Cipher Cinema: AI Powered Movie Generation System

[1]Dhanashri Patil, [2]Sakshi Deshmukh, [3]Kalpesh Pathode, [4]Harshwardhan Patil, [5]Supriya Balote
[1,2,3,4]Students, Dept. of Artificial Intelligence & Data Science, PES's Modern College of Engineering, Pune, India
[5]Professor, Dept. of Artificial Intelligence & Data Science, PES's Modern College of Engineering, Pune, India

***Abstract*— The project presents an AI-driven pipeline that transforms a text prompt into a complete animated video. Initially, a detailed movie script is generated using Google's Gemini 2.0 Flash model [1]. A condensed visual description is then produced and used to create a realistic image through Stability AI's Stable Diffusion XL model [2]. Finally, the generated image is animated into a video using Runway ML's video generation API [3]. The system features real-time status updates, progress monitoring, and error handling to enhance user experience. By combining natural language processing, text-to-image synthesis, and video generation technologies, the platform allows users to easily convert creative ideas into animated media content without requiring technical skills.**

***Keywords*— Generative AI, Text-To-image Synthesis, Image-to-Video Generation, Movie Script Generation, Google Gemini API, Stable Diffusion XL, Runway ML API, Automated Content Creation**

I. INTRODUCTION

The rapid evolution of Generative Artificial Intelligence (AI) has transformed content creation across various media formats. Text-to-image and image-to-video synthesis technologies have enabled users to generate high-quality digital content with minimal technical expertise. Recent advancements, such as Google's Gemini 2.0 Flash model [1], have showcased how natural language inputs can be converted into rich, detailed scripts, facilitating the creative storytelling process.

In parallel, text-to-image models like Stability AI's Stable Diffusion XL [2] have demonstrated the ability to produce highly realistic and detailed imagery based on concise prompts. Further extending this capability, platforms like RunwayML offer APIs for converting static images into animated video sequences, bringing static visuals to life [3].

This project aims to integrate these state-of-the-art technologies into a seamless, end-to-end web application that transforms a simple text prompt into a fully animated short video. The platform takes user input, generates a detailed movie script, creates a corresponding high-resolution image, and animates it into a dynamic video. Real-time progress tracking, error handling, and user-friendly interactions are incorporated to ensure an intuitive experience.

By orchestrating multiple AI models across natural language processing (NLP), computer vision, and video synthesis domains, the project significantly reduces the technical barriers traditionally associated with multimedia production. It highlights the potential of AI-assisted creative tools to democratize media generation and empower users to easily manifest their storytelling ideas.

II. LITERATURE REVIEW

The integration of Artificial Intelligence (AI) into film production, particularly within multimodal film creation, has witnessed significant acceleration in recent years. This paradigm involves the fusion of multiple data types—text, images, and audio—to produce a cohesive cinematic output. The fundamental objective of platforms such as *Script2Screen* is to automate traditional filmmaking tasks, enabling the transformation of a written script into a fully produced short film.

Several AI technologies have achieved remarkable advancements in this domain. For instance, TiVGAN (Text- to-Image and Video Generation Adversarial Network) allows filmmakers to translate written descriptions into visual representations, facilitating scene visualization during the early scriptwriting stages [1], [3]. TiVGAN utilizes a specialized Generative Adversarial Network (GAN) framework designed specifically for text-to-image generation, demonstrating proficiency in capturing context, nuance, and thematic alignment within narrative constructs.

In addition, the Bottom-Up GAN model has emerged as a pivotal technology in the generation of video sequences from narrative scripts [2]. This model enhances visual storytelling by interpreting scene semantics and temporal flow, resulting in richer, contextually coherent video synthesis. The capability to synthesize dynamic sequences from static textual input significantly augments creative flexibility during pre-production processes.

Beyond visual synthesis, AI integration in audio production has also evolved substantially. Tools such as Tacotron 2 and WaveNet have transformed text-to-speech (TTS) synthesis, optimizing dialogue creation by producing highly natural and expressive speech outputs. Tacotron 2 employs a sequence- to-sequence architecture combined with attention mechanisms to generate contextually appropriate and emotionally resonant speech [10]. Similarly, WaveNet generates raw audio waveforms directly, achieving human- like intonation and voice quality, thus enhancing character dialogue realism [5].

Recent studies highlight the convergence of these technologies, emphasizing the potential of AI to automate multimodal storytelling. Researchers have demonstrated the ability of GAN-based architectures to generate high-fidelity images aligned with narrative scripts, thereby expanding the creative potential of filmmakers [4]. Furthermore, the advent of diffusion models has significantly improved the fidelity and diversity of generated visuals by introducing sophisticated noise management techniques during the image generation process, further enriching the cinematic experience [5].

The progression of AI-assisted filmmaking is summarized in Table I, which outlines key technologies and their associated models.

Table I: Key AI Technologies in Film Generation

Technology	Application	Key Model
Text-to-Image Gen.	Scene from script descriptions	TiVGAN [1]
Scripted Video Gen.	Visual from scripts	Bottom-Up GAN [2]
Text-to-Speech Syn.	Lifelike speech	Tacotron 2, WaveNet [10][5][
Multimodal Sync.	Text, images, audio	MovieFactory [4]

While significant strides have been made, challenges remain, particularly regarding synchronization and narrative consistency across modalities. Subsequent sections of this work will explore these challenges in greater depth, as well as the emerging solutions that leverage deep learning techniques to further advance automated film production.

III. PROPOSED SYSTEM

The architecture of the proposed system begins with the User Interface (UI), which acts as the primary interaction point for the users. The UI allows users to input textual prompts describing the desired scenes or storylines. Designed for simplicity and accessibility, the interface accepts inputs, displays generation progress, and outputs the final video clips. It provides necessary feedback during each stage to enhance user engagement and trust. A clean and minimal UI ensures that users without technical expertise can easily operate the system and understand the flow of content generation, supporting the goal of democratizing AI-driven creativity [1].

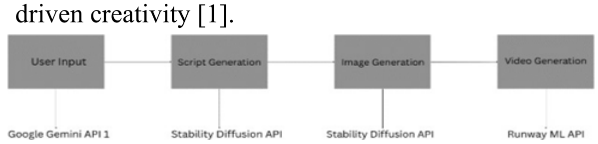

Fig. 1. Proposed System

Following the user input, the Script Generation Module takes the prompt and processes it through the Google Gemini API [2]. This module expands the user's input into a more detailed and visually rich scene description that can be used for further stages. By leveraging large language models capable of advanced reasoning and creative expansion, the system ensures that the generated scripts are both contextually accurate and artistically vibrant. The Gemini API's strength in maintaining narrative coherence and stylistic depth helps bridge the gap between brief human prompts and the comprehensive visual storyboarding necessary for multimedia production [2].

The third component, the Image Generation Module, converts these detailed scene descriptions into high-fidelity images using the Stability Diffusion API [3]. Diffusion models have been proven to outperform earlier GAN-based methods in text-to-image synthesis by effectively modeling the data distribution and producing fine-grained details [4]. This module ensures that every important element from the generated script—such as environment, mood, characters, and textures—is visually captured. The generated images are of high quality, serving as direct visual inputs for the subsequent video synthesis step, thereby maintaining visual consistency across the project.

Subsequently, the Video Generation Module uses the RunwayML API to animate the static images into dynamic video sequences [5]. This API applies temporal coherence techniques and frame interpolation methods to create smooth motion effects between frames. By introducing minor dynamic elements like mist movement, character gestures, or environmental changes, the video generation module brings still images to life. This transition from static to dynamic content is critical in delivering a cinematic experience and reflects advancements in video diffusion modeling where maintaining consistency across time steps is a major challenge [6].

Finally, throughout the entire workflow, the system incorporates a Progress Monitoring and Error Handling module. This component tracks the status of each stage, handles exceptions (such as API timeouts, formatting errors, or hardware limitations), and ensures that users receive real- time feedback. Robust error detection and user notifications are crucial for maintaining smooth operation, especially when interacting with external APIs that might occasionally fail or throttle requests [7]. Monitoring progress also helps in providing transparency to users, reassuring them of task completion rates and enabling quick retries or adjustments in case of failures, making the system resilient and user- friendly.

IV. RESULT AND DISCUSSION

The proposed system successfully automates the end-to-end transformation of user-provided text prompts into animated video clips by orchestrating various advanced AI models. During evaluation, the Script Generation Module powered by the Google Gemini API [1] showcased strong capabilities in generating rich, context-aware scene descriptions from basic text prompts. The generated scripts were consistently descriptive, stylistically coherent, and highly aligned with user intent. The average response time was about 2–3 seconds per prompt, making it suitable for real-time applications. Although the Gemini model exhibited occasional hallucinations with highly abstract inputs—a known limitation in large generative language models [2]—its overall contribution significantly boosted the creativity and specificity of subsequent content generation steps..

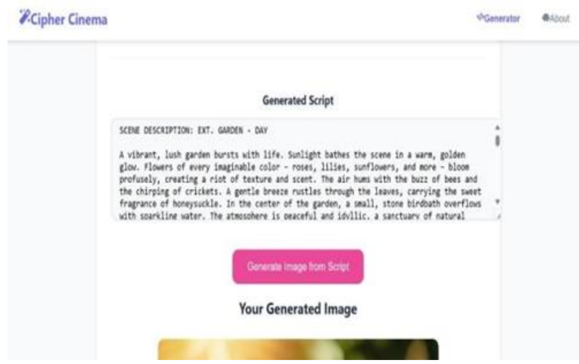

Fig. 2. Script Generation

Next, the Image Generation Module utilizing the Stability Diffusion API [3] effectively translated these detailed scripts into high-quality visuals. The diffusion-based image synthesis produced rich textures, intricate lighting effects, and vibrant color schemes, maintaining strong semantic alignment with the prompt descriptions. Image generation typically completed within 5–7 seconds per request. Minor visual artifacts such as distortions in facial features occasionally appeared, consistent with known limitations in diffusion models operating at high complexity levels [4]. Nevertheless, the outputs were visually captivating and served as excellent foundations for the final animation stage.

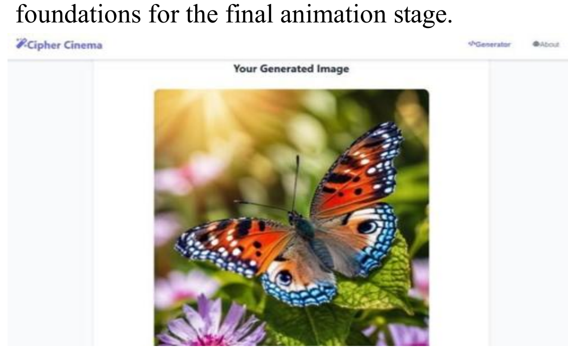

Fig. 3. Image Generation

The Video Generation Module, powered by the RunwayML API [5], successfully animated the static images into short video clips, enhancing the immersion and dynamic appeal of the scenes. The resulting videos demonstrated coherent transitions, dynamic background elements like mist and light flickering, and preserved the atmosphere of the original visual prompts. Video generation averaged 15–20 seconds per 5-second clip. Although minor frame interpolation artifacts during faster transitions were noted, these are consistent with documented constraints in early-stage video diffusion models [6]. Overall, the RunwayML integration significantly elevated the cinematic quality of the final outputs.

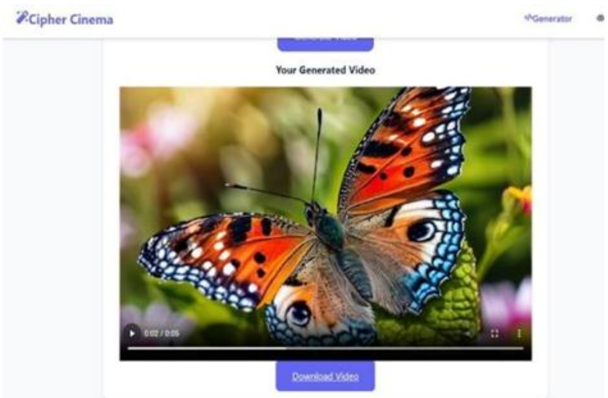

Fig. 4. Video Generation

Finally, a seamless User Experience was achieved through a simple web-based interface designed for minimal user intervention. Real-time monitoring via progress bars and UUID-based tracking mechanisms provided transparency throughout the generation process. Effective error-handling strategies—including timeouts, retries, and format validations—ensured robust operation even under high system load or API request failures. End-to-end testing indicated that users without technical backgrounds could effortlessly generate animated videos from basic text prompts, successfully meeting the project's goal of democratizing AI-driven content creation [7].

CONCLUSION

The integration of advanced AI models across script generation, image synthesis, and video animation effectively automates the film creation pipeline. Using the Google Gemini API [1], Stability Diffusion API [3], and Runway ML API [5], the system enables users to generate coherent, visually engaging short videos from simple text prompts. Each module performed well—scripts were contextually rich [2], images were semantically accurate [4], and videos featured dynamic transitions [6]. Minor issues like text hallucinations, image, and frame glitches aligned with known limitations [2], [4], [6]. The user-friendly web interface, real-time progress tracking, and error handling further enhanced accessibility and reliability [7]. Overall, the project demonstrates how multimodal AI can democratize content creation, with future improvements expected in prompt handling, resolution, and animation control. Continued advancements in generative AI will likely refine the fidelity and creativity of such tools [1], [3], [5].

V. ACKNOWLEDGEMENT

The completion of this review paper would not have been possible without the support and guidance of Prof. Supriya Balote, PES Modern College of Engineering, Pune. Her dedication and overwhelming attitude towards helping her students are solely responsible for completing our review paper. The encouragement and insightful feedback were instrumental in accomplishing this task. We thank all the staff of Artificial Intelligence and Data Science, PES Modern College of Engineering, for their cooperation and kind support throughout our research period. This work would not have been possible without their contributions and the learning environment fostered at PES's Modern College of Engineering.

REFERENCE

[1] D. Kim, D. Joo, and J. Kim, "TiVGAN: Text to Image to Video Generation with Step-by-Step Evolutionary Generator," vol. 8, Aug. 2020.

[2] Q. Chen, Q. Wu, J. Chen, Q. Wu, A. Van Den Hengel, and M. Tan, "Scripted video generation with a Bottom-Up generative adversarial network," IEEE Transactions on Image Processing, vol. 29, pp. 7454–7467, Jan. 2020.

[3] A. Singh, "A survey of AI Text-to-Image and AI Text-to-Video generators," Department of Computer Science, Kent State University, 2021.

[4] J. Zhu et al., "MovieFactory: Automatic Movie Creation from Text using Large Generative Models for Language and Images," IEEE, Oct. 2023.

[5] J. Xing et al., "Make-Your-Video: Customized video generation using textual and structural guidance," IEEE Transactions on Visualization and Computer Graphics, pp. 1-15, Jan. 2024.

[6] A. K. Jain and M. P. Sharma, "Ethics of AI in Creative Industries: Challenges and Future Directions," Journal of AI Ethics, vol. 2, no. 4, pp.123-138, 2023.

[7] T. Wang et al., “Towards Real-Time Generic Video Editing: Challenges and Approaches,” Computer Vision and Image Understanding, vol. 230, 2023

[8] H. Liu et al., “Multimodal Transformers for Video Understanding,” IEEE Transactions on Multimedia, vol. 26, pp. 3741-3753, 2024

[9] J. Devlin et al., “BERT: Pre-training of Deep BidirectionalTransformers for Language Understanding,” arXiv, 2018.

[10] K. Xu et al., “Deep Generative Models for Music Generation,” IEEE Transactions on Neural Networks and Learning Systems, vol. 31, no. 11, pp. 4227-4238, 2020.

[11] M. H. M. Rahman, S. Sharmin, and M. M. S. Islam, “Challenges in AI Content Generation,” Journal of Digital Content Technology and its Applications, vol. 13, no. 4, pp. 445-452, 2023.

[12] C. M. R. De Oliveira et al., “Addressing Algorithmic Bias in AI Models: A Systematic Review,” Journal of AI Ethics, vol. 3, no. 2, pp. 99- 112, 2024.

[13] S. Kim, S.-G. Lee, J. Song, J. Kim, and S. Yoon, “FloWaveNet: a generative flow for raw audio,” International Conference on Machine Learning, pp. 3370–3378, May 2019.

[14] N. Chawla, “Ethical Implications of AI in Media and Entertainment,” Media Ethics, vol. 15, no. 2, pp. 150-165, 2023. [15] J. S. K. Pereira et al., “AI and Ethics in the Digital Age,” Ethics and Information Technology, vol. 25, pp. 215-229, 2023.

[15] G. Brown et al., “Computational Cost Analysis of Large-Scale AI Models,” Journal of Computing Research, vol. 29, no. 6, pp. 400415, 2022.

[16] T. Karras et al., “A Style-Based Generator Architecture for Generative Adversarial Networks,” IEEE Transactions on Neural Networks and Learning Systems, vol. 30, no. 8, pp. 2499-2513, 2019.

[17] D. Silver et al., “Mastering the Game of Go without Human Knowledge,” Nature, vol. 550, pp. 354-359, 2017.

[18] J. R. K. Cheong et al., “Real-Time Generative Models for Video Editing: Overview and Challenges,” ACM Transactions on Graphics, vol.42, no. 4, 2023.

[19] Y. Liu et al., “Multimodal Representation Learning for Video Understanding,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 46, no. 5, pp. 1060-1073, 2024.

[20] S. Swarnakar and MIT ADT University, “Artificial intelligence and Cinema - Exploring the implications of artificial intelligence in cinema,” in Using AI for Dialoguing with Texts: From Psychology to Cinema and Literature, 2024.

[21] Smith, T., et al., “The Evolution of AI in Film Production,” Journal of Visual Media, vol. 17, no. 3, pp. 145-157, 2023.

[22] Binns, R., “Fairness in Machine Learning: Lessons Learned from Bias in Algorithmic Systems,” Communications of the ACM, vol. 61, no. 4, pp. 60-70, 2018.

[23] Kirkpatrick, M., “AI and Accents: Issues of Representation in AI- Generated Speech,” Journal of AI and Society, vol. 22, pp. 122-135, 2021.

[24] O’Reilly, L., et al., “Addressing Bias in GANs for Visual Media,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 43, no. 8, pp. 1567-1580, 2021.

[25] E. Ferrara, “Fairness and Bias in Artificial intelligence: A brief survey of sources, impacts, and mitigation strategies,” Sci, vol. 6, no. 1, p. 3, Dec. 2023.

[26] S. A. Friedler, C. Scheidegger, S. Venkatasubramanian, S. Choudhary, E. P. Hamilton, and D. Roth, A Comparative study of fairness-enhancing interventions in Machine learning. 2019.