AniTales: End-to-End Multimodal Story Generation Through Natural Language Prompting (Student Abstract)

Mrigendra Agrawal[1], **Yunze Xiao**[2]
[1]The University of Queensland
St Lucia QLD 4072, Australia
[2]Carnegie Mellon University
mrigendra.agrawal@student.uq.edu.au, yunzex@cs.cmu.edu

Abstract

We present AniTales, a system designed to generate multimodal visual novels from natural language prompts. Our system integrates large language models for story generation, diffusion models for character art, and text-to-speech for voice acting. This paper describes the system's architecture and presents findings from a pilot user study. We evaluated the system with general users (n=10) and domain experts (n=5), focusing on usability, coherence, and visual consistency. General users reported high usability (SUS: 84/100) and strong character-dialogue consistency (4.2/5), along with an average score of 82/100 for their intention to continue using the platform. These initial results suggest AniTales is a promising approach for bridging the gap between text-based AI storytelling and end-to-end multimedia content creation.

AniTales — https://anitales.chat

Introduction

AI storytelling platforms like AI Dungeon and Character.AI show significant demand for AI based interactive media, yet most experiences remain text only. Visual novels offer an ideal testbed for evaluating AI-driven multimodal storytelling. A visual novel is an interactive story presented on screen with character sprites placed over backgrounds, dialogue shown in a text box, as well as background music and voice acting.

We present AniTales, which turns natural language prompts into complete, playable visual novels. Prior tools are fragmented, require manual asset creation, produce inconsistent characters across scenes, or fail to tie visuals and voice to the emotional moment. AniTales provides an end to end, coherence focused pipeline that automatically generates consistent character art using FLUX.1 dev Style LoRA (Labs 2024), emotionally aligned voice acting, and branching interactivity that maintain continuity across modalities. In a user study with general users (n=10), AniTales achieved high usability (SUS 84 out of 100) and strong ratings for visual and narrative consistency (4.2 out of 5). Beyond entertainment, it enables rapid, customizable scenarios for education and language learning, democratizing multimedia creation.

Figure 1: AniTales generation example

Researchers have been working on turning stories into pictures and videos. Early work made datasets to convert text into image sequences (Huang et al. 2016). Later systems tried to make stories scene by scene, but maintaining character consistency proved challenging (Li et al. 2019). Newer work makes longer stories and lets users add to them (Maharana, Hannan, and Bansal 2022; Yang et al. 2024), but characters still don't stay consistent.

Some systems let users make interactive stories with images, dialogue, and choices. These work with multiple characters but need lots of user input to keep things consistent (Gong et al. 2023). AI-powered systems do more work automatically (Shen et al. 2025). However, character identity still breaks down across scenes, making stories feel disconnected. Other research tries new ways to keep character identity consistent across different scenes (Tewel et al. 2024; Akdemir and Yanardag 2024; Avrahami et al. 2024), but these focus on individual image generation rather than complete interactive experiences. NarrativePlay (Zhao et al. 2024) generates images based on context, but they are still not consistent across scenes.

AniTales differentiates itself by generating complete visual novels from start to finish. It handles story structure, creates all the art automatically, and keeps characters looking right across all scenes. This means less work for authors while keeping everything consistent. We also ran user studies to test how well our system works.

System Design

AniTales uses a modular pipeline that coordinates multiple AI models to produce cohesive multimodal stories. The *Story Generation Module* processes user prompts with LLMs (Gemini, OpenRouter) and returns a structured JSON containing speaker identification, dialogue, background prompts, and crucially, emotion_tags for each utterance. Users can choose style, voices, and safety level. An AI moderator reviews the prompt and only compliant content proceeds.

For *Character Portrait Generation*, FLUX.1 dev with Style LoRA (Labs 2024) creates initial portraits from text. After users accept a portrait, sprite expressions are generated by prompting FLUX.1 Kontext to render the same character with specific expressions (e.g., neutral, angry, happy).

The *Voice Synthesis Pipeline* uses MiniMax Speech 02 to render dialogue. This module is directly guided by the emotion_tags from the JSON script to ensure vocal prosody follows the emotional context. Audio files are stored for reuse.

A *Scene Composition Engine* assembles interactive scenes, using the JSON's speaker and emotion tags to synchronize text, audio, and the correct character sprite. The *Branching and Export Module* inserts branch points at any scene with continuation conditioned on dialogue history. Projects export as JSON or video.

Results

We ran two complementary studies. The general user study recruited ten Prolific participants familiar with Character.AI and similar roleplay services; all were avid gamers (80% play weekly or more) and all had completed at least one visual novel; demographics were 18–34 (80%) with an even gender split (50% male, 50% female). A separate expert study involved five domain experts (developers, visual novel readers, and artists) who used the same questionnaire. Each participant created one or two ten-scene stories, designed two or three characters, played the result, and then completed ratings and open-ended feedback.

Table 1 summarizes the metrics. General users reported high narrative coherence and visual consistency, with SUS 84 ± 9.7 indicating solid usability; experts were more critical of logic yet rated visuals highest. General users rated their intention to continue using the platform at 82/100 on average. Voice quality was the only metric below four in both cohorts, marking the clearest area for improvement. Future work will focus on longer-form coherent story generation and more modes of interaction, improved voice libraries and emotional prosody for more natural conversation, and platform expansion to mobile and tablet, followed by a larger-scale evaluation.

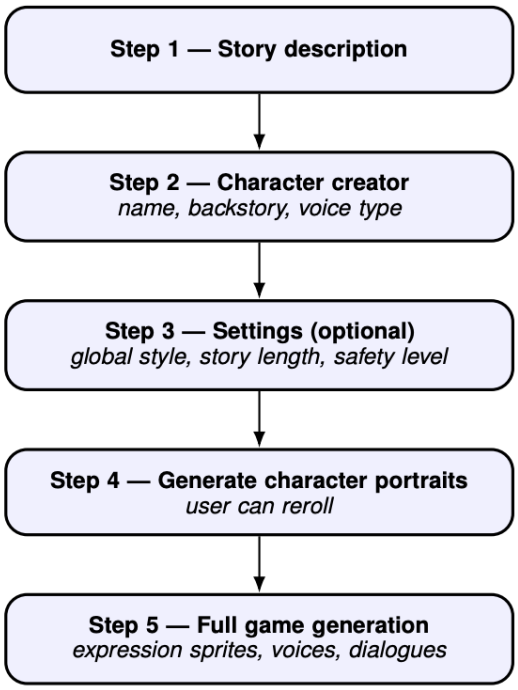

Figure 2: AniTales end to end pipeline

Metric	General Users (n=10)	Experts (n=5)
Logical Consistency (1–5)	4.4 ± 0.7	3.4 ± 1.1
Character–Dialogue Match (1–5)	4.2 ± 0.79	4.0 ± 0.0
Voice Suitability (1–5)	3.6 ± 0.84	3.6 ± 0.9
Visual Consistency (1–5)	4.2 ± 0.78	4.8 ± 0.44
Likelihood to Recommend (1–10)	8.2 ± 1.1	6.4 ± 2.87
Overall Enjoyment (1–5)	4.3 ± 0.48	3.8 ± 1.0
SUS (0–100)	84 ± 9.7	81 ± 13.2

Table 1: Evaluation metrics for general users and experts (five-point scale except where noted).

Ethical Considerations

AniTales includes an AI moderator implemented as a separate LLM call to screen prompts and outputs. Prolific participants were compensated at €12 per hour; expert evaluators participated voluntarily.

References

Akdemir, K.; and Yanardag, P. 2024. Oracle: Leveraging mutual information for consistent character generation with loras in diffusion models. *arXiv preprint arXiv:2406.02820*.

Avrahami, O.; Hertz, A.; Vinker, Y.; Arar, M.; Fruchter, S.; Fried, O.; Cohen-Or, D.; and Lischinski, D. 2024. The chosen one: Consistent characters in text-to-image diffusion models. In *ACM SIGGRAPH 2024 conference papers*, 1–12.

Gong, Y.; Pang, Y.; Cun, X.; Xia, M.; He, Y.; Chen, H.; Wang, L.; Zhang, Y.; Wang, X.; Shan, Y.; et al. 2023. Inter-

active story visualization with multiple characters. In *SIGGRAPH Asia 2023 Conference Papers*, 1–10.

Huang, T.-H.; Ferraro, F.; Mostafazadeh, N.; Misra, I.; Agrawal, A.; Devlin, J.; Girshick, R.; He, X.; Kohli, P.; Batra, D.; et al. 2016. Visual storytelling. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Human language technologies*, 1233–1239.

Labs, B. F. 2024. FLUX. https://github.com/black-forest-labs/flux.

Li, Y.; Gan, Z.; Shen, Y.; Liu, J.; Cheng, Y.; Wu, Y.; Carin, L.; Carlson, D.; and Gao, J. 2019. Storygan: A sequential conditional gan for story visualization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6329–6338.

Maharana, A.; Hannan, D.; and Bansal, M. 2022. Storydall-e: Adapting pretrained text-to-image transformers for story continuation. In *European conference on computer vision*, 70–87. Springer.

Shen, X.; et al. 2025. StoryGPT-V: Large Language Models as Consistent Story Visualizers. In *CVPR*.

Tewel, Y.; Kaduri, O.; Gal, R.; Kasten, Y.; Wolf, L.; Chechik, G.; and Atzmon, Y. 2024. Training-free consistent text-to-image generation. *ACM Transactions on Graphics (TOG)*, 43(4): 1–18.

Yang, S.; Ge, Y.; Li, Y.; Chen, Y.; Ge, Y.; Shan, Y.; and Chen, Y. 2024. Seed-story: Multimodal long story generation with large language model. *arXiv preprint arXiv:2407.08683*.

Zhao, R.; Zhang, W.; Li, J.; Zhu, L.; Li, Y.; He, Y.; and Gui, L. 2024. Narrativeplay: An automated system for crafting visual worlds in novels for role-playing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 23859–23861.