iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** V **Month of publication:** May 2025

DOI: https://doi.org/10.22214/ijraset.2025.70567

www.ijraset.com

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

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

GenNarrate: AI-Powered Story Synthesis with Visual and Audio Outputs

Dr.Manimala S, Ananya U, Shreeram S R, Sudhiksha B A, Sanjana N
Department of Computer Science and Engineering JSS Science and Technology University, Mysuru, India

Abstract: ***The emergence of generative artificial intelligence has redefined the boundaries of digital content creation, particularly in the domain of computational storytelling. This paper presents GenNarrate, a modular, multi-modal generative AI system engineered to synthesize coherent narratives augmented with corresponding visual and auditory elements. The architecture leverages advanced machine learning models, including LLaMA2 for text generation, DALL·E for image synthesis, and a combination of Google Text-to-Speech (GTTS) and AudioLDM for expressive audio narration and sound design. GenNarrate facilitates user-driven content generation by accepting configurable parameters-such as genre, tone, character elements, and desired multimedia outputs-through an interactive front-end interface. These inputs are orchestrated through a Flask-based backend pipeline, which integrates the constituent modules and produces downloadable outputs comprising narrated stories, image-enhanced documents, and synchronized audio tracks. The proposed system demonstrates a novel approach to narrative automation, emphasizing cross-modal coherence, scalability, and personalization. This study further situates GenNarrate within the broader context of AI-enhanced storytelling technologies, offering comparative insights with existing open-source models such as GPT-3 and Stable Diffusion. Potential applications are explored across educational content delivery, therapeutic interventions, creative industries, and interactive media. The findings underscore the transformative potential of multi-modal AI systems in facilitating immersive, user-centric storytelling experiences, while also identifying avenues for future development in real-time interaction, fine-grained customization, and adaptive content generation.***

Index Terms: Generative AI, Multimodal Storytelling, LLaMA2, DALL·E, Text-to-Speech, AudioLDM, Natural Language Generation, Narrative Synthesis, Human-AI Collaboration.

I. INTRODUCTION

The rise of generative artificial intelligence has significantly reshaped content creation, especially in the realm of storytelling. Models capable of generating coherent text, realistic images, and expressive audio have opened new avenues for creative expression. However, most existing solutions treat these modalities in isolation-tools for text generation, image synthesis, and audio narration often function independently, requiring manual integration to produce cohesive multimedia content. This lack of a unified framework limits accessibility and diminishes the potential for fully immersive, AI-generated narratives.

GenNarrate addresses this gap by offering an integrated, multi-modal storytelling platform. It enables users to input story parameters-such as genre, tone, and desired outputsthrough an intuitive interface, and leverages a pipeline of advanced AI technologies to synthesize the narrative. LLaMA2 is used for generating structured and coherent text, DALL·E handles the visual storytelling, while GTTS and AudioLDM create layered audio outputs including narration and ambient sound. These components are orchestrated through a modular backend, ensuring cross-modal consistency and automation.

By converging these technologies, GenNarrate facilitates the creation of rich, personalized storytelling experiences with minimal user effort. The platform's design not only addresses a clear technological gap but also presents broad applicability across education, entertainment, therapy, and digital media production. This paper outlines the system's architecture, implementation, and research foundation, and demonstrates how GenNarrate advances the possibilities of AI-driven narrative generation.

II. LITERATURE REVIEW

The advancement of generative artificial intelligence has led to significant progress in text, image, and audio generation models. While several studies have contributed to the individual development of these modalities, the literature reveals a clear gap in platforms that integrate all three cohesively for immersive storytelling. This section surveys key research contributions that inform the architecture and implementation of GenNarrate, focusing on language modeling, image generation, audio synthesis, and multi-modal integration.

Fotedar et al. [1], in their paper *"Storytelling AI: A Generative Approach to Story Narration"*, presented a detailed survey of generative techniques for interactive storytelling. Their work emphasized the role of user input in guiding narrative progression and discussed challenges such as coherence maintenance and narrative branching. The study highlighted the importance of balancing creativity with control in automated story systems, which GenNarrate builds upon by offering userdefined inputs to direct AI-generated narratives.

Kuznetsov et al. [2], through *"Storytelling through Deep Learning"*, explored the application of deep neural networks in automated story creation. The authors discussed model architectures capable of learning narrative structures and semantic flow, and identified limitations in creativity and thematic coherence. GenNarrate addresses these gaps by integrating genre and tone constraints into its LLaMA2-powered text generation module.

Ramesh et al. [3] introduced *DALL·E* in their landmark paper *"Zero-Shot Text-to-Image Generation"*, showcasing a model capable of generating diverse and semantically accurate images from textual prompts. They demonstrated how transformer-based diffusion techniques allow fine-grained visual synthesis. GenNarrate extends this capability by translating key narrative elements into DALL·E prompts to produce scene-specific illustrations.

Aaron van den Oord et al. [4] proposed *WaveNet*, a powerful generative model for raw audio synthesis, capable of producing natural-sounding speech from text. Their model significantly outperformed traditional concatenative and parametric TTS systems in terms of realism and prosody. Although WaveNet is computationally intensive, its expressive capabilities have informed our system's voice generation strategies, particularly in integrating GTTS with ambient audio.

AudioLDM, introduced by Liu et al. [5], in *"AudioLDM: Text-to-Audio Generation with Latent Diffusion Models"*, explored how diffusion models can synthesize high-quality, diverse audio from text. The model captures background environments and acoustic scenes, offering a leap forward in audio-text alignment. In GenNarrate, AudioLDM is used to enrich voice narration with ambient soundscapes, thereby enhancing story immersion.

Cardona-Rivera and Roberts [6], in *"Controlling Narrative Time in Interactive Storytelling"*, examined mechanisms to influence user engagement through temporal structuring. They proposed frameworks for time manipulation within digital narratives, useful in enhancing immersion. GenNarrate inherits this principle by breaking down longform outputs into chapters and aligning image and audio transitions accordingly.

Radford et al. [7], in *"Learning Transferable Visual Models From Natural Language Supervision"*, introduced CLIP, a model that learns to associate textual and visual semantics through contrastive learning. While CLIP is not used directly in GenNarrate, its approach to multimodal alignment influenced how we extract image prompts from text content.

Goodfellow et al. [8], in *"Generative Adversarial Networks"*, presented the foundational concept of adversarial training for generative models. Their technique inspired numerous advances in image synthesis, though challenges such as mode collapse remain. GenNarrate relies instead on diffusion-based models, which are more stable for our multimodal pipeline.

Suraj et al. [9] published *"A Survey on the State of the Art in Audio Generation Models"*, offering an extensive review of techniques like vocoders, autoregressive models, and GANbased audio synthesis. They also discussed trade-offs between quality and computational overhead. GenNarrate benefits from these insights to choose a lightweight GTTS-based TTS approach, optimized using pydub for post-processing.

Khan et al. [10], in *"StoryGenAI: An Automatic GenreKeyword Based Story Generation"*, proposed a novel framework that generates narrative text conditioned on genre and keyword cues. Their model achieved improved coherence and relevance in short-form storytelling. GenNarrate similarly leverages genre-conditioned inputs but scales the concept by adding visuals and audio to the generated text.

III. SYSTEM REQUIREMENTS AND ANALYSIS

A. *Functional Requirements*

The GenNarrate system is architected to fulfill a series of functional objectives aimed at enabling a seamless, multimodal storytelling experience. Each component plays a critical role in transforming user-defined inputs into narrative-rich media outputs.

1) *Text Generation:* The system must accept structured user input-such as genre, tone, number of chapters, and character archetypes-and generate coherent, contextually appropriate narratives. This functionality is powered by LLaMA2, a stateof-the-art transformer-based large language model capable of producing human-like text. The model is fine-tuned for story synthesis and supports control over structure, pacing, and thematic progression.

2) *Image Synthesis:* The platform must be capable of identifying key narrative scenes and generating high-quality illustrations that match the story context. For this, GenNarrate utilizes DALL·E. This model translates textual promptsextracted from the generated story-into visually coherent and semantically aligned images.
3) *Audio Generation:* Audio output is an essential component of immersion. The system must generate expressive narration using GTTS (Google Text-to-Speech), and layer it with ambient soundscapes created via AudioLDM. The Pydub library is employed for audio composition and temporal alignment, resulting in a synchronized and engaging auditory experience.
4) *Integrated Output and Frontend:* The final output must be compiled into downloadable formats, including an illustrated storybook (PDF) and a synchronized audio file. The system must allow users to interact through a responsive web-based interface developed using React.js and Material-UI. The frontend allows for customization of inputs and previews of generated media, while the Flask-based backend handles request processing and inter-module communication through API endpoints.

B. *Non-Functional Requirements*

In addition to its core capabilities, GenNarrate must adhere to a series of non-functional requirements that ensure the platform is responsive, and consistent across modalities.

1) *Performance and Latency:* Low-latency interaction is crucial for a responsive user experience. The architecture minimizes computational overhead by decoupling modules and using asynchronous API calls for inter-process communication. Caching mechanisms and GPU acceleration for inference further optimize runtime performance.
2) *Cross-Modal Consistency:* The generated text, visuals, and audio must align in tone, mood, and narrative structure. To ensure this, the system enforces consistent prompt extraction methods and shares metadata (e.g., chapter segmentation, theme tags) across all modules.
3) *Reliability and Fault Tolerance:* The system is designed for robust execution, with failover mechanisms for API errors and fallback logic for missing outputs (e.g., when an image fails to generate, the system inserts a placeholder). Logging and monitoring are performed using Prometheus and Grafana to identify and address faults in real-time.
4) *Usability and Accessibility:* GenNarrate's interface must be intuitive and accessible across devices. Accessibility features such as keyboard navigation, ARIA labels, and screen reader compatibility are supported, ensuring inclusivity for a diverse user base.

C. *System Analysis*

GenNarrate's architecture consists of three primary generative modules-text, image, and audio-interlinked through a lightweight Flask-based backend. User inputs are collected via the frontend and passed to the backend, which then delegates processing tasks to the appropriate AI services.

The text generation module is powered by Meta's LLaMA2, chosen for its balance of coherence, creativity, and computational efficiency. Compared to alternatives like GPT3, LLaMA2 offers greater control over output length and structure, making it better suited for story synthesis.

The image synthesis module leverages both DALL·E and Stable Diffusion. DALL·E excels in scene creativity and textimage relevance, while Stable Diffusion provides enhanced resolution and compositional accuracy. The system selects the model dynamically based on user preferences or image complexity.

For audio synthesis, the system combines GTTS for narration and AudioLDM for ambient sound and music. GTTS ensures fast and natural-sounding voice synthesis, whereas AudioLDM is used to interpret and generate rich audio backgrounds based on narrative keywords and themes.

These modules communicate via Flask APIs, and data exchange is facilitated by shared metadata structures. A central control logic handles orchestration, error recovery, and output assembly. The output artifacts are rendered as a downloadable illustrated eBook and synchronized audio file, delivered to users through the React-based interface.

Overall, the design prioritizes modularity, scalability, and cross-modal coherence, establishing GenNarrate as a unified platform for AI-powered storytelling.

IV. SYSTEM DESIGN

A. *Architectural Overview*

GenNarrate adopts a modular architecture composed of independently functioning yet interlinked components, each dedicated to one of the three generative modalities-text, image, and audio. The system is divided into five core modules:

1) Input Module: Provides a web-based interface (React.js with Material UI) for users to configure story parameters, including genre, tone, number of scenes, and image/audio preferences. These inputs are transmitted to the backend as structured JSON through Flask APIs.
2) Text Generation Module: The LLaMA2 model is employed for generating the story narrative. Prompt engineering techniques are applied to control tone, structure, and genre alignment. Testing involves tuning decoding parameters such as temperature and top-p to optimize narrative creativity and coherence. The output is segmented into chapters or scenes using natural language processing tools for downstream use.
3) Image Generation Module: Scene-specific prompts derived from the generated text are sent to the DALL·E model. The system constructs prompts by identifying visual descriptors (e.g., setting, atmosphere, objects) from each scene. These prompts are fed to DALL·E to produce high-quality, semantically aligned images.
4) Audio Generation Module: This module converts the full text into natural-sounding narration using Google Text-to-Speech (GTTS), and enriches it with background music generated via AudioLDM. Narration and music tracks are integrated using audio editing techniques to produce a cohesive and immersive audio experience.
5) Output Compilation Module: Text, images, and audio are compiled into deliverables including a PDF storybook and MP3 audio file. A LaTeX-based template formats the text and embeds the generated visuals, while audio is finalized into a downloadable track. These are served to the user through the frontend interface.

B. Workflow Design

The user initiates the workflow by submitting a story prompt and configuration options through the frontend. This data is processed in the backend, beginning with the invocation of the LLaMA2 model for text generation. Custom prompt templates and inference parameters guide the model to produce genrespecific, structured narratives.

Once generated, the narrative is parsed to identify distinct scenes or chapters. From each scene, representative lines are extracted to form prompts for the DALL·E module, which returns a corresponding image for each key segment of the story.

In parallel, the complete story text is passed to the Audio Generation Module, which consists of two distinct subcomponents: narration generation using GTTS and background music generation using AudioLDM. The GTTS engine transforms the story into speech, while AudioLDM produces ambient audio tracks aligned with the mood and setting of each scene.

- *Background Music Generation and Integration:* Background music is synthesized using the AudioLDM model, which takes scene-level mood descriptors extracted from the story text (e.g., ”tense forest atmosphere”, ”joyful celebration”) as input prompts. These are mapped to sound themes and passed to the model for generation. The resulting ambient tracks are then processed using the Pydub library to normalize volume and align them temporally with the GTTS-generated narration.

The integration step involves merging the narration and background music into a single synchronized audio file. Pydub handles voice-over alignment, fade-ins/outs, and timing adjustments to ensure smooth transitions between scenes. This process creates an immersive listening experience where narration and music are thematically coherent and temporally synced.

C. Output Compilation

In the final stage, all generated media are compiled. The text and images are embedded into a structured PDF using LaTeX, while the combined audio narration is exported in MP3 format. These outputs are bundled and made available through the user interface for download. The architecture ensures cross-modal coherence, modularity, and scalability, supporting dynamic, multi-sensory storytelling through a unified system.

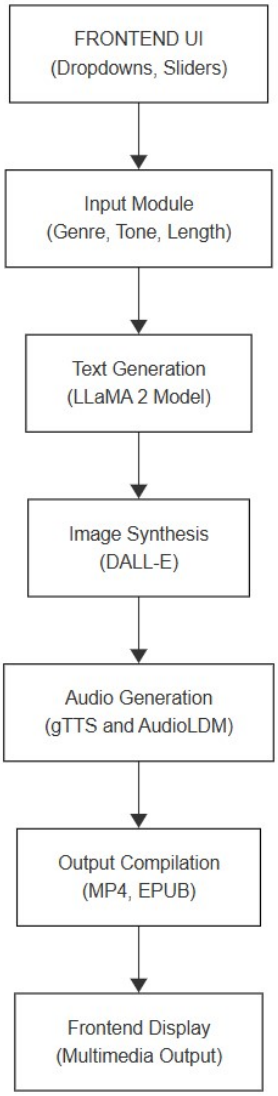


Fig. 1: System Architecture of GenNarrate illustrating the flow from user input to multi-modal output.

V. SYSTEM IMPLEMENTATION

The implementation of GenNarrate follows a modular and service-oriented architecture, where each core functiontext generation, image synthesis, audio narration, and user interaction-is realized through a combination of specialized AI models and supporting Python-based infrastructure. The backend is implemented in Flask and manages inter-module communication, while the frontend is developed using React.js to provide an interactive user interface.

A. Text Generation

The core of GenNarrate's text synthesis is powered by Meta's LLaMA2 large language model, accessed via a dedicated API endpoint. The user's input parameters-such as genre, tone, and story length-are used to dynamically generate prompts that guide the narrative generation process. Prompt engineering plays a critical role in controlling the structure and creativity of the output. Genre-specific templates are used to introduce key themes and settings, while custom tokens
(e.g., <scene>, <dialogue>) are embedded to enforce story flow and segmentation.
The LLaMA2 API is queried with adjusted decoding parameters such as temperature, top-k, and top-p sampling to balance creativity with coherence. The generated text is post-processed and parsed using SpaCy and NLTK libraries to detect sentence boundaries and extract scene-level metadata. These segments are passed downstream to both the image and audio modules to ensure modal alignment.

B. Audio Generation

Audio generation in GenNarrate is implemented as a twostep process: narration synthesis and background music generation. The narration is produced using Google Text-to-Speech (GTTS), which converts the story text into natural-sounding audio. The voice configuration, speech rate, and pitch are adjusted based on user-selected tone and emotion settings. For example, suspenseful tones may employ slower and deeper narration, while comedic tones use faster and brighter delivery.

To enhance immersion, background music is generated for each scene using AudioLDM. Scene descriptions extracted during text parsing (e.g., ”dark forest”, ”epic battle”) are transformed into prompts for the AudioLDM model. These ambient audio segments are then integrated with the narration using the Pydub library. The integration process includes normalization, trimming, volume balancing, and time alignment to ensure smooth transitions across scenes. The final output is a single MP3 file that combines narration and ambient music in a cohesive audio experience.

C. Image Generation

For visual synthesis, GenNarrate uses OpenAI’s DALL·E 3 model to create contextually relevant illustrations that correspond to key story events. The image prompts are constructed from scene-level summaries extracted from the story using natural language processing. Each prompt is enriched with visual descriptors (e.g., ”sunset”, ”cyberpunk city”, ”medieval knight in armor”) derived from character and setting metadata. The prompts are submitted to the DALL·E API, which returns high-resolution images. These are post-processed using Python Imaging Library (PIL) to standardize size, format, and resolution before being embedded into the final storybook PDF. The image-to-scene mapping is stored in a structured index to ensure proper ordering during output compilation.

D. Backend and Frontend Integration

The system is orchestrated using a Flask-based backend server that exposes RESTful API endpoints to manage all module interactions. The backend handles user session management, prompt parsing, model invocation, and output caching. It coordinates the flow of data between the text, image, and audio modules and maintains a shared metadata layer to ensure consistency.

The frontend is implemented using React.js with Material UI components. It features a form-based input panel where users can select genre, tone, number of images, and whether to include narration. Asynchronous API calls are used to send data to the backend and poll for generation status. Once all outputs are ready, the UI renders previews of the text and images, along with download buttons for the storybook (PDF) and audio file (MP3).

This tightly coupled yet modular implementation allows for efficient, real-time processing and delivery of personalized multi-modal storytelling content.

VI. FUTURE IMPROVEMENTS

While GenNarrate demonstrates a robust foundation for AI-driven multi-modal storytelling, several enhancements are planned to improve scalability, performance, and user experience. One key area of development involves reducing model loading and inference latency, particularly in cloud-deployed environments. The current implementation of LLaMA2 and DALL·E models, though accurate and expressive, incurs significant computational overhead when scaling to serve multiple concurrent users. Optimizations will include model quantization, asynchronous task handling, and server-side GPU acceleration to improve responsiveness under load.

Scalability is also a critical focus, particularly in anticipation of wider public deployment. Moving forward, containerized deployment using Docker and orchestration via Kubernetes will allow for dynamic resource allocation, autoscaling, and fault-tolerant API management across different regions and cloud providers.

Another major area of improvement lies in the integration of scene-aware video generation. Instead of limiting outputs to static images, future versions of GenNarrate will support video synthesis by interpolating between keyframes generated by DALL·E and using lip-syncing or animated avatars for character narration. This would transform the current storybook format into an animated video with synchronized narration and ambient soundtracks, offering a more immersive and cinematic user experience.

Additionally, the current system supports English-language input and narration exclusively. As part of future localization efforts, GenNarrate will be extended to support multilingual input and output generation, including translation of text and synthesis of speech in regional accents and languages. This will involve integrating multilingual transformer variants for text and exploring TTS engines like Coqui and multilingual GTTS models.

Improvements in user interactivity are also being planned, such as real-time content previews, dynamic prompt regeneration based on user feedback, and adaptive storytelling that adjusts in response to user corrections or emotions. Finally, to support on-device generation and offline use cases, lightweight transformer architectures (e.g., distilled LLMs or mobilefriendly diffusion variants) will be explored, allowing GenNarrate to function efficiently even in low-resource environments.

VII. CONCLUSION

The GenNarrate platform represents a significant advancement in the domain of AI-powered content creation by unifying three complex modalities-text generation, image synthesis, and audio narration-into a seamless, user-centric storytelling experience. Unlike traditional story generators or single-modal generative tools, GenNarrate leverages state-of-the-art models such as LLaMA2, DALL·E, GTTS, and AudioLDM to construct richly immersive, personalized narratives based on userdefined parameters.

The implementation achieves high cross-modal coherence by synchronizing inputs and outputs through a modular backend architecture built with Flask and a responsive frontend developed in React.js. By employing prompt engineering, automated text segmentation, scene extraction, and audio layering, the system ensures that each generated story is logically structured, visually engaging, and emotionally resonant.

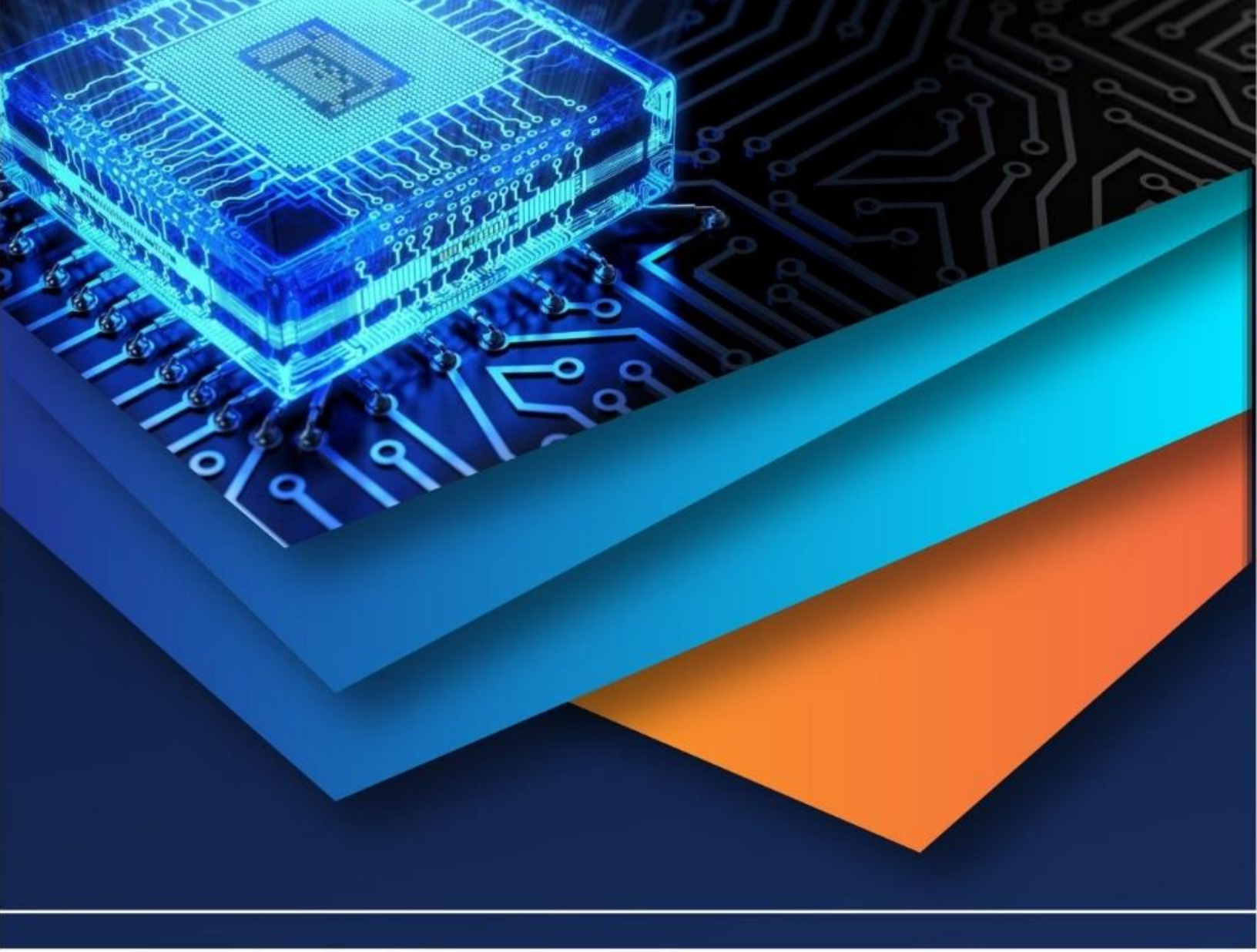
Beyond technical execution, GenNarrate illustrates the potential of AI as a co-creative agent. Its applications span a broad spectrum-from educational storytelling tools and therapeutic media to entertainment and content marketingdemonstrating its versatility across domains. The platform's interactive interface, automated pipeline, and downloadable output formats offer an end-to-end solution that democratizes multimedia storytelling for users with no technical background.

Although current capabilities are limited to static images, monolingual narration, and cloud-based deployment, future iterations of GenNarrate aim to incorporate multilingual support, video generation, neural voice cloning, and real-time content feedback. These enhancements will further extend the system's usability, accessibility, and expressive range.

In conclusion, GenNarrate exemplifies the power of generative AI when modular design, cutting-edge models, and thoughtful user experience come together. It sets a new benchmark in automated storytelling systems and lays the groundwork for continued research in multi-modal narrative synthesis, human-AI co-creativity, and adaptive content generation.

REFERENCES

[1] S. Fotedar et al., "Storytelling ai: A generative approach to story narration," CEUR Workshop Proceedings, vol. 2794, 2021. [Online]. Available: https://ceur-ws.org/Vol-2794/paper4.pdf
[2] G. Kuznetsov et al., "Storytelling through deep learning," CEUR Workshop Proceedings, 2019. [Online]. Available: https://ceur-ws.org/ Vol-2794/paper4.pdf
[3] Ramesh et al., "Zero-shot text-to-image generation," arXiv preprint arXiv:2102.12092, 2021. [Online]. Available: https://arxiv.org/abs/2102. 12092
[4] van den Oord et al., "Wavenet: A generative model for raw audio," arXiv preprint arXiv:1609.03499, 2016. [Online]. Available: https://arxiv.org/abs/1609.03499
[5] H. Liu et al., "Audioldm: Text-to-audio generation with latent diffusion models," arXiv preprint arXiv:2301.12503, 2023. [Online]. Available: https://arxiv.org/abs/2301.12503
[6] R. E. Cardona-Rivera and D. L. Roberts, "Controlling narrative time in interactive storytelling," in International Conference on Interactive Digital Storytelling, 2012. [Online]. Available: https: //www.researchgate.net/publication/221456514
[7] Radford et al., "Learning transferable visual models from natural language supervision," arXiv preprint arXiv:2103.00020, 2021. [Online]. Available: https://arxiv.org/abs/2103.00020
[8] Goodfellow et al., "Generative adversarial networks," arXiv preprint arXiv:1406.2661, 2014. [Online]. Available: https://arxiv.org/abs/1406. 2661
[9] Suraj et al., "A survey on the state of the art in audio generation models," arXiv preprint arXiv:2005.00341, 2021. [Online]. Available: https://arxiv.org/abs/2005.00341
[10] L. P. Khan et al., "Storygenai: An automatic genre-keyword based story generation," IEEE Xplore, 2023. [Online]. Available: https://ieeexplore.ieee.org/document/10183482



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY