Analyzing Multimodal Interaction Strategies for LLM-Assisted Manipulation of 3D Scenes

Junlong Chen*
University of Cambridge

Jens Grubert †
Coburg University of Applied Sciences

Per Ola Kristensson ‡
University of Cambridge

2025 IEEE Conference Virtual Reality and 3D User Interfaces (VR) | 979-8-3315-3645-9/25/$31.00 ©2025 IEEE | DOI: 10.1109/VR59515.2025.00045

Figure 1: Example workflow for scene editing with our proposed ASSISTVR technique. Left: The user adopts the *bulk modification* strategy to select all blue objects in the original scene to modify their appearance together. Middle: The user adopts the *incremental exploration* strategy to modify the appearance on individual objects. Right: The final scene matches the target scene of one of the tasks in our empirical user study.

ABSTRACT

As more applications of large language models (LLMs) for 3D content in immersive environments emerge, it is crucial to study user behavior to identify interaction patterns and potential barriers to guide the future design of immersive content creation and editing systems which involve LLMs. In an empirical user study with 12 participants, we combine quantitative usage data with post-experience questionnaire feedback to reveal common interaction patterns and key barriers in LLM-assisted 3D scene editing systems. We identify opportunities for improving natural language interfaces in 3D design tools and propose design recommendations. Through an empirical study, we demonstrate that LLM-assisted interactive systems can be used productively in immersive environments.

Index Terms: Virtual reality, large language models, 3D scene editing.

1 INTRODUCTION

Large Language Models (LLMs) have gained popularity in assisting task completion in immersive environments. LLMs provide various advantages to improve interaction experience in virtual and augmented reality, such as improving task completion efficiency [36], democratizing VR content creation for non-expert users [15], and improving expressiveness while reducing the user's perceived workload [26]. However, the introduction of LLMs in interaction tasks such as scene editing can also pose barriers and adversely affect the interaction experience due to the current limitations of LLMs and its capability to integrate with 3D scene content. Examples of these barriers include transparency and explainability [29] reflected through user trust in the system, as well as appropriate error handling and timely user feedback [15].

*e-mail: jc2375@cam.ac.uk
†e-mail: jens.grubert@hs-coburg.de
‡e-mail: pok21@cam.ac.uk

We investigate the following research questions:

- **RQ1:** What common interaction patterns do users exhibit when they adopt an LLM-assisted multimodal interactive system to complete interaction tasks such as scene editing in VR?
- **RQ2:** Do LLM-assisted multimodal interactive systems pose interaction barriers and how do users work around them?

As speech-and-pointing interfaces have been widely studied in existing virtual and augmented reality research [7, 49, 50, 24], we have created the Advanced Speech Support and Interactive System for Virtual Reality (ASSISTVR), which integrates LLMs with speech and pointing interaction techniques. This system serves as an example for multimodal interactive systems in general and informs our understanding of the above two research questions. ASSISTVR uses an off-the-shelf Microsoft Azure Conversational Language Understanding (CLU) Service and GPT-4o to handle user queries. We use this system to study the effects of LLMs on user behavior patterns in scene editing tasks through an empirical user study with 12 participants. Specifically, we focus on whether user interaction with such LLM-assisted interactive systems reveal certain high-level *interaction strategies*, and if so, was the LLM-assisted interactive system able to assist participants in identifying more efficient interaction strategies with very limited external guidance. We also examine whether the system poses any *interaction barriers*, and suggest design approaches to overcome these. Through this study, we extract observations on user performance and interaction patterns, and provide design implications for future LLM-assisted interactive systems for immersive 3D content and possibly general interactive systems which involve LLMs.

This paper contributes to existing literature by analyzing interaction patterns and strategies through an *exploratory* study. We deliberately chose not to engage in a comparative study since prior work [11, 15, 26] have proposed systems which apply LLMs to immersive content, and the capabilities of LLMs advance in a very rapid speed. Instead of making a technical contribution, this paper provides insights on observed user strategies and behavioral patterns, which generalizes to different types of interactive systems involving LLMs.

2642-5254/25/$31.00 ©2025 IEEE
DOI 10.1109/VR59515.2025.00045

2 Related Work

2.1 Scene Editing and Multimodal Interaction in XR

Rakkolainen et al. [35] reviews recent advances in multimodal interaction technologies for extended reality (XR) content, pointing out how XR technologies introduce new interaction concepts and play an important role in addressing accessibility barriers. Similar views were proposed by Spittle et al. [41], who suggests that multimodal interaction facilitates selection and manipulation tasks. In 3D editing tasks in virtual reality (VR), a combined gesture and speech interface can perform on par with a unimodal interface of a radial menu in terms of promoting creativity, usability, and presence [50].

Williams et al. [44] report on an elicitation study of speech, gesture, and multimodal speech and gesture interactions in unconstrained object manipulation tasks in augmented reality. Zhou et al. [49] found that participants preferred to use the same gesture for one and two-object manipulation in the same task, and revealed associations between speech patterns and gesture strokes during 3D object manipulation. Rodriguez et al. [37] studied natural unimodal and multimodal interaction techniques for 3D sketching in virtual reality.

Plopski et al. [33] reviewed gaze interaction and eye tracking research in XR and outlined how eye gaze has been applied to enhance user interaction with virtual content and interface design. Multimodal interactive systems such as GazePointAR [27] also demonstrate the possibility of leveraging eye gaze and pointing gestures to provide contextual information for speech queries.

2.2 Large Language Models for Extended Reality

A plethora of recent research in AI and XR has focused on different aspects, including accessibility and inclusion [23, 8], privacy [8], 3D content generation [19, 28], and general applications [15, 31, 11, 40]. Ma et al. [30] reviews integration of LLMs with 3D spatial data as 3D-LLMs and applications. Recent work [20, 21, 3] has further explored how LLMs can assist agents in altering the physical 3D world in various ways.

In terms of 3D content editing, LLM-assisted systems such as LLMR [11] demonstrate a wide range of possible applications in XR, including world creation, multimodal interaction, scene editing, scene query, and integration with other external plugins, platforms, and sensors. DreamCodeVR [15] is an AI-powered tool for generating code in VR applications during runtime to modify the appearance and behavior of elements in a 3D scene. Prior work has also studied LLM prompting for immersive content. Roberts et al. [36] show that prompt-based methods can accelerate in-VR level editing and become an integrated part of the gameplay. Aghel Manesh et al. [2] used a Wizard of Oz elicitation study to examine the implicit expectations of users when they prompt generative AI agents to create interactive virtual scenes.

2.3 Interaction Pattern Analysis

Interaction analysis is an important part of human-computer interaction (HCI) research. Wright et al. [45] proposed the resources model to analyze human-computer interaction as distributed cognition, where interaction strategies play a crucial role in bringing resources in use to generate actions. Scholz et al. [38] proposed a model to study user behavior and interaction patterns in online news forums while Guo et al. [17] studied interaction modes and user agency in human-LLM collaboration tasks. Beyan et al. [6] conducted a human-human interaction analysis and identified interaction patterns and behaviors such as nonverbal cues which resulted in effective performance. These interaction patterns are often uncovered through log analysis [42] or audio and video analysis [22].

Interaction patterns have also been studied within the context of extended reality. To support the analysis of interaction patterns, symbolic event visualization methods have been proposed by Rabasahl et al. [34]. Feit et al. [13] and Foy et al. [14] studied ten-finger typing on a physical keyboard and mid-air typing in virtual reality respectively, and summarized common typing behaviors as interaction patterns. Dudley et al. [12] studied the performance envelopes of four alternative text input strategies in virtual reality to provide design implications for novel text entry systems.

3 Methodology

LLM systems have evolved from text-based interaction [10] to vision-language models [43], which support multimodal text and images, to general-purpose multimodal LLMs [46] that support any combination of text, image, video, and audio as inputs and outputs. For immersive 3D environments, while multimodal interactive systems assisted by LLMs have been proposed [11, 25, 15], there is still need to investigate their effects on user behavior and interaction patterns. We have designed AssistVR to provide an integrated speech-and-pointing 3D editing system.

Through a scene editing user study, we gather quantitative usage data and qualitative feedback from post-experience questionnaires. The study is approved by the research ethics committee in the Department of Engineering at the University of Cambridge. Collectively, these results help us to identify main interaction strategies as well as reoccurring interaction patterns. Through post-hoc analysis of the study data, we identify key barriers in user interaction with LLM-assisted interactive systems in virtual reality and propose design implications for future LLM-assisted interactive systems.

Apparatus. To study user behavior and patterns when interacting with LLM-assisted 3D scene editing systems, we designed AssistVR. An outline of the system workflow is provided in Figure 2. The system leverages large language models such as Microsoft Azure Conversational Language Understanding (Azure CLU, as shown in grey) and GPT-4 Omni (GPT-4o, as shown in blue) [32, 1] to process user natural language input. Azure CLU extracts intents and key entities from the user natural language input, and GPT-4o catches all exceptions which cannot be handled by the Azure CLU classifier to provide speech instructions to the user.

In the 'Training Phase' of Azure CLU, representative utterance data of possible user speech input samples are labelled with intents (such as 'Select', 'Deselect', 'Modify', 'Undo', or other intents) and key entities (such as 'Object of Interest', 'Original Color', 'Original Material', 'Target Color', and 'Target Material'), and are used to finetune the default model (2022-09-01 training configuration) provided by the Azure CLU service. Upon training the model, the utterances with labeled intents and entities are adjusted to iteratively improve model performance. The final model with an F1 score of 92.73% on intent classification was deployed. There is no training phase for GPT-4o.

In the 'Deployment Phase' of Azure CLU and GPT-4o in Figure 2, the system first uses the Azure Speech Recognition service to recognize user speech, then uses the Azure CLU model to classify the recognized speech input into different intents and extracts key entities from the user input. If the intent is classified as 'Select', 'Deselect', 'Modify', or 'Undo', the system executes post-processing scripts in Unity to perform object selection, object deselection, color/material modification, or actions to undo the previous color/material editing step. Following De La Torre et al. [11], the system generates a scene graph in JSON format to represent content in the 3D scene. If the intent does not fall under these four categories, the user natural language input, an instructions prompt (providing context about the scene editing task, available colors, and available materials), and a JSON file containing the scene graph of the current 3D scene are sent to the cloud-based GPT-4o model via Application Programming Interface (API) calls. GPT-4o subsequently generates a natural language response, which is then synthesized into speech and sent to the user.

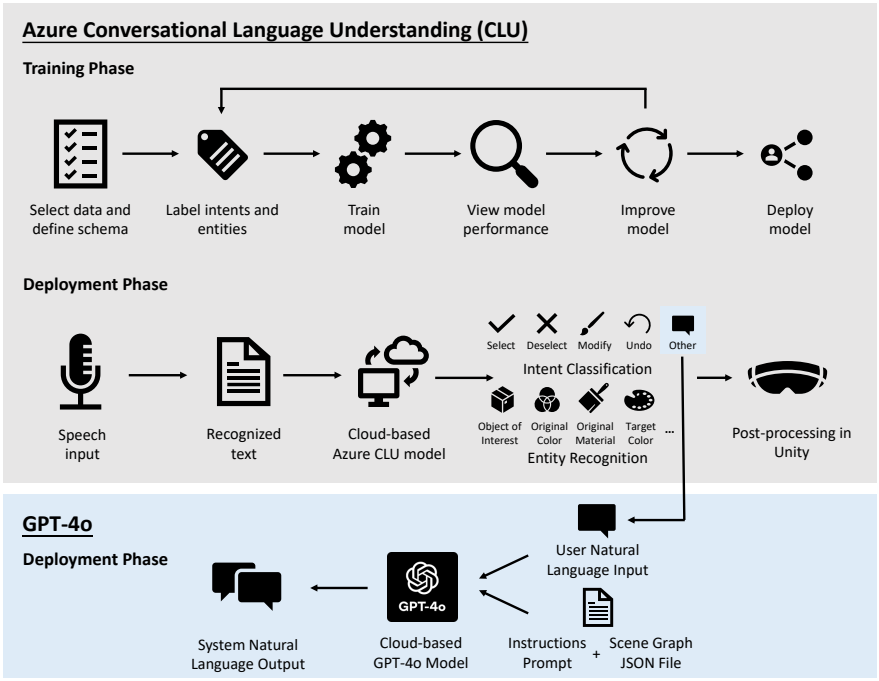

Figure 2: Workflow of the ASSISTVR system designed for the study. In the training phase, only Azure Conversational Language Understanding (CLU) is involved. The developer labels a number of utterances with intents and entities, and finetunes the Azure CLU model. The model is iteratively improved based on performance metrics. In the deployment phase, Azure CLU classifies user speech input into different intents. If the intent falls under the 'Select', 'Deselect', 'Modify', or 'Undo' categories, further post-processing steps to modify the scene are conducted in Unity. If the intent does not fall under these categories, the user speech input and a text file containing the instructions prompt and scene graph of the current scene are sent to GPT-4o, which generates a natural language response synthesized into speech for the user.

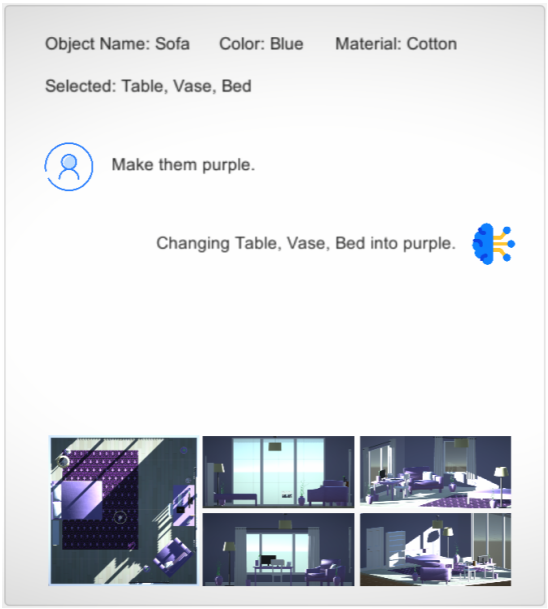

Figure 3: Example of the draggable panel. The panel shows that the current object hit by the raycast is the 'Sofa' with 'Blue' color and 'Cotton' material. Currently, objects 'Table', 'Vase', and 'Bed' are selected. The user says, "Make them purple." The system responds, "Changing Table, Vase, Bed into purple," and modifies the color of the selected objects. At the bottom of the panel, screenshots of the target scene at different angles are shown.

Apart from the speech-based interaction powered by Azure CLU and GPT-4o, the system also includes other interaction modalities including selection/deselection of virtual objects via raycast, and a draggable panel to provide feedback by displaying from top to bottom the name, color, and material property of the current object hit by the right raycast, the list of names of all currently-selected objects, the recognized user speech input, the natural language output from the system, as well as screenshots of the target scene at the bottom. An example screenshot of the draggable panel is provided in Figure 3.

Participants wore an Oculus Quest 2 headset and held the right controller during the study. The headset was connected to a Windows 10 laptop PC (Intel i5-9300H CPU, 16GB memory, and GTX 1050 graphics card) via an Oculus link cable. Scenes were implemented with Unity 3D (Version 2022.3.15f1) and publicly available resources[1] on Unity Asset Store.

Participants. We recruited 12 participants (6 male and 6 female) aged between 22 and 35 ($M = 26.3, SD = 3.8$). Around 62.5% of participants were familiar with VR, 50% of participants were familiar with speech recognition systems, and around 67% of participants were familiar with 3D modelling or design software. All participants understood and spoke English, and 50% reported themselves as native English speakers. None of the participants reported any form of disability.

Task. The task involves matching the original indoor scene to a target scene based on a combination of natural language instructions and image instructions. Here, this scene editing task was chosen because referencing tasks and object manipulation tasks are

[1]Source: https://assetstore.unity.com/packages/3d/environments/interior-house-assets-urp-257122.

considered as canonical interactions commonly used to evaluate interaction techniques in VR [4, 5, 47], and both tasks are encapsulated in our editing task. The scene and task are designed such that there are a number of objects which can be referenced with a common color/material property (such as 'blue' or 'cotton'), and one special object (the carpet) whose target appearance can only be inspected visually by the user. The user is not instructed how to reference the pattern of the carpet verbally at the beginning of the task. The purpose of including this special object in the task is to simulate cases when objects are difficult to reference for the user and to study whether LLM-assisted systems can aid participants in referencing these objects with higher perplexity.

Task type A involves making all blue objects in the original scene into grey and making all cotton objects in the original scene into leather, see Figure 1 (right) for the target scene. Task type A also involves editing the carpet into white pattern, but this requirement was given by showing images instead of natural language descriptions. Similarly, task type B involves making all blue objects purple, making all leather objects cotton, and making the carpet into purple pattern, see subfigure "T: Target" in Figure 5 for the target scene. The target pattern of the carpet was also given via images and not natural language. During the study, the order of task type A and task type B was counterbalanced across all 12 participants. After rearranging the order of task type A and task type B, the tasks were delivered as Task 1 and Task 2, with Task 1 preceding Task 2.

Additional instructions for participants during Task 1 were to explore the list of all available colors and materials and to find the most efficient way to modify color and material. An additional instruction for Task 2 was to modify the scene based on the most efficient way participants found in Task 1. This reason for providing these additional instructions is because we are interested in finding whether LLM-assisted systems help participants obtain any performance improvement, and if so, how the improvement is reflected through the change in interaction strategies and patterns. The complete verbal instructions for both tasks and both task types are provided in the Online Appendix.

Procedure. After filling out a consent form and demographics questionnaire, participants were briefed about the study procedure, which involved asking participants to edit the color and material of various objects in a VR living room scene. Participants were told that the system supported a list of colors such as red, orange, and yellow and a list of materials such as plastic. Here, only a few examples were given, and the complete list of colors and materials were not given to the user. Users were briefed about the main functions (raycast selection, speech, and assistive panel) of the system, as well as a high-level introduction of the types of supported speech commands (Select/Deselect, Commands to modify appearance, Commands to undo, and Query commands). Participants were not taught about the exact phrases used to elicit these commands. Participants were instructed to think aloud during the study. Prior to starting the two tasks, participants were given some time to familiarize themselves with the system in a practice trial. Participants were encouraged to try using speech and raycast to select and edit the color and material of a few objects, or try asking some questions on system usage and the current status of the scene.

After the practice trial, the goal of Task 1 was introduced to participants by showing them images of the original scene and target scene to match the carpet appearance and by giving them verbal instructions. All participants were exposed to both task type A and B in Task 1 and Task 2.

In Task 1, participants further explored the scene and attempted the editing task. Participants also tried to figure out how to use the system efficiently, including which functions to use and what speech commands worked well. Participants were not allowed to obtain additional information from the study moderator but were allowed to ask the system. The task ended when participants were satisfied that the scene matched the target appearance. Participants were then asked to complete a post-experience questionnaire including open-ended questions, SUS [9], NASA-TLX [18], and UEQ-S [39] questionnaires for Task 1 and take a 5-minute break.

Next, participants were given instructions for Task 2, which included task type A or B, as well as modifying the scene based on the most efficient way they found in Task 1. After receiving the instructions for Task 2, participants modified the scene and stopped when they were satisfied that the scene matched the target appearance. Participants completed a similar post-experience questionnaire for Task 2, and gave final comments on which features they liked/disliked based on their experience throughout the entire study. The entire study lasted for around an hour. At the end of the study, participants were thanked for their participation and remunerated.

4 Results

Observations and quantitative data from the user study revealed several common patterns in user behavior. These findings are organized and presented below as overall performance, interaction patterns and interaction barriers. Here, significance tests do not serve to conduct comparisons between different system or interfaces, but instead serve as a tool to indicate how well users can learn to use the system over time.

4.1 Overall Performance

Task Completion Quality. As different participants achieved different goal states which match the target scene appearance to different extents, we consider the difference between the color and material of all objects in the current scene and the color and material of all objects in the target scene as the number of **Remaining Elemental Editing Steps (REES)**, a metric to quantify user progress and task completion quality in the scene editing task. Mathematically, it is defined as:

$$REES = \sum_{i=1}^{n} \left[\mathbb{1}(o_{i(c)} \neq t_{i(c)}) + \mathbb{1}(o_{i(m)} \neq t_{i(m)}) \right], \qquad (1)$$

where $\mathbb{1}(x \neq y)$ is the indicator function. It satisfies $\mathbb{1}(x \neq y) = 1$ if $x \neq y$ and $\mathbb{1}(x \neq y) = 0$ if $x = y$. o_i represents object i in the scene at the current state, and t_i represents the target appearance of object i at the final target state. The parameter c represents the color property, while m represents the material property. There are in total n objects in the scene.

This final REES metric measures how close the final state of the scene is compared to the target scene, with a lower final REES value indicating a closer match to the target scene and higher task completion quality. A Wilcoxon Signed-Rank test revealed a **significant difference ($W = 3, p < .05, |r| = .73$) in the final REES** between TASK1 ($M = 4.58, SD = 4.72$) and TASK2 ($M = 1.83, SD = 3.69$), suggesting that participants were able to match the target scene significantly better in Task 2 compared with their performance in Task 1. Please note, that while it can be expected that participants' performance improves over time, the scale of this improvement (60.0% reduction in final REES on average) can indicate that users can adopt quickly to the multimodal editing system.

Task Completion Time. Another measure for task completion is the time taken for each participant to edit the scene to match the target appearance. A Wilcoxon Signed-Rank test revealed a **significant difference ($W = 3, p < .05, |r| = .87$) between the completion time** of TASK1 ($M = 11.2$ minutes, $SD = 4.9$) and TASK2 ($M = 5.7$ minutes, $SD = 4.0$), suggesting that participants completed Task 2 in a significantly shorter amount of time. We are interested in task completion time as this metric helps to inform us on the efficiency in carrying out the tasks.

Combining the results for task completion quality and task completion time, we observe that participants were able to match the

Table 1: Participant comments organized by themes in the post-experience questionnaire.

Theme	Sub-Theme	Participant Comments
System Ease of Use	Response efficiency	*"It was incredibly quick. I like how efficient it was. I did not need to select anything, which made it really easy. I just told the system what to do and only had to use 5 commands."* (P6)
	Design redundancies	*"I very much liked the technique. The speech recognition is much faster than selecting objects with raycast, but I can still use the raycast to check the object properties."* (P8)
	Straightforward to use	*"It was efficient and straightforward to use once the commands were known."* (P11)
	Help and support	*"I like how in the end I asked the system how can I change the color & pattern of the carpet and it shows many examples of the exact commands that I could say. And I tried to communicate with it and complete the task in the end and I feel like the examples that the system gave was helpful."* (P12)
Multimodal Interaction	Benefits of raycast	*"Raycast enabled precision control, when I want to select, deselect a specific object that I did not know its natural name."* (P7)
	Benefits of speech	*"The speech recognition is much faster ... but I can still use the raycast to check the object properties."* (P8)
	Other possible modalities	*"Maybe if the system is also able to factor in my gaze or selections to provide further context. E.g., if I am looking at a particular book and/or selected it, the voice command should factor this in."* (P7)
User Agency	Unexpected response	*"I only dislike it when I find the system not responding the way I thought it would. E.g., I selected a particular book and asked it to change book to purple, but the system changed all the books."* (P7)
	User adaptations	*"Try to avoid complicating the system ... I prefer to give short and clear instructions and complete the task step by step."* (P4)
User Trust	Ways of establishing trust	*"Before starting the task, I saw visually that the system was changing colors correctly for an object I had selected. I trusted it to select all objects of a certain color at once because it seemed to correctly know the color of every object. Bulk-selecting and changing seemed trustworthy and the most efficient."* (P6)
	Cases of a lack in trust	*"If things are identical, I feel confident multiselecting them using voice command. However, if things are not entirely identical, I feel better selecting and changing them one by one, so that I don't mis-select any item that I didn't mean to."* (P2)
Level of Feedback	Visual feedback via panel	*"The panel shows what I said and what the response will be, so I can confirm if the recognition is correct or not and try again if the recognition is wrong."* (P8)
	Other visual feedback	*"It would be nice to view a sample of the colours when listing them - especially for the patterns."* (P1)

scene significantly closer to the target scene in a significantly shorter amount of time in Task 2 after familiarizing with the system in Task 1. High standard deviations in the results also suggest that different individuals can have a vastly different performance.

4.2 Post-Experience Questionnaire Findings

Participants provided task load ratings on mental demand, physical demand, temporal demand, performance, effort, and frustration from a scale of 1 to 10 using the unweighted version of the NASA-TLX questionnaire [18]. A Wilcoxon signed rank test revealed that the **overall task load rating of TASK1 ($M = 3.97, SD = 1.08$) was significantly higher ($W = 4, p < .05, |r| = .78$) than that of TASK2 ($M = 3.35, SD = .98$)**.

Results from the System Usability Scale (SUS) [9] of TASK1 and TASK2 yielded a higher average SUS score in Task 2 compared with Task 1. However, a Wilcoxon signed rank test did not reveal a significant difference ($W = 9.5, p = .073, |r| = .57$) between the SUS ratings of TASK1 ($M = 72.1, SD = 15.5$) and TASK2 ($M = 75.2, SD = 14.9$).

Results from the short version User Experience Questionnaire (UEQ-S) [39] show that TASK2 attains a higher average overall score ($M = 1.70, SD = .81$) compared with TASK1 ($M = 1.50, SD = .59$), but Wilcoxon signed rank tests did not reveal a significant difference ($W = 13, p = .083, |r| = .52$). For the subcategories of the UEQ-S ratings, no significant differences were found in the PRAGMATIC quality ($W = 14.5, p = .199, |r| = .41$) between TASK1 ($M = 1.50, SD = .80$) and TASK2 ($M = 1.79, SD = 1.09$) or the HEDONIC quality ($W = 0, p = .089, |r| = 1.0$) between TASK1 ($M = 1.50, SD = 1.06$) and TASK2 ($M = 1.60, SD = 1.04$).

Participants also provided descriptions of the most efficient strategy they found, as well as open comments about the system. Ten out of twelve participants were able to find an efficient strategy of bulk-editing object properties by interacting with the system without additional external assistance by the end of the study. Participants who did not find the bulk modification strategy described their strategy as follows. P2 said, "For identical items such as blue walls, blue vases and leather pillows, I tend to use voice command to change their colours/material...For non-repetitive items such as the pen holder and keyboard, I just selected and changed them individually one by one." Meanwhile, P12 commented, "Because I found selecting multiple objects at the same time [being] troublesome, I directly ask[ed] the speech system to help chang[e] the color of multiple objects." These two participants either found it more reassuring to change individual objects (P2) or did not find the supported command or workflow to select multiple objects with the same property first and then use another command to modify the appearance of all selected objects (P12). A thematic analysis [16] on open comments about the system revealed the following trends, which provide further insights on the observed behaviors and interaction patterns. For the following themes, examples of quotes from individual participants are provided in Table 1, while participant conversation histories are provided in the Online Appendix.

System ease of use. Participants appreciated how easy and efficient it was to complete the scene editing task once they knew how to phrase the commands and which strategy to adopt. Dialogues between participants and ASSISTVR revealed that participants found the system useful in providing suggestions to help them find the list of all supported colors, materials, supported speech commands, as well as the efficient bulk modification strategy.

Multimodal Interaction. In the post-experience questionnaire, participants appreciated how different interaction modalities including speech and raycast worked together to facilitate scene editing tasks. Some participant comments also demonstrated that multimodal interaction is not limited to speech and raycast but can instead incorporate a broader range of interaction modalities in future systems.

User Agency. Participants commented how sometimes the system did not respond to speech input as they expected, which negatively affected their sense of agency over the system. During the study, participants were aware of gaining control of their actions and sought to improve agency by choosing appropriate strategies.

User Trust. Post-experience comments revealed that some participants sought simple ways to verify that the system was processing speech commands correctly before bulk-editing the scene. The

difference in user trust in the system also likely led them to choose different interaction strategies.

Level of feedback. Participants appreciated how the system provided an adequate amount of visual feedback via the draggable panel and voice feedback through synthesized speech (see quotes under 'System ease of use'). Participants also commented how it would be helpful if they received more visual feedback on the list of colors and materials, in addition to their names.

4.3 Interaction Patterns

The study revealed how participants preferred to iteratively modify the color and material of individual objects in the scene to match the target appearance, or to select a group of objects with a common feature, and change their color/material using a single voice command. Figure 4 plots the number of remaining elemental editing steps for all 12 participants with respect to elapsed time.

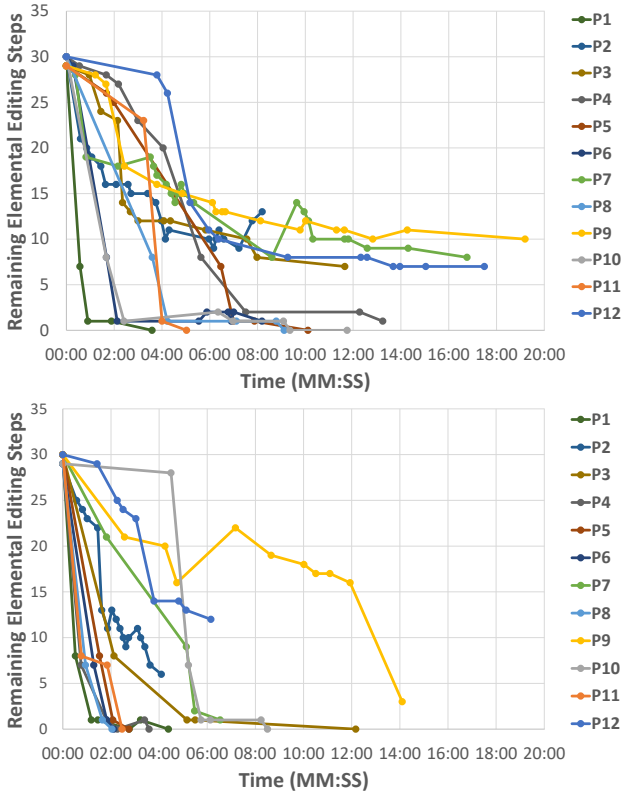

Figure 4: Number of remaining elemental editing steps to match the target scene in Task 1 (top figure) and Task 2 (bottom figure). The horizontal axis is denoting relative time in minutes and seconds.

As shown in Figure 4, in Task 1, in which participants are asked to find the most efficient way to edit the scene to match the target appearance, P2, P3, P7, P9, and P12 preferred to make incremental edits to individual objects. Similarly in Task 2, in which participants are asked to edit the scene based on the most efficient method they found, P2, P9 and P12 also preferred to modify the scene iteratively. We define this high-level scene editing strategy as:

Incremental Exploration (IE) This strategy emphasizes visual inspection of individual object properties and combines raycast selection or speech selection of single objects by their names and modifying object appearance using speech commands, or direct modification (without explicit selection) of individual object appearance through speech commands.

In Task 1, P1, P6, P8, P10, P11 used speech commands to select a group of objects via their common color or material property and used a single voice command to bulk edit their appearance. This strategy is also observed in Task 2 within the behavior of more participants including P1, P3, P4, P5, P6, P7, P8, P10, and P11. We define this high-level interaction strategy as:

Bulk Modification (BM) This strategy uses speech to select a group of objects with a shared color/material property, then uses speech to bulk modify their appearance. In this strategy, there is not explicit involvement of visual inspection of individual object properties.

It is important to note that the interaction strategy of a certain user can change over time. For example in Task 1, P4 started the task with *incremental exploration*, then adopted *bulk modification*, before returning to *incremental exploration*. Therefore, we visualize how interaction strategies have changed (if any) over the course of time in Task 1 and Task 2 for each participant in Figure 5. Based on these interaction patterns, we make the following observations:

Color modification tends to precede material modification in IE and BM. In Task 1, among all 12 participants, 7 edited color before editing material (P1, P3-P6, P8, P10) while none edited material before editing color. The remaining 5 participants did not exhibit a strong preference on editing a certain property before another (P2, P7, P9, P11, P12). In Task 2, 8 participants edited color before editing material (P1, P3-P6, P8, P11, P12) while none edited material before editing color. The remaining 4 participants did not exhibit a strong preference on the editing sequence (P2, P7, P9, P10). This trend in editing sequence regardless of the high-level strategy employed demonstrates how the majority of participants drew attention to the more distinguishable visual features such as object colors and edited these features first in comparison with less distinguishable visual features such as object material.

Carpet tends to be edited last. Prior to the study, participants were instructed to match the appearance of the carpet to the appearance shown in an image of the target scene. Participants were not explicitly told how to modify the carpet appearance, or how to reference the target appearance of the carpet. In comparison, the remaining objects were given an explicit target color (grey or purple). The carpet represents objects which are difficult to edit verbally, and the study results revealed that in Task 1, 9 out of 12 participants (P1, P4–P9, P11, P12) chose to edit the carpet last. In Task 2, 9 out of 12 participants (P1, P3–P8, P10, P12) edited the carpet after editing the remaining objects. The results show that in speech-based interfaces, users tend to edit objects with a clear goal state such that the speech commands are easy to enunciate.

Total time spent on incremental exploration tends to exceed the time spent on bulk modification. Figure 6 (left) provides a box plot of the time spent on incremental exploration and the time spent on bulk modification for all 12 participants. Wilcoxon Signed-Rank tests indicate that **the total minutes spent on the incremental exploration strategy** ($M = 9.47, SD = 5.22$) **in Task 1 is significantly greater** ($W = 2, p < .05, |r| = .92$) **than the minutes spent on the bulk modification strategy** ($M = 1.78, SD = 1.62$). For Task 2 however, the difference between the time taken on incremental exploration ($M = 3.38, SD = 2.87$) and bulk modification ($M = 2.36, SD = 1.70$) was not significantly different ($W = 28, p = .424, |r| = .23$).

A larger percentage of time was spent on Bulk Modification in Task 2 compared to Task 1. Figure 6 (middle) provides box plots of the percentage of time spent on incremental exploration and bulk modification for each participant. Wilcoxon Signed-Rank tests indicate that **the percentage of time spent on bulk modification for each participant significantly increased** ($W = 2, p < .05, |r| =$

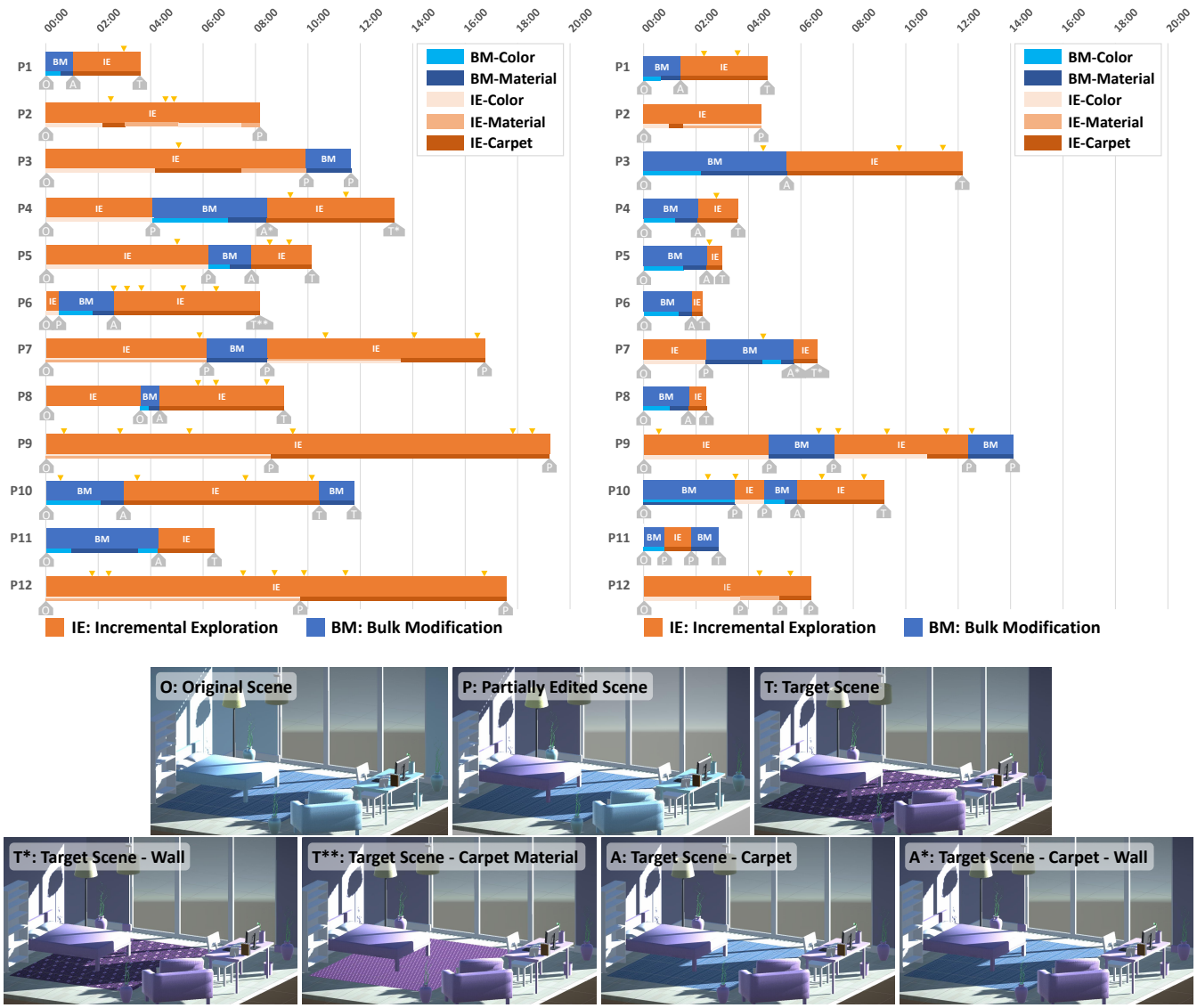

Figure 5: Interaction strategies adopted by different users across the duration of Task 1 (top left) and Task 2 (top right). Time on the horizontal axis is displayed in the format MM:SS (minutes:seconds). Triangles above the timeline of each participant indicate user queries. The main timeline bar for each participant indicates the high-level strategy employed (IE: Incremental Exploration, or BM: Bulk Modification). The secondary timeline bar below the main timeline bar indicates the low-level strategy employed, namely Bulk Modify Color (BM-Color), Bulk Modify Material (BM-Material), Color Editing with Incremental Exploration (IE-Color), Material Editing with Incremental Exploration (IE-Material), or Carpet Editing with Incremental Exploration (IE-Carpet). Grey tags below the timeline bars represent the current scene status (O: Original scene. T: Target scene. P: Partially-edited scene.) The target scene can be the grey scene in Figure 1 (right), or the purple scene shown here. Among partially-edited scenes, some scenes occur frequently and are labelled explicitly. These include: T*: In addition to T, one of the walls received an extra edit, REES=1. T**: In addition to T, the carpet material is incorrect, REES=1. A: Color/material changed for all objects except the carpet, REES=1. A*: In addition to A, one of the walls received an extra edit, REES=2. Part of the secondary timeline for P8 in Task 1 is blank because only selections instead of edits occurred. Examples of these scenes are provided below the timeline.

.81) **in Task 2** ($M = .465, SD = .279$) **compared with Task 1** ($M = .194, SD = .190$). This suggests that users are likely to have learned about the efficiency of the bulk modification strategy and prefer to spend more time on it.

More queries were posed during IE. Figure 6 (right) provides box plots of the number of queries posed during incremental exploration ($M = 3.17, SD = 2.08$) and bulk modification ($M = .08, SD = .29$) for all participants in Task 1 and the number of queries posed during incremental exploration ($M = 1.25, SD = 1.36$) and bulk modification ($M = .42, SD = .67$) for all participants in Task 2. Wilcoxon Signed-Rank tests indicate that **significantly more queries were posed during incremental exploration, as compared to bulk modification in both Task 1** ($W = 0, p < .05, |r| = .87$) **and Task 2** ($W = 2.5, p < .05, |r| = .76$).

Queries did not necessarily guide participants to find the BM strategy. P2, P9, and P12 who did not try the bulk modification strategy in Task 1 posed 3, 6, and 7 queries respectively, but only P9 shifted to a combination of the incremental exploration strategy and bulk modification strategy in Task 2, while P2 and P12 continued with the incremental exploration strategy and were not successful in matching the scene to the target appearance.

Participants who tried BM in Task 1 achieved high performance in Task 2. P1, P3 to P8, P10, and P11 tried the bulk mod-

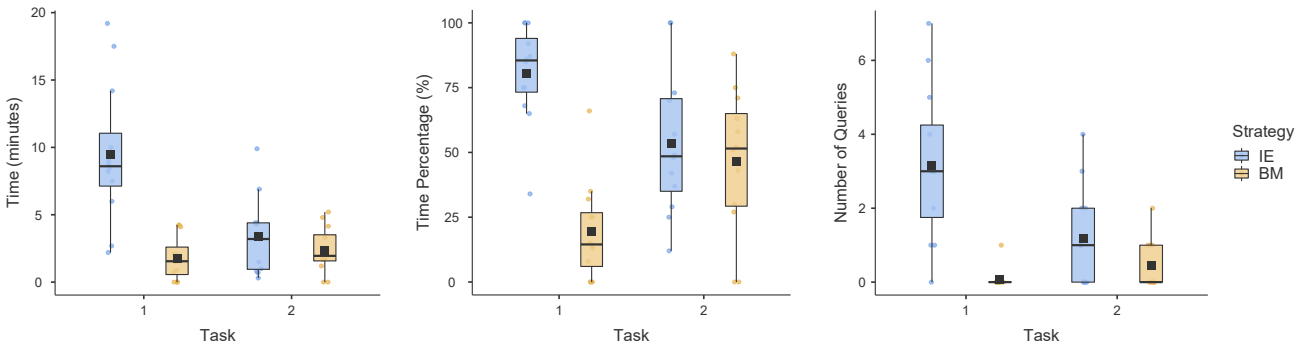

Figure 6: Box plots of the time spent in minutes for all participants on the incremental exploration (IE) strategy and the bulk modification (BM) strategy (left), the percentage of time spent on both strategies for all participants in Task 1 and Task 2 (middle), and the number of queries posed during both strategies for all participants in Task 1 and Task 2 (right). Black squares indicate the mean values.

ification strategy in Task 1. All participants were able to complete Task 2 to match exactly the target scene (P1, P3-P6, P8, P10, P11) or match sufficiently close (T*) to the target scene (P7). Wilcoxon Signed-Rank tests on the task completion time of these participants also revealed a **significantly shorter** ($W = 3, p < .05, |r| = .78$) **completion time of Task 2 compared with Task 1**.

4.4 Interaction Barriers

The study also revealed certain interaction barriers which adversely affected the completion quality of scene editing tasks or the completion efficiency of the tasks.

Speech Recognition/Processing Issues. The misrecognition and processing errors of certain words by the system required participants to repeat their queries multiple times, which ultimately resulted in a delay in task completion. Interaction barriers under this category can be due to a recognition error from the Microsoft Azure Speech Recognition service, or due to a processing error in the misrecognition of user intents or key entities by Azure CLU or an error occurred when matching key entities to the GameObjects or textures in Unity during the Unity post-processing step.

Feedback Clarity. In the user study, participants were often confused when the system failed to respond to user speech input according to user expectations. In such circumstances, the system does not always provide clear feedback on why a command did not work or instruct participants on how to phrase it correctly. For queries that were not categorized as 'Other', the Azure CLU model did not have the capability to provide feedback to users. For speech input processed by GPT-4o, the model lacked enough contextual information on the current status of each object (such as whether they are selected or not) in the scene and could not provide enough feedback. Participants also commented how visual feedback could be further improved by, for example, adding images to describe colors and materials in addition to text, or adding a progress bar to indicate that the LLM is processing the user query.

Command Phrasing. Some participants had difficulty finding the correct phrasing for some commands, especially for changing patterns or materials. This is because the training data of the Azure CLU model only labelled commands with a certain sentence structure as selection or editing commands. Commands with different phrasing are processed by GPT-4o, but the model often replied that it did not have the capability to select or modify objects, which resulted in confusion among participants. This example shows how special considerations should be included in the GPT prompt to instruct LLMs to incorporate information from other sources, such as information directly from the 3D scene or the Azure CLU model. This will guide users to find the correct command instead of providing a misleading response to state that the system is incapable of completing the selection or editing task.

5 Discussion

This study highlights the promising potential of LLM-assisted interactive systems in guiding users towards more efficient multimodal interaction strategies, thereby improving user performance in typical interaction tasks such as scene editing in virtual reality.

5.1 LLM-Assisted Systems for VR

Immersive environments present distinct spatial, multimodal, and latency requirements. For example, we found that users' reliance on both voice and visual feedback in the scene exposed limitations and areas of future work for such LLM-assisted systems to align natural language suggestions with visual cues in the 3D spatial context. This emphasizes the need to go beyond LLM applications for 2D interfaces to incorporate spatial reasoning capabilities in LLM-assisted systems for immersive content, and ensure alignment in different channels of natural language and visual information.

Additionally, our findings reveal how the immersive nature of VR amplifies certain characteristics in human-AI interaction. Hallucinated outputs from LLMs become more disruptive in immersive environments compared with traditional graphical user interfaces due to the higher cognitive load and conflicting visual information from the user's immediate spatial environment. These findings highlight the importance of addressing well-known issues of LLM applications, such as trust, multimodality, and uncertainty in the immersive domain.

5.2 User Performance and Interaction Patterns

User Performance. Performance indicators, such as the number of remaining elemental editing steps and task completion time reported in Section 4.1, reveal how user performance significantly improved in Task 2 compared with Task 1. First, the study exemplifies the impact of choosing the correct interaction strategy on task completion quality. While **431.3%** more time was used on *incremental exploration* as compared to *bulk modification* in Task 1, *bulk modification* resulted in a **66.38%** reduction in the remaining elemental scene editing steps compared with the *incremental exploration* strategy in Task 1. Second, the performance indicators reveal how LLM-assisted interactive systems helped to guide users to select better interaction strategies which result in improved performance. In several cases for P12, P2, P6, and P5, the LLM-assisted scene editing system was able to give constructive feedback in response to user queries on the supported speech commands, available colors, available materials, and the most efficient way supported by the system for users to complete the scene editing task, with examples provided in the Online Appendix. However, the absence of a baseline (e.g., task completion without LLM assistance) limits our ability to definitively attribute these gains to the LLM, and results should be treated with caution.

Interaction Patterns. Results from Section 4.3 revealed certain interaction patterns, offering insights on how individuals interact with LLMs in immersive environments. For example, users consistently prioritized visually salient features and objects with clear target states, suggesting that spatial and visual affordances play a significant role to guide AI-supported decision making. These findings highlight the need for LLM-assisted systems in immersive environments to align LLM suggestions with the user's focus and provide visual previews to align verbal commands with 3D outcomes.

The study also revealed that even when Task 1 and Task 2 explicitly encourage participants to find the most efficient way of scene editing, two out of twelve participants displayed rigid interaction behavior, adhering to a limited set of strategies regardless of system feedback. This observation reveals a potential barrier when users are encouraged to explore more optimal modes of interaction, and demonstrates the need for LLM-assisted systems to guide users towards exploiting system capabilities efficiently.

5.3 Implications

Results from this study shed light on design implications for future LLM-assisted interactive systems. While the study is conducted in a VR environment, design implications are applicable to 3D content applications in general, and even possibly applicable to general interactive systems where LLMs are involved. Based on results from the study, we formulate design implications as follows.

First, effective use of multimodal input is critical for improving the user experience for LLM-assisted interactive systems. While multimodality has been widely-researched for immersive technology, our findings emphasize the need for dynamic integration of inputs such as controller, gesture, gaze, and verbal commands to support embodied interaction. For example, users frequently used verbal commands for high-level tasks but relied on controller input to finetune certain selections. This combination of high-level verbal commands combined with detailed-level controller or gesture input presents an opportunity for LLM-assisted systems to support more intuitive interactions in immersive environments and allows the system to disambiguate or refine user input which would otherwise be imprecise when only speech input is supported. This corroborates findings from Liao et al. [29] who state that interaction with LLM systems with only the natural language modality can be easily affected by subtle language cues.

Second, the design of LLM-assisted interactive systems should place special considerations on fostering user trust and improving user agency. Trust and agency are fundamental to LLM-assisted interactions. However, these concepts are expressed differently in VR due to its immersive nature. For example, users such as P2 expressed more trust in the incremental exploration strategy as it provided more visual confirmation. When P12 did not receive feedback on her command to select all blue objects, her sense of agency and control over the system through speech commands was affected. These observations inform future designs to allow users to preview the results of LLM-generated actions and incorporate reversible actions in addition to 2D visual displays such as the draggable panel implemented in AssistVR to reinforce a sense of control.

Finally, LLM-assisted interactive systems should implement measures to convey the fundamental uncertainties that emerge from LLM interaction, such as hallucination. While the risk of hallucinations is well-recognized in AI research, our findings indicate that their impact is magnified in immersive environments due to their spatial and interactive nature. For example, hallucinated scene information can result in inconsistent responses. LLMs involved directly in scene editing (instead of attached to Unity post-processing scripts like the Azure CLU pipeline) can also result in erroneous spatial edits. These errors can all severely disrupt task flow and user immersion. To mitigate this, LLM-assisted systems for immersive content editing should explicitly convey uncertainty in LLM outputs through visual indicators (such as color-coded confidence levels) or audio cues. While LLMs inherently have limitations in hallucinating information [48], it is important to signpost to the user when such information can be inaccurate or incomplete.

6 Conclusion and Future Outlook

This work has provided an analysis of user interaction patterns and strategies with LLM-assisted interactive systems through an example scene editing task in virtual reality. As evidenced by the results in Section 4.1, LLM-assisted interactive systems have the potential to guide users to find more effective and efficient interaction strategies and improve task performance with very limited external guidance. Results from the post-experience questionnaire corroborate findings in prior work [26, 15] on the strengths of LLM-assisted interactive systems for immersive content in perceived workload, usability, and user experience. Based on post-experience questionnaire comments, we summarize design considerations for LLM-assisted interactive systems in terms of multimodal interaction, user trust, user agency, and appropriate feedback to cope with uncertainty and hallucination. We also summarize interaction patterns such as the fact that visually distinguishable features tend to be edited first, and objects with an obscure goal state tend to be edited last. Interaction patterns further reveal how participants were able to improve their strategy through interaction with the system. Based on these qualitative and quantitative observations, we proposed a set of design implications for LLM-assisted interactive systems.

Novelty effects possibly inflated usability perceptions and thus results have to be treated with caution and we encourage replication efforts. We also acknowledge limitations in our task design. For example, our tasks do not fully capture or analyze ambiguous user input cases typical of speech interfaces. Our tasks also do not take into account complex high-level editing requirements from the user. The current tasks on color and material modification and the AssistVR system presented in the paper lack the versatility to cater to various design requirements, which will be addressed in future work. Nevertheless, this study provides a promising outlook for LLM-assisted interactive systems and provides a reference for future work on interaction analysis of LLM-assisted systems. We envision that these interaction pattern findings and design implications will be applicable to LLM-assisted interactive systems in general to guide a broad range of future designs in VR and beyond.

Supplemental Materials

Supplementary materials can be found in the Online Appendix at https://osf.io/2hmnd/. This includes: (1) The utterance training data for Azure CLU, the prompts for GPT-4o, (2) Verbal instructions given to participants before the practice trial, and before both tasks and both task types, and (3) Selected conversation histories between participants and AssistVR grouped into themes.

Acknowledgments

Junlong Chen is supported by the China Scholarship Council and Cambridge Trust.

References

[1] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*, 2023. 2

[2] S. Aghel Manesh, T. Zhang, Y. Onishi, K. Hara, S. Bateman, J. Li, and A. Tang. How People Prompt Generative AI to Create Interactive VR Scenes. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference*, pp. 2319–2340, 2024. 2

[3] M. Ahn, A. Brohan, N. Brown, Y. Chebotar, O. Cortes, B. David, C. Finn, C. Fu, K. Gopalakrishnan, K. Hausman, et al. Do As I Can, Not As I Say: Grounding Language in Robotic Affordances. *arXiv preprint arXiv:2204.01691*, 2022. 2

[4] F. Argelaguet and C. Andujar. A survey of 3D object selection techniques for virtual environments. *Computers & Graphics*, 37(3):121–136, 2013. 4

[5] J. Bergström, T.-S. Dalsgaard, J. Alexander, and K. Hornbæk. How to Evaluate Object Selection and Manipulation in VR? Guidelines from 20 Years of Studies. In *proceedings of the 2021 CHI conference on human computing systems*, pp. 1–20, 2021. 4

[6] C. Beyan, A. Vinciarelli, and A. D. Bue. Co-Located Human–Human Interaction Analysis Using Nonverbal Cues: A Survey. *ACM Computing Surveys*, 56(5):1–41, 2023. 2

[7] R. A. Bolt. "Put-that-there" Voice and gesture at the graphics interface. In *Proceedings of the 7th annual conference on Computer graphics and interactive techniques*, pp. 262–270, 1980. 1

[8] E. Bozkir, S. Özdel, K. H. C. Lau, M. Wang, H. Gao, and E. Kasneci. Embedding Large Language Models into Extended Reality: Opportunities and Challenges for Inclusion, Engagement, and Privacy. In *Proceedings of the 6th ACM Conference on Conversational User Interfaces*, pp. 1–7, 2024. 2

[9] J. Brooke et al. SUS-A quick and dirty usability scale. *Usability evaluation in industry*, 189(194):4–7, 1996. 4, 5

[10] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, eds., *Advances in Neural Information Processing Systems*, vol. 33, pp. 1877–1901. Curran Associates, Inc., 2020. 2

[11] F. De La Torre, C. M. Fang, H. Huang, A. Banburski-Fahey, J. Amores Fernandez, and J. Lanier. LLMR: Real-time prompting of interactive worlds using large language models. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pp. 1–22, 2024. 1, 2

[12] J. Dudley, H. Benko, D. Wigdor, and P. O. Kristensson. Performance Envelopes of Virtual Keyboard Text Input Strategies in Virtual Reality. In *2019 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 289–300. IEEE, 2019. 2

[13] A. M. Feit, D. Weir, and A. Oulasvirta. How We Type: Movement Strategies and Performance in Everyday Typing. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pp. 4262–4273, 2016. 2

[14] C. R. Foy, J. J. Dudley, A. Gupta, H. Benko, and P. O. Kristensson. Understanding, Detecting and Mitigating the Effects of Coactivations in Ten-Finger Mid-Air Typing in Virtual Reality. In *Proceedings of the 2021 CHI conference on Human Factors in Computing Systems*, pp. 1–11, 2021. 2

[15] D. Giunchi, N. Numan, E. Gatti, and A. Steed. DreamCodeVR: Towards Democratizing Behavior Design in Virtual Reality with Speech-Driven Programming. In *2024 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*, pp. 579–589. IEEE, 2024. 1, 2, 9

[16] G. Guest, K. M. MacQueen, and E. E. Namey. Introduction to Applied Thematic Analysis. *Applied Thematic Analysis*, 3(20):1–21, 2012. 5

[17] J. Guo, V. Mohanty, J. H. Piazentin Ono, H. Hao, L. Gou, and L. Ren. Investigating Interaction Modes and User Agency in Human-LLM Collaboration for Domain-Specific Data Analysis. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pp. 1–9, 2024. 2

[18] S. G. Hart and L. E. Staveland. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In P. A. Hancock and N. Meshkati, eds., *Human Mental Workload*, vol. 52 of *Advances in Psychology*, pp. 139–183. North-Holland, 1988. doi: 10.1016/S0166-4115(08)62386-9 4, 5

[19] K. He, A. Lapham, and Z. Li. Enhancing Narratives with SayMotion's text-to-3D animation and LLMs. In *ACM SIGGRAPH 2024 Real-Time Live!*, pp. 1–2. 2024. 2

[20] W. Huang, P. Abbeel, D. Pathak, and I. Mordatch. Language Models as Zero-Shot Planners: Extracting Actionable Knowledge for Embodied Agents. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, eds., *Proceedings of the 39th International Conference on Machine Learning*, vol. 162 of *Proceedings of Machine Learning Research*, pp. 9118–9147. PMLR, 17–23 Jul 2022. 2

[21] W. Huang, F. Xia, T. Xiao, H. Chan, J. Liang, P. Florence, A. Zeng, J. Tompson, I. Mordatch, Y. Chebotar, et al. Inner Monologue: Embodied Reasoning through Planning with Language Models. *arXiv preprint arXiv:2207.05608*, 2022. 2

[22] A. Jebeli, L. K. Chen, K. Guerrerio, S. Papparotto, L. Berlin, and B. J. Harden. Quantifying the Quality of Parent-Child Interaction Through Machine-Learning Based Audio and Video Analysis: Towards a Vision of AI-assisted Coaching Support for Social Workers. *ACM Journal on Computing and Sustainable Societies*, 2(1):1–21, 2024. 2

[23] L. Jiang, M. Phutane, and S. Azenkot. Beyond Audio Description: Exploring 360° Video Accessibility with Blind and Low Vision Users Through Collaborative Creation. In *Proceedings of the 25th international ACM SIGACCESS conference on computers and accessibility*, pp. 1–17, 2023. 2

[24] E. Kaiser, A. Olwal, D. McGee, H. Benko, A. Corradini, X. Li, P. Cohen, and S. Feiner. Mutual Disambiguation of 3D Multimodal Interaction in Augmented and Virtual Reality. In *Proceedings of the 5th International Conference on Multimodal Interfaces*, pp. 12–19, 2003. 1

[25] M. Konenkov, A. Lykov, D. Trinitatova, and D. Tsetserukou. VR-GPT: Visual Language Model for Intelligent Virtual Reality Applications. *arXiv preprint arXiv:2405.11537*, 2024. 2

[26] R. Kurai, T. Hiraki, Y. Hiroi, Y. Hirao, M. Perusquia-Hernandez, H. Uchiyama, and K. Kiyokawa. MagicItem: Dynamic Behavior Design of Virtual Objects with Large Language Models in a Consumer Metaverse Platform. *arXiv preprint arXiv:2406.13242*, 2024. 1, 9

[27] J. Lee, J. Wang, E. Brown, L. Chu, S. S. Rodriguez, and J. E. Froehlich. GazePointAR: A Context-Aware Multimodal Voice Assistant for Pronoun Disambiguation in Wearable Augmented Reality. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pp. 1–20, 2024. 2

[28] X.-L. Li, H. Li, H.-X. Chen, T.-J. Mu, and S.-M. Hu. DIScene: Object Decoupling and Interaction Modeling for Complex Scene Generation. In *SIGGRAPH Asia 2024 Conference Papers*, SA '24. Association for Computing Machinery, New York, NY, USA, 2024. doi: 10.1145/3680528.3687589 2

[29] Q. V. Liao and J. W. Vaughan. AI Transparency in the Age of LLMs: A Human-Centered Research Roadmap. *arXiv preprint arXiv:2306.01941*, pp. 5368–5393, 2023. 1, 9

[30] X. Ma, Y. Bhalgat, B. Smart, S. Chen, X. Li, J. Ding, J. Gu, D. Z. Chen, S. Peng, J.-W. Bian, et al. When LLMs step into the 3D World: A Survey and Meta-Analysis of 3D Tasks via Multi-modal Large Language Models. *arXiv preprint arXiv:2405.10255*, 2024. 2

[31] G. Manfredi, U. Erra, and G. Gilio. A Mixed Reality Approach for Innovative Pair Programming Education with a Conversational AI Virtual Avatar. In *Proceedings of the 27th International Conference on Evaluation and Assessment in Software Engineering*, pp. 450–454, 2023. 2

[32] OpenAI. *Hello GPT-4o*, May 2024. Available at `https://openai.com/index/hello-gpt-4o`. 2

[33] A. Plopski, T. Hirzle, N. Norouzi, L. Qian, G. Bruder, and T. Langlotz. The Eye in Extended Reality: A Survey on Gaze Interaction and Eye Tracking in Head-worn Extended Reality. *ACM Computing Surveys (CSUR)*, 55(3):1–39, 2022. 2

[34] S. Rabsahl, T. Satzger, S. Kalamkar, J. Grubert, and F. Beck. Symbolic Event Visualization for Analyzing User Input and Behavior of Augmented Reality Sessions. 2023. 2

[35] I. Rakkolainen, A. Farooq, J. Kangas, J. Hakulinen, J. Rantala, M. Turunen, and R. Raisamo. Technologies for Multimodal Interaction in Extended Reality—A Scoping Review. *Multimodal Technologies and Interaction*, 5(12):81, 2021. 2

[36] J. Roberts, A. Banburski-Fahey, and J. Lanier. Steps towards prompt-based creation of virtual worlds. *arXiv preprint arXiv:2211.05875*, 2022. 1, 2

[37] R. Rodriguez, B. T. Sullivan, M. D. Barrera Machuca, A. U. Batmaz, C. Tornatzky, and F. R. Ortega. An Artists' Perspectives on Natural Interactions for Virtual Reality 3D Sketching. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pp. 1–20,

2024. 2

[38] F. Scholz, T. E. Kolb, and J. Neidhardt. Classifying User Roles in Online News Forums: A Model for User Interaction and Behavior Analysis. In *Adjunct Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization*, pp. 240–249, 2024. 2

[39] M. Schrepp, A. Hinderks, and J. Thomaschewski. 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. 4, 5

[40] J. Song, B. Wang, Z. Wang, and D. K.-M. Yip. From Expanded Cinema to Extended Reality: How AI Can Expand and Extend Cinematic Experiences. In *Proceedings of the 16th International Symposium on Visual Information Communication and Interaction*, pp. 1–5, 2023. 2

[41] B. Spittle, M. Frutos-Pascual, C. Creed, and I. Williams. A Review of Interaction Techniques for Immersive Environments. *IEEE Transactions on Visualization and Computer Graphics*, 29(9):3900–3921, 2022. 2

[42] J. R. Trippas, S. F. D. Al Lawati, J. Mackenzie, and L. Gallagher. What do Users Really Ask Large Language Models? An Initial Log Analysis of Google Bard Interactions in the Wild. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2703–2707, 2024. 2

[43] M. Tsimpoukelli, J. L. Menick, S. Cabi, S. Eslami, O. Vinyals, and F. Hill. Multimodal Few-Shot Learning with Frozen Language Models. *Advances in Neural Information Processing Systems*, 34:200–212, 2021. 2

[44] A. S. Williams and F. R. Ortega. Understanding Gesture and Speech Multimodal Interactions for Manipulation Tasks in Augmented Reality Using Unconstrained Elicitation. *Proceedings of the ACM on Human-Computer Interaction*, 4(ISS):1–21, 2020. 2

[45] P. C. Wright, R. E. Fields, and M. D. Harrison. Analyzing Human-Computer Interaction as Distributed Cognition: The Resources Model. *Human-Computer Interaction*, 15(1):1–41, 2000. 2

[46] S. Wu, H. Fei, L. Qu, W. Ji, and T.-S. Chua. NExT-GPT: Any-to-Any Multimodal LLM. *arXiv preprint arXiv:2309.05519*, 2023. 2

[47] D. Yu, T. Dingler, E. Velloso, and J. Goncalves. Object Selection and Manipulation in VR Headsets: Research Challenges, Solutions, and Success Measurements. *ACM Computing Surveys*, 2024. 4

[48] Y. Zhang, Y. Li, L. Cui, D. Cai, L. Liu, T. Fu, X. Huang, E. Zhao, Y. Zhang, Y. Chen, et al. Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models. *arXiv preprint arXiv:2309.01219*, 2023. 9

[49] X. Zhou, A. S. Williams, and F. R. Ortega. Eliciting Multimodal Gesture+Speech Interactions in a Multi-Object Augmented Reality Environment. In *Proceedings of the 28th ACM Symposium on Virtual Reality Software and Technology*, pp. 1–10, 2022. 1, 2

[50] C. Zimmerer, E. Wolf, S. Wolf, M. Fischbach, J.-L. Lugrin, and M. E. Latoschik. Finally on Par?! Multimodal and Unimodal Interaction for Open Creative Design Tasks in Virtual Reality. In *Proceedings of the 2020 International Conference on Multimodal Interaction*, pp. 222–231, 2020. 1, 2