.09245, 2020.

[14] Y. Tian et al., "A Good Image Generator Is What You Need for High-Resolution Video Synthesis," arXiv preprint arXiv: 2104.15069, 2021.

[15] D. Kim, D. Joo, and J. Kim, "TiVGAN: Text to Image to Video Generation with Step-by-Step Evolutionary Generator," arXiv preprint arXiv: 2009.02018, 2020.

[16] Karessli, N., Akata, Z., Schiele, B., & Bulling, A. (2017). Zero-Shot Image Classification using Human Gaze as Auxiliary Information. In Proceedings of the IEEE Conference

[17] Kumar, A., Eslami, S. M. A., Rezende, D., Garnelo, M., Viola, F., Lockhart, E., & Shanahan, M. (2019). Consistent generative query networks for future frame prediction in videos. arXiv preprint arXiv:1807.02033.

[18] Brooks, T., & Barron, J. T. (2019). Generating motion blur from unblurred photos using neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 6840-6848. doi:10.1109/cvpr.2019.06840

[19] Kim, D., Joo, D., & Kim, J. (2020). TiVGAN: Text-to-image to-video generative adversarial network. IEEE Access, 8, 153113-153122. doi:10.1109/access.2020.2986494

[20] Aldausari, N., Sowmya, A., Marcus, N., & Mohammadi, G. (2022). Review of video generative adversarial networks (GANs) models. ACM Computing Surveys, 55(2), Article 30. doi:10.1145/3487891.

[21] Meadows, J., Zhou, Z., & Freitas, A. (2022). PhysNLU: A tool for evaluating natural language understanding in physics. In Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), 4904-4912. doi:10.18653/lrec-2022-4904.

[22] Yang, R., Srivastava, P., & Mandt, S. (2022). Video creation using diffusion probabilistic models. arXiv preprint arXiv:2203.09481.

[23] Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., & Fleet, D. J. (2022). Video generation using video diffusion models. arXiv preprint arXiv:2204.03409.

[24] Hong, W., Ding, M., Zheng, W., Liu, X., & Tang, J. (2022). CogVideo: Large-scale pretraining for transformer-based text-to-video generation. arXiv preprint arXiv:2205.15868.

[25] Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., Zhang, S., Hu, Q., Yang, H., Ashual, O., Gafni, O., Parikh, D., Gupta, S., & Taigman, Y. (2022). Make-a-Video:Text-to-Video Generation without Text-Video Data. arXiv preprint arXiv:2209.14792.

[26] Ho, J., Chan, W., Saharia, C., Whang, J.,Gao, R., Gritsenko, A., Kingma, D. P., Poole, B., Norouzi, M., Fleet, D. J., & Salimans, T. (2022). Imagen Video: High-Definition Video Generation using Diffusion Models. arXiv preprint arXiv:2210.02303.

[27] Wang, Z. J., Montoyo, E., Munechika, D.,Yang, H., Hoover, B., & Chau, D. H. (2022). DiffusionDB: A sizable prompt gallery dataset for text-to-image generative models. arXiv preprint arXiv:2210.11890.

[28] Khachatryan, L., Movsisyan, A., Tadevosyan, V., Henschel, R., Wang, Z., Navasardyan, S., & Shi, H. (2023). Text2Video- Zero: Zero-Shot Video Generation using Text-to Image Diffusion Models. arXiv preprint arXiv:2303.13439.

[29] Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., & Lee, H. (2016). Text-to-image synthesis using generative adversarial networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), doi:10.1109/cvpr.2016.296.

[30] Brade, S., Wang, B., Sousa, M., Oore, S., & Grossman, T. (2023). Promptify: Interactive prompt exploration for text to-image generation. arXiv preprint arXiv:2304.09337.