Grenze International Journal of Engineering and Technology, January Issue

TaleWeave AI: Automated Visual Storytelling with Generative AI

Ranjana Jadhav[1], Suraj Baviskar[2], Aayush Bhadbhade[3], Shravani Chunkhade[4], Jayesh Bairagi[5] and Aarya Chavan[6]
[1-6]Dept. of Information Technology, Vishwakarma Institute of Technology, Pune, India
Email: ranjana.jadhav@vit.edu, suraj.baviskar23@vit.edu, aayush.bhadbhade232@vit.edu, shravani.chunkhade23@vit.edu, jayesh.bairagi23@vit.edu, aarya.chavan23@vit.edu

***Abstract*— The rapid advancement of Generative Artificial Intelligence (GenAI) has enabled novel applications in creative content generation. This paper introduces TaleWeave AI, an AI-powered web application designed to transform user-provided keywords into coherent, multi-paragraph narrative stories, with each paragraph subsequently illustrated by AI-generated images. TaleWeave AI integrates Large Language Models (LLMs), specifically Google Gemini, for narrative synthesis and offers a choice of image generation models, including Stable Diffusion or the Gemini Vision API, for visual enhancement. The system features a FastAPI backend for processing and a React frontend for an interactive user experience, facilitating a seamless workflow from keyword input to the final visually enriched story. We present the system architecture, methodology, and qualitative results demonstrating its capability to produce contextually relevant text and image pairings. TaleWeave AI aims to democratize creative storytelling, providing an accessible tool for applications in education, entertainment, and content creation. While acknowledging current limitations in visual consistency and subjective quality assessment, this work highlights the potential of combining state-of-the-art AI models for automated visual narrative generation.**

***Index Terms*— Generative Artificial Intelligence (GenAI), Story Generation, Text-to-Image Synthesis, Large Language Models (LLMs), Visual Storytelling.**

I. INTRODUCTION

Storytelling, a fundamental aspect of human communication and culture, serves as a powerful medium for education, entertainment, and the conveyance of complex ideas. In recent years, the advent of advanced Generative Artificial Intelligence (GenAI) has opened unprecedented avenues for augmenting and automating creative processes, including narrative generation and visual illustration. Large Language Models (LLMs) such as the GPT series and Google's Gemini family have demonstrated remarkable capabilities in generating coherent and contextually relevant text from minimal prompts. Concurrently, generative vision models, particularly diffusion models like Stable Diffusion, have achieved significant breakthroughs in synthesizing high-fidelity images from textual descriptions, transforming the landscape of digital art and content creation. This confluence of advanced AI technologies presents a fertile ground for developing novel tools that can democratize and enhance the storytelling experience.

Grenze ID: 01.GIJET.12.1.55

The exploration of AI in narrative and visual co-creation has garnered increasing research interest. Studies such as Ali and Parikh [1] have empirically demonstrated that AI-generated visuals can significantly enhance human creativity in storytelling, although they also highlighted potential challenges in maintaining narrative coherence with initial prompts. Building on this, researchers have explored more integrated systems. For instance, Chen et al. [2] proposed an interactive pipeline for visual story co-creation driven by user-specified emotions and keywords, leveraging diffusion models for image synthesis. Similarly, Sun et al. [3] introduced "1001 Nights," an AI-native game employing GPT-4 for dynamic text generation and Stable Diffusion for corresponding visualizations, where narrative elements directly influence game mechanics. More comprehensive frameworks have also emerged, such as MovieFactory by Zhu et al. [4], which aims to automate the creation of full movies from text using LLMs and generative models for both visual and audio synthesis, and StoryImager by Tao et al. [5], which focuses on ensuring coherence and realism in sequential story images through a unified bidirectional generation framework. While these works highlight significant progress in AI-driven narrative generation and visual storytelling, there remains a need for accessible, user-friendly applications that seamlessly integrate keyword-to-story generation with granular, paragraph-level visual illustrations, empowering a broader range of users in their creative endeavors.

In this paper, we introduce TaleWeave AI, an AI-powered web application designed to transform user-provided keywords into coherent, multi-paragraph narrative stories, with each paragraph subsequently illustrated by AI-generated images. TaleWeave AI aims to provide a fluid and intuitive platform for AI-assisted creativity, leveraging the power of contemporary LLMs, such as Google Gemini models, for narrative generation and state-of-the-art image synthesis techniques, including Stable Diffusion and the Gemini API, for visual enhancement. The system is architected with a FastAPI backend for robust AI task management and a React frontend for an interactive user experience, facilitating a seamless workflow from initial keyword input to the final visually-enriched story.

The primary contributions of this work are threefold: first, the design and implementation of a novel, fully integrated web application, TaleWeave AI, that combines keyword-driven story generation with paragraph-specific AI image illustration. Second, a demonstration of a practical pipeline that effectively utilizes leading generative AI models (Gemini and Stable Diffusion/Gemini Vision) for a cohesive text-and-image storytelling output. Third, an exploration of the potential of such a system to aid in various applications, including the creation of children's books, educational content, and entertainment, thereby lowering the barrier to creating visually rich narratives.

II. Literature Survey

The intersection of artificial intelligence, narrative generation, and visual synthesis has become a vibrant area of research, driven by rapid advancements in generative models. This section reviews key prior works that inform the development of systems like TaleWeave AI, focusing on AI-assisted storytelling, visual content generation, and integrated text-to-visual narrative frameworks.

A. Early Explorations in AI-Assisted Creativity

Early explorations into AI-assisted creativity highlighted the potential of generative visuals to augment human storytelling. Ali and Parikh [1] explored the influence of AI-generated visuals on human creative storytelling. Their study employed an experimental design where participants tasked with writing stories were either provided with visual aids generated by models such as VQGAN and BigGAN or worked without such aids (control group). The findings suggested a significant enhancement in perceived creativity, originality, and visualizability of stories produced with visual assistance, alongside increased user engagement. Notably, VQGAN emerged as the preferred visual generation tool among those tested. However, the research also highlighted a potential drawback: while AI visuals appeared to stimulate divergent thinking, they concurrently risked hindering convergent thinking, as participants utilizing visuals demonstrated less effective integration of the initial textual prompts. This observation underscores both the promise of AI in augmenting creative writing and the critical need to ensure that visual aids cohesively support rather than overshadow the core narrative.

B. Interactive and Controllable Visual Storytelling Systems

Subsequent research has focused on creating more interactive and controllable visual storytelling systems. Chen et al. [2] introduced an interactive visual story co-creation pipeline focused on user-guided narrative development through the integration of emotions and keywords. Their methodology involved a dual-component system: one for narrative generation utilizing fine-tuned BERT-based models for emotion prediction and T5-

based models for sentence creation, conditioned by Plutchik's emotion model, story context, and keywords; and another for image generation employing diffusion models like Disco Diffusion, with subsequent object recognition to extract keywords for iterative story enhancement. The findings, supported by metrics such as BLEU and BERTScore, indicated that explicitly prompting with emotions and keywords led to demonstrable improvements in the coherence and consistency of generated stories compared to baseline models reliant solely on story context. However, the research also acknowledged limitations, including the sub-optimal accuracy of their emotion classifier and challenges in generating specific types of imagery with Disco Diffusion without detailed artist references.

C. AI in Interactive Entertainment and Gaming

The application of generative AI has also extended into interactive entertainment and gaming. Sun et al. [3] introduced "1001 Nights," an AI-driven narrative game that explores the concept of "language as reality." In this co-creative storytelling experience, players guide an LLM-driven King (powered by GPT-4, employing LLM reasoning for story consistency) to narrate tales where specific keywords spoken by characters materialize as tangible in-game weapons for the protagonist, Shahrzad. The game world is dynamically visualized using Stable Diffusion, with a pixelization model applied to maintain a consistent art style, thereby blurring the lines between narrative and in-game reality. The authors propose the term "AI-Native games" for such experiences where Generative AI is fundamental to the core mechanics and existence of the game. While "1001 Nights" showcases an innovative fusion, challenges in evaluating player engagement and ensuring meaningful story contributions were noted.

D. Comprehensive Multi-Modal Generation Frameworks

More ambitious frameworks aim for comprehensive multi-modal generation. Zhu et al. [4] presented MovieFactory, a framework designed for the fully automated generation of multi-scene, high-definition movies with accompanying audio from textual inputs. Their approach leverages ChatGPT for script expansion and a two-stage extension of Stable Diffusion for video synthesis, complemented by an audio retrieval system. MovieFactory demonstrated superior performance on metrics like Frechet Video Distance (FVD) and CLIP similarity. However, its reliance on retrieved audio sidesteps the complexities of de novo audio synthesis.

Ensuring coherence in sequences of generated visuals is a critical challenge. Tao et al. [5] proposed StoryImager, a unified and efficient framework for coherent story visualization and completion. It employs a Storyboard-based Generation method with a Target Frame Masking Strategy and a decomposed Frame-Story Cross Attention Module built upon diffusion models. This approach enables bidirectional synthesis and addresses both visualization and completion tasks, outperforming previous models in image fidelity and story consistency. StoryImager's focus is on storyboard-style image sequences from given captions and their structural completion.

While these advancements demonstrate significant strides in AI-driven storytelling and visualization, TaleWeave AI aims to provide a distinct contribution by offering a user-friendly web application that directly translates user keywords into multi-paragraph narratives, each uniquely illustrated. This approach emphasizes ease of use and granular visual support for each narrative segment, catering to applications requiring rapid, AI-assisted creation of visually enriched stories.

III. METHODOLOGY

The TaleWeave AI system is designed as a web-based application employing a client-server architecture to facilitate the generation of narrative stories from user-defined keywords, complemented by AI-generated visual illustrations for each paragraph. The overall workflow, from user input to the final visually enriched story, is depicted in Fig. 1. The system comprises a frontend user interface, a backend processing engine, and interfaces to external generative AI services.

A. System Overview

The frontend, developed using React, provides an intuitive interface for user interaction, including keyword submission and the display of generated content. It communicates with the backend via RESTful APIs. The backend, implemented using FastAPI (Python), orchestrates the core logic of story and image generation, input validation, and interaction with generative AI models. The system leverages Google's Gemini models (Pro or Flash via the google.generativeai library) for narrative generation and offers a choice between Stable Diffusion or the Gemini API for image synthesis, configurable via environment variables.

B. Narrative Generation Process

The narrative generation process is initiated when the user submits a set of keywords through the frontend interface (Fig. 1, Step 1).

- Request Handling and Validation: The frontend transmits the keywords to the backend via a POST request to the /api/story/generate-with-images endpoint (Fig. 1, Step 2). Upon receiving the request, the FastAPI backend first validates the input keywords to ensure they meet predefined criteria (e.g., non-empty, appropriate length) (Fig. 1, Step 3).
- Story Synthesis: Validated keywords are then passed to the selected LLM (Google Gemini Pro or Flash). The LLM generates a coherent, multi-paragraph narrative story based on these keywords (Fig. 1, Step 4). The interaction with the Gemini API incorporates error handling and retry logic, managed by the tenacity library, to enhance robustness.
- Paragraph Segmentation: Once the complete story text is generated, the backend processes it to segment the narrative into individual paragraphs (Fig. 1, Step 5). This segmentation is crucial for the subsequent paragraph-specific image generation. The backend then returns the story text, along with its paragraph-wise breakdown, to the frontend.

C. Visual Illustration Module

Upon receiving the paragraph-segmented story, the frontend initiates the visual illustration process for each paragraph asynchronously.

- Paragraph-Specific Image Generation: For each paragraph, the frontend sends a separate POST request to the /api/story/generate-image endpoint on the backend. This request includes the text content of the specific paragraph.
- Image Synthesis: The backend, upon receiving a paragraph, utilizes the configured image generation model (Stable Diffusion or Gemini Vision API) to synthesize a visual illustration that corresponds to the semantic content of that paragraph (Fig. 1, Step 6). The generated image is returned to the frontend as a base64-encoded string. This asynchronous per-paragraph approach allows for a more responsive user experience, as images can be generated and displayed progressively.

D. Frontend User Interface and Workflow

The React-based frontend is responsible for rendering the generated content and managing user interactions.

- Content Display: The frontend displays the generated story paragraphs. As images are received from the backend, they are rendered side-by-side with their corresponding paragraphs (Fig. 1, Steps 7 and 8). uuid is employed to uniquely identify and manage each story section (paragraph and its associated image) during this process.
- User Experience Enhancements: Loading states are implemented to provide feedback to the user during story and image generation. Toast notifications, managed by the Sonner library, are used to convey status updates or errors. Styling is handled by TailwindCSS, with icons provided by Lucide Icons.
- Story Export: Finally, the user is provided with options to copy the generated story text to the clipboard or download the entire story as a text file.

IV. Results and Discussion

To evaluate the effectiveness of TaleWeave AI, we conducted a series of qualitative assessments based on a diverse set of keyword inputs. The system's performance was judged on two primary criteria: the narrative coherence of the generated stories and the contextual relevance of the AI-generated images to their corresponding paragraphs. For this evaluation, we utilized Google Gemini Pro for story generation and Stable Diffusion for image synthesis.

In one representative example, the input keywords were "enchanted forest, ancient ruins, hidden treasure." TaleWeave AI generated a four-paragraph narrative detailing the journey of a young adventurer named Elara. The first paragraph described her entry into the mystical forest, the second detailed her discovery of the vine-covered ruins, the third narrated the moment she found a treasure chest, and the final paragraph concluded with her reflections on the adventure. The AI-generated images for each paragraph visually corresponded to these narrative beats, depicting a lush forest, crumbling stone structures, a glowing treasure chest, and Elara looking thoughtfully at the horizon, respectively. This demonstrates the system's ability to generate a coherent story arc and provide relevant visual support.

In another test case with the keywords "futuristic city, flying cars, robot assistant," the system produced a story about a detective solving a case in a neon-lit metropolis. The images successfully captured the described sci-fi aesthetic, with visuals of sleek, airborne vehicles and a humanoid robot companion. The narrative followed a logical progression, and the images enhanced the futuristic atmosphere.

However, limitations were also observed. In some instances, while the images were contextually relevant at a high level, they lacked specific details mentioned in the text. For example, a paragraph describing a "red-haired protagonist" might be accompanied by an image of a protagonist with a different hair color. This indicates a challenge in ensuring fine-grained visual consistency. Furthermore, the subjective quality of the generated images varied, with some appearing more polished and aesthetically pleasing than others. These limitations highlight the ongoing challenges in text-to-image synthesis, particularly in controlling specific visual attributes and maintaining consistent quality.

Despite these limitations, TaleWeave AI successfully demonstrates the potential of combining LLMs and generative vision models for automated visual storytelling. The system provides a valuable tool for rapid prototyping of creative ideas and can serve as a source of inspiration for writers, artists, and educators. Future work will focus on improving visual consistency through more advanced prompt engineering and potentially incorporating user feedback loops for refining image generation. We also plan to explore the integration of additional generative models and expand the system's capabilities to include more complex narrative structures and interactive elements.

V. Conclusion

This paper introduced TaleWeave AI, an AI-powered web application that automates the creation of visually enriched narratives from user-provided keywords. By integrating Large Language Models for story generation and text-to-image models for visual illustration, TaleWeave AI provides a seamless and accessible tool for creative content creation. Our qualitative evaluations demonstrate the system's ability to produce coherent stories with contextually relevant images, highlighting its potential in education, entertainment, and digital art. While we acknowledge the current limitations in visual consistency and subjective image quality, this work represents a significant step toward democratizing visual storytelling. Future research will focus on enhancing the system's capabilities, improving the quality and consistency of generated content, and exploring new applications for this technology.

Acknowledgment

The authors wish to thank the Department of Information Technology at Vishwakarma Institute of Technology for their support and resources. This work was supported by the institutional research facilities and computational resources provided by VIT, Pune.

References

[1] S. Ali and R. Parikh, "The Impact of AI-Generated Visuals on Human Creative Storytelling," in Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, 2022, pp. 1–14.

[2] Y. Chen, Z. Wu, and Y. Liu, "An Interactive Visual Story Co-creation Pipeline via Emotion and Keyword Control," in Proceedings of the 30th ACM International Conference on Multimedia, 2022, pp. 1-5.

[3] Z. Sun, Y. Wang, and J. Li, "1001 Nights: An AI-Native Narrative Game," in Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems, 2023, pp. 1-8.

[4] H. Zhu, J. He, and J. Li, "MovieFactory: An End-to-End Framework for Movie Generation from Text," arXiv preprint arXiv:2306.05275, 2023.

[5] Y. Tao, Z. Li, and J. Feng, "StoryImager: A Unified and Efficient Framework for Coherent Story Visualization and Completion," arXiv preprint arXiv:2305.10328, 2023.

[6] R. Jadhav, S. Baviskar, A. Bhadbhade, S. Chunkhade, J. Bairagi, and A. Chavan, "TaleWeave AI: Automated Visual Storytelling with Generative AI," in 2024 International Conference on Communication, Information and Computing Technology (ICCICT), 2024, pp. 1-6.