

“Computer, Generate!” – Investigating User-Controlled Generation of Immersive Virtual Environments

Carina LIEBERS^{a,1}, Niklas PFÜTZENREUTER^b, Jonas AUDA^b,
Uwe GRUENEFELD^b, and Stefan SCHNEEGASS^a

^aUniversity of Duisburg-Essen, Schuetzenbahn 70, 45127 Essen, Germany

^bGenerIO.ai, Schuetzenbahn 70, 45127 Essen, Germany

Abstract. For immersive experiences such as virtual reality, explorable worlds are often fundamental. Generative artificial intelligence looks promising to accelerate the creation of such environments. However, it remains unclear how existing interaction modalities can support user-centered world generation and how users remain in control of the process. Thus, in this paper, we present a virtual reality application to generate virtual environments and compare three common interaction modalities (voice, controller, and hands) in a pre-study ($N = 18$), revealing a combination of initial voice input and continued controller manipulation as best suitable. We then investigate three levels of process control (all-at-once, creation-before-manipulation, and step-by-step) in a user study ($N = 27$). Our results show that although all-at-once reduced the number of object manipulations, participants felt more in control when using the step-by-step approach.

Keywords. Generative AI, Virtual Reality, Human-Controlled Scene Generation

1. Introduction

The Holodeck from Star Trek[®] enables users to generate realistic environments by expressing their desires with simple voice commands next to manual controls. As artificial intelligence (AI) progresses, generating individualized virtual worlds becomes more feasible, as virtual reality (VR) allows users to create and experience such worlds. However, individualized environments are too complex to describe with short voice commands. The question remains of how to give humans control over the generation.

To empower users to generate virtual worlds, we must understand the extent to which they need control during the generation process. User-controlled generation includes many aspects, from object creation to adjusting sizes, textures, and positions to broader environmental elements. To reduce complexity, we break the generation down into individual objects (i.e., the building blocks of the environment). Nevertheless, users still have to manage object creation, type, placement, aesthetics, animation, and interactivity. Although in-depth customization offers great autonomy, it may also be inefficient. Thus, the question remains of how different control flows affect efficiency and user experience.

Although existing literature explored user-driven virtual environment design, it remains unclear which control flow enhances user’s control. Prior works investigated gen-

¹Corresponding Author: Carina Liebers, carina.liebers@uni-due.de

eration through object catalog selection [1,2,3] and adjustments of spatial [4,2] and appearance [4,3] attributes, but these often do not generalize well. Furthermore, none assessed user control for generating virtual environments.

To investigate human control in virtual environment generation, we designed a VR testbed, including the input modalities (controller, hand, or voice input) and generation steps (creation, movement, scaling, coloring, and deletion). In a pre-study, we assess which input modalities support the generation steps best, revealing a combination of initial voice and subsequent controller input. Using these modalities, we examine how humans can control the generation process to meet their expectations, using the control strategies all-at-once, creation-before-manipulation, and step-by-step (**RQ1**). While all-at-once reduced object manipulations, participants felt higher control using step-by-step.

Contribution Statement. We propose a VR test application for environment generation, including controller, hand, and voice input for the generation steps. Using the most suitable input modality for each generation step from a pre-study ($N = 18$), we conduct a user study ($N = 27$) to compare three control flows (all-at-once, creation-before-manipulation, and step-by-step), assessing user control during generation.

2. Related Work

Since we aim to assess the optimal control flow for generating virtual environments within VR, our research is grounded in input modalities and scene-generation methods for VR applications.

2.1. Object Interaction Techniques

To create a virtual world, for instance for level design, users need to interact with the system and the generated objects. These interactions encompass object manipulations, such as generating and deleting objects, and spatial actions involving positioning and scaling objects.

Since gestures and speech constitute language [5], approaches like "Put-That-There" [6] combine gesture and speech to spatially manipulate objects. Participants particularly preferred speech commands such as "move" for translation, but gestures for rotation tasks [7]. For object translation and rotation in VR, research has shown that hand-held controllers tend to offer higher performance, usability, and user preference compared to hand-tracking solutions [8]. Further, 3D interfaces performed better than speech interfaces for positioning and rotating during furnishing tasks in VR [9]. In the context of scaling operations, gesture-speech-based interactions are more effective than grasping interactions [10]. Specifically, when gesture and speech were compared, gestures were rated as the superior modality for scaling objects [7]. For object creation and deletion as well as text input, speech commands have been found to outperform all other input modalities [7,9]. Thus, the input modality depends on the specific tasks [11,12].

Controller-based or gesture-based interactions are typically preferred for translation and rotation tasks, while speech is often preferred for non-spatial operations. However, it is important to note that many of these studies primarily focused on Augmented Reality (AR) environments. As our work involves creating subjective virtual environments that

encompass both spatial and non-spatial operations, we conducted a pre-study to evaluate which input modalities are best suited for our specific use case in VR.

2.2. User-Assisted Scene Generation Using Machine Learning

Machine Learning (ML) methods have been widely used for the creation of virtual scenes and objects in VR and AR environments. These methods typically utilize images or videos as input to generate 3D models and scenes [13]. In the following, we provide an overview of related research that focuses on user-assisted virtual scene and object generation, and ways in which users can interact with generative methods.

To create a virtual world within VR, users primarily select objects to design the environment. They can create a world through catalog selection [1,2,3] or scanning physical environments [14,15]. Environment building methods vary. Some allow direct object placement via controllers [14] or combine voice and hand input for placement [4]. Others first modify the terrain before object manipulation utilizing hybrid wand/tablet interface [16]. *VR Safari Park* [17] offers a novel approach, letting users add virtual blocks to a world tree, representing entities, supporting especially novices with an effective overview. Environment generation from text input starts with extracting the objects via voice [18] or text input [19,20,21,22]. To translate requests into actual scenes, the objects are used to create a semantic scene graph to generate sub-scenes from 3D databases, which are later adjusted to match user intentions [19]. Alternatively, scenes can be formed by considering inferred object relations [21,20], scanning real-world environments [15,23], or comparing input against annotated datasets and reference images [22].

To generate 3D models for building virtual scenes, current approaches compute multiple images of an object. Utilizing AI systems, especially GANs (Generative Adversarial Networks), facilitates the generation of images from a textual description. They excel in generating images using unlabeled training data [24], yet diffusion models surpassed them in deriving images from text, offering better scene complexity and image quality [25,26]. In previous works, users could control the object generation using voice input, adjusting attributes like style, shape, and categories of the objects generated using pre-trained text-to-image model [27]. User-driven control of ML techniques has primarily focused on image generation. One example is GANzilla [28], which lets users highlight areas of interest, analyze clusters, and explore alternate image generation outcomes. In this work, we aim to assess user control strategies for generating a virtual environment through generating 3D objects using the Point-E generative network [29].

3. Generating Immersive Environments

To investigate how to give users control over virtual environment generation, we first define the scope of our generation procedure. Then, we describe the selected input modalities and control flows and our VR environment as a testbed for the evaluation.

3.1. Overview of the Generation Procedure

Immersive environments consist of numerous elements, from scene-specific settings such as lighting and weather conditions to individual objects' details and appearances [1]. To reduce the complexity of the generation, we perceive immersive environments as a

symphony of individual objects existing in the environment. We consider general scene aspects by using the example of a virtual garden with a fixed configuration (i.e., preset lighting and weather conditions). For the generation process of the objects, we identified a set of five different *generation steps* that can be used to create or manipulate an object. In the first step, the object itself is created in step creation. Afterwards, the objects can be manipulated in a spatial manner resulting in the steps: movement and scaling. Last, the appearance of an object can be changed in the steps coloring and deletion.

3.2. Input Modalities

Existing literature suggests controller is effective for Grab-and-Place tasks [8], hand has lower accuracy for precise tasks [8,30] but being more intuitive for users [31,30] and voice input offering easy usage for simple command-based tasks [9] while being used in more application cases [32]. Given the variety of tasks during environment generation, it remains unclear which *input modality* facilitates the control process best: controller input, hand input, and voice input.

Controller Input. To create an object, we use a virtual keyboard in VR, motivated by Weiß et al. [9]. Users can then translate, rotate, or scale the object either by pointing at or grabbing the object and performing the wished transformation, as suggested by Whitlock et al. [33]. For coloring, we employ a spray gun metaphor, and deletion is achieved by grabbing the object and pressing a button.

Hand Input. The hand input is designed similarly to controller interaction. Users create objects using a keyboard, move, and scale objects by grabbing them, color objects using a spray gun, and delete objects by pressing a button. To select a letter, grab an object, activate the spray gun, or press buttons in the graphical user interface (GUI), users perform the pinch gesture by touching the thumb with the index finger.

Voice Input. Voice input allows speaking a specific command, which is transferred into the desired action. We mapped each generation step to keywords to improve request recognition when using natural language. Following Bolt's method [6], we integrated voice commands with pointing gestures to select the intended object or specify a location.

3.3. Control Flows

A crucial aspect is understanding what temporal order (referred to as control flow) users want to generate the immersive worlds. We define three *control flows*: all-at-once, creation-before-manipulation, and step-by-step. While the Holodeck enables environment generation with a single voice command (realized as all-at-once), an alternative is to individually create each object and modify it directly after creating (referred to as step-by-step). We further introduce a mixed method (creation-before-manipulation), combining both. We aim to assess:

RQ1: Which control flow for generating virtual environments supports users best to achieve results that fulfill their goal?

All-At-Once. A detailed initial voice input allows the generation of the entire environment. The request includes all objects and their attributes like location, rotation, scale, and appearance. Users can make manual post-generation adjustments.

Creation-Before-Manipulation. The mixed control flow involves initially requesting all environment objects at once, with subsequent adjustments made in separate steps. It allows pausing object creation, offering users greater flexibility.

Step-By-Step. This approach involves creating and manipulating each object once at a time. Users first generate a single object through a distinct request and then apply spatial and appearance adjustments, similar to De Leon et al. [1].

3.4. Implementation of the VR Testbed

Following, we describe the developed VR environment and the input modalities. We used Unity3D [34] for deployment on the Meta Quest 2 [35]. To realize the controller and hand input, we used the Meta Interaction SDK [36], and for voice input, we used the Wit.ai speech-to-text interface [37]. Next, we describe the VR environment as well as the applied generation and processing technologies.

The Virtual Garden Environment. We designed a garden environment next to a white house, surrounded by a wooden fence measuring 4x4 meters. To enhance realism, trees are positioned outside of the garden under a mostly sunny sky. We include two interaction panels hanging from the veranda: one panel displays the ongoing task and its duration, and provides a button to confirm task finalization. An additional interaction panel is present for controller input and hand input.

Voice Recognition. In voice mode, users could activate voice commands using the controller's B button. The system recorded voice until a decreased volume signaled speech completion, or for a maximum of 20 seconds. All recordings are visualized as feedback. Given the possibility of non-native English speakers participating, recordings were translated to English using the deep-translator package [38] in Python. Afterward, Wit.ai's speech-to-text transcribed the speech, determining its intent and identifying entities like objects or colors. A command executes if recognized with over 80% confidence. A successful recognition triggered audio feedback.

Object Generation. Object requests are processed by the Point-E generative network [29], which provides point clouds of the objects integrated using REST. We generate meshes from these point clouds for colliders to ensure accurate object interaction. The point cloud and its mesh are loaded into the garden during runtime. With that, we enable users to create every imaginable object without being limited to an existing database.

4. Pre-study: Input Modalities for Environment Generation

Given the variety of tasks involved in generating environments, it's unclear which input modality best facilitates the control process, as related work primarily focuses on specific tasks [12,11] and AR environments. Thus, in our pre-study, we aim to identify the optimal input modality for the generation steps in VR. We compare the independent variables *input modalities* with three levels: controller input, hand input, and voice input, and *tasks groups* with three levels: creation, spatial manipulations (movement and scaling), and appearance manipulations (coloring and deletion). This research sought to enhance understanding of VR interaction design, particularly which input modalities best support user engagement and task efficiency in a virtual environment.

We conducted a controlled, within-subjects laboratory study to explore the suitability of three input modalities for the different task groups (see Figure 1). During the study, our participants should generate objects of varying complexity (creation), move and scale




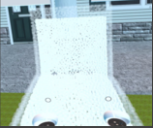
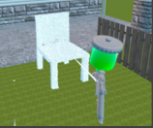
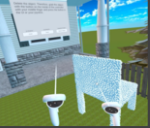





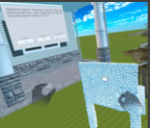






Generation Steps Input Modality	Creation	Spatial Manipulations		Appearance Manipulations	
		Movement	Scaling	Coloring	Deletion
 Controller					
 Hand					
 Voice					

Figure 1. Tasks in the Pre-Study: We compared the input modalities (controller, hand, and voice) for the generation steps (creation, movement, scaling, coloring, and deletion), categorized into three task groups: creation, spatial manipulation, and appearance manipulation.

objects to predetermined specifications (spatial manipulation), change object colors, and delete them (appearance modification) with all input modalities. The study incorporated a Latin square design to counterbalance the input modality and task group order. We measured the following dependent variables: *task completion time (TCT)*, *spatial accuracy*, *usability*, *user experience*, *perceived TCT*, *expectation conformity* and the *preferred modality*, alongside qualitative feedback through semi-structured interviews after each task group. Details of the variables are outlined in [Table 1](#).

To determine the required participant number, we conducted an a priori power analysis using G*Power for a repeated-measures within-factors analysis of variance (ANOVA), given $\alpha = .05$ and effect size $f = .4$. Following our experimental study design, we set the number of groups to 5 and the number of measurements to 5. G*Power 3.1.9.7 suggested a total sample size of 15 ($\lambda = 24, F = 2.61$). Due to our randomization, we invited 18 participants for our study. Of the volunteers (15 identified as male, 3 identified as female, and 0 as non-binary), all except one were right-handed, whose median age was 26.5 years ($SD = 4.34, Min = 19, Max = 35$). In terms of VR experience, 1 participant reported to have “never” used VR, 5 participants used VR “once a year”, 4 “once a month”, 7 “once a week”, and 1 participant used VR “daily”. For voice control experience, 2 participants reported to have “never” used voice control, 5 participants used Voice “once a year”, 6 “once a month”, 1 “once a week”, and 4 participants used voice control “daily”. Conducted in a room with a $4m \times 4m$ open space, we ensured a safe and controlled setting for the VR interaction. To ensure the participants’ privacy, we only recorded pseudonymized data. Our local ethics committee approved the study.

Voice Input and Controller Input Are Preferred for Creation Our quantitative data shows no significant difference in *TCT* among the input modalities for creation. Despite this, participants perceived hand input as slower than voice input (*Perceived TCT*, $t(4.66) = -3.42, p = .005 < .01$), especially with virtual keyboards. Regarding *usability*, voice input significantly outperformed hand input ($t(4.99) = -4.10, p < .001$), whereas participants often referred to the hand input’s accuracy issues as drawbacks.

Voice input was praised for bypassing manual input despite occasional inaccuracies. Participants stated that controller interaction was more cumbersome than a voice query for creation, although they enabled faster typing. Voice input also surpassed hand input in pragmatic *user experience* ($t(5.33) = -2.96, p = .017 < .05$) and outperformed controller input in the overall scores ($t(5.33) = -2.58, p = .041 < .05$). Our participants' modality rankings mirrored these insights, favoring controller input ($t(4.67) = -3.43, p = .005 < .01$) and voice input ($t(4.67) = 4.93, p < .001$) over hand input for creation between the input modalities. These trends indicate a preference for voice and controller input during creation tasks.

Controller Input Outperforms Hand and Voice Interaction for Spatial Manipulations
For spatial manipulations, voice input was less accurate than controller input ($t(3.8) = -6.08, p < .001$) and hand input ($t(3.8) = -3.48, p = .004 < .01$) for translation and rotation, and many participants found precise modifications challenging. In *expectation conformity*, controller input led to significantly more expected outcomes than voice input ($t(3.01) = 9.33, p < .001$). The usability scores display higher usability for controller input than voice input ($t(4.22) = 5.62, p < .001$). Comparing hand input and voice input, hand input received higher pragmatic *User Experience Questionnaire (UEQ)* ($t(4.01) = 3.12, p = .011 < .05$) and *preferred modality* scores ($t(3.55) = -5.36, p < .001$). While participants indicated technological limitations for voice and hand inputs in the interviews, hand input was preferred for intuitive interaction. In summary, controller input displayed an advantage for spatial manipulations, as evidenced by qualitative feedback and higher accuracy in translation and rotation. It outperformed hand ($t(4.22) = 3.22, p = .009 < .01$) and voice inputs ($t(4.22) = 5.62, p < .001$) in System Usability Scale (SUS) and pragmatic UEQ scores (hand input: $t(4.01) = 3.39, p = .005 < .01$), voice input: ($t(4.01) = 6.5, p < .001$)).

Controller Input and Voice Input Are Both Suitable for Appearance Manipulations
Contrary to our expectations, controller input was significantly faster than hand input ($t(9.8) = -10.25, p < .001$) and voice input ($t(9.8) = -7.88, p < .001$) in appearance manipulations. Although we found no significant difference in TCT between hand input and voice input, participants perceived hand input as slower than voice input (*perceived TCT*). Hand input further received lower *usability* scores than controller input ($t(3.87) = 5.11, p < .001$) and voice input ($t(3.87) = -6.03, p < .001$). Participants faced challenges with the pinch gesture interfering with the appearance manipulations for hand input. Voice input ($t(4.14) = -4.94, p < .001$) and controller input ($t(4.14) = 4.22, p < .001$) outperformed hand input in *user experience* and were strongly favored (*preferred modality*, controller input: ($t(3.88) = -5.67, p < .001$), voice input: ($t(3.88) = 6.70, p < .001$)). Between them, controller input excelled in efficiency (*TCT*), while voice input was also highly valued for its *user experience*.

5. Human Control in the Generation of Virtual Environments

We assess which *control flow* for generating virtual environments provides the most control over the generation to fulfill their initial vision (**RQ1**), focusing on the temporal user engagement in the generation process, ranging from an early to late user intervention. Following, we outline the study details and its findings. We used the same apparatus as in the pre-study (see [Section 4](#)) and obtained approval from our local ethics committee.

Table 1. Variables and Their Measurement for Evaluating *Input Modalities* in the Pre-Study.

Measurement	Calculation	Measurement	Calculation
Task completion time (TCT)	Time of first to last task-related action	Perceived TCT	7-Point Likert items: "I was able" "to accomplish the tasks fast."
Spatial Accuracy	Translation (positional offset), rotation (angular distance), scale (scale vector distance)	Expectation Conformity	7-Point Likert items: "The outcome" "of the interactions was as expected."
Usability	System Usability Scale (SUS) questionnaire [39]	Preferred Modality	Modality rating from best to worst
		User Experience	User Experience Questionnaire (UEQ) Short [40]

5.1. Study Design

We conducted a within-subject design, examining the independent variable *control flow*, with three characteristics: all-at-once, creation-before-manipulation, and step-by-step. We hypothesized that a more automated process (e.g. all-at-once) would likely diminish the interactions users need to shape their environment. Fewer interactions could lead to higher usability and allow users to realize their visions faster. Additionally, more interactions could lead to prolonged adjustment times, potentially resulting in users feeling fatigued, bored, or stressed, potentially negatively affecting their user experience during the generation. Moreover, given the extended adjustment period and possible negative feelings, we assumed a less automated approach (e.g. step-by-step) would result in a less satisfying scene outcome. Thus, we hypothesized:

H1.1 All-at-once leads to higher usability than step-by-step.

H1.2 Step-by-step leads to a worse user experience compared to all-at-once.

H1.3 Step-by-step leads to a less satisfying scene outcome compared to all-at-once.

During the study, participants generated a virtual garden with five objects based on three scene templates to negate the influence of object types, colors, and positions. We adopted a Latin square design to counterbalance scene task and control flow, leading to nine configurations. Our measured dependent variables can be found in Table 2. Following the study tasks, we conducted semi-structured interviews for qualitative feedback.

5.2. G*Power Analysis

To determine our required sample size, we conducted an a priori power analysis for a repeated-measures within-factors ANOVA, given $\alpha = .05$ and effect size $f = .4$. Following our study design, we set the number of groups to 3 and measurements to 3. G*Power 3.1.9.7 suggested a total sample size of 18 ($\lambda = 17.28$, $F = 3.32$), which we exceeded by half due to non-existing references for similar studies.

5.3. Participants

We recruited 27 participants (21 male, 6 female, 0 non-binary) for our study of which 17 also took part in the pre-study. All except one were right-handed, and 28.0 years old in the median ($SD = 3.97$, $Min = 19$, $Max = 35$). The participant self-reported to have used

VR 3 times "never", 6 times "once a year", 7 times "once a month", 10 times "once a week", and 1 times "daily". They stated to have used voice control, 3 times "never", 12 times "once a year", 5 times "once a month", 3 times "once a week", and 4 times "daily".

5.4. Procedure

Before the study, we briefed participants about the goals and procedure, emphasizing their right to withdraw at any time, and informed them of the recording of the TCT and task correctness. After they provided written consent, the study began. During the study, the participants generated a virtual garden based on three provided templates (see Figure 2), which included pre-defined objects with distinct positions and colors, using the control flows. They first entered a training scene for each control flow to familiarize themselves with the generation process. Afterward, participants moved to the actual task, followed by filling in the SUS, UEQ, and an individual questionnaire before progressing to the next control flow. The study concluded with a semi-structured interview to capture qualitative feedback. An entire session took 45 minutes per participant.

5.5. Study Tasks

We utilized the best input modalities for the generation steps from the pre-study: voice for creation and controller input for the subsequent spatial and appearance adjustments.

Step-By-Step. Participants use voice input to create one object at a time and then adjust its color and position with the controller. The process halts after each creation to prevent simultaneous creations until the next object is initiated.

Creation-Before-Manipulation. Participants use a single voice command to initiate multiple object creations, which appear one after another. They can pause the process to adjust an object's appearance or position, allowing step-wise adjustment of all objects.

All-At-Once. Participants generate the entire environment with one voice command, including objects, colors, and positions. The generation process remains uninterrupted until completion. Afterward, participants can make manual adjustments.

Table 2. Variables and Their Measurement for Evaluating Control Flows

Measurement Calculation		Measurement	7-Point Likert Items
TCT	Duration of the study task	Perceived Control	"I felt like I had full control over the generation process."
Task Correctness	Correctly created, placed, and colored objects	Adjustment Need	"It was never necessary to adjust the scene."
Manipulation Time	Mean object manipulation time	Perceived Success	"I have accomplished all my goals."
Manipulation Number	Number of object manipulations	Perceived Suitability	"The intervention possibilities were absolutely appropriate."
Usability	System Usability Scale (SUS) questionnaire [39]	Perceived Engagement	"I felt very engaged in the generation process."
User Experience	User Experience Questionnaire (UEQ) Short [40]	Temporal Intervention	"I had the feeling that I could intervene at any time."
		Perceived Effectivity	"I found the input method very effective."



(a) Garden view from left, distant corner.

(b) Garden view from right, close corner.

Figure 2. During the study, the participants were asked to generate a garden based on a garden template (1). They could view the template within the study scene (2). A) Shows a participant’s view of the built garden from one edge, B) from the opposite side.

5.6. Results

Quantitative Analysis For each variable, we first tested for homogeneity and normality. If met, we proceeded with an ANOVA for parametric data, else we first performed an Aligned Rank Transform (ART). Following, we only report significant results from post hoc tests with Bonferroni correction.

We measured a **manipulation number** of 13 (IQR=16.5) in all-at-once, 18 (IQR=16.0) in creation-before-manipulation, and 24 (IQR=10.0) in step-by-step. We investigate the effect of *control flow* on the manipulation number of our participants, and found a significant difference with a one-way ANOVA ($F(2, 52) = 10.10, p < .001, \eta_p^2 = .28$). We found a significant difference for spatial manipulation between step-by-step and all-at-once ($t(3.88) = -4.47, p < .001$), and between creation-before-manipulation and all-at-once ($t(3.88) = -2.64, p = .033 < 0.05$). We can conclude that all-at-once leads to a lower manipulation number than the step-by-step and creation-before-manipulation.

Our participants rated the **perceived control** of all-at-once with 5 (IQR=3.0), creation-before-manipulation with 6 (IQR=1.5), and step-by-step with 6 (IQR=1.5). With one-way ANOVA, we found a significant difference of *control flow* ($F(2, 52) = 3.88, p = .027 < .05, \eta_p^2 = .13$) on the participants’ perceived control of the generation. Post hoc analysis revealed a significant difference between step-by-step and all-at-once ($t(5.81) = -2.58, p = .038 < .05$). Thus, we can conclude that our participants perceived higher control using step-by-step than using all-at-once.

Qualitative Analysis Using the thematic analysis [41], two researchers analyzed the interview responses. We iteratively refined codes grouped them into topics and identified four main themes: Application Design and User Experience, Advantages/Disadvantages Input Modalities, Differences of Control Flows, and Use Cases of Generation.

Advantages/Disadvantages Input Modalities. Many participants (15) found voice input effective. However, challenges, such as the need for clear or loud statements without pausing during voice commands, often resulted in aborted or incorrect outcomes. A few noted task complexity, mentioning difficulties with longer keywords for positioning.

For the controller, feedback focused on its ease of use and button allocation. 6 participants found the spray gun easy to use, while 9 felt manual object positioning intuitive.

Application Design and User Experience. Many participants (10) enjoyed the application and found the VR environment creation engaging, "*The system was great to use. You felt like you could do whatever you wanted*" (P21). However, participants reported technical issues, especially with voice input, including objects not spawning or being adjusted as requested. They further had difficulties in selecting objects that were closely positioned with controller input. 6 participants wished for a different control flow, suggesting creating single objects with their appearance attributes using voice input followed by manual spatial manipulation.

Control Flow Differences. Although appreciating the idea, our participants expressed all-at-once required a high mental effort due to submitting long sentences. It required envisioning the entire scene before making the request, including object names, colors, and positions. Despite these challenges, most participants found all-at-once efficient. They felt it sped up the generation process by setting colors and placements in advance, reducing their overall effort, "*It was definitely the most efficient and easiest to use.*" (P21, All-at-once). Many participants (9) praised its innovation and expressed excitement when their requests were executed accurately. For step-by-step, several participants (7) found it less convenient, more time intensive, and less effective due to the manual adjustments. Although they perceived a higher workload for step-by-step, many participants (10) found it more reliable and predictable, resulting in higher felt control. One stated having a better overview of what he was doing, while another emphasized fewer errors. Some viewed creation-before-manipulation as having a higher workload. Others found it more efficient, facilitating adjustments while awaiting object generation, "*I could already say several things, I already let it work and continued with the next object or checked another*" (P19, Creation-before-manipulation). Due to making manual adjustments, a few participants perceived greater control, similar to step-by-step.

Use Cases of Generation. Discussing potential applications, many participants envisioned home-related uses like interior or garden design (19). Architectural, urban planning, and video game scene designs were also suggested, with several referencing the Sims™ game. Overall, participants saw potential in all control flows. For tasks demanding precision, they favored step-by-step for its control. Creation-before-manipulation was seen as efficient, saving time while allowing subsequent adjustments. For scenarios with time constraints, where the primary goal is not creation, they considered all-at-once for its time-saving and streamlined process.

5.7. Discussion

Following, we discuss the findings of our assessment. While the progress of generative AI might impact input modality selection, generation speed, and visual results, we consider control flows of virtual environment generation as largely unaffected.

All-At-Once: Efficiency in Time-Sensitive Tasks We hypothesized all-at-once to enhance usability by reducing user interactions, based on literature [42,43]. However, our findings did not support **H1.1**, though they did highlight its efficiency in reducing manipulations. Participants found manual adjustments in other methods lead to an increased workload, as supported by Kieras et al. [44], and appreciated all-at-once when time is prioritized over detail. Despite fewer manipulations, user experience scores did not in-

crease, potentially due to voice input's limitations and the mental effort of verbalizing concepts, supporting prior studies [6,45,7,9]. With no significant differences in usability, TCT, and task correctness across control flows, all seem suitable for generating virtual environments. Still, all-at-once's efficiency might benefit quick design iterations, like indoor or architectural planning, enabling faster testing of various designs.

Step-By-Step: Preferred for Fine-Granular Tasks We hypothesized that increased interactions in step-by-step might impair user experience by adding cognitive load [44] and extending adjustment times, negatively impacting user experience [43,42] and scene quality. Contrary, our findings did not support **H1.2** and **H1.3**. However, participants found that step-by-step offered enhanced control compared to all-at-once, particularly for precise spatial tasks, and noted reduced mental workload through iterative adjustments. Thus, all-at-once's drawbacks, particularly the cognitive demands and limited voice input affordance, reduced perceived control, in line with Myers et al. [46]. Further, participants preferred step-by-step for its ease of use and tactile feedback from controllers, aiding in precise object placement. Thus, it might be especially suitable for precise tasks where control and accuracy are primary, while time is less critical, such as fine-grained customization of gaming environments or room configurations for marketing.

Control Flow Variants Participants suggested alternative control flows, like combining voice input for object creation and controller input for spatial adjustments. Further, they wished for customizable processes based on task and personal preferences, in line with related work supporting adjustments of user interfaces to user preferences [47,48]. When used in adaptive systems, personal preferences could be recognized after some time and integrated directly into personal settings regarding the control flow. This finding further suggests exploring other mixed control flow combinations to enhance user satisfaction and efficiency while facilitating user adjustments to support their individual needs.

6. Conclusion

We investigated how humans can control the generation of virtual immersive environments in VR by assessing three control strategies (all-at-once, creation-before-manipulation, or step-by-step). To do so, we first designed a VR testbed. In a preliminary study, we utilized it to evaluate the input modalities (controller, hand, and voice input) for different generation tasks. Combining voice for object creation and controller for spatial and appearance manipulation was the most effective. Thus, we employed voice for creation and controller for spatial and appearance manipulation to evaluate the three control flows regarding their level of control over the generation process. We found that generating the environment all-at-once resulted in fewer object manipulations, which points to enhanced efficiency – a sentiment echoed by our participants. However, participants felt more control over the generation process when using step-by-step because the cognitive demands at each step were manageable, facilitating more fine-grain adjustments with fewer errors. Our findings suggest that the control strategy should consider the application's focus and the user's preference for efficiency or control.

Acknowledgements. This work is funded by the German Federal Ministry of Education and Research (01IS21068B).

Declaration of Conflicting Interests. The authors declare no conflicts of interest.

References

- [1] De Leon JDO, Tavas RP, Aranzanso RA, Atienza RO. Genesys: A Virtual Reality scene builder. 2016 IEEE Region 10 Conference (TENCON). 2016:3708-11. doi:10.1109/TENCON.2016.7848751.
- [2] Eroglu S, Stefan F, Chevalier A, Roettger D, Zielasko D, Kuhlen TW, et al. Design and Evaluation of a Free-Hand VR-based Authoring Environment for Automated Vehicle Testing. 2021 IEEE Virtual Reality and 3D User Interfaces (VR). 2021:1-10. doi:10.1109/VR50410.2021.00020.
- [3] Beever L, Pop S, John NW. LevelEd VR: A virtual reality level editor and workflow for virtual reality level design. 2020 IEEE Conference on Games (CoG). 2020:136-43. doi:10.1109/CoG47356.2020.9231769.
- [4] Barot C, Carpentier K, Collet M, Cuella-Martin A, Lanquepin V, Muller M, et al. The Wonderland Builder: Using storytelling to guide dream-like interaction. 2013 IEEE Symposium on 3D User Interfaces (3DUI). 2013:201-2. doi:10.1109/3DUI.2013.6550248.
- [5] McNeill D, editor. *Gesture and Thought*. University of Chicago Press; 2005. Available from: <https://www.degruyter.com/document/doi/10.7208/9780226514642/html>. doi:10.7208/9780226514642.
- [6] Bolt RA. "Put-That-There": Voice and Gesture at the Graphics Interface. (Proceedings of the 7th Annual Conference on Computer Graphics and Interactive Techniques. 1980:262-70. doi:10.1145/800250.807503.
- [7] Williams AS, Garcia J, Ortega F. Understanding Multimodal User Gesture and Speech Behavior for Object Manipulation in Augmented Reality Using Elicitation. IEEE transactions on visualization and computer graphics. 2020;26(12):3479-89. doi:10.1109/TVCG.2020.3023566.
- [8] Masurovsky A, Chojecki P, Runde D, Lafci M, Przewozny D, Gaebler M. Controller-Free Hand Tracking for Grab-and-Place Tasks in Immersive Virtual Reality: Design Elements and Their Empirical Study. *Multimodal Technologies and Interaction*. 2020;4(4):91. doi:10.3390/mti4040091.
- [9] Weiß Y, Hepperle D, Sieß A, Wolfel M. What User Interface to Use for Virtual Reality? 2D, 3D or Speech-A User Study. 2018 International Conference on Cyberworlds CW. 2018:50-7. doi:10.1109/CW.2018.00021.
- [10] Piumsomboon T, Altimira D, Kim H, Clark A, Lee G, Billingham M. Grasp-Shell vs gesture-speech: A comparison of direct and indirect natural interaction techniques in augmented reality. 2014 IEEE International Symposium on Mixed and Augmented Reality ISMAR. 2014:73-82. doi:10.1109/ISMAR.2014.6948411.
- [11] Lee M, Millingham M. A Wizard of Oz study for an AR multimodal interface. Proceedings of the 10th International Conference on Multimodal Interfaces. 2008:249-56. doi:10.1145/1452392.1452444.
- [12] Lee M. *Multimodal Speech-Gesture Interaction with 3D Objects in Augmented Reality Environments*. University of Canterbury. Department of Computer Science and Software Engineering; 2010. doi:10.26021/2223.
- [13] Wang M, Lyu XQ, Li YJ, Zhang FL. VR content creation and exploration with deep learning: A survey. *Computational Visual Media*. 2020;6(1):3-28. doi:10.1007/s41095-020-0162-z.
- [14] Ipsita A, Li H, Duan R, Cao Y, Chidambaram S, Liu M, et al. VRFromX: From Scanned Reality to Interactive Virtual Experience with Human-in-the-Loop. In: Kitamura Y, Quigley A, Isbister K, Igarashi T, editors. *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM; 2021. p. 1-7. doi:10.1145/3411763.3451747.
- [15] Liebers C, Pfitzenreuter N, Prochazka M, Megarajan P, Furuno E, Löber J, et al. Look Over Here! Comparing Interaction Methods for User-Assisted Remote Scene Reconstruction. *CHI '24 Extended Abstracts on Human Factors in Computing Systems*. 2024:1-8. doi:10.1145/3613905.3650982.
- [16] Wang J, Leach O, Lindeman RW. DIY World Builder: An immersive level-editing system. 2013 IEEE Symposium on 3D User Interfaces (3DUI). 2013:195-6. doi:10.1109/3DUI.2013.6550245.
- [17] Ichikawa S, Takashima K, Tang A, Kitamura Y. VR Safari Park: A Concept-Based World Building Interface Using Blocks and World Tree. Proceedings of the 24th ACM Symposium on Virtual Reality Software and Technology. 2018:1-5. doi:10.1145/3281505.3281517.
- [18] Seversky LM, Lijun Y. Real-Time Automatic 3D Scene Generation from Natural Language Voice and Text Descriptions. Proceedings of the 14th ACM International Conference on Multimedia. 2006:61-4. doi:10.1145/1180639.1180660.
- [19] Ma R, Patil AG, Fisher M, Li M, Pirk S, Hua BS, et al. Language-driven synthesis of 3D scenes from scene databases. *ACM Transactions on Graphics*. 2018;37(6):1-16. doi:10.1145/3272127.3275035.

- [20] Abrami G, Henlein A, Kett A, Mehler A. Text2SceneVR: Generating Hypertexts with VAnnotatoR as a Pre-processing Step for Text2Scene Systems. Proceedings of the 31st ACM Conference on Hypertext and Social Media. 2020:177-86. doi:10.1145/3372923.3404791.
- [21] Chang A, Savva M, Manning C. Semantic Parsing for Text to 3D Scene Generation. Proceedings of the ACL 2014 workshop on semantic parsing. 2014:17-21.
- [22] Goloujeh AM, Smith J, Magerko B. Explainable CLIP-Guided 3D-Scene Generation in an AI Holodeck. Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment. 2022;18(1):276-8. Available from: <https://ojs.aaai.org/index.php/AIIDE/article/view/21973>. doi:10.1609/aiide.v18i1.21973.
- [23] Liebers C, Megarajan P, Auda J, Stratmann TC, Pflingsthorst M, Gruenefeld U, et al. Keep the Human in the Loop: Arguments for Human Assistance in the Synthesis of Simulation Data for Robot Training. Multimodal Technologies and Interaction. 2024;8(3):18. doi:10.3390/mti8030018.
- [24] Creswell A, White T, Dumoulin V, Arulkumaran K, Sengupta B, Bharath AA. Generative Adversarial Networks: An Overview. IEEE Signal Processing Magazine. 2018;35(1):53-65. doi:10.1109/MSP.2017.2765202.
- [25] Dhariwal P, Nichol A. Diffusion Models Beat GANs on Image Synthesis. Advances in Neural Information Processing Systems. 2021;34:8780-94. Available from: https://proceedings.neurips.cc/paper_files/paper/2021/file/49ad23d1ec9fa4bd8d77d02681df5cfa-Paper.pdf.
- [26] Gu S, Chen D, Bao J, Wen F, Zhang B, Chen D, et al. Vector Quantized Diffusion Model for Text-to-Image Synthesis. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2022:10696-706.
- [27] Jain A, Mildenhall B, Barron JT, Abbeel P, Poole B. Zero-Shot Text-Guided Object Generation With Dream Fields. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2022:867-76.
- [28] Evirgen N, Chen XA. GANzilla: User-Driven Direction Discovery in Generative Adversarial Networks. Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology. 2022:1-10. doi:10.1145/3526113.3545638.
- [29] Nichol A, Jun H, Dhariwal P, Mishkin P, Chen M. arXiv, editor. Point-E: A System for Generating 3D Point Clouds from Complex Prompts. arXiv; 2022. doi:10.48550/arXiv.2212.08751.
- [30] Khundam C, Vorachart V, Preeyawongsakul P, Hosap W, Noël F. A Comparative Study of Interaction Time and Usability of Using Controllers and Hand Tracking in Virtual Reality Training. Informatics. 2021 Sep;8(3):60. Number: 3 Publisher: Multidisciplinary Digital Publishing Institute. Available from: <https://www.mdpi.com/2227-9709/8/3/60>. doi:10.3390/informatics8030060.
- [31] Juan MC, Elexpuru J, Dias P, Santos BS, Amorim P. Immersive virtual reality for upper limb rehabilitation: comparing hand and controller interaction. Virtual Reality. 2023;27(2):1157-71. doi:10.1007/s10055-022-00722-7.
- [32] Seaborn K, Miyake NP, Pennefather P, Otake-Matsuura M. Voice in human-agent interaction: A survey. ACM Computing Surveys (CSUR). 2021;54(4):1-43.
- [33] Whitlock M, Harnner E, Brubaker JR, Kane S, Szafir DA. Interacting with Distant Objects in Augmented Reality. In: 2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR); 2018. p. 41-8. doi:10.1109/VR.2018.8446381.
- [34] Unity Technologies. Unity3D; 2024. <https://unity.com/de>, last accessed on April 11, 2024.
- [35] Meta. Meta Quest 2; 2024. <https://www.meta.com/de/quest/products/quest-2/>, last accessed on April 11, 2024.
- [36] Meta. Interaction SDK; 2024. <https://developer.oculus.com/documentation/unity/unity-isdk-interaction-sdk-overview/>, last accessed on April 11, 2024.
- [37] Meta. Wit.ai; <https://wit.ai/>, last accessed on April 11, 2024.
- [38] Baccouri N. deep-translator; 2020. <https://deep-translator.readthedocs.io/en/latest/README.html>, last accessed on April 11, 2024.
- [39] Brooke J. SUS-A quick and dirty usability scale. Usability evaluation in industry. 1996;189(3):189-94.
- [40] Schrepp M, Hinderks A, Thomaschewski J. Design and evaluation of a short version of the user experience questionnaire (UEQ-S). International Journal of Interactive Multimedia and Artificial Intelligence, 4 (6), 103-108. 2017.
- [41] Braun V, Clarke V. Using thematic analysis in psychology. Qualitative Research in Psychology. 2006;3(2):77-101. doi:10.1191/1478088706qp063oa.
- [42] Nielsen J. Usability engineering. Morgan Kaufmann; 1994.

- [43] Benyon D. Designing interactive systems: A comprehensive guide to HCI, UX and interaction design. Pearson; 2014.
- [44] Kieras D, Polson PG. An approach to the formal analysis of user complexity. *International journal of man-machine studies*. 1985;22(4):365-94.
- [45] Schmandt C, Hulteen EA. The Intelligent Voice-Interactive Interface. *Proceedings of the 1982 conference on Human factors in computing systems*. 1982:363-6. doi:10.1145/800049.801812.
- [46] Myers C, Furqan A, Nebolsky J, Caro K, Zhu J. Patterns for how users overcome obstacles in voice user interfaces. In: *Proceedings of the 2018 CHI conference on human factors in computing systems*; 2018. p. 1-7.
- [47] Shneiderman B. Designing for fun: how can we design user interfaces to be more fun? *interactions*. 2004;11(5):48-50.
- [48] Greenberg S, Carpendale S, Marquardt N, Buxton B. *Uncovering the Initial Mental Model. Sketching User Experiences: The Workbook*. Elsevier: Hoboken, NJ, USA; 2012.