AIsop: Exploring Immersive VR Storytelling Leveraging Generative AI

Elia Gatti* Daniele Giunchi† Nels Numan‡ Anthony Steed§

University College London, UK

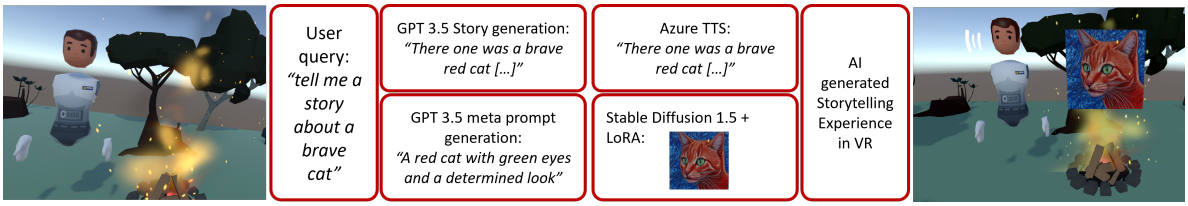

Figure 1: AIsop: AI-generated storytelling in VR.

Abstract

We introduce AIsop, a system that autonomously generates VR storytelling experiences using generative artificial intelligence (AI). AIsop crafts unique stories by leveraging state-of-the-art Large Language Models (LLMs) and employs Text-To-Speech (TTS) technology for narration. Further enriching the experience, a visual representation of the narrative is produced through a pipeline that pairs LLM-generated prompts with diffusion models, rendering visuals for clusters of sentences in the story. Our evaluation encompasses two distinct use cases: the narration of pre-existing content and the generation of entirely new narratives. AIsop highlights the myriad research prospects spanning its technical architecture and user engagement.

Keywords: large language model, VR, storytelling, generative AI

Index Terms: Human-centered computing—Virtual reality Human-centered computing—Natural language interfaces

1 Introduction

Virtual Reality (VR) has emerged as a powerful medium for storytelling, offering immersive experiences that traditional mediums cannot replicate. The evolution of VR technology and its associated content, though not as rapid as anticipated, has shown consistent growth [2]. This paper focuses on the exploration of storytelling within the VR domain. Notable early examples of VR storytelling, such as "Son of Jaguar" and "Dear Angelica" [6] exemplify how narrative elements are effectively interwoven with visual representations in virtual environments, demonstrating the unique capabilities of VR in enhancing the storytelling experience. AI's role in storytelling extends across multiple facets, ranging from assisting narrators in maintaining coherence in the main plot and its subplots [8], to predicting characters' emotional arcs [1]. In interactive mediums like computer games, AI has been instrumental in developing behaviors for Non-Playable Characters (NPCs) [4], as well as in generating dynamic NPC dialogues [3]. The emergence of Large Language Models (LLMs) has marked a significant leap forward in generating written content. These models, trained on extensive datasets, can emulate patterns characteristic of human writing, often producing outputs nearly indistinguishable from human-authored texts [12]. LLMs' role in collaborative storytelling has been previously explored [13], yielding results engaging and reminiscent of human creativity. We connect the world of VR entertainment and AI storytelling by presenting AIsop. AIsop represents an initial exploration of VR storytelling, utilizing LLMs and image generation models for content creation. The system produces a composite output that includes story content, narration, and visual representation. Its framework, based on the Ubiq [5, 9] platform, is highly modular, enabling seamless integration of various third-party models including Text-To-Speech (TTS), image, and speech generation. This flexibility allows for the creation of tailored storytelling experiences. A unique aspect of AIsop is its use of prompt engineering to generate narrative content and corresponding visual 'meta-prompts concurrently.' The exploratory nature of AIsop is illustrated through a two-phase user study, which assesses the impact of AI-generated visuals on enhancing narrative engagement and user experience within VR.

2 User Experience Walkthrough

We introduce a VR storytelling approach that combines the natural interactive capabilities of LLMs with the immersive environment of VR to enhance user engagement with the narrative. The VR experience starts with users at a virtual camping site, greeted by an AI agent's avatar beside a bonfire. Users are invited to suggest a story theme, for example, *"A story about a large green tortoise and a cherry tree"*. Following this input, the AI agent commences the storytelling session. As the narrative unfolds, static images correlating with the story's events materialize mid-air around the bonfire. This bonfire serves a dual purpose: it acts as a narrative device to justify the appearance of images. It helps guide the user's attention, addressing potential issues with attentional focus in VR. Development of AIsop began with creating an immersive VR scene using the Unity engine. A key feature within this scene is the visual effect of images emerging from the bonfire. This was achieved by designing a floating, transparent projection surface above the bonfire, onto which 2D images generated by our pipeline are displayed. The integration of Ubiq and its extension Ubiq-Genie enabled the creation of a networked collaborative experience relying on off-the-shelf generative AI models.

3 Prompt Engineering

Since LLMs can produce varied texts based on query construction, additional context is often provided to align the LLM with the task. To craft the prompt most suitable for generating stories, a taxonomy of prompt modifiers like subject terms and style modifiers was used to steer LLM outputs. We focused on creating prompts with dual objectives: generating storytelling content and meta-prompts for

*e-mail: elia.gatti@ucl.ac.uk
†e-mail: d.giunchi@ucl.ac.uk
‡e-mail: nels.numan@ucl.ac.uk
§e-mail: a.steed@ucl.ac.uk

image creation. The final prompt template integrated user topics and standardized LLM outputs and avoided using proper nouns or character descriptions to aid image generation. The prompt also included specific structures for meta-prompts to trigger image creation during narration. Fifteen different prompts were tested against criteria like story length, distinction between story and meta-prompts, and clarity for image generation. The most successful prompt followed a structured format and was refined to ensure uniform style, using techniques like repetitions and specific style references.

4 Image Generation

Diffusion models are becoming more prevalent in the digital art scene, notably in creating comics and graphic novels. However, a significant challenge arises in generating consistent representations of characters across varied contexts. Addressing this issue involves methods such as prompt engineering and more advanced techniques like fine-tuning the models. Specific strategies like Dreambooth [11] have been employed to train models with select images to achieve consistent visual representations and create diverse scenes. In character design consistency, Low-Rank Adaptation (LoRA) models have demonstrated promising results [7]. These models train on limited, specialized image datasets to achieve greater uniformity in character depiction. This research utilizes Stable Diffusion 2 [10], enhanced with LoRA weights, to generate uniform character images. The generation process involves textual input from GPT 3.5 and focuses on maintaining visual consistency throughout the narrative.

5 User Studies

Experiment 1: *Evaluating the Influence of AI-Generated Imagery* This experiment enlisted 12 volunteers to evaluate the impact of images generated by AIsop. Participants were exposed to two narrative formats: "audiobook mode" (audio only) and "picture mode" (audio with images). Their assessment focused on the quality of images, relevance to the story, and presentation timing. The findings indicate that integrating visual aids notably enhances the listening experience, fostering deeper immersion and aiding comprehension. Participants generally appreciated the timing and story relevance of the images. Anecdotal responses suggested improvements in attention and story processing when images were incorporated. However, some discrepancies in the depiction of characters were observed.

Experiment 2: *Comprehensive Assessment of AIsop's VR Storytelling Experience* In this study, seven individuals with experience in VR and storytelling engaged with AIsop. The evaluation encompassed the VR environment, user interaction, immersion, embodiment, and the storytelling process. The simplicity of the VR environment was well-received, though there was a desire for more interactive elements, enhanced AI agent engagement, sound effects, and 3D visuals. High immersion levels were reported during storytelling sessions, but the lack of full-body representation led to a low sense of embodiment. The coherence of AI-crafted narratives was noted, though some narratives were perceived as linear. The TTS voice was criticized for its robotic nature, with suggestions for dynamic voice modulation and pacing. Lastly, the system's potential for educational and entertainment purposes was acknowledged, and there was enthusiasm for its continued development.

6 Conclusion and Future Work

Our early exploration of AI-integrated VR storytelling has demonstrated the potential to deliver immersive and engaging experiences. The feedback from users was predominantly positive, with the AI-generated stories and accompanying imagery being commended for their relevance and timeliness. Future enhancements include:

- **Optimizing Storytelling**: Enhancing the GPT model's storytelling capabilities could further engage users. Techniques such as annotated datasets, graph-based structures, or preference learning could refine the LLM's storytelling proficiency.
- **Visual Representation Enhancement**: Broadening the "Template characters" range and investigating novel approaches for consistent character portrayal would enrich the visual experience. Additionally, optimizing the timing and method of image presentation could elevate the user experience.
- **Advanced TTS Voices**: Integrating Long-Form TTS voices specifically tuned for storytelling could significantly improve narrative expressiveness and user engagement.
- **Interactive and Collaborative Storytelling**: AIsop's Ubiq-based framework for social VR could enable collaborative storytelling, fostering interaction between multiple users and AI entities.
- **Testing New Integrations**: AIsop's modular design facilitates the experimentation with new features, such as semantically linked sounds or alternative image generation models, broadening its applicability and enhancing user experience.

References

[1] F. Brahman and S. Chaturvedi. Modeling protagonist emotions for emotion-aware storytelling. *arXiv preprint arXiv:2010.06822*, 2020. 1

[2] J. Bucher. *Storytelling for virtual reality: Methods and principles for crafting immersive narratives*. Routledge, 2017. 1

[3] M. Cavazza and F. Charles. Dialogue generation in character-based interactive storytelling. In *AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, vol. 1, pp. 21–26, 2005. 1

[4] C. R. Fairclough and P. Cunningham. Ai structuralist storytelling in computer games. In *Proceedings of the International Conference on Computer Games: Artificial Intelligence, Design and Education*, vol. 5. University of Wolverhampton Press Reading, 2004. 1

[5] S. J. Friston, B. J. Congdon, D. Swapp, L. Izzouzi, K. Brandstätter, D. Archer, O. Olkkonen, F. J. Thiel, and A. Steed. Ubiq: A system to build flexible social virtual reality experiences. In *Proceedings of the 27th ACM Symposium on Virtual Reality Software and Technology*, VRST '21. Association for Computing Machinery, New York, NY, USA, 2021. doi: 10.1145/3489849.3489871 1

[6] J. Gutierrez, D. Eisenmann, C. Curtis, K. Bodyfelt, J. Anderholm, C. Curtis, S. Stafford, K. Dart, C. Cellucci, T. Latzko, et al. Behind the headset: the making of google spotlight stories'" son of jaguar"," sonaria", and oculus story studio's" dear angelica". In *ACM SIGGRAPH 2017 Production Sessions*, pp. 15–15. 2017. 1

[7] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. Lora: Low-rank adaptation of large language models, 2021. 2

[8] M. Kreminski, M. Dickinson, N. Wardrip-Fruin, and M. Mateas. Loose ends: a mixed-initiative creative interface for playful storytelling. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, vol. 18, pp. 120–128, 2022. 1

[9] N. Numan, D. Giunchi, B. Congdon, and A. Steed. Ubiq-genie: Leveraging external frameworks for enhanced social vr experiences. In *2023 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, pp. 497–501, 2023. doi: 10.1109/VRW58643.2023.00108 1

[10] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-Resolution Image Synthesis with Latent Diffusion Models. In *CVPR2022*, pp. 10674–10685. IEEE, New Orleans, LA, USA, June 2022. doi: 10.1109/CVPR52688.2022.01042 2

[11] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22500–22510, June 2023. 2

[12] T. J. Sejnowski. Large language models and the reverse turing test. *Neural computation*, 35(3):309–342, 2023. 1

[13] H. Shakeri, C. Neustaedter, and S. DiPaola. Saga: Collaborative storytelling with gpt-3. In *Companion Publication of the 2021 Conference on Computer Supported Cooperative Work and Social Computing*, pp. 163–166, 2021. 1