Special Issue on Generative AI for Computer Graphics

HoloJig: Interactive Spoken Prompt Specified Generative AI Environments

Llogari Casas, *3FINERY, Edinburgh Napier University,United Kingdom*

Samantha Hannah, *3FINERY, Edinburgh Napier University,United Kingdom*

Kenny Mitchell, *3FINERY, Edinburgh Napier University, United Kingdom*

*Abstract—**HoloJig** offers an interactive, speech-to-VR, virtual reality experience that generates diverse environments in real-time based on live spoken descriptions. Unlike traditional VR systems that rely on pre-built assets, HoloJig dynamically creates personalized and immersive virtual spaces with depth-based parallax 3D rendering, allowing users to define the characteristics of their immersive environment through verbal prompts. This generative approach opens up new possibilities for interactive experiences, including simulations, training, collaborative workspaces, and entertainment. In addition to **speech-to-VR** environment generation, a key innovation of HoloJig is its progressive visual transition mechanism, which smoothly dissolves between previously generated and newly requested environments, mitigating the delay caused by neural computations. This feature ensures a seamless and continuous user experience, even as new scenes are being rendered on remote servers.*

HoloJig brings a novel interactive *speech-to-VR* approach to virtual remote spaces defined by spoken prompts to generate customized virtual environments in real-time. Traditional Virtual Reality (VR) systems often rely on pre-built assets and static environments, limiting the flexibility and creativity available to users. HoloJig allows users to verbally describe their desired space, which is then realized through generative AI technologies, providing a personalized virtual environment to their vocalized preference. With live *speech-to-text* processing, HoloJig translates verbal input into parameters that define the characteristics of the virtual environment. These environments are rendered with a depth-based parallax 3D shader projection method in high resolution using image-based generative AI models, ensuring that users can collaborate in settings that are as diverse and unique as their imaginations. The system importantly incorporates a visual transition mechanism that progressively dissolves between virtual scenes, allowing for seamless environment changes without disrupting the continuity of the experience.

Introduction

Much explored in the narrative fantasies of *Star Trek*, Dolgoff and Roddenberry's *Holodeck* [6] promises to transport us to any place of our choosing without physically moving, just by telling vocally where we want to be. This vision represents the ultimate flexibility and immersion, where users can describe their desired environment and have it rendered interactively, enabling limitless possibilities for exploration, collaboration, and entertainment. In this context, we introduce HoloJig, an innovative VR platform that bridges the gap between traditional static VR experiences and the flexible, on-demand environments of a holodeck. Unlike conventional VR systems, HoloJig allows users to generate custom virtual spaces based on speech-to-text spoken prompts, enabling the creation of highly interactive and personalized environments generated via diffusion models that render dynamic, high-resolution virtual landscapes in real time (fig. 1).

A key feature of HoloJig is its ability to handle the delay caused by the computational process of environment generation. Instead of requiring users to wait for the full depth-based parallax stereo-3D VR rendering, the system provides an intermediate low-resolution preview that gradually transitions into the

Digital Object Identifier 10.1109/MCG.2025.3553780

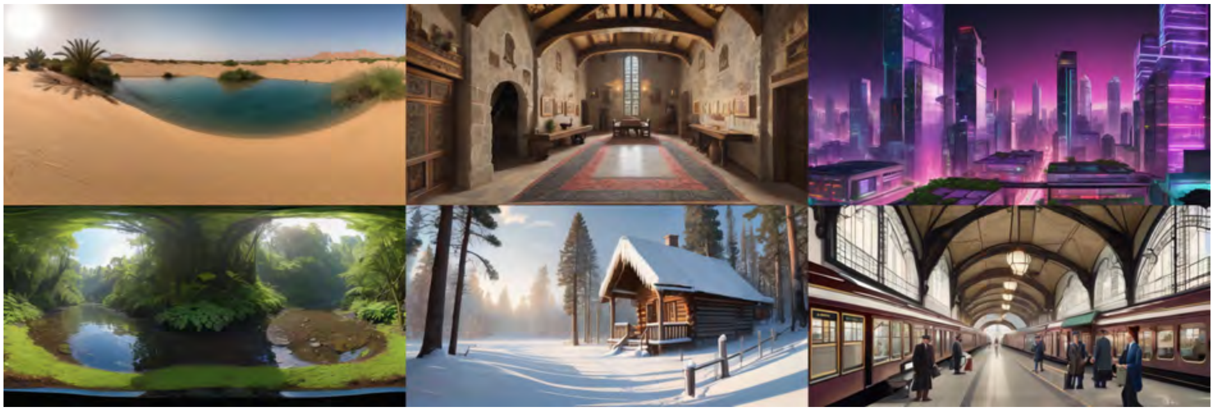

FIGURE 1. Illustration of the wide range of worlds available on-demand to any novice user, showcasing the diverse and dynamic environments that *HoloJig* can create for on-demand in VR for remote collaborative experiences.

final, high-quality environment. This progressive dissolve mechanism ensures that users remain engaged and immersed, without the disruption caused by long load times or abrupt scene changes.

HoloJig opens up new possibilities across various domains, including interactive training simulations, virtual collaborative workspaces, entertainment and educational environments. Users can verbally describe the environment they need, whether it's a simulated urban landscape for training purposes, a calming natural scene for relaxation, or a futuristic setting for creative exploration, and have it generated on demand. This flexibility not only enhances the user experience but also transforms how we think about virtual spaces and their role in our daily lives. By allowing users to create and interact with environments in real time, HoloJig brings us closer to the vision of a true *Holodeck*, where virtual worlds can be as diverse and dynamic as human imagination. In this regard, *HoloJig* introduces:

- An online interactive *speech-to-VR* method for gathering spoken virtual world descriptions into textual formatted prompts for live relay to generative AI remote server processes and subsequent realization in VR of the received generated environment data.
- The delay of computation and retrieval of each generative AI environment is performed with a visual transition between current and next virtual environments, shown continuously as a visually fading dissolve of the first scene into a progressive refinement of the new scene, through early low resolution initial environment image generation until the new scene has been fully composed and received.
- The delivered VR data is formed as a color image with inferred depth formed in either parallax depth relief rendered planes or fully immersive depth panoramas from generation *LatLong* maps. The corresponding generated depth imagery permits sufficient movement degrees of freedom through re-projected rendering to be suitable for a shared remote environment in VR.

Dictated VR Scene Generation

HoloJig presents the practical realization of language guided environment generation verbally [5] and vocally [7], for the purpose of responsively providing a place for remote collaborative experience in VR using our proposed method in real-time. Through the VR headset's microphone, users can describe their desired virtual environment. These voice inputs are captured and sent to the speech-to-text module, which utilizes *Google Cloud's Chirp* [1] speech-to-text conformer model (figure 3). This module converts the spoken language into text by streaming the audio to the cloud service, where it is processed using natural language processing models to achieve accurate transcription. Similarly to generative augmented mirror embodiment [14], the resulting text is then parsed to extract relevant parameters that will define the virtual environment.

The environment generation module utilizes a stable diffusion model [3], accessed through the Stable Diffusion Web UI API [12], to translate the spoken prompts into high-resolution 3D landscapes. In order to access a widest available variety of generated imagery we selected the *DreamShaper XL* model, which excels at generating detailed and diverse environments. The system retrieves intermediate images at each itera-

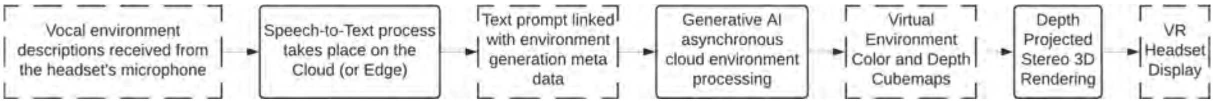
Vocal environment descriptions received from the headset's microphone
Speech-to-Text process takes place on the Cloud (or Edge)
Text prompt linked with environment generation meta data
Generative AI asynchronous cloud environment processing
Virtual Environment Color and Depth Cubemaps
Depth Projected Stereo 3D Rendering
VR Headset Display

FIGURE 2. Live processing stages of the *HoloJig* platform.

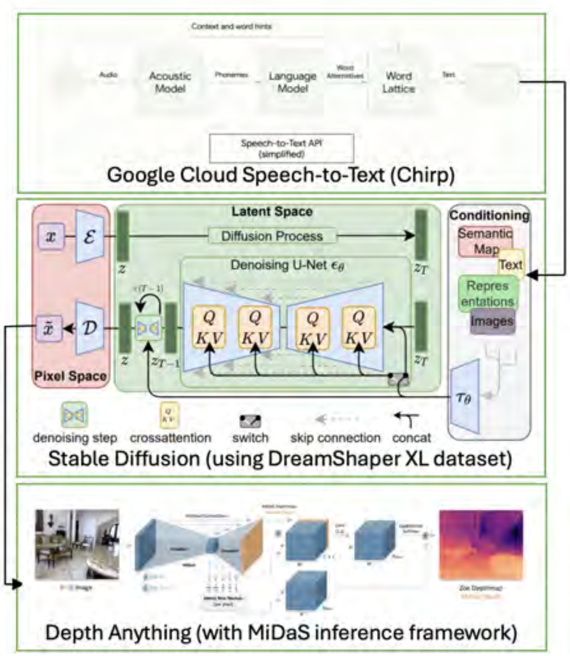

FIGURE 3. Integration of end-to-end speech prompt to immersive 3D environment in real-time, composing *Chirp*, *Stable Diffusion* and *Depth Anything* for speech to text, text to image and image to depth respectively.

tion during the rendering process, with a total of 20 iterations per generation. This is managed through the intermediate image saving extension [13], allowing continuous real-time feedback on the generation process as each iteration progresses. For that, textual prompts are sent to the model through the API, defining the parameters of the virtual environment to be generated. These prompts capture user inputs such as the desired scene, lighting conditions and atmospheric elements. After receiving a prompt, the system begins generating the environment by progressively refining the image over the 20 iterations. At each single interval, an intermediate image is retrieved and displayed to the user via cross fading with the prior staged image, offering a low-resolution preview of the environment. This approach ensures a continuous presence and the ability to begin experiencing the new virtual environment early in the process while the final high-resolution version is being generated.

To ensure a smooth visual transition between environments, *HoloJig* employs a progressive dissolve transition as described by Pointecker et al [4], but initialized almost instantaneously with a low detail version of the generated environment for an in progress visual transition before the final version. As the new environment is generated, the system gradually blends with a cross fade of the low detail generated environment image frame with the previously existing one, until the fully detailed environment generation is completed and received. This process minimizes any jarring visual disruptions and is carefully managed to maintain immersion and continuity, enhancing the overall aesthetic and experiential quality of the virtual space [2].

TABLE 1. Stage timings in environment generation

Stage of Process	Time (Seconds)
Speech-to-Text Conversion	2.5
AI Environment Generation	20.0
Initial Low-Resolution Rendering	2.0
Full-Resolution Scene Rendering	0.5
Total Time	**25.0**

Once the final high-resolution image is generated after 20 iterations, *HoloJig* leverages the *DepthAnything* model [11] to generate its corresponding depth map for the virtual environment (figure 3). This is crucial for maintaining immersion, as it provides the basis for realistic parallax effects and spatial cues as users move through the virtual world. The depth map is then integrated into the virtual reality system, allowing for realistic interactions with the environment, including accurate lighting, shadow casting and object placement relative to the user's position.

As measured in Table 1, the entire process of generating a virtual environment in *HoloJig*, from sending the initial request to receiving the completed scene, typically takes around 25 seconds. As depicted in figure 2, it begins with the user describing their desired environment through spoken prompts captured via the VR headset's microphone. This spoken input is swiftly transcribed into text using a speech-to-text conversion module powered by Google Cloud's API. This transcription process incurs low delay, typically taking around 2.5 seconds, depending on the complexity of the spoken prompt. Once the textual description is

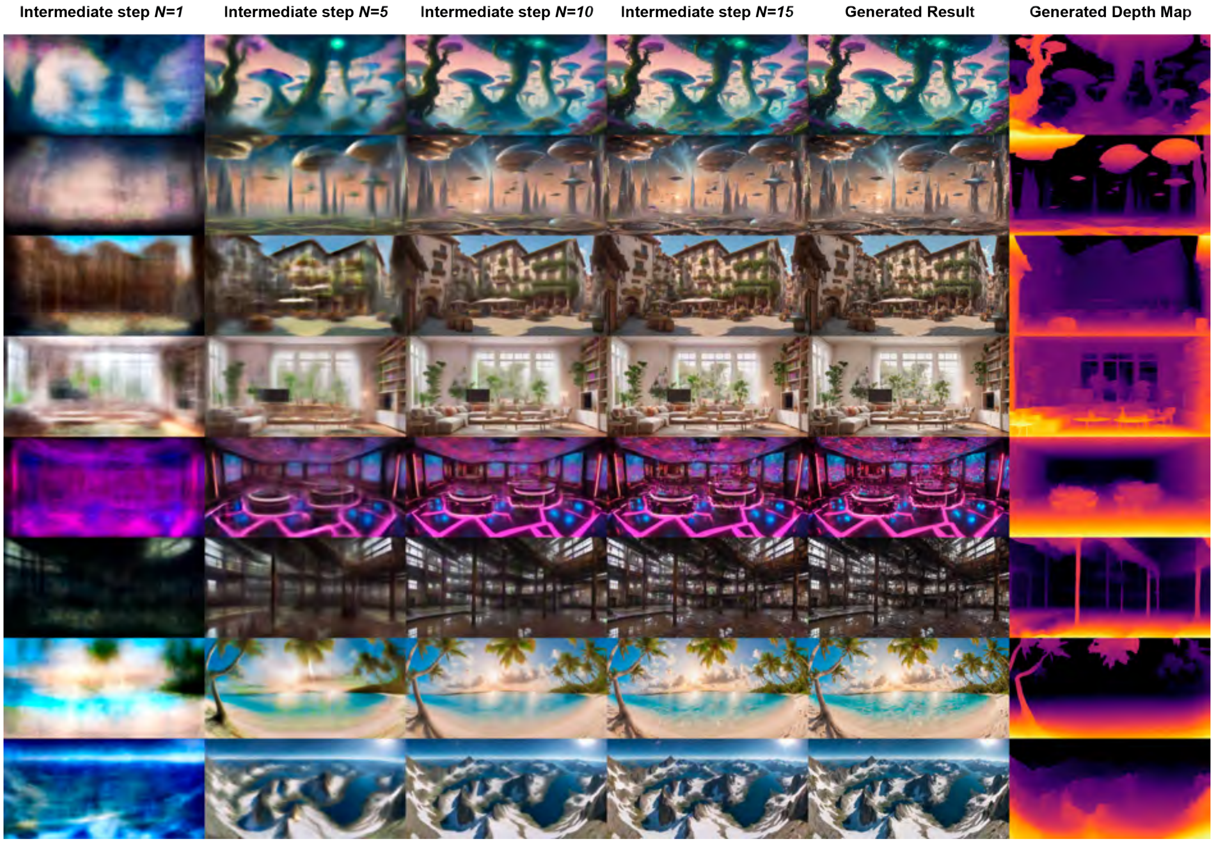

FIGURE 4. Illustration of the process for generating environments with real-time access to live previews of the images. Intermediate samples are displayed at computation steps 1, 5, 10, 15, and the final output. A corresponding depth map is generated using *DepthAnything* [11], providing additional spatial information of depth-based parallax 3D rendering in VR.

available, it is sent to the generative AI model. This image generation process is the most time-intensive stage, consuming around 20 seconds to generate the high-resolution environment. To ensure uninterrupted user immersion, an initial low-resolution version of the environment is displayed within approximately 2 seconds of generation. This low-resolution preview helps maintain engagement until the final high-resolution version is ready with subsequent progressive refinement steps displayed smoothly with cross fading over the duration of approximately 500 milliseconds for each refinement stage. The final step, which involves rendering the fully detailed environment, is nearly instantaneous, taking less than a second to complete. The entire process ensures a seamless transition from input to output, with minimal disruption to the immersive experience.

The generated environments can be experienced in two distinct ways, offering users the choice between a more observational (observation deck-like) or fully immersive (360 degree FOV) engagement. They can interact with the environment as a window to the world, where they remain in a virtual space while observing another environment through a portal-like interface, maintaining a sense of separation while still exploring new settings. Alternatively, they can opt for a fully immersive experience, where the entire virtual environment dynamically adapts to the generated scene, making them feel completely transported to a new world with a heightened sense of presence. In practical terms, a portal view approach affords access of a wider variety of generative AI image models and some performance improvements. However, this flexibility to choose either mode of immersive VR ensures that users can tailor their level of immersion to best suit their needs, preferences and specific use cases.

Background music is curated based on the prompt, ensuring that the audio complements the generated environment in terms of mood, atmosphere and thematic coherence. In our experiment we prepared con-

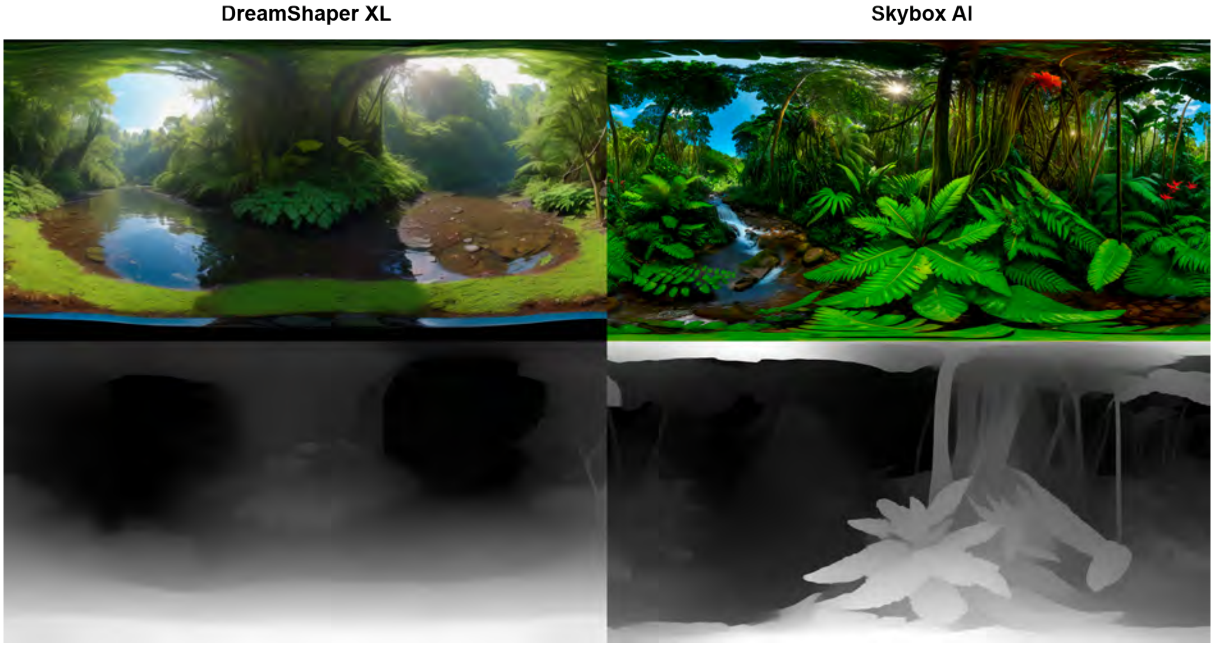

FIGURE 5. Comparison of generated *LatLong* map with corresponding depth image using *DreamShaper XL* and *Skybox AI* for the prompt: *"A lush jungle clearing surrounded by dense tropical vegetation, with towering trees, thick vines and ferns creating a natural barrier"*. The generated images depicted here are 360-degree projections, which can be represented as a full skybox, enhancing their usability in immersive VR applications.

textual music to preselected prompt themes. We expect an automation of this curation can be applied for example with a *text-to-music* retrieval [19] or generation approach according to system performance needs. In the later music generation case, a cross fading of part-generated musical bars may provide an continuous audio musical environment without disjoint interruptions in sound. Overall,this approach enhances immersion by reinforcing the emotional tone of the scene and guiding user perception, creating a more cohesive audiovisual experience. The integration of curated music with visual elements plays a crucial role in establishing a sense of presence, making the virtual environment more engaging and dynamic.

Progressive Refinement Evaluation

The environment generation process combines real-time generative AI with progressive rendering to continuously deliver highly immersive virtual spaces. To ensure minimal disruption and maintain immersion, we retrieve and display intermediate samples at key computation steps during the generation process. Figure 4 depicts steps 1, 5, 10 and 15 of each one leading up to the final output at step 20.

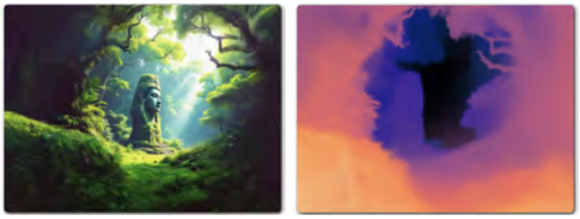

FIGURE 6. For visual comparison only: World Labs AI jungle scene given only color and depth final frame outputs [18]. We nonetheless can see similar quality of color scene richness and depth accuracy is provided through World Labs AI, as with our use of *DreamShaper XL* and *Skybox AI* models.

To evaluate the versatility of the system, we conducted a comparison using two different models, *DreamShaper XL*[17] and *Skybox AI*[16] as depicted in figure 5. Both models generated 360-degrees *LatLong* maps and corresponding depth images based on this detailed scene description. Upon visual inspection, the results produced appeared largely equivalent in terms of the quality of visual details, with each model accurately capturing the scene's lush greenery, natural light interplay and the subtle variations in color. A

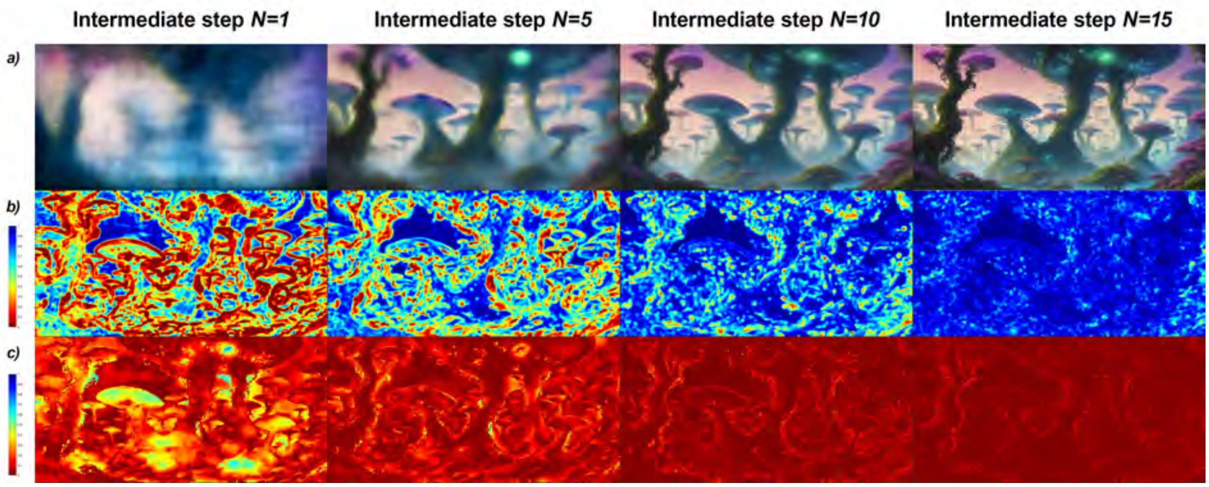

FIGURE 7. SSIM and PSNR visualizations across intermediate steps. (a) Generated images refine progressively with increasing *N*. (b) SSIM heatmaps show increasing structural similarity towards the final state. (c) PSNR heatmaps indicate decreasing noise error to the resulting signal, with an increased rate of improvements occurring in the early steps.

further visual only comparison can be found in figure 6. Despite their similarities, using multiple models provides a distinct advantage in terms of flexibility and adaptability as depending on the specific needs of the virtual scene, different models can be chosen to best suit the application. Moreover, the depth images generated by both models were comparable in their ability to represent spatial relationships within the scene. This consistency allows for a seamless and immersive environment generation regardless of the chosen generative model.

The progressive refinement of generated images over increasing intermediate steps (N) demonstrates a clear improvement in perceptual quality, as evidenced by both visual analysis in figure 7 and quantitative metrics in figure 8. At $N = 1$, the generated image is highly blurred, lacking recognizable structures. By $N = 5$, silhouettes of the target structures emerge, though fine details remain ambiguous. At $N = 10$, the image exhibits well-defined shapes, and by $N = 15$, it achieves a high level of detail with enhanced texture and shading, closely resembling the final reference image. This improvement is naturally quantitatively supported by the sequences of SSIM measures, which increase steadily from approximately 0.4 at $N = 1$ to near-optimal values beyond $N = 15$, confirming a growing structural similarity to the reference. The PSNR also exhibits a consistent upward trend, surpassing 30 dB at later stages, suggesting a significant reduction in perceptual distortion. Meanwhile, the MSE decreases rapidly within the first 5–10 steps, indicating that most structural errors are corrected early, with diminishing returns in later iterations. The combined interpretation of SSIM, PSNR, and MSE suggests that the most substantial perceptual changes occur within the first 10 steps, after which refinements primarily enhance texture and fine details. Beyond $N = 15$, the gains in visual quality become marginal, as structural fidelity has nearly converged with the reference image. This indicates that while early iterations establish the fundamental composition, later steps serve to enhance realism and perceptual acuity.

Further from a perceptual standpoint, the transition from an abstract, unstructured form to a recognizable image follows a nonlinear trajectory. In the early stages, perception is dominated by broad color and shape distributions, which may evoke abstract impressions rather than concrete recognition. As N increases, identifiable structures begin to emerge, leading to a shift in perception from ambiguity to clarity. This process aligns with human visual cognition, where global features are processed first, followed by finer details as more information becomes available. Although current processing occurs at a slower pace than human visual perception rates, which expect this approach holds as system performance improves in the future. Additionally, the perception of transition is influenced by the balance between structure formation and texture refinement. While initial steps create a perceptible scene, later steps provide depth, sharpness, and realism. The subjective experience of this transition aligns with the observed SSIM and PSNR trends, where significant

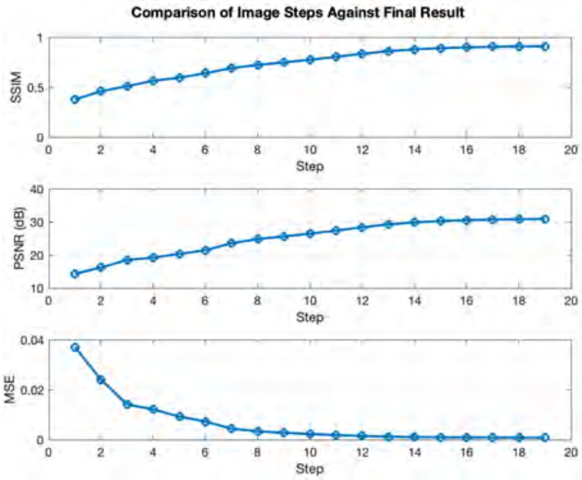

FIGURE 8. Quantitative evaluation of progressive image refinement over sequential steps. The SSIM (top) shows a steady increase, indicating smoothly improving structural similarity progressing towards the reference transitioned result image. The PSNR (middle) measure also rises consistently, suggesting reduced perceptual distortion over time, whilst the MSE (bottom) decreases sharply in early steps, signifying rapid error correction in the early steps of the transition. Overall, preferring perceptual continuity of the transition over the MSE numeric error distance.

perceptual shifts occur early, followed by refinements that contribute to an enhanced sense of realism.

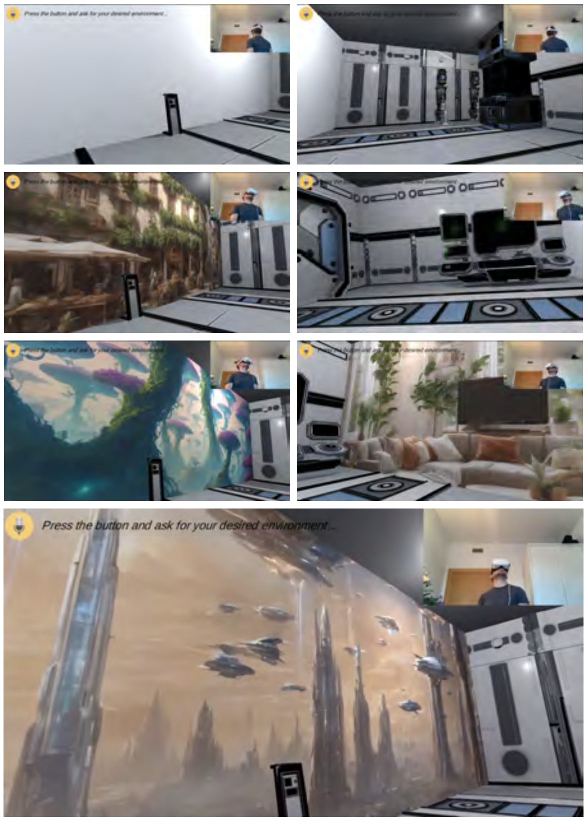

FIGURE 9. Immersive HoloJig application presented as a SciFi room with one wall replaced with the requested depth parallax generated virtual environment.

Discussion and Conclusion

We have presented an interactive voice commanded VR environment generation method. The progressive rendering mechanism in HoloJig is designed to address the latency inherent in generating high-resolution virtual environments. By leveraging an iterative refinement approach, the system continuously updates the scene with increasingly detailed representations, ensuring that users remain engaged throughout the transition. This method not only improves continuity of immersion but also mitigates the discomfort often associated with abrupt scene changes in VR. A more detailed evaluation of these benefits can be developed through undertaking a formal perceptual user-study in future work.

Beyond technical performance, our preliminary user feedback has highlighted the importance of maintaining a coherent sense of presence during environmental transitions. Informal testing with early adopters revealed that users favored the progressive rendering method over abrupt environment swaps, citing reduced disorientation and a more fluid experience. Some users noted that while the initial low-resolution preview was beneficial in maintaining continuity, they desired greater control over the transition speed, suggesting a potential adaptive approach that adjusts the refinement rate based on user preferences or context.

As depicted in figure 9, the ability to generate custom environments dynamically opens new possibilities for interactive virtual experiences. One key application of our platform is in the performing arts, particularly in dance. Dancers and choreographers can use the platform to create virtual spaces that match the mood, theme, and artistic intent of their performances. By simply describing the desired setting, *HoloJig* can instantly generate a tailored environment, allowing dancers to rehearse and perform in spaces that enhance their creative expression. This also provides the opportunity for remote collaboration, where dancers in different locations can share the same virtual stage, co-creating environments that respond dynamically to their movements and choreography. Beyond rehearsals and performances, *HoloJig* also offers significant benefits for dance education and training. Instructors can use the platform to immerse students in a variety of practice settings, such as a virtual theater or a public square, helping them adapt to different performance environments. Real-time feedback and motion capture integration ensure that both students and instructors can focus on refining technique while interacting with highly customizable, engaging virtual spaces. This flexibility enables a deeper connection between movement and environment, opening new avenues for creative learning and artistic exploration. In addition, the system's integration with real-time motion capture frameworks, such as *DanceGraph* [10], would ensure high-fidelity performance and synchronization, creating a responsive and lifelike experience that adapts seamlessly to user input and interaction.

Our system is not without limitations. With a depth based image parallax approach in the portal viewport mode of VR rendering (see figure 9) larger offsets of view points result in stretching texels that would otherwise reveal multi-layered depth image pixels, and as such requires an inpainting approach for these regions or otherwise limiting the strength of the parallax effect. In the 360 depth panorama mode of VR these artifacts are somewhat more apparent. A further improvement may be developed to reflect the generated environment image-based lighting emission from the viewed portal scene into the holodeck observation space.

In a prior series of works on immersive light field video rendering for VR, such as *IRIDiuM* [8], [9], 360 degree depth panoramas have been employed similarly to provide high quality stereo 3D view point rendering with depth based projection of color cubemap texel elements. HoloJig can be seen as a first-step of integration of those works with generative AI image based methods. In addition, the popular trend of *Guassian Splatting* and similar visual primitive representations optimized for VR [15] can be applied to *text-to-3D* [20] and subsequently have promise as an avenue for more fine grained spoken control of generative AI 3D environments in future research.

Although the current system provides smooth transitions between scenes, ongoing research into more efficient generative models could further reduce latency and improve real-time responsiveness. The platform's ability to generate interactive virtual spaces based on natural language input can be extended to domains such as education, healthcare, and remote collaboration. For instance, in therapeutic settings, real-time environmental customization could be used to support mental health interventions, offering calming or stimulating virtual environments tailored to user needs. In education, virtual classrooms or simulations could be dynamically adjusted based on spoken inputs, enhancing learning experiences [14].

Acknowledgments

This project has received funding from European Union's Horizon 2020 research and innovation programme under grant agreement number 101017779.

References

1. Gulati, A., Qin, J., Chiu, C.-C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y., Pang, R (2020). Conformer: Convolution-augmented transformer for speech recognition. Proceedings of Interspeech 2020. https://doi.org/10.21437/Interspeech.2020-3015
2. Feld, N., Bimberg, P., Weyers, B., & Zielasko, D. (2023, April). Keep it simple? evaluation of transitions in virtual reality. In Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems (pp. 1-7). https://doi.org/10.1145/3544549.3585811
3. Po, R., Yifan, W., Golyanik, V., Aberman, K., Barron, J. T., Bermano, A., & Wetzstein, G. (2024, May). State of the art on diffusion models for visual computing. In Computer Graphics Forum (Vol. 43, No. 2, p. e15063). https://doi.org/10.1111/cgf.15063
4. Pointecker, F., Friedl-Knirsch, J., Jetter, H. C., & Anthes, C. (2024, May). From real to virtual: Exploring replica-enhanced environment transitions along the reality-virtuality continuum. In Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (pp. 1-13). https://doi.org/10.1145/3613904.3642844
5. Smith, J., Anaraki, N. A. T., Goloujeh, A. M., Khosla, K., & Magerko, B. (2021). Towards an AI Holodeck: Generating Virtual Scenes from Sparse Natural Language Input. In the Joint Proceedings of the AIIDE 2021 Workshops co-located with 17th AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment. https://ceur-ws.org/Vol-3217/paper4.pdf
6. StarTrek.com Staff (2014). Meet The Man Behind The Holodeck. https://www.startrek.com/news/meet-the-man-behind-the-holodeck-part-1
7. Yang, Y., Sun, F. Y., Weihs, L., VanderBilt, E., Herrasti, A., Han, W., & Clark, C. (2024). Holodeck: Language guided generation of 3d embodied ai environments. In Proceedings of the IEEE/CVF Conference on Com-

puter Vision and Pattern Recognition (pp. 16227-16237). https://doi.org/10.48550/arXiv.2312.09067

8. Koniaris, B., Huerta, I., Kosek, M., Darragh, K., Malleson, C., Jamrozy, J., & Mitchell, K. (2016). IRIDiuM: Immersive rendered interactive deep media. In ACM SIGGRAPH 2016 VR Village (pp. 1-2). https://doi.org/10.1145/2929490.2929496
9. Koniaris, C., Kosek, M., Sinclair, D., & Mitchell, K. (2018). Compressed animated light fields with real-time view-dependent reconstruction. IEEE Transactions on Visualization and Computer Graphics, 25(4), 1666-1680. https://doi.ieeecomputersociety.org/10.1109/TVCG.2018.2818156
10. Sinclair, D., Ademola, A. V., Koniaris, B., & Mitchell, K. (2023, May). DanceGraph: A complementary architecture for synchronous dancing online. In 36th International Computer Animation Social Agents (CASA) 2023. https://api.semanticscholar.org/CorpusID:259100629
11. Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., & Zhao, H. (2024). Depth Anything: Unleashing the power of large-scale unlabeled data. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 10371-10381). https://doi.org/10.48550/arXiv.2401.10891
12. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 10684-10695). https://doi.ieeecomputersociety.org/10.1109/CVPR52688.2022.01042
13. *AIUlkesh*, *heloss*, Colburn, J. (2024). Stable Diffusion Intermediate Images Exporter. https://github.com/AIUlkesh/sd_save_intermediate_images
14. Casas, L., & Mitchell, K. (2024, December). Structured Teaching Prompt Articulation for Generative-AI Role Embodiment with Augmented Mirror Video Displays. In The 19th ACM SIGGRAPH International Conference on Virtual-Reality Continuum and its Applications in Industry (pp. 1-7). https://doi.org/10.1145/3703619.3706049
15. Murray, A., Mitchell, S., Bradley, A., Waite, E., Ross, C., Jamrozy, J., & Mitchell, K. (2022). Generating real-time detailed ground visualisations from sparse aerial point clouds. ACM SIGGRAPH European Conference on Visual Media Production. http://researchrepository.napier.ac.uk/Output/2950589
16. Chivers, K., Takahashi, D. (2024) Blockade Labs improves quality for ai generated 3d art for 360 degree apps. Skybox AI. https://venturebeat.com/ai/blockade-labs-improves-quality-for-ai-generated-3d-art-for-360-degree-apps/
17. *Lykon* (2024). DreamShaper XL: Model Report. Open Laboratory Model Repository for Stable Diffusion https://openlaboratory.ai/models/dreamshaper-xl
18. Li, F.-F., Wiggers, K (2024, December) World Labs AI can generate interactive 3D scenes from a single photo, TechCrunch. https://techcrunch.com/2024/12/02/world-labs-ai-can-generate-interactive-3d-scenes-from-a-single-photo/. https://www.worldlabs.ai/blog.
19. Doh, S., Won, M., Choi, K., Nam, J. 2023 Toward Universal Text-To-Music Retrieval, In proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). https://github.com/seungheondoh/music-text-representation
20. Chen, Z., Wang, F., Wang, Y., & Liu, H. (2024). Text-to-3d using gaussian splatting. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 21401-21412). https://doi.ieeecomputersociety.org/10.1109/CVPR52733.2024.02022

Biographies

Llogari Casas is a scientist and entrepreneur with a Ph.D. in Augmented Reality, specializing in Mixed Reality and Artificial Intelligence. His work focuses on developing innovative AR/AI technologies that bridge virtual and real-world experiences. With several patents and numerous publications in leading venues, Llogari has contributed to advancements in real-time graphics and immersive user interaction.

Samantha Hannah is a technical artist and animator with several years of professional experience 3D development, including AR and VR. Her expertise includes 3D modeling and animation, as well as creating textures and materials for real-time applications.

Kenny Mitchell is a professor of video games technology with Edinburgh Napier University. He is an outstanding scientist and practitioner in the creative industries with a career spanning 4 decades of games, graphics, robotics and computer vision. He is co-founder of 3FINERY LTD as part-time CTO delivering a new way to help people be together and play through mixed reality experiences.