From Prompt to Metaverse: User Perceptions of Personalized Spaces Crafted by Generative AI

Simin Yang
Hong Kong University of Science and Technology
Hong Kong, China
syangcj@connect.ust.hk

Yuk Hang Tsui
Hong Kong University of Science and Technology
Hong Kong, China
yhtsui@connect.ust.hk

Xian Wang
Hong Kong Polytechnic University
Hong Kong, China
xiann.wang@connect.polyu.hk

Ahmad Alhilal
Hong Kong University of Science and Technology
Hong Kong, China
aalhilal@connect.ust.hk

Reza Hadi Mogavi
University of Waterloo
Waterloo, Canada
rhadimog@uwaterloo.ca

Xuetong Wang
Hong Kong University of Science and Technology
Hong Kong, China
xwangdd@connect.ust.hk

Pan Hui*
Hong Kong University of Science and Technology (Guangzhou)
Guangzhou, China
panhui@ust.hk

ABSTRACT

Generative artificial intelligence (AI) has revolutionized content creation. In parallel, the Metaverse has emerged to transcend the constraints of our physical reality. While Generative AI has a multitude of exciting applications for the fields of writing, coding, and graphic design, its usage to personalize our virtual space has not yet been explored. In this paper, we investigate the application of Artificial Intelligence Generated Content (AIGC) to personalize our virtual spaces and enhance the metaverse experience. To this end, we present a pipeline to enable users to customize their virtual spaces. Moreover, we explore the hardware resources and latency required for personalized spaces, as well as user acceptance of the AI-generated spaces. Comprehensive user studies follow extensive system experiments. Our research evaluates users' perceptions of two generated spaces: *panoramic images* and *3D virtual spaces*. According to our findings, users have shown a great interest in 3D personalized spaces, and the practicality and immersion of 3D space generation tools surpass panoramic space generation tools.

CCS CONCEPTS

• **Human-centered computing** → **Virtual reality**; *Natural language interfaces*; **Text input**; **User interface programming**.

KEYWORDS

AI-Generated Content; Generative Artificial Intelligence; Metaverse; Virtual Reality; HCI; Virtual Spaces; Personalization

*Pan Hui is also affiliated with Hong Kong University of Science and Technology, and University of Helsinki, Finland

CSCW Companion '24, November 9–13, 2024, San Jose, Costa Rica

ACM ISBN 979-8-4007-1114-5/24/11
https://doi.org/10.1145/3678884.3681897

ACM Reference Format:
Simin Yang, Yuk Hang Tsui, Xian Wang, Ahmad Alhilal, Reza Hadi Mogavi, Xuetong Wang, and Pan Hui. 2024. From Prompt to Metaverse: User Perceptions of Personalized Spaces Crafted by Generative AI. In *Companion of the 2024 Computer-Supported Cooperative Work and Social Computing (CSCW Companion '24), November 9–13, 2024, San Jose, Costa Rica.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3678884.3681897

1 INTRODUCTION

The advent of the *Metaverse* has given rise to a realm that intricately blends physical and digital spaces, transcending the limitations of the physical world through mixed reality [10, 37]. This convergence is made possible by an array of technologies, including augmented reality (AR), virtual reality (VR), and computer vision [35]. Recently, *Generative AI* has emerged to create virtual content [1, 11]. It uses algorithms that generate authentic textures, shapes, and movements based on real-world data and user feedback [5, 30]. Consequently, Generative AI can generate lifelike avatars and virtual objects that interact seamlessly with users [5–7, 17, 32]. Hence, Generative AI has the potential to revolutionize the metaverse by facilitating personalized and engaging virtual experiences [27]. For example, Text2Light [6] generates panoramic images from text input, simulating an immersive 3D space experience on XR devices, while Text2Room [17] creates 3D environments via text input, enabling users to personalize their space within XR devices. Unlike the traditional methods of building metaverse spaces, which are labor-intensive [12], Generative AI streamlines the creation process, significantly facilitating its development [27]. However, there remains considerable room for improvement in algorithms for generating 3D objects and scenes.

The Metaverse was initially conceived as a conceptualized utopia for individuals, where the personalization of virtual spaces is a particularly important aspect [23]. Driven by advancements in generative AI, which offer the potential to accelerate the development of the Metaverse and mitigate excessive resource consumption, we explore the use of Artificial Intelligence Generative Content (AIGC) tools to create personalized spaces within the Metaverse.

In this study, our objective is to answer the following research questions: **RQ1)** How close is AIGC from the realization of the

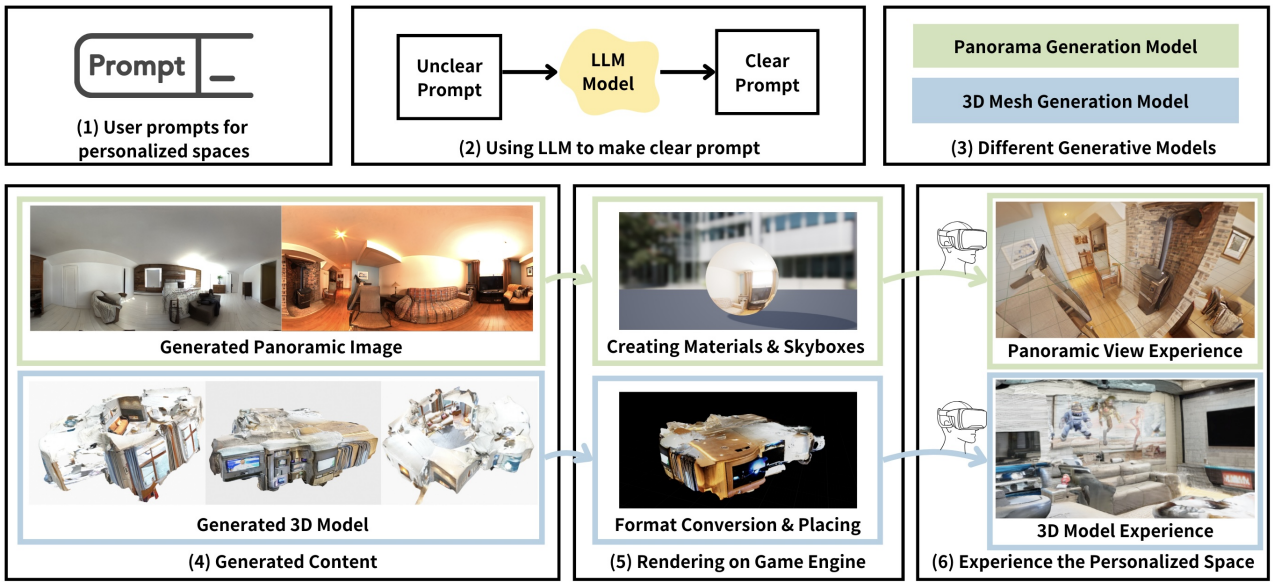


Figure 1: AI-generated personalized virtual space. The whole experience of the users is as follow: (1) Users input a prompt; (2) an LLM refines this prompt; (3) two AI tools then generate panoramic images and 3D models; (4) the content is obtained; (5) Unreal Engine processes this content, converting panoramic images into skyboxes and formatting and positioning 3D models; (6) users experience these personalized virtual spaces in VR.

dynamic generation of personalized spaces for individuals?, and **RQ2)** What is the user perception of AIGC's customizability and capability to provide immersive and personalized experience? Therefore, we design a pipeline as a proof of concept, as illustrated in Figure 1, which applies generative AI to facilitate the design of semi-personalized virtual spaces. In particular, we use AIGC tools and Unreal Engine to create tailored 3D models, taking inspiration from commonly encountered rooms in everyday life, such as bedrooms, meditation rooms, and offices. Afterward, we identify key technical factors, particularly computational power and response time, influencing user satisfaction within the metaverse. Finally, we investigate the impact of Generative AI and various visual representations, such as panoramas and 3D models on user engagement and the overall metaverse experience. Our main contributions are: (1) We develop a pipeline for creating personalized virtual spaces using Generative AI; (2) In our preliminary study, we thoroughly researched the required computing resources and settings for generating semi-customized metaverse spaces; (3) We assess the effectiveness of Generative AI and diverse visual representations in enhancing user engagement and immersion.

We hope that our findings can assist HCI researchers and practitioners in exploring the potential of generative AI to create personalized virtual spaces and enhance user experience in the metaverse.

2 BACKGROUND AND RELATED WORKS

AI-generated 3D content has grown increasingly popular in recent years due to advancements in machine learning and computer graphics. Based on the type of input and guidance, AI-generated 3D content can be categorized into three primary directions: text-to-3D, image-to-3D, and 3D-to-3D generation [38].

Text-to-3D generation involves the creation of 3D models or scenes based on natural language descriptions. Early work in this area includes ShapeNet [4], a large-scale repository of 3D models, which facilitated the development of deep learning-based methods for 3D content generation. Recent approaches have integrated transformer models, such as GPT [28] and BERT [9], to better understand natural language inputs and generate more accurate 3D representations. Notable examples include Text2Shape [5], which combines a GAN-based architecture with transformers to generate 3D shapes, and DreamFusion [26], which proposes solving these problems with a pre-trained text-to-2D model.

Image-to-3D generation refers to the creation of 3D content from 2D images or image sequences. Pioneering work includes the 3D ShapeNets [33] and 3D-R2N2 [8], which applied convolutional neural networks (CNNs) to single and multiple views of objects, respectively, for 3D reconstruction. More recent approaches have employed generative adversarial networks (GANs), such as Pix2Vox [34], and volumetric representations like Occupancy Networks [22] to improve the quality and accuracy of generated 3D models. Additionally, Image2Mesh [25] introduces a method to generate 3D meshes from 2D images using a novel deep learning architecture.

3D-to-3D generation involves the transformation, completion, or manipulation of existing 3D models. Early approaches, such as AtlasNet [15], focused on reconstructing 3D shapes from incomplete data using deep learning. More recent methods have expanded upon this foundation to include unsupervised and supervised techniques for various 3D generation tasks. Examples include PointFlow [36], which uses a normalizing flow-based architecture for 3D point

cloud generation, and FlowNet3D [20], which addresses the task of 3D scene flow estimation.

The field of AI-generated 3D content has witnessed significant advancements in recent years, with various deep learning techniques applied to text-to-3D, image-to-3D, and 3D-to-3D generation tasks. This study focuses on text-to-3D generation and delves into three scenarios pertaining to virtual reality work, with particular emphasis on three commonly encountered environments that play a crucial role in people's everyday lives. These environments comprise the living room [21], meditation room [18], and office [16], each chosen for its representation of distinct aspects of the human experience. Specifically, these environments encompass leisure and socialization, mental health and self-care, and work and collaboration, respectively. We explore the factors contributing to their widespread appeal and analyze how they manifest in virtual reality applications.

3 IMPLEMENTATION

3.1 Virtual Space Generation

We provide three frequently encountered scenarios for users to choose from: a living room, an office, and a meditation space. Users select a scene that interests them, opening their imagination and allowing them to write down their thoughts. To produce panoramas, we employ a text-conditioned global sampler to generate a low-resolution holistic image based on the input prompt. Given the limited content and resolution of this image, a structure-aware local sampler is utilized to synthesize local patches, resulting in marginally higher-resolution images. Subsequently, the super-resolution inverse tone mapping operator (SR-iTMO) is used to produce a high-resolution image, which is transformed into a panorama and imported into the Unreal Engine for display. For 3D room mesh generation, images from various preset camera poses are initially produced based on the input prompt using Stable Diffusion. These images are then refined with the aid of the IronDepth model [2], generating additional images with marginally different angles. The mesh is formed and refined simultaneously throughout the image generation process. Upon generating a sufficient number of images, extra poses are sampled to fill any gaps in the mesh, thereby enhancing mesh quality and completing the room mesh. The polishing process loops through the mesh and refines the mesh based on a set of adjustable parameters. Based on the parameters, the mesh is refined multiple times and causes huge latency. The completed mesh is saved as a 3D polygonal model and converted into a point cloud format for importation into the Unreal Engine.

3.2 Presentation for Users

We utilize Unreal Engine (UE) version 4.24.3[1] to render the generated panoramas and 3D models, enabling users to experience immersive virtual spaces on XR devices like the Oculus Quest 2[2]. For panorama rendering, we create a new material using the panoramic image and map it onto a sphere, providing users with a 3D experience within the spherical environment. Our model generates 3D spaces stored in the PLY format. To enable the rendering of these PLY models on XR devices, we opt to convert the PLY format models into PTS format models. Subsequently, we employ point cloud visualization techniques to display the models effectively. To visualize point cloud data, we use the LiDAR Point Cloud Sample[3] plugin. We then adjust the size of the point cloud until the gaps between the points are eliminated, shaping the point cloud into a square to sequentially compensate for any missing patches.

4 EXPERIMENTAL PROTOCOL

Our goals are to: (1) explore the extent to which AIGC is nearing the realization of dynamically generating personalized spaces, taking into account the present computational technology accessible to the general population; (2) investigate the user perspective of AIGC's capability to generate personalized spaces that offer a deeply immersive experience. Therefore, we design a pipeline to generate virtual spaces from prompts. We then use the pipeline to generate panoramic images and 3D mesh-based spaces from the same prompt. Finally, we conduct a more comprehensive user experience study on the preferred technological formats among users.

4.1 System Pipeline

Figure 1 demonstrates the pipeline to create and experience 3D space from textual input. Initially, users input their desired personalized scenarios in the form of text. Secondly, we analyze these texts using the Large Language Model (LLM). Thirdly, we use two models to generate personalized spaces in two forms: one is the creation of 3D models [17], and the other is the generation model of panoramic images [6]. We then integrate both forms of generated content into XR devices for an immersive user experience. Regarding the 3D mesh models, we presented them in a point cloud format. As for panoramic images, we map them into a sphere as textures to serve as a skybox, providing users with a comprehensive environmental experience. For more details on the pipeline, please refer to Figure 1.

4.2 Setup for Latency Evaluation

Our first objective is to assess the proximity of AIGC towards achieving the dynamic generation of personalized spaces. Hence, we carry out several experiments to evaluate the runtime latency of space generation across multiple devices. These devices include two PCs and a laptop, each equipped with different graphics cards: GTX1080, GTX1080 Ti, and RTX3080 Ti, respectively. We employ CPU offloading as a means of overcome the memory limitations of the graphics cards. While this approach introduces slightly additional runtime, it lowers the minimum CUDA memory requirement. We use the pipeline to generate three indoor environments: a living room, a meditation room, and an office room. This helps us assess runtime latency across various representative spaces with different content complexities.

4.3 User study:

Our second objective is to assess AIGC's capability to generate personalized spaces that offer both personalization and immersive experiences. Therefore, we conduct user experiments to assess various aspects of user experience, including presence, pragmatic

[1] https://www.unrealengine.com/en-US/blog/unreal-engine-4-24-releaseds
[2] https://www.meta.com/quest/products/quest-2/
[3] https://www.unrealengine.com/marketplace/en-US/product/lidar-point-cloud-sample

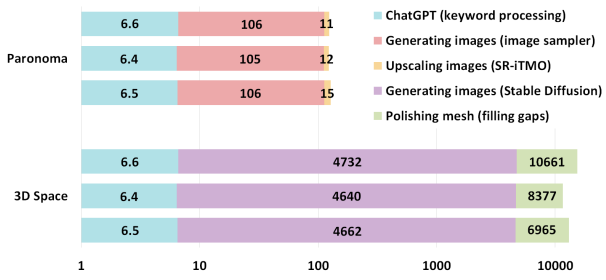


Figure 2: Latency decomposition of the pipeline for generating panoramic images and 3D mesh on NVIDIA RTX 3080 Ti.

quality, hedonic quality, and the overall quality of the generated environments.

4.3.1 Study design. We divided the user study into two parts. In the first part, we compared participants' perceptions of two different generation techniques. In the second part, based on the preferred technique identified in the first part of the user study, we investigated users' perceptions of personalized spaces. In both parts of the user study, we conducted user studies utilizing the three most common types of spaces encountered in daily life: the living room, meditation room, and office.

In the first user study, we design a 2×3 within-subject experiment. Participants were invited to experience two presentations: the generated 3D mesh (PLY) and panoramic images. We construct three scenarios in a randomized order, with the sequence of scenarios determined by a Latin square design to minimize potential order effects. The objective of our user study is to evaluate various aspects of user immersion and experiences comprehensively. To achieve this, participants provided self-reported feedback after each scenario via standardized questionnaires, specifically User Experience Questionnaire (UEQ) [3], and Igroup Presence questionnaire (IPQ) [29].

In the second experiment, we investigated participants' perception of personalized spaces using the technique preferred by users. Each participant experienced the personalized space generated based on their own prompt using XR devices and subsequently completed the User Experience Questionnaire (UEQ) [3].

4.3.2 Participants: We recruited local volunteers who agreed to participate in our study and provided informed consent. All participants had normal or corrected-to-normal vision, ensuring that they were able to interact with our system without visual impairment. In the first user study, our sample consisted of 12 individuals, evenly divided between male and female participants, with ages ranging from 20 to 50 years ($M = 31.38, SD = 8.17$). In the second user study, we collected 21 samples, ages ranging predominantly from 20 to 30 ($M = 25.43, SD = 1.76$), and females comprising 42.8% of the sample.

5 RESULTS

We first present the results related to the runtime latency of virtual space generation, detailed in Sections 5.1 and 5.2. We then explore the findings of user perception of the generated virtual scenes in Section 5.3.

5.1 Pipeline Characterization

Figure 2 illustrates the pipeline's latency decomposition to generate panoramic images and 3D mesh from the input prompt on NVIDIA RTX 3080 Ti. This includes the latency for refining the prompt using the ChatGPT API (mean of approximately 6.5 seconds) in both space types. This time encompasses the reception, generation, and response phases. For panorama generation, the image sampler takes approximately 106 seconds to generate a panoramic image, while the super-resolution inverse tone mapping operator (SR-iTMO) process takes approximately 15 seconds to upscale the image and enhance its visual quality. For generating a 3D mesh, Stable Diffusion takes approximately 78 minutes to generate a high-resolution image, while refining and enhancing the visual quality of the 3D mesh requires an additional 140 minutes for polishing and hole-filling. It is noteworthy that the latency for user input is not included in our analysis, as this latency is influenced by individual factors such as typing proficiency and speaking speed in VR environments. Additionally, audio-to-text runtime latency is not presented as it operates in real time.

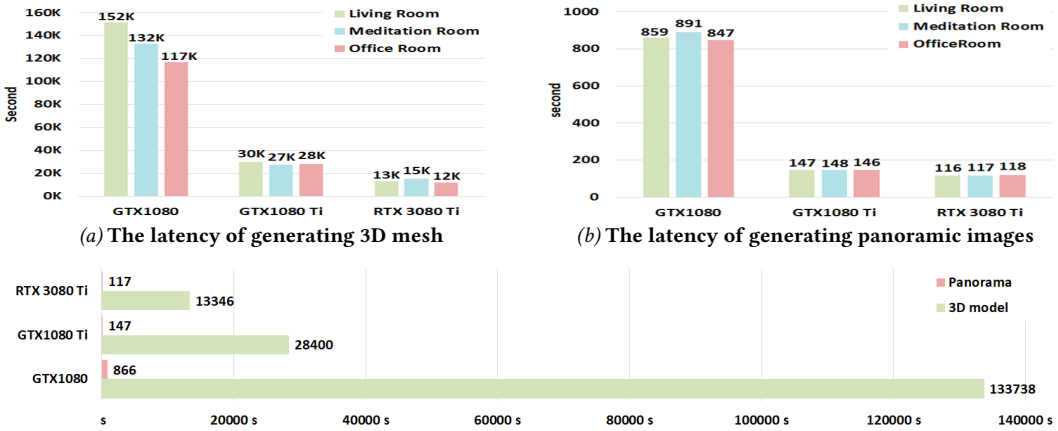
Overall, the generation of 3D meshes incurs substantial runtime latency due to the extensive amount of information required for mesh creation and subsequent polishing.

5.2 Performance Per Computing Hardware Settings

Figures 3a and 3b depict the cumulative latency involved in generating 3D meshes and panoramic images, respectively, for three representative indoor environments: the living room, meditation room, and office room. Although the latency for 3D mesh generation of the meditation room is lower than that of the living room on GTX 1080, it takes the longest time on RTX 3080 Ti. This discrepancy can be attributed to the varying content complexity of the spaces generated by AIGC, even when prompted with the same input. This discrepancy also holds in the panoramic generation of the three environments across the tested GPUs. Figure 3c depicts the accumulative latency for generating a panoramic image and a 3D mesh from the user input prompt. The generation time of panoramic images takes approximately 2 minutes on RTX 3080Ti, 2.6 minutes on GTX 1080 Ti, and 14 minutes on GTX 1080. This latency increases significantly in 3D mesh generation due to the reason discussed in subsection 5.1. The generation time of the 3D mesh takes approximately 4 hours, 7.9 hours, and 37 hours on RTX 3080 Ti, GTX 1080 Ti, and GTX 1080, respectively. This generation latency for panorama and 3D mesh decreases exponentially with increased computational power.

With the availability of robust hardware support, panoramic generation using AIGC appears to be a more compelling and feasible choice for dynamic and personalized content generation.

AIGC shows promise as a feasible and viable choice for (semi-)dynamically generating panoramic images. On high-end PCs equipped with cutting-edge GPUs (RTX 4090 Ti and newer), the generation of panoramic images can be completed in a few seconds. On the contrary, despite the availability of powerful GPUs in the market for



(a) **The latency of generating 3D mesh**

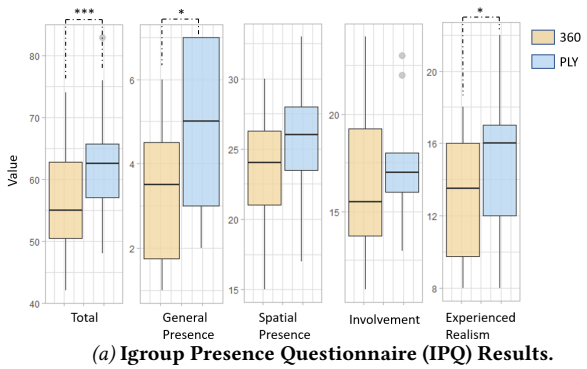
(b) **The latency of generating panoramic images**

(c) **Comparison of average generation time of panorama and 3D spaces over three hardware setups**

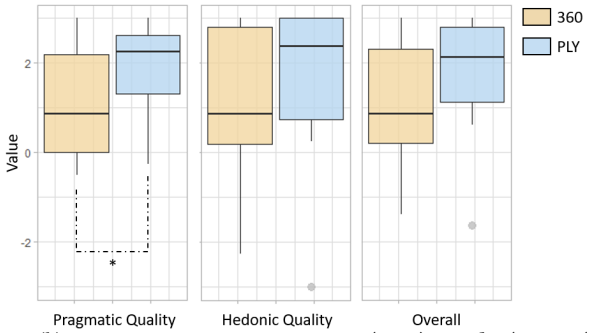
Figure 3: Latency of panorama and 3D mesh generation across different GPUs for three customized spaces.

average individuals, the dynamic generation of 3D meshes remains a formidable challenge. High-end PCs do provide notable performance enhancements, but they still fall short of (semi-)dynamically generating 3D mesh.

5.3 User Perception and Comparison of Two Technical Approaches



(a) **Igroup Presence Questionnaire (IPQ) Results.**



(b) **User Experience Questionnaire (UEQ) Results ($P < .05$(*)).**

Figure 4: Perception of panorama and 3D mesh. (a) IPQ results ($P < .05$(*), $P < .01$(**), $P < .001$(***)), (b) UEQ results ($P < .05$(*)).

In this section, we present findings regarding various aspects of user experience, encompassing presence, pragmatic quality, hedonic quality, and the overall quality of the generated environments. Additionally, we compare two technical approaches: 3D mesh (PLY) and panoramic images.

Igroup Presence questionnaire (IPQ). The paired-sample t-tests revealed significant differences between the two conditions for the total IPQ score and its four subscales. For the total IPQ score, participants experienced a significantly higher sense of presence in *PLY* ($M = 62.92, SD = 9.74$) compared to *360* ($M = 56, SD = 10.09$), t(11) = -5.16, p < .001. Similarly, the General Presence subscale indicated a significant improvement in *PLY* ($M = 4.83, SD = 1.85$) compared to *360* ($M = 3.42, SD = 1.92$), t(11) = -2.24, p < .05. No significant differences were found for the Spatial Presence subscale, with scores in *PLY* ($M = 25.67, SD = 4.31$) and *360* ($M = 23.5, SD = 4.68$), t(11) = -1.98, p = .073. Involvement also did not show significant differences in *PLY* ($M = 17.42, SD = 2.78$) compared to *360* ($M = 16.58, SD = 3.99$), t(11) = -0.88, p = .400. Lastly, the Experienced Realism subscale showed significantly higher scores in *PLY* ($M = 15.00, SD = 4.02$) than in *360* ($M = 12.83, SD = 3.51$), t(11) = -2.44, p < .05. These results suggest that the virtual 3D scenes (*PLY*) substantially enhances users' sense of presence in particular General Presence and Experienced Realism subscales compared to the panoramic images.

User Experience Questionnaire (UEQ). Pair-sample t-tests were performed to analyze the differences in the short version of the UEQ results between the two technologies, PLY and 360, for 12 participants. The analyses focused on Pragmatic Quality, Hedonic Quality, and Overall Quality.

Significant differences were found only in the Pragmatic Quality scores between the two technologies. Participants reported higher Pragmatic Quality scores for the PLY technology ($M = 1.92, SD = 1.07$) compared to the 360 technology ($M = 1.13, SD = 1.30$), t(11) = -2.27, p < .05. However, no significant differences were observed for the Hedonic Quality (PLY: $M = 1.63, SD = 1.79$; 360: $M = 1.13, SD = 1.80$; t(11) = -1.10, p = .29) and Overall Quality scores

(PLY: $M = 1.77, SD = 1.38$; 360: $M = 1.13, SD = 1.40$; t(11) = -1.72, p = .11) between the two technologies. These findings suggest that the PLY technology demonstrates superior Pragmatic Quality compared to the 360 technology, while the Hedonic and Overall Quality experiences were similar for both technologies.

Preference. Upon experiencing both technologies and three scenarios associated with each technology, all 12 participants self-reported a preference for the scenarios generated by the PLY technology. Regarding the preferences for the three distinct life scenes generated, the Living Room scene (N = 6) was favored more than the Office Room (N = 3) and Meditation Room (N = 3) scenes. These findings indicate a unanimous inclination toward 3D mesh-generated scenarios, with the Living Room scene emerging as the most preferred option among the three life scenes evaluated.

The findings underscore the efficacy of 3D mesh generation to enhance users' sense of presence and user experience compared to panoramic images. The 3D mesh demonstrates superior pragmatic quality, while hedonic and overall quality are comparable to panoramic images.

5.4 Follow-up User Study

In this experiment, we collected 21 samples using the users' preferred 3D generation technique, resulting in statistically significant and representative findings.

Table 1: User Experience Questionnaire (UEQ) Results of the follow-up User Study

UEQ Scales (Mean and Variance)		
c	**M**	**SD**
Attractiveness	↑1.365	0.45
Perspicuity	↑1.440	0.57
Efficiency	↑1.369	0.62
Dependability	→0.655	0.80
Stimulation	↑1.286	0.46
Novelty	↑1.286	0.81

User Experience Questionnaire (UEQ). The results of the UEQ are shown in Table 1. Values between -0.8 and 0.8 represent a more or less neutral evaluation of the corresponding scale; values > 0.8 represent a positive evaluation and values < -0.8 represent a negative evaluation. The results showed that our system received positive feedback in the five subscales of the UEQ questionnaire, Attractiveness, Perspicuity, Efficiency, Stimulation and Novelty, and only in the subscale Dependability did the results show less than neutral evaluation, with a mean value below 0.8.

6 DISCUSSION

The purposes of this paper are (1) to evaluate the extent to which AIGC (AI-generated content) can realize the dynamic generation of personalized spaces, and (2) to explore the user perspective regarding the quality of immersion and personalization of these generated spaces. To accomplish this, we have developed a pipeline that harnesses state-of-the-art generative AI and AIGC tools. This pipeline facilitates the creation of 3D meshes for virtual 3D spaces within the Metaverse, as well as the generation of 2D panoramic images. These specific spaces serve as immersive environments for end users and form the focus of our investigation.

In pursuit of our first objective, we examined the runtime latency for generating both panoramic images and 3D meshes. We observed a significant exponential decrease in latency when transitioning from 10-series graphics cards to 20 and 30-series graphics cards. For instance, the RTX 3080 Ti demonstrates the ability to generate panoramic images in approximately 2 minutes, and this trend persists with enhanced versions (Ti). The introduction of RTX 40 series graphics cards further reduces the generation time to a few seconds for panoramas, and it is anticipated to reach a few hundred milliseconds with the upcoming RTX 50 series graphics cards. However, despite these advancements, the runtime latency for 3D mesh generation remains in the order of hours when upgrading from 10-series to 30-series graphics cards. Despite the potential reduction in latency to a few dozen minutes with RTX 40 series graphics cards, achieving semi-dynamic 3D mesh generation remains a considerable challenge for AIGC, indicating that it is still far from being realized. In pursuit of our second objective, we delved into the user perception of immersion quality and space personalization. The findings of our user study reveal that the generated 3D mesh provides an enhanced sense of presence and user experience compared to panoramic images. Specifically, participants experienced a significantly higher sense of presence in the 3D mesh scenarios, particularly in the general presence, and experienced realism subscales. Additionally, the 3D mesh demonstrates superior pragmatic quality, while hedonic and overall quality are similar to panoramic images.

These findings indicate that despite the quicker generation of panoramic images, the superior immersive and personalized experience of 3D meshes outweighs the latency concerns, significantly enhancing user engagement in virtual worlds.

7 CONCLUSION

This paper presents a framework to personalize spaces within the Metaverse. By inputting user prompts, the framework generates personalized spaces in the form of panoramas and 3D meshes. We conducted extensive experiments and a user study to evaluate the runtime latency and user perception of these generated spaces. Our experiments covered a variety of hardware and settings, while our user study examined the user perception of the generated panoramic images and 3D mesh spaces.Our findings revealed that while AIGC presents promising advancements in Metaverse personalization, the computation and runtime overhead remain significant challenges for the average user. However, users generally appreciate the experiences offered by 3D virtual spaces.

ACKNOWLEDGEMENTS

This research was supported in part by a grant from the Guangzhou Municipal Nansha District Science and Technology Bureau under Contract No.2022ZD01 and the MetaHKUST project from the Hong Kong University of Science and Technology (Guangzhou).

REFERENCES

[1] Ali Abdari, Alex Falcon, and Giuseppe Serra. 2023. Metaverse Retrieval: Finding the Best Metaverse Environment via Language. In *Proceedings of the 1st*

International Workshop on Deep Multimodal Learning for Information Retrieval. 1–9.
[2] Gwangbin Bae, Ignas Budvytis, and Roberto Cipolla. 2022. IronDepth: Iterative Refinement of Single-View Depth using Surface Normal and its Uncertainty. arXiv:2210.03676 [cs.CV]
[3] Simone Borsci, Stefano Federici, and Marco Lauriola. 2009. On the dimensionality of the System Usability Scale: a test of alternative measurement models. *Cognitive processing* 10 (2009), 193–197.
[4] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. 2015. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012* (2015).
[5] Kevin Chen, Christopher B Choy, Manolis Savva, Angel X Chang, Thomas Funkhouser, and Silvio Savarese. 2019. Text2shape: Generating shapes from natural language by learning joint embeddings. In *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part III 14*. Springer, 100–116.
[6] Zhaoxi Chen, Guangcong Wang, and Ziwei Liu. 2022. Text2light: Zero-shot text-driven HDR panorama generation. *ACM Transactions on Graphics (TOG)* 41, 6 (2022), 1–16.
[7] Zezhou Cheng, Menglei Chai, Jian Ren, Hsin-Ying Lee, Kyle Olszewski, Zeng Huang, Subhransu Maji, and Sergey Tulyakov. 2022. Cross-modal 3d shape generation and manipulation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part III*. Springer, 303–321.
[8] Christopher B Choy, Danfei Xu, JunYoung Gwak, Kevin Chen, and Silvio Savarese. 2016. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*. Springer, 628–644.
[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).
[10] Haihan Duan, Jiaye Li, Sizheng Fan, Zhonghao Lin, Xiao Wu, and Wei Cai. 2021. Metaverse for Social Good: A University Campus Prototype. In *Proceedings of the 29th ACM International Conference on Multimedia* (Virtual Event, China) *(MM '21)*. Association for Computing Machinery, New York, NY, USA, 153–161. https://doi.org/10.1145/3474085.3479238
[11] Ahmed Elhagry. 2023. Text-to-Metaverse: Towards a Digital Twin-Enabled Multimodal Conditional Generative Metaverse. In *Proceedings of the 31st ACM International Conference on Multimedia*. 9336–9339.
[12] Social Europe. February 07, 2022. The Metaverse is a labour issue. https://www.socialeurope.eu/the-metaverse-is-a-labour-issue.
[13] C Frasson et al. 2021. A framework for personalized fully immersive virtual reality learning environments with gamified design in education. In *Novelties in Intelligent Digital Systems: Proceedings of the 1st International Conference (NIDS 2021), Athens, Greece*, Vol. 338. 95.
[14] Guo Freeman, Samaneh Zamanifard, Divine Maloney, and Alexandra Adkins. 2020. My body, my avatar: How people perceive their avatars in social virtual reality. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–8.
[15] Thibault Groueix, Matthew Fisher, Vladimir G Kim, Bryan C Russell, and Mathieu Aubry. 2018. A papier-mâché approach to learning 3d surface generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 216–224.
[16] E Jeffrey Hill, Maria Ferris, and Vjollca Märtinson. 2003. Does it matter where you work? A comparison of how three work venues (traditional office, virtual office, and home office) influence aspects of work and personal/family life. *Journal of vocational behavior* 63, 2 (2003), 220–241.
[17] Lukas Höllein, Ang Cao, Andrew Owens, Justin Johnson, and Matthias Nießner. 2023. Text2Room: Extracting Textured 3D Meshes from 2D Text-to-Image Models. *arXiv preprint arXiv:2303.11989* (2023).
[18] Jon Kabat-Zinn, Daniel Siegel, Thich Nhat Hanh, and Jack Kornfield. 2011. *The mindfulness revolution: Leading psychologists, scientists, artists, and meditation teachers on the power of mindfulness in daily life*. Shambhala Publications.
[19] Jacob Kritikos, Georgios Alevizopoulos, and Dimitris Koutsouris. 2021. Personalized virtual reality human-computer interaction for psychiatric and neurological illnesses: a dynamically adaptive virtual reality environment that changes according to real-time feedback from electrophysiological signal responses. *Frontiers in Human Neuroscience* 15 (2021), 596980.
[20] Xingyu Liu, Charles R Qi, and Leonidas J Guibas. 2019. Flownet3d: Learning scene flow in 3d point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 529–537.
[21] Cheryl Mattingly. 2013. Moral selves and moral scenes: narrative experiments in everyday life. *Ethnos* 78, 3 (2013), 301–327.
[22] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. 2019. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4460–4470.
[23] Dimitris Mourtzis, Nikos Panopoulos, John Angelopoulos, Baicun Wang, and Lihui Wang. 2022. Human centric platforms for personalized value creation in metaverse. *Journal of Manufacturing Systems* 65 (2022), 653–659.
[24] Silvia Francesca Maria Pizzoli, Ketti Mazzocco, Stefano Triberti, Dario Monzani, Mariano Luis Alcañiz Raya, and Gabriella Pravettoni. 2019. User-centered virtual reality for promoting relaxation: an innovative approach. *Frontiers in psychology* 10 (2019), 479.
[25] Jhony K Pontes, Chen Kong, Sridha Sridharan, Simon Lucey, Anders Eriksson, and Clinton Fookes. 2019. Image2mesh: A learning framework for single image 3d reconstruction. In *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part I 14*. Springer, 365–381.
[26] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. 2022. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988* (2022).
[27] Hua Xuan Qin and Pan Hui. [n. d.]. Empowering the Metaverse with Generative AI: Survey and Future Directions. ([n. d.]).
[28] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.
[29] Thomas Schubert, Frank Friedmann, and Holger Regenbrecht. 2001. The experience of presence: Factor analytic insights. *Presence: Teleoperators & Virtual Environments* 10, 3 (2001), 266–281.
[30] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. 2022. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792* (2022).
[31] Angelica G Thompson-Butel, Christine T Shiner, John McGhee, Benjamin John Bailey, Pascal Bou-Haidar, Michael McCorriston, and Steven G Faux. 2019. The role of personalized virtual reality in education for patients post stroke—a qualitative case series. *Journal of Stroke and Cerebrovascular Diseases* 28, 2 (2019), 450–457.
[32] Tengfei Wang, Bo Zhang, Ting Zhang, Shuyang Gu, Jianmin Bao, Tadas Baltrusaitis, Jingjing Shen, Dong Chen, Fang Wen, Qifeng Chen, et al. 2022. Rodin: A Generative Model for Sculpting 3D Digital Avatars Using Diffusion. *arXiv preprint arXiv:2212.06135* (2022).
[33] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 2015. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1912–1920.
[34] Haozhe Xie, Hongxun Yao, Xiaoshuai Sun, Shangchen Zhou, and Shengping Zhang. 2019. Pix2vox: Context-aware 3d reconstruction from single and multi-view images. In *Proceedings of the IEEE/CVF international conference on computer vision*. 2690–2698.
[35] Minrui Xu, Wei Chong Ng, Wei Yang Bryan Lim, Jiawen Kang, Zehui Xiong, Dusit Niyato, Qiang Yang, Xuemin Sherman Shen, and Chunyan Miao. 2022. A full dive into realizing the edge-enabled metaverse: Visions, enabling technologies, and challenges. *IEEE Communications Surveys & Tutorials* (2022).
[36] Guandao Yang, Xun Huang, Zekun Hao, Ming-Yu Liu, Serge Belongie, and Bharath Hariharan. 2019. Pointflow: 3d point cloud generation with continuous normalizing flows. In *Proceedings of the IEEE/CVF international conference on computer vision*. 4541–4550.
[37] Riyan Yang, Lin Li, Wensheng Gan, Zefeng Chen, and Zhenlian Qi. 2023. The Human-Centric Metaverse: A Survey. In *Companion Proceedings of the ACM Web Conference 2023*. 1296–1306.
[38] Chaoning Zhang, Chenshuang Zhang, Sheng Zheng, Yu Qiao, Chenghao Li, Mengchun Zhang, Sumit Kumar Dam, Chu Myaet Thwal, Ye Lin Tun, Le Luang Huy, et al. 2023. A Complete Survey on Generative AI (AIGC): Is ChatGPT from GPT-4 to GPT-5 All You Need? *arXiv preprint arXiv:2303.11717* (2023).

A FROM PROMPTS TO VISUAL REALMS

Figure 5 illustrates three different prompts (first column) to generate virtual content for three spaces, living room, office, and meditation room. Accordingly, our AIGC-based prototype generates the corresponding 3D models (second column), and panoramic images (fourth column). When wearing a VR headset to experience in VR, the second column showcases the appearance of a 3D model.

To generate the visual realms, we design a pipeline that employs cutting-edge AIGC tools. This pipeline enables the generation of immersive panoramas and versatile 3D spaces, laying the foundation for a wide array of applications and experiences within the rapidly growing metaverse landscape. This breakthrough holds

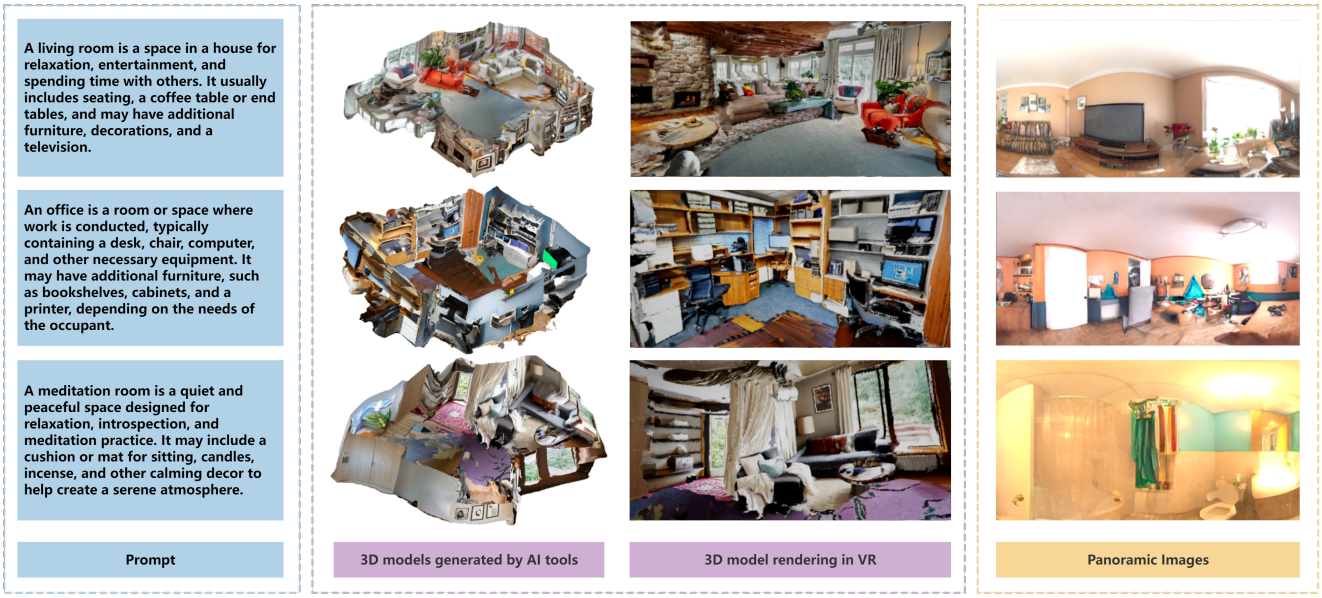


Figure 5: Illustration of generated 3D model representation and its panoramic counterpart using the same input prompts.

great promise for enhancing user engagement and shaping the future of virtual worlds. The following sections discuss our endeavors for dynamic and personalized virtual experience:

A.1 Tailored Experience

Users can input keywords, such as "living room, warm color, big bed," through a virtual keyboard or audio input in the VR space. ChatGPT formulates the input prompt, while Text2room and Text2light models generate corresponding panoramic images and 3D meshes. This enables the creation of personalized spaces for tailored user experiences, which are crucial for virtual social activities, psychotherapy, education, and applications promoting relaxation and personal well-being in virtual environments [13, 14, 19, 24, 31]. Our findings of the user study reveal that such personalized spaces enhance immersion and users' sense of presence, especially in virtual 3D spaces.

A.2 Dynamic Content Generation

The primary technical challenge lies in the computational power limitations, particularly for dynamically generating personalized virtual 3D spaces. Our evaluation of runtime and latency breakdown indicates that the majority of time is spent on 3D mesh generation, while other components can operate in real time. Our investigation using multiple GPUs demonstrates that stronger computational power can significantly reduce rendering time. However, current graphic cards accessible to average users are insufficient for semi-dynamic 3D mesh generation. Conversely, AIGC shows promise in generating near real-time panoramic images.

A.3 Quality-latency Trade-off

The visual quality of generated content is vital for a compelling user experience. However, the latency breakdown reveals that image upscaling in panoramic image generation and mesh polishing in 3D space generation contribute to significant runtime latency. The latency is directly influenced by output quality, which depends on factors like the scale of super-resolution and the desired mesh quality. Customizing the visual quality and generation latency allows for prioritizing user objectives, leading to an enhanced overall experience.

A.4 Towards Higher Visual Quality

Further exploration can be conducted to assess the potential of advanced rendering techniques, higher fidelity 3D models, and diverse environmental settings to enhance the visual quality of the generated spaces. This is particularly crucial for immersive user experiences, especially within virtual 3D spaces. Investigating the implications of these findings in different application contexts, such as virtual reality gaming, training simulations, and remote collaboration, would be valuable in optimizing configurations for high and ultra-high resolutions to create immersive environments.