© MAY 2026 | IRE Journals | Volume 9 Issue 11 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I11-171833

Personalized Image-to-Audio Bedtime Story Generation Using Multimodal AI with User Profiling

KRUTIKA SUSHIL NIKUMBH[1], DR. PRAKASH KENE[2]
[1]*Master of Computer Application P.E.S. Modern College of Engineering, Pune, India*
[2]*Associate Professor, MCA Department P.E.S. Modern College of Engineering, Pune, India*

Abstract- ***Bedtime stories help children improve imagination, communication skills, and emotional connection with parents. However, generating personalized and engaging bedtime stories daily can be challenging for parents. Existing AI-based storytelling systems mainly rely on text inputs, resulting in generic narratives that lack contextual relevance and emotional adaptation. This paper presents a multimodal framework that transforms an input image into a personalized bedtime audio story using user profile attributes such as age, preferences, and mood. The system combines image understanding, generative language models, and text-to-speech techniques to produce context-aware and emotionally adaptive narratives with calming themes. The proposed approach improves storytelling by integrating visual context and personalization, making stories more engaging and meaningful. This work highlights the potential of multimodal AI in enhancing bedtime routines and supporting child well-being.***

Keywords- ***Multimodal AI, Personalized Storytelling, Image-to-Audio, Text-to-Speech, User Profiling***

I. INTRODUCTION

Storytelling is an essential activity that contributes to children's creativity, emotional development, and communication skills. Bedtime stories, in particular, help children relax and strengthen the emotional bond between parents and children. Beyond stimulating creative thinking, such narratives actively contribute to the advancement of cognitive faculties and linguistic competence in young learners

In modern lifestyles, parents often face time constraints, making it difficult to create new and engaging stories regularly. As a result, there is a growing need for automated systems that can assist in storytelling while maintaining emotional and contextual relevance.

Artificial intelligence has enabled the development of automated storytelling systems that generate narratives using text-based prompts. However, such systems tend to generate redundant and standardized narratives that fall short in terms of individual customization and emotional resonance. They also fail to incorporate real-world elements from a child's surroundings, such as familiar objects or experiences. In order to overcome these shortcomings, the current study presents a combined modality-based system that incorporates visual inputs alongside individual user characteristics to generate personalized bedtime stories in audio form. This approach aims to make storytelling more meaningful, engaging, and adaptive to individual user needs.

With the increasing adoption of AI in daily life, an increasing expectation has emerged for systems capable of not just cognitive processing but also affective responsiveness. Personalized storytelling represents a meaningful application of AI where technology directly enhances human experiences and relationships.

In recent years, the integration of artificial intelligence into everyday applications has transformed the way users interact with digital systems. Across domains ranging from content suggestion platforms to conversational agents, tailoring experiences to individual users has emerged as a critical driver in enhancing user experience.

However, narrative generation platforms have yet to fully harness this potential, particularly in intimate settings like pre-sleep routines where affective connection holds great significance.

By incorporating multimodal inputs and user-centric design, storytelling can be made more meaningful

and interactive. This creates a compelling case for developing platforms that go beyond mere content production and instead dynamically shape output in response to the user's situational and emotional context.

II. LITERATURE REVIEW

2.1 LITERATURE SURVEY

LeCun et al. (2015). This study provided a thorough examination of neural learning approaches applied to pattern identification and data interpretation. These techniques have seen broad application across computer vision and intelligent systems. That said, the study does not specifically target narrative generation or user-centered customization aspects.

Vaswani et al. (2017). The authors proposed the transformer architecture, which significantly improved sequence modeling and text generation tasks. This approach laid the foundation for modern language models used in storytelling systems. Despite its impact, the architecture offers no mechanism to handle combined data types or adapt outputs to individual users.

Brown et al. (2020). This landmark work introduced large-scale language models capable of generating human-like text. The study demonstrated the effectiveness of transformer-based architectures in natural language generation. However, such models frequently yield one-size-fits-all results and are not natively designed to accommodate individualization or cross-modal data fusion.

Sharma et al. (2021). This study explored automated story generation using natural language processing techniques. The authors developed a system capable of generating narratives from textual prompts, demonstrating the potential of AI in creative storytelling. However, the narratives produced were predominantly uniform in nature, without adaptation to user characteristics like developmental stage, individual interests, or emotional disposition.

Rao and Mishra (2021). This study focused on text-to-speech (TTS) systems for converting textual content into natural and expressive audio. The authors highlighted improvements in speech quality and user experience. Despite its effectiveness in audio generation, the system operated independently and did not influence the content or emotional tone of the generated stories.

Reddy and Banerjee (2021). This research examined user profiling techniques for personalization in intelligent systems. The authors demonstrated how user attributes such as preferences and behavior could improve engagement and relevance. However, the scope of their study was confined to content suggestion platforms and was not adapted for use in narrative generation contexts.

Jurafsky and Martin (2022). The authors presented foundational concepts in speech and language processing, covering both natural language processing and speech synthesis. Their work provides essential techniques for building storytelling systems. However, it does not focus on integrating multiple modalities or creating personalized storytelling experiences.

Kulkarni and Chatterjee (2022). The authors introduced an image-to-text storytelling framework using convolutional and recurrent neural networks. Their system successfully generated descriptive narratives from visual inputs, capturing objects and scene context effectively. However, the approach lacked personalization and did not adapt stories based on user-specific preferences or emotional states.

Verma and Bansal (2023). This work focused on improving narrative coherence through sequence-based deep learning models. The authors proposed techniques to enhance the logical flow and structure of generated stories. While their approach improved storytelling quality, it remained limited to text-based inputs and did not consider multimodal data such as images or user profiles.

Ahmed and Kapoor (2024). This paper proposed a multimodal AI framework that integrates text, image, and audio data for content generation. The study highlighted the advantages of combining multiple data modalities for better contextual understanding. Nevertheless, the framework lacked a specific focus on child-centric applications and did not incorporate emotional adaptation in storytelling.

Beyond the works reviewed above, a number of scholars have underscored the value of fusing diverse AI methodologies to advance the capabilities of automated content creation. While advancements in deep learning and natural language processing have significantly improved text generation, the lack of integration with visual These findings collectively suggest that while notable strides have been achieved within separate fields, an integrated approach that brings these techniques together remains largely absent to achieve better performance and user satisfaction.

2.2 RESEARCH GAP

Most existing storytelling systems focus on individual components such as text-based story generation, image-based narration, or text-to-speech conversion. While these approaches are effective in their respective domains, they operate independently and do not provide a unified solution.

Additionally, current systems generate generic outputs and do not consider user-specific factors such as age, preferences, and emotional state. The absence of personalization and emotional adaptation limits their effectiveness, especially in applications like bedtime storytelling.

Consequently, developing a cross-modal system that brings together visual data interpretation, individualized user modeling, and contextual story creation becomes imperative to create personalized and emotionally adaptive storytelling experiences.

III. EXISTING SYSTEM

Conventional narrative generation platforms predominantly depend on written prompts as their primary source of input. These systems use natural language processing techniques to create stories based on user prompts. However, the generated content is often generic and lacks personalization.

Systems that operate on visual data can extract scene descriptions and basic story elements from images, yet they remain incapable of adjusting their output according to user-defined parameters such as maturity level, personal interests, or current emotional condition.

Text-to-speech systems convert textual content into audio, improving accessibility and user experience. However, they operate independently and do not contribute to the story generation process.

Thus, existing systems fail to provide a unified solution that integrates personalization, visual context, and emotional adaptation in storytelling.

3.1 LIMITATIONS OF EXISTING SYSTEM

Despite advancements in AI-based storytelling, existing systems face several limitations. Text-based systems generate repetitive and generic stories, lacking personalization and contextual relevance. Image-based systems can extract visual information but fail to adapt narratives according to user-specific attributes.

Additionally, the majority of existing platforms overlook affective dimensions including a user's current emotional state and content sensitivity thresholds, both of which are indispensable for child-oriented pre-sleep narrative experiences. The lack of integration between different components such as image processing, narrative generation, and audio synthesis further reduces the effectiveness of these systems.

The aforementioned gaps collectively point toward the necessity of building a cohesive and individually adaptive narrative generation system.

In addition to these limitations, current platforms similarly struggle with expandability and flexibility, as they are predominantly built for broad storytelling use cases and remain unresponsive to the distinct needs of individual users.

Furthermore, the absence of multimodal integration reduces the ability of these systems to generate context-rich narratives. As a result, the overall storytelling experience remains limited and less engaging for users.

IV. PROPOSED SYSTEM

The proposed system introduces a multimodal approach for generating personalized bedtime stories

© MAY 2026 | IRE Journals | Volume 9 Issue 11 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I11-171833

by integrating image input, user profile data, and AI-based narrative generation.

The system takes an image as input, such as a child's drawing or a familiar object, and extracts relevant visual features including objects and contextual elements. Along with this, user profile information such as age, preferences, and mood is considered to personalize the story.

A generative language model then creates a story by combining visual context and user-specific inputs. The generated narrative is designed to be simple, engaging, and emotionally suitable for bedtime.

Finally, the story is converted into audio using text-to-speech technology, providing a complete and soothing storytelling experience.

This framework aims to bridge the gap between generic AI storytelling and personalized, context-aware narrative generation. By integrating multiple modalities, the system enhances both the quality and relevance of generated stories.

A complete illustration of the suggested system's operational sequence is depicted in Fig 1.

Fig 1. Proposed System Architecture

4.1 TRADITIONAL VS PROPOSED SYSTEM

A comparison between traditional storytelling systems and the proposed system is presented below. Traditional storytelling systems primarily rely on text-based inputs, whereas the proposed system incorporates image inputs along with user profile data to enhance contextual understanding. In conventional approaches, the generated narratives are often generic and lack personalization. In contrast, the proposed system produces personalized and context-aware stories tailored to individual users.

Furthermore, traditional systems do not consider emotional adaptation, which limits their effectiveness in applications such as bedtime storytelling. The proposed system addresses this limitation by generating calming and mood-based narratives. Additionally, traditional systems typically provide only text or basic audio output, whereas the proposed system delivers expressive and engaging audio storytelling.

This comparison clearly demonstrates the effectiveness and improvements offered by the proposed approach.

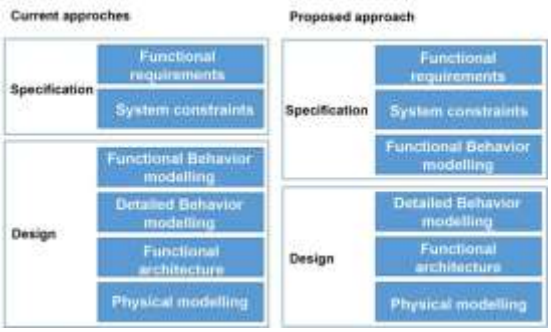

Fig 2: Traditional vs Proposed System Comparison

V. MULTIMODAL STORYTELLING FRAMEWORK

5.1 MULTIMODAL AI IN STORYTELLING

Multimodal AI involves the integration of different data types such as images, text, and audio to generate meaningful outputs. In storytelling, this approach enhances narrative quality by incorporating visual context along with textual generation.

In contrast to conventional single-input platforms that process only written content, cross-modal narrative systems facilitate the creation of situationally relevant and captivating stories. This improves the relevance and creativity of the stories.

The integration of multiple modalities allows the system to better understand real-world scenarios and generate more natural and meaningful narratives. This significantly improves user engagement relative

© MAY 2026 | IRE Journals | Volume 9 Issue 11 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I11-171833

to conventional approaches that depend on a single data source.

5.2 STORY GENERATION PIPELINE

The storytelling process follows a structured pipeline:

- Image Analysis: Extracting objects and scene details from the input image.
- User Profile Integration: Incorporating user-specific attributes such as age, preferences, and mood.
- Narrative Generation: Creating a story using a language model based on combined inputs.
- Audio Conversion: Converting the generated story into speech using text-to-speech technology.

This pipeline ensures a systematic and personalized storytelling process.

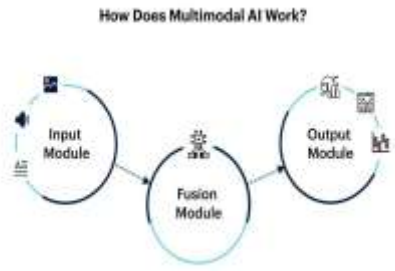

Fig 3: Story Generation Pipeline

VI. PERSONALIZED STORY GENERATION MODEL

The personalized story generation model focuses on adapting narratives based on user-specific characteristics. It considers multiple factors such as age, preferences, mood, and sensitivity to ensure that the generated story is appropriate and engaging. The system adjusts the linguistic sophistication and narrative structure in alignment with the developmental stage of the user, while also incorporating favourite themes, characters, or settings based on individual preferences. Additionally, the tone of the story is modified depending on the user's mood to create either a calming or engaging narrative. Sensitivity is also taken into account to avoid content that could be unsuitable or distressing for young audiences.

Through the combination of these variables, the platform produces narratives that carry greater personal relevance and appropriateness for each specific user. The personalization process significantly enhances user engagement by making stories more relatable and context-aware. This approach is particularly beneficial for children, as it creates a sense of familiarity and emotional connection, which is essential for effective storytelling.

Fig 4: User Personalization Flow

VII. ADAPTIVE STORY GENERATION APPROACH

Traditional storytelling systems focus only on generating content without considering emotional impact. The proposed approach introduces adaptability by modifying narratives based on user context.

For example, bedtime stories are designed with calming language and positive endings. The system can also simplify narratives for younger users and adjust emotional tone based on mood.

This methodology guarantees that narrative generation extends beyond mere automation to encompass affective sensitivity and a strong orientation toward individual user needs.

7.1 SYSTEM COMPONENTS DESCRIPTION

The presented architecture is composed of several interconnected modules that function collaboratively to produce individually tailored pre-sleep narratives. Each component plays a specific role in the overall workflow.

- Image Processing Module: This unit handles the interpretation of the provided visual content and retrieves significant attributes such as objects,

colors, and scene context. These features form the basis of story generation.

- User Profiling Module: This module collects and processes user-specific information such as age, preferences, and mood. It ensures that the generated story is specifically customized to match each user's unique profile.
- Story Generation Module: This module uses a language model to generate a narrative by combining visual features and user data. The generated story is designed to be simple, engaging, and emotionally appropriate.
- Text-to-Speech Module: This module converts the generated story into audio using speech synthesis techniques. It ensures that the output is clear, expressive, and suitable for bedtime listening.

The coordination between these components enables the system to produce a complete and personalized storytelling experience.

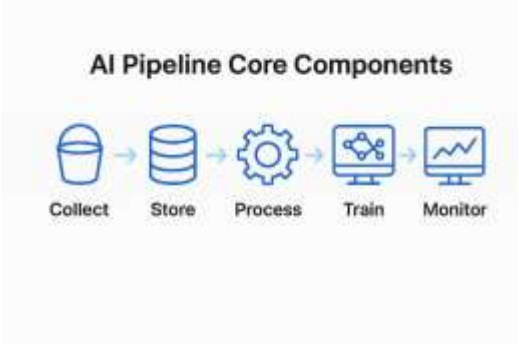

Fig 5: System Components Architecture

7.2 ADVANTAGES OF PROPOSED SYSTEM

The presented framework delivers a range of benefits that distinguish it clearly from conventional narrative generation methods. It generates personalized stories based on user preferences and contextual inputs, making each narrative unique and relevant. The use of multimodal inputs enhances user engagement by incorporating both visual and textual information into the storytelling process. Moreover, the system produces affectively responsive and soothing story content, making it particularly well-suited for pre-sleep scenarios. It also improves accessibility by delivering stories in audio format, allowing users to listen rather than read. Furthermore, the architecture accommodates future expansion and diverse use cases, positioning it as both an extensible and accessible platform compared to existing approaches.

VIII.USER EXPERIENCE AND ETHICAL CONSIDERATIONS

Personalized storytelling systems must ensure a safe and positive user experience. Since the system uses user data such as preferences and emotional state, privacy and data security are important considerations.

The generated content must be appropriate for children, avoiding harmful or sensitive topics. Ensuring fairness, transparency, and responsible AI usage is essential for building trust in such systems.

8.1 EXPECTED OUTCOMES AND BENEFITS

The proposed system is expected to improve storytelling quality by generating personalized and context-aware narratives. Unlike traditional systems, it enhances user engagement by incorporating familiar visual elements and user preferences.

The use of multimodal inputs allows the system to produce more meaningful and relatable stories. Additionally, the integration of emotional adaptation makes the storytelling process more effective for bedtime applications, helping children relax and feel comfortable.

Overall, the system offers significant improvements in terms of personalization, engagement, and usability compared to existing approaches.

IX. CONCLUSION and FUTURE SCOPE

Current AI-based storytelling systems are limited in their ability to generate personalized and emotionally adaptive content, often producing generic narratives that lack real-life context and user relevance. We propose a multimodal storytelling framework that integrates image input, user profile data, and audio generation to create personalized bedtime stories. This approach combines visual understanding, narrative generation, and text-to-speech to deliver engaging and context-aware storytelling experiences.

Future work can focus on improving emotional understanding through real-time mood detection, enhancing story diversity, and incorporating interactive elements. By moving from generic content

generation to personalized and context-driven storytelling, the present work endeavors to strengthen children's involvement with stories, nurture their psychological health, and enrich nightly pre-sleep practices.

The proposed framework demonstrates how artificial intelligence can be applied in a meaningful and human-centric way. By focusing on personalization and emotional adaptability, the system goes beyond traditional automation and contributes to improving user experience and well-being.

The proposed system can be further enhanced by integrating real-time emotion detection using facial expressions or voice inputs. Additionally, interactive storytelling features can be introduced where children can influence the direction of the story.

Future improvements may also include multilingual story generation and advanced voice modulation techniques to create more immersive storytelling experiences.

REFERENCES

[1] S. Sharma, A. Verma, and R. Singh, “AI-Based Story Generation Using Natural Language Processing,” International Journal of Computer Applications, vol. 183, no. 42, pp. 15–20, 2021.

[2] R. Verma and S. Bansal, “Sequence Modeling Techniques for Story Generation,” in Proceedings of the IEEE International Conference on Artificial Intelligence, 2023, pp. 120–125.

[3] A. Kulkarni and S. Chatterjee, “Image-to-Text Generation Using Deep Learning,” International Journal of Computer Vision, vol. 130, no. 5, pp. 1120–1135, 2022.

[4] K. Rao and A. Mishra, “Neural Text-to-Speech Systems: A Review,” IEEE Transactions on Audio, Speech, and Language Processing, vol. 29, pp. 230–245, 2021.

[5] A. Reddy and R. Banerjee, “User Profiling for Personalized Recommendation Systems,” Journal of Intelligent Systems, vol. 30, no. 2, pp. 210–220, 2021.

[6] T. Brown et al., “Language Models are Few-Shot Learners,” in Advances in Neural Information

Processing Systems (NeurIPS), vol. 33, pp. 1877–1901, 2020.

[7] A. Vaswani et al., “Attention Is All You Need,” in Advances in Neural Information Processing Systems (NeurIPS), 2017, pp. 5998–6008.

[8] Y. LeCun, Y. Bengio, and G. Hinton, “Deep Learning,” Nature, vol. 521, no. 7553, pp. 436–444, 2015.

[9] D. Jurafsky and J. H. Martin, Speech and Language Processing, 3rd ed., Pearson, 2022.

[10] A. Ahmed and P. Kapoor, “Multimodal AI for Content Generation: A Comprehensive Study,” IEEE Access, vol. 12, pp. 34567–34580, 2024.