Design Considerations of Voice Articulated Generative AI Virtual Reality Dance Environments

Llogari Casas, Kenny Mitchell
3FINERY LTD
Edinburgh Napier University
United Kingdom
llogari@3finery.com
kenny@3finery.com

Monica Tamariz
Edinburgh Napier University
Heriot-Watt University
United Kingdom
m.tamariz@hw.ac.uk

Samantha Hannah, David Sinclair, Babis Koniaris, Jessie Kennedy
Edinburgh Napier University
United Kingdom
S.Hannah@napier.ac.uk
d.sinclair@napier.ac.uk
b.koniaris@napier.ac.uk
J.Kennedy@napier.ac.uk

Abstract

We consider practical and social considerations of collaborating verbally with colleagues and friends, not confined by physical distance, but through seamless networked telepresence to interactively create shared virtual dance environments. In response to speech recognition textual language prompts our, *HoloJig* system performs according to Dolgoff and Roddenberry's science fiction language guided *Holodeck* narrative environment generation. Here instead realized for presentation through virtual reality headsets to create dance halls, clubs, concerts or more abstract spaces.

CCS Concepts: • **Computing methodologies → Mixed / Virtual Reality**.

Keywords: Virtual Reality, Artificial Intelligence, Generative AI, Human-Computer Interaction, Social Media, Ethics

ACM Reference Format:
Llogari Casas, Kenny Mitchell, Monica Tamariz, and Samantha Hannah, David Sinclair, Babis Koniaris, Jessie Kennedy. 2024. Design Considerations of Voice Articulated Generative AI Virtual Reality Dance Environments. In *Proceedings of Dancing in the Holodeck: Generative AI and the Future of Remote Collaboration (GenAI in UGC Workshop, CHI24).* ACM, New York, NY, USA, 3 pages. https://doi.org/XXXXXXX.XXXXXXX

GenAI in UGC Workshop, CHI24, Honolulu, HI, USA,

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

Figure 1. Partners immersed with their avatars represented in an online *Plant World* fantasy dance space generated interactively with speech-to-text and generative AI imaging optimized for virtual reality headset display.

1 Introduction

For centuries, dance has served as a powerful means of human expression that goes beyond cultural boundaries promoting connection through shared movement and emotions. Nevertheless, the traditional limitations of physical space and distance have often constrained collaborative dance practices, hindering the ability of dancers to connect and express themselves freely with remote collaborators.

Much explored in the narrative fantasies of Dolgoff and Roddenberry's *Holodeck* [2, 8] we further approach the practical realization of language guided environment generation [10], here instead optimized for the purpose of responsively providing a place for dance partners to co-experience their articulated setting in real-time in virtual reality.

Our *HoloJig*[1] system initially provides a means of interactive creation of 3D environments around the participants in virtual reality in response to verbal text prompts.

We begin with a brief outline of the system's methods as a context, then we follow with the design considerations in

[1]With 'jig' meaning both dance and a guiding machine tool, the name intends a solution for guided generation of Holodecks for dancing.

the categories of practicality and scalability, user experience friction, social factors, trust and safety and bias.

2 Practical Generative AI Environments

The vision of synchronized online dance collaboration has long been hampered by the limitations of networked body pose sharing, introducing latency that disrupts the critical real-time nature of dancing as tackled in the *DanceGraph* system [7]. With *HoloJig* we empower personalized and interactive environments that transform remote dance collaborations into a user specified dynamic and immersive experience. This allows dancers to explore a diverse range of settings, adapting landscapes and atmospheres to suit the theme, mood, and cultural context of their performance. In that sense, dancers can become virtual architects, manipulating elements like lighting, scenery, props, and atmospheric effects to their preference. Through the use of speech-to-text prompt generation, dancers can describe their ideal settings verbally, with spoken words that subsequently translate into immersive virtual landscapes (see figure 1). Similarly to *The EXPERIENCE Project* [9], we utilize a fine-tuned version of a high-resolution image synthesis latent diffusion model [4], originally trained on the massive LAION-5B dataset [6], to produce generative environments. The resultant albedo and depth maps are projected into cube map space and rendered with depth parallax to provide left and right eye stereoscopic frames efficiently for low-latency high refresh rate virtual reality headset displays.

Determining which dance partner has control to establish the immersive scene to co-experience with others presents some alternatives, which we draw from related VR collaborative animation authoring work of Pan et al [3], with

- a *master* controller assigned among the group to vocally summarize a prompt following verbal agreement
- a *most recently changed* rule of the last prompt issued among any participant
- an averaged LLM summarized prompt generated among the collection of suggested prompts from participants

Further additional context of an image sketch or selected photo can steer the resulting generative spatial environment imagery. The generation of cube image and depth maps is not free of cost for large scale context delivery networks (less than dense 3D model generated content however), which benefits from partial or lower resolution preview generation until dance partners each signal agreement of the outcome place in which to dance together.

3 User Experience Friction

Changing environment prompts naively resulting in instant teleportation to radically new spaces can be jarring to participants and could be a negative actor exploit to cycle spaces rapidly to cause nausea and discomfort, so we consider smoothly animated transitions to new spaces and limited change frequency to address this.

Often with current large language model understanding, image generation control and more generally the limitation of specificity of natural language, frustration is encountered where the realization of the generated space is mismatched with the user's mental vision, and ensuing with multiple iterations of prompt refinements until fatigued.

4 Social and Safety

When dancers use virtual environments to the advantage of their artistic vision, issues concerning cultural sensitivity and ethical implications arise, specifically regarding risks of appropriation and misrepresentation of cultural elements.

Distributing enjoyable dance experiences on wider social networks may attract negative reactions in the context of the current debate of use of AI art generation methods versus human artist empowerment. A proportion of users may prefer to generate environments only to share and dance with AI driven dance partner avatars, perhaps building confidence with the new experience before connecting with other human driven dance partner avatars. In our experiences with online dancing communities [5], whilst dance avatars are appreciated to be recognised by handles, their IRL [1] persona is often preferred to be kept anonymous.

With an unbounded scope for placing online participant in generated environments there is much risk of exposure to adversely conceive images. In one view, the textual prompts that define the generated environments' appearance can be moderated automatically with low computational overhead than image or video moderation algorithmic methods permitting somewhat more assurance of the HoloJig generative AI scheme. However, even with effective moderation processes, such systems can be circumvented, so a fallback *safehouse* environment must be accessible to users at any time.

5 Future More Principled Models

With *HoloJig* we provide a glimpse of a future where generative AI media surrounds your digital experience. However, such techniques built upon visual generative image synthesis are singularly limited to just that of just enough visual representation to satisfy the particle large language model's training criteria and consequently the visuals don't bear up to close inspection but instead show many artifacts of seams and out of place patterns. Further, there is no underlying understanding of the generated models to be informed of physical, aural or haptic properties, which could perform further authenticity and consistency of the resulting shared online spaces to inhabit. And so, this is also an area of interest of our further work.

Acknowledgments

This project has received funding from European Union's Horizon 2020 research and innovation program under grant agreement No. 101017779 for project CAROUSEL, dancing online with AI.

References

[1] Masshuda Glencross, Kenny Mitchell, Jassim Happa, Anthony Steed, Noelle Martin, and Moya Kate Baldry. 2022. X-IRL risks: Identifying privacy and security risks in inter-reality attacks and interactions. *IEEE VR Workshop* (2022). https://x-irl-risks.uqcloud.net/

[2] Dooley Murphy. 2023. While We Wait for the Holodeck; or, How Agency in VR Only Tells Half a Story. *The Velvet Light Trap* 91, 1 (2023), 65–70.

[3] Ye Pan and Kenny Mitchell. 2020. Group-Based Expert Walkthroughs: How Immersive Technologies Can Facilitate the Collaborative Authoring of Character Animation. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. 188–195. https://doi.org/10.1109/VRW50115.2020.00041

[4] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, Los Alamitos, CA, USA, 10674–10685. https://doi.org/10.1109/CVPR52688.2022.01042

[5] Ian Sachs, Vivek Virma, Sean Palmer, Anindita Ghosh, Markus Laattala, Mubasir Kapadia, Victor Zordan, Tom Sonacki, and Kenny Mitchell. 2023. Expressive Avatar Interactions in Online Co-Experiences. *ACM SIGGRAPH Frontiers Workshop* (2023).

[6] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. 2022. LAION-5B: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems* 35 (2022), 25278–25294. arXiv:2210.08402 [cs.CV]

[7] David Sinclair, Adeyemi Ademola, Floyd Chitalu, Babis Koniaris, and Kenny Mitchell. 2023. DanceGraph: A Complementary Architecture for Synchronous Dancing Online. In *Proceedings of the 36th International Computer Animation and Social Agents (CASA) 2023* (Limassol, Cyprus). https://api.semanticscholar.org/CorpusID:259100629

[8] StarTrek.com. 2014. Meet The Man Behind The Holodeck. https://www.startrek.com/news/meet-the-man-behind-the-holodeck-part-1. [Online; accessed 20-February-2024].

[9] Gaetano Valenza, Mariano Alcaniz, Antonio Luca Alfeo, Matteo Bianchi, Vladimir Carli, Vincenzo Catrambone, Mario CGA Cimino, Gabriela Dudnik, Andrea Duggento, Matteo Ferrante, Claudio Gentili, Jaime Guixeres, Simone Rossi, Nicola Toschi, and Virginie van Wassenhove. 2023. The EXPERIENCE Project: Unveiling Extended-Personal Reality Through Automated VR Environments and Explainable Artificial Intelligence. In *2023 IEEE International Conference on Metrology for eXtended Reality, Artificial Intelligence and Neural Engineering (MetroXRAINE)*. 757–762. https://doi.org/10.1109/MetroXRAINE58569.2023.10405613

[10] Yue Yang, Fan-Yun Sun, Luca Weihs, Eli VanderBilt, Alvaro Herrasti, Winson Han, Jiajun Wu, Nick Haber, Ranjay Krishna, Lingjie Liu, Chris Callison-Burch, Mark Yatskar, Aniruddha Kembhavi, and Christopher Clark. 2023. Holodeck: Language Guided Generation of 3D Embodied AI Environments. *The IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023). arXiv:2312.09067 [cs.CV]