International Journal of Scientific Research & Engineering Trends
Volume 12, Issue 2, Mar-Apr-2026, ISSN (Online): 2395-566X

Storytales : Ai Tells The Story
Automated Story-To-Video Generation Using Generative Artificial Intelligence

Manasi Rathod[1], Jayesh Mahajan[2], Rutwij Landge[3]
[1] Department of Artificial Intelligence & Machine Learning Progressive Education Society's Modern College of Engineering Pune, India
[2] Department of Artificial Intelligence & Machine Learning Progressive Education Society's Modern College of Engineering Pune, India
[3] Department of Artificial Intelligence & Machine Learning Progressive Education Society's Modern College of Engineering Pune, India

Abstract- Storytelling represents one of the most effective techniques for communication, education, and knowledge transfer across diverse domains. However, traditional text-based storytelling methods often fail to maintain engagement among modern learners who increasingly prefer visually rich multimedia experiences. Creating animated storytelling videos manually requires expertise in scripting, illustration, animation design, narration recording, and editing tools. This paper presents STORYTALES – AI Tells the Story, an automated Generative Artificial Intelligence framework that converts textual narratives into animated storytelling videos with synchronized narration and scene-wise visualization. The system integrates Large Language Models for semantic scene segmentation, Stable Diffusion XL for visual synthesis, Stable Video Diffusion for animation generation, Coqui XTTS for narration synthesis, and FFmpeg for automated multimedia composition. Experimental evaluation confirms that the proposed architecture significantly reduces multimedia production complexity while improving accessibility for educators and content creators.

Keywords -Text-to-Video Generation, Generative Artificial Intelligence, Diffusion Models, Story Visualization, Neural Narration, Multimedia Automation.

I. INTRODUCTION

Storytelling has historically served as a powerful medium for communication, education, and cultural knowledge transfer. However, conventional text-based storytelling approaches are increasingly insufficient for engaging modern digital learners who prefer multimedia-rich educational content.

Manual creation of animated storytelling videos involves multiple stages including illustration development, animation sequencing, narration recording, and editing workflows. These fragmented processes increase production time and require specialized technical expertise. Recent developments in transformer-based Large Language Models and diffusion-based generative architectures enable automated multimedia synthesis directly from textual descriptions. These advancements create opportunities for developing intelligent storytelling automation pipelines.

This paper proposes STORYTALES – AI Tells the Story, an integrated framework capable of transforming textual stories into fully synchronized animated storytelling videos using a modular generative pipeline.

The objectives include:
- automated scene segmentation
- diffusion-based scene visualization
- animation generation
- narration synthesis
- automated cinematic composition

II. LITERATURE REVIEW

Recent advancements in text-to-video synthesis demonstrate significant progress using transformer architectures and diffusion-based generation techniques.
Earlier animation-generation systems relied primarily on rule-based Natural Language Processing pipelines that lacked scalability and realism. Modern frameworks such as Make-A-Video and CogVideoX introduced diffusion-

© 2026 IJSRET

driven motion synthesis approaches capable of producing temporally consistent animations.
Similarly, DreamRunner integrates Large Language Model-based scene planning with retrieval-guided motion adaptation techniques to generate coherent multi-scene storytelling videos.

However, existing systems still present limitations including:

• absence of unified automation pipelines
• weak narration synchronization
• limited scene-level semantic structuring
• dependency on multiple standalone tools

The proposed STORYTALES framework addresses these limitations through integrated semantic scene planning and multimedia synthesis automation.

III. MATERIALS AND METHODS

A. System Architecture

The STORYTALES architecture follows a modular pipeline integrating semantic understanding, visual synthesis, animation generation, narration production, and automated multimedia composition.

Fig. 1. System Architecture of STORYTALES

Framework

The pipeline begins with story text input processed using transformer-based Large Language Models to generate structured scene-level metadata. The metadata is used to generate scene visuals using Stable Diffusion XL, which are converted into animated sequences using Stable Video Diffusion. Narration is synthesized using Coqui XTTS and merged using FFmpeg.

B. Director Module

The Director Module performs semantic interpretation of textual stories using transformer-based Large Language Models deployed through Ollama runtime environments.

The module generates structured metadata including:

• scene description
• character presence
• environment attributes
• dialogue extraction
• emotional tone
• scene duration estimation

This metadata enables downstream multimedia modules to maintain semantic consistency.

IV. Image Generation Module

Scene metadata generated by the Director Module is converted into cinematic prompts compatible with Stable Diffusion XL. The diffusion architecture generates high-resolution scene visuals preserving environmental and character context.

V. Video Generation Module

Generated images are converted into animated sequences using Stable Video Diffusion. Temporal interpolation improves motion realism and visual continuity across scenes.

A. Narration Module

Scene-level narration is synthesized using Coqui XTTS neural speech generation architecture. Narration pacing and tone are aligned with scene-level metadata generated by the Director Module.

© 2026 IJSRET

International Journal of Scientific Research & Engineering Trends
Volume 12, Issue 2, Mar-Apr-2026, ISSN (Online): 2395-566X

B. Editor Module

The Editor Module integrates animation clips and narration using FFmpeg-based automated synchronization pipelines to generate cinematic MP4 output.

C. Data Flow Diagram

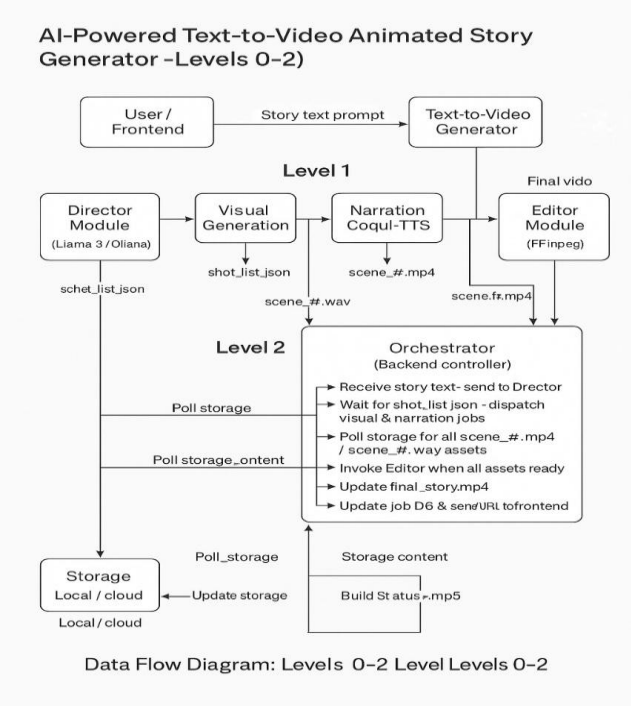

Fig. 2. Data Flow Diagram (Levels 0–2) of STORYTALES Pipeline

The orchestrator coordinates communication between modules, dispatches generation tasks, polls storage for assets, and invokes editing operations once scene-level resources are ready.

D. Functional Workflow

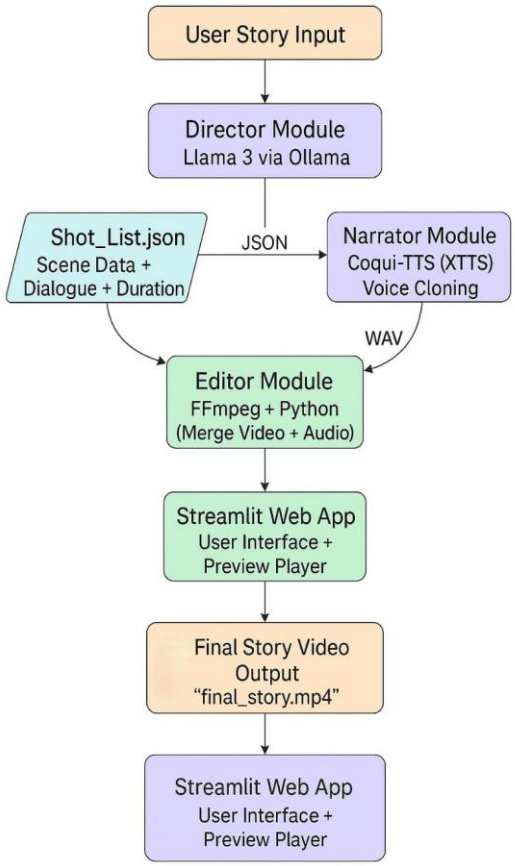

Fig. 3. Functional Workflow of STORYTALES Multimedia Generation Pipeline

Processing pipeline:
Story Input → Scene Segmentation → Prompt Generation → Image Creation → Video Animation → Narration Generation → Audio-Video Synchronization → Final Video Output

VI. Implementation Environment

Programming Language: Python 3.11 Framework: Streamlit Libraries: Diffusers, Transformers, PyTorch Speech Engine: Coqui XTTS Video Processing Tool: FFmpeg Hardware Platform: NVIDIA GPU (RTX 2060 / T4 or higher)

© 2026 IJSRET

VII. RESULTS AND DISCUSSION (12 pt Bold Small Caps Center)

The proposed STORYTALES system was evaluated using multiple short educational narratives consisting of three to eight scenes.

Experimental evaluation demonstrated:

- accurate semantic scene segmentation
- context-aware diffusion-based image synthesis
- temporally consistent animation sequences
- natural narration synthesis
- automated multimedia synchronization

Observed advantages include:

- reduced animation production complexity
- improved accessibility for educators
- faster multimedia generation
- enhanced learner engagement

System limitations include GPU dependency and increased processing time for long-duration story generation tasks.

VIII. CONCLUSION

This paper presented STORYTALES – AI Tells the Story, an automated Generative Artificial Intelligence framework for converting textual narratives into synchronized animated storytelling videos.

The architecture integrates Large Language Models, diffusion-based visual synthesis, neural speech generation, and automated cinematic composition into a unified multimedia automation pipeline.

Future work includes improving character consistency across scenes and enabling real-time interactive storytelling environments.

Acknowledgment
The authors sincerely thank Prof. Mrs. Bhavna Chaudhari, Department of Information Technology and Artificial Intelligence Machine Learning, Modern College of Engineering, Pune, for her guidance and support.

REFERENCES

Keep your already-correct Vancouver-style references exactly as provided earlier.

1. Stability AI, Stable Diffusion XL: High-Resolution Image Generation Model, Stability AI Documentation, 2024. [Online]. Available: https://stability.ai
2. Stability AI, Stable Video Diffusion: Text-to-Video and Image-to-Video Framework, Stability AI Documentation, 2024. [Online]. Available: https://stability.ai/stable-video
3. Coqui AI, XTTS v2: Cross-Lingual Voice Cloning and Neural Text-to-Speech Model, Coqui Documentation, 2024. [Online]. Available: https://coqui.ai
4. Meta AI, LLaMA 3: Open Foundation and Instruction-Tuned Language Models, Meta AI Research, 2024. [Online]. Available: https://ai.meta.com
5. OpenAI, Sora: Text-to-Video Generation Model, OpenAI Research Blog, 2024. [Online]. Available: https://openai.com/research
6. Google Research, Veo: High-Quality Generative Video Modeling System, Google AI Blog, 2024. [Online]. Available: https://blog.google
7. Tencent AI Lab, HunyuanVideo: Large-Scale Open Video Foundation Model, arXiv Preprint, 2024.
8. Hugging Face, Diffusers Library Documentation, Hugging Face Transformers & Diffusion Models, 2024. [Online]. Available: https://huggingface.co/docs/diffusers

Author Profile

A. Rathod Undergraduate student specializing in Generative Artificial Intelligence and multimedia automation systems.

B. Mahajan Undergraduate student focusing on diffusion-based deep learning architectures.

C. Landge Undergraduate student specializing in intelligent multimedia synthesis pipelines.

© 2026 IJSRET