Embodied AI for Virtual Space Design

Author: Linghang Cai

Instructors: Sam Creely, Mashinka Firunts Hakopian

Class: GMDP-603-01 Critical Practice 1

1. Abstract

Embodied interaction means a content-aware and multi-modality approach to achieve real and natural experience in the field of human computer interaction (HCI). This approach is a common trusted way to design free, natural and immersive experiences.

The interaction in the real world is often engaged with multi-dimensional information communication, which is less discussed by the traditional and relatively linear screen-based interaction. However, the advancement of VR/XR technology and spatial computing has driven significant research into embodied interaction [1, 3, 4] because it requires designers to think more than a “click” function, but including verbal, gestural and bodily inputs.

For example, traditional screen-based interactions rely on mouse and keyboard inputs, converting 2D information into 3D environments. This method often requires users to learn complex operation patterns, such as dragging, zooming, and rotating, which demand significant effort to learn 3D modeling. However, new VR-based modeling tools now offer a more intuitive and natural approach to space modeling in VRs. These tools primarily use gestures and controllers as inputs, user can have more freedom and natural interaction when creating 3d model and virtual space, but they did not fully account for the complex and diverse ways users naturally behave when designing and building spaces, these type of 3d modeling software often come with 2d design principle,basically brings 2D interface to 3D environments.

This opens the door to designing a novel interaction system with embodied interaction principles. It is natural for users to combine body movements, gestures, and verbal expressions to describe and conceptualize imaginary scenes during spatial design. Therefore, such a system should allow multiple input modalities to create a natural and efficient user experience.

This paper explores the feasibility of constructing an embodied system for virtual space design. It focuses on integrating body movement, gestural, and verbal inputs with human-AI collaborative principle to enable natural, intuitive interactions. By examining these possibilities, the paper aims to provide a foundation for embodied VR interaction and human-AI collaboration, advancing the design of immersive and adaptive virtual environments.

Keywords

Embodied Interaction, Human-AI Collaboration, Virtual Space Design, Virtual Reality, Mixed Reality, Spatial Computing, Multimodal Inputs, Body Movement, Gesture Recognition, Verbal Input, Environment-Aware Systems, User-Centric Interaction

2. Introduction

VR provides users with immersive experiences and richer interactions across multiple modalities, enhancing user engagement and satisfaction [3–5, 8, 22]. But then what? Why is this engagement important? Why is embodied interaction and physically engaged experiences worth the significant effort to research? How does it fundamentally differ from existing forms of interaction? And what transformative changes can it bring to our future?

Some might even ask, in an era of highly advanced screen-based interaction, where we can achieve almost anything through a small electronic device—from work, life, to entertainment—is the pursuit of VR, AR, immersive interaction, tangible, and embodied interactions truly meaningful? A worker's day might begin with checking messages on their phone, pass through watching short videos to kill time during a commute, proceed to working on electronic documents at the office, and end with watching videos in bed. Flat, screen-based interaction has seemingly fulfilled all our needs. So why chase after something different?

Let us rewind to childhood. During a festival or birthday, we might have wished for magical powers—to control everything, alter the weather, manipulate objects and water, see static characters in newspapers come alive, or converse with fictional characters in books. In these imaginings, our interaction with the magical world and objects starts with ourselves. It is direct and seamless, clearly not mediated by a two-dimensional screen. Critics might argue that the limitations of existing technology have shaped current interactions and applications; if better technologies were available, people would naturally gravitate toward more direct and natural interactions. After all, when interaction design was first explored, screen-based interfaces quickly became dominant.

But here lies the emerging question: have people adapted to 2D screens because they are close to perfection? Or is it because, like many other things in this era, they have always existed and been accepted, integrating into our cognition? Is the translation process of controlling a cursor on a vertical screen via a mouse truly as natural and flawless as it seems? My mother, who has not received higher education and grew up in a household without access to most electronic devices, still finds keyboard and mouse input difficult to adapt to. In contrast, smartphones—frequently appearing in her world—are the only electronic devices she uses often. For her, the logic of chat applications is simple to grasp: open the phone, click the icon, and hold down a button to call me. However, she often struggles with the camera position or the

auto-adjustment of layouts when holding the phone horizontally. For a long time, I had to flip my own phone upside down to ensure the image displayed correctly on her end.

She certainly does not use any modeling software or AR/VR devices, yet she has strong ideas about decorating the family shop. She decides where to place the reception counter, where to display clothing, which type of lighting to use, and even creates a window display for nighttime viewing. Such ideas are common in modern commercial design, but these decisions are typically reserved for well-trained designers. In reality, the right to compute and design interactions is concentrated in the hands of the elite, while others pay for "the right to use."

Returning to my research in interaction design, I often observe a degree of misalignment between designers and users, resulting in usability and intent issues. What seems straightforward and intuitive to designers often requires users to spend disproportionately more time learning. While users can adapt within a lab setting, in real-world scenarios many—like my mother—face greater difficulties, often abandoning the effort altogether. This abandonment has more significant consequences than imagined, as it implies that technology excludes certain groups from the digital world, effectively imposing a kind of "technological trade embargo."

Interacting with technology should not be a one-sided process where humans must learn and adapt to counterintuitive tools (even though humans have a remarkable ability to learn and adapt). Instead, technology should learn to understand people. *Computation is intentional* [5]. As a mediator between developers, designers, and users, digital products should not merely be "usable" or "learnable." They should be designed with a more empathetic and human-centered perspective—not just thinking about "how technology can improve," but "how technology can better understand humans."

This is the foundational motivation behind my research into embodied interaction. Criticism of 2D interfaces is not about their efficiency or functionality, but about how they confine people to a screen, prioritizing technology over human-centered design. As Dourish has emphasized, embodied interaction does not merely focus on whether an interaction is embodied; it considers how interaction arises from an embodied, human-centered perspective. It is inherently multimodal, environmental-aware, and beyond 2D and screen-based interaction.

2.1 Digital Creation Tools

As people move from the physical world into the digital space, an ever-expanding array of tools for education, media, entertainment, industrial manufacturing, and workplace productivity appears. Over time, people have adapted to using digital tools, learned to communicate with or via them in ways that computers can understand.

People who are so used to these tools, regardless of their different applications, can serve as extensions of the body, transforming their inputs into modifications of content. For example, a hammer shifts from being an external object—“present-at-hand”—to becoming an integrated part of the body—“ready-to-hand”—once mastered [7]. In this state of fluency, users focus not on the tool itself but on their distant intention: the deformation of metal, the precise strike, or the outcome of the task. This seamless integration between tool and intention exemplifies a natural interaction, where the tool bridges the gap between capability and desire in a good way.

The evolution of digital tools has changed how designers create, transitioning from traditional methods (like pencil and canvas) to digital software. Today, spatial designers and architects commonly mostly use 3D modeling software such as Blender to design spaces, and platforms like Unity and Unreal Engine to create interactivity. These tools offer powerful functionalities—including object manipulation, material application, and rendering scenes from a human perspective—to effectively communicate their designs with the audiences.

However, these tools, like Blender and Rhino, rely heavily on screen-based inputs like the mouse and keyboard, which means translating 2D inputs into 3D environments, a process that requires users to learn and adapt to complex translation rules [5]. The learning curve is sharp, involving not just an understanding of design, but also the knowledge of computer graphics algorithms, such as mesh, vertex and shader manipulation. This approach limits how naturally users can interact with and express their spatial ideas, making it particularly challenging for beginners.

The problems are related to both the system design principle and current technique limitation, for example, understanding users’ intentionality—the link between what they want to achieve and how they act—is always a crucial area of research [14]. As digital tools become increasingly complex, this natural relationship often breaks down. Current screen-based tools for spatial design impose significant learning barriers that disrupt the intuitive connection between intention

and action. Recognizing and addressing these challenges is key to designing the next generation of digital tools [5, 9].

2. How Computer Graphics Algorithms Impact User Interaction

The complexity of mainstream modeling software is rooted in the foundational principles of computer graphics and the algorithmic design decisions [19]. Tools like SketchUp and Rhino, for instance, adopt different modeling paradigms—polygon-based versus NURBS-based modeling—that impact how users interact with and create geometry [2]. Rhino, using NURBS, allows for smooth, precise surfaces by manipulating control points, while SketchUp relies on polygonal meshes. Such differences not only shape the user experience but also dictate how accessible the tool is to new users. For example, users using SketchUp often find it's easier to build solid objects (a type of closed geometry), because users can join and bridge surfaces by drawing on the vertices, while in Rhino users have to manually join different surfaces.

This computational foundation imposes significant learning efforts, especially on novice users, by requiring them to understand abstract concepts that do not naturally map to physical-world interactions. As a result, while these tools are powerful, they are often cumbersome and inaccessible, leading designers to reconsider how users might engage more directly with 3D environments. Tools like blender, spline and even figma are trying to provide a light and easy way to reduce the learn and use effort by empowering improved algorithms and more accessible user interface design.

3. Enter the Era of Immersive Technologies

The advent of immersive technologies like AR and VR presents an opportunity to rethink traditional screen-based spatial design paradigms. AR and VR offer more direct and embodied ways for users to interact with 3D environments, using their bodies, gestures, and movements naturally, instead of indirect input methods like the mouse and keyboard [13, 19, 23]. This kind of embodied interaction can help bridge the gap between the physical and virtual interactions by aligning more naturally with users' intuitive understanding of spatial relationships and object manipulation.

Yet, despite their promise, current VR-based modeling tools are limited. Many VR tools still mimic screen-based interactions, relying on single inputs from gestures or controllers that are translated into the same old 2D paradigms. This new layer of interaction risks adding further complexity rather than simplifying the design process. To unlock the true potential of AR and

VR, we need to develop a new interaction paradigm that enables users to not merely hide behind the interface but active participants in the virtual space.

4. Challenges and Opportunities in Embodied Interaction

Embodied interaction allows users to use their natural senses and bodily movements to interact with digital environments. This form of interaction involves not only creating realistic representations of physical environments but also developing interaction systems grounded in real-world knowledge.

For example, the concept of coupling—aligning user intention with the system's capabilities. When people are driving, they use wheels and accelerator pedals to control the direction and speed, which should be integrated seamlessly to get them to their destination. What's more, the coupled system should allow users to use them by different intentions, like racing, uphilling that related to the position change activities. Coupling is particularly relevant in ensuring that embodied interactions maintain a natural flow, supporting creativity and usability.

However, challenges remain. Current interactions are still limited by technological capabilities and the lack of a deep understanding of human behaviors. These problems are widespread over today's HCI field, when we are moving forward to the era of more natural interactive systems. When incorporating AI, for example, there are difficulties for generative models to interpret and align with users' intentions. To overcome these obstacles, we need a paradigm shift in how we think about interaction design, from building tools to building the extension of human capabilities, focusing on providing a more intuitive and embodied user experience.

5. AI, VR and System Design

While existing AI-assisted design tools leveraging machine learning and generative AI aim to solve complex design challenges holistically, they struggle with usability, particularly in understanding nuanced human intentions. By proposing a new interaction paradigm that integrates AR/VR technologies with embodied interaction principles, we can lay the groundwork for more natural, intuitive, and powerful design tools.

This new paradigm aims to use the user's body as an active design tool—engaging directly with virtual space, leveraging AI to interpret multimodal inputs (such as gestures, body movements, and speech), and coupling these inputs with design intentions to create seamless spatial interactions. With such a system, designers could achieve a more direct connection between

their conceptual ideas and their spatial representations, resulting in a workflow that not only reduces learning curves but also enhances creative freedom.

6. Research Goals

The aim of this research is to explore the design of AR/VR tools for spatial design with AI, focusing on developing guidelines and paradigms that support embodied interaction as a natural approach. The research will examine the role of AI in enhancing human imagination and the creative process, and enabling multimodal interaction systems that are closely aligned with users' intentions. Ultimately, this work seeks to bridge the gap between traditional design principles and the new opportunities afforded by immersive technologies and AI, transforming spatial computing into a more user-centric, intuitive design practice.

3. Literature Review

3.1 Virtual Environments Creation

Virtual environments (VEs) represent a critical intersection of technology, perception, and human-computer interaction. Defined broadly as immersive, interactive systems that simulate physical spaces or constructs, VEs leverage multiple sensory modalities—visual, auditory, and haptic—to create synthetic experiences that are believable and engaging [6, 15]. The concept of virtual environments is rooted in the principles of interactivity and immersion. Immersion pertains to the system's ability to deliver extensive and vivid sensory input, including inclusivity (excluding real-world stimuli), surrounding capabilities (panoramic experiences), and vividness (realistic renderings). Proprioceptive feedback further enhances immersion through accurate tracking of user movements and their alignment with virtual representations. Presence, a psychological correlate of immersion, reflects the user's sense of "being there" within the VE. Research has emphasized that higher degrees of immersion are positively correlated with a greater sense of presence. This dynamic underpins the utility of VEs for tasks requiring realistic simulations, such as training or therapeutic interventions.

The feeling of “being there” is important, how users engage physically and cognitively in a VE can be measured by the factor involvement, adaptation, immersion, sensory fidelity and the quality of interface [20]. It aligns with the principle of embodied interaction by focusing on how the user's sensory and physical inputs affect their experience.

Immersion is a technological attribute (e.g., fidelity of graphics, responsiveness of controls), while presence describes the subjective psychological experience of users within the VE[15, 24]. VEs are used in telemedicine, training simulations, entertainment, and collaborative tasks, offering cost-effective and flexible alternatives to physical environments[6].

Research emphasizes hybrid systems that integrate VEs with AI and multi-user platforms to enhance interaction and presence, aligning with practical applications like e-therapy and collaborative learning [12].

3.2 Embodied Interaction

Embodied cognition views the body as central to perception, action, and thought, influencing how humans interact with and understand technology. Phenomenology emphasizes the first-person perspective of the moving body as critical for meaningful interaction design [4, 5]. Gestural interfaces, motion-sensing games, and virtual environments use body interactions for immersive and intuitive user experiences. However, the novelty often wears off without meaningful and context-rich interactions. Embodied interaction bridges the physical world with digital systems, focusing on tools and methods that align with human bodily capacities and social practices [3–5, 8, 13].

Advances in tangible and embodied interactions are enhancing AR/VR technologies by making them more intuitive. The notion of variable coupling in designing tools can help align user intentions and interactions in a more embodied manner, resulting in more natural experiences [24]. Embodied interaction is instrumental in enhancing immersion by allowing users to utilize natural behaviors (like movement and gestures) to interact with virtual spaces. This significantly boosts the sense of presence and engagement. From previous research, gestural and bodily action can reduce cognitive load and enhance communication [10]. Also, when users in the interaction, they are not just doing, or through a planned behavior, they are doing while thinking, and the intentionality is a keypoint to explore, it connects what they want and what is done [5, 10].

Embodied interaction provides richer contextual information, which helps make systems more adaptable and intuitive. However, there are challenges in designing VR-based modeling tools that can effectively understand the user's body language, intentions, and gestures, making embodied interaction more than just an immersive experience but a practical tool for spatial design.

Body interaction highlights the role of human physicality—gesture, movement, and spatial awareness—in shaping technology-mediated experiences. This approach stems from the embodied cognition paradigm, which argues that cognition is deeply rooted in bodily experiences [10, 11]

3.3 Human-AI Interaction

AI has increasingly become a transformative force in creative processes, particularly in AR/VR environments, by using functionalities such as AI-assisted design and generative AI. These technologies augment the designer's ability to create, offering tools that support visualization, material generation, and contextual understanding. For example, ControlNet provides finer control in image generation, allowing users to condition diffusion models with specific image inputs like edge maps, human poses, and segmentation maps, thereby aligning generated outputs more closely with users' imagery [22], which highlights the growing role of AI in enhancing user understanding.

While AI technologies have proven valuable, interpreting user intentions remains a significant challenge. Many generative systems rely on training data and pre-existing rules, limiting their ability to adapt to nuanced or evolving user goals. Research into Explainable AI (XAI) emphasizes the importance of creating systems that offer transparent decision-making and intuitive feedback loops, enabling users to trust and guide AI outputs effectively [25]. For instance, generative AI models often fail to balance creativity with user control, leading to unpredictable outcomes [17], if it can provide the chain of its thinking, the user could gain more control of the generative tasks. These challenges underscore the need for more adaptable algorithms and systems.

The integration of AI into creative workflows increasingly positions these systems as collaborative assistants rather than mere tools. Studies have demonstrated the effectiveness of systems like BrainFax in facilitating designerly co-creation through generative AI, offering functionalities such as sketch-based modifications and iterative ideation [17]. By embedding AI into established workflows, tools like these enhance the speed and depth of creative exploration. Moreover, systems employing multimodal interaction (e.g., integrating gestures, text, and voice) promise a more holistic understanding of user intent, thereby fostering a seamless human-AI partnership [16, 21].

Emerging paradigms in human-AI collaboration emphasize embodied interaction and co-creativity. Instead of viewing AI as a passive executor, these paradigms conceptualize AI as an active co-creator, working alongside humans to enhance creativity and problem-solving. For example, leveraging AI in spatial design not only allows designers to manipulate objects using gestures but also empowers the AI to interpret and predict user actions, aligning system

capabilities with human intentions [16]. These frameworks prioritize adaptability and contextual awareness, ensuring that AI systems can assist across different stages of the design process, from ideation to implementation.

Explainability remains a cornerstone of effective human-AI collaboration. Systems must provide clear, actionable feedback to users, especially in creative contexts where ambiguity often reigns. Research into XAI suggests that systems capable of offering local explanations (specific to a given task) and global explanations (outlining overarching system logic) are better positioned to support collaborative endeavors [25]. This dual-layered approach helps bridge the gap between human intuition and machine reasoning, enabling users to navigate and refine the creative process more effectively.

The future of AI in collaborative design lies in its ability to adapt dynamically to user needs, leveraging contextual and embodied cues to guide its outputs. Advances in generative AI models, such as integrating post-human design principles that treat AI as an equal creative agent, promise to redefine the boundaries of human-machine interaction. By focusing on explainability, adaptability, and multimodal interaction, these systems can unlock new possibilities for creativity and innovation in design [16].

4. Research Questions

The advent of AR/VR technologies, combined with the integration of embodied interaction principles, offers an opportunity to rethink how creative tools can align with human intention and imagination. These tools, operating at the intersection of physical and virtual spaces, challenge conventional paradigms of design that have long been tethered to the limitations of screens and 2D interfaces. By grounding this exploration in human-computer interaction (HCI) principles, this study seeks to transcend the constraints of current tools and propose frameworks that prioritize intuitive, multimodal interactions.

4.1 The Challenge of Defining "Natural" Interaction

The concept of natural interaction is contentious and often lacks a rigorous definition. Some argue that no interaction is inherently natural—any action, given sufficient learning and practice, can feel intuitive, some studies show that users adapt to complex systems through learned behavior, ultimately developing a sense of comfort through repetition.

However, insights from cognitive science challenge this view, emphasizing the role of innate spatial and temporal cognition in shaping our interactions with the world. Humans, for instance, universally use spatial metaphors such as "close" or "far" to describe relationships, in other words, mental positions [26], indicating that our understanding of space is deeply embedded in our cognitive frameworks. This innate knowledge allows individuals to intuitively navigate and interpret spatial environments, making it a critical element for designing embodied interactions.

In virtual spaces, this duality—between learned behaviors and innate intuition—complicates efforts to define what is "natural." Users often perceive familiar interactions as intuitive, even if they have required significant effort to learn. Conversely, interactions grounded in innate spatial cognition may initially seem unfamiliar but can offer a more seamless and universally accessible foundation for design, but it still requires learning efforts.

4.2 Bridging Spatial Cognition and Interaction Design

To understand what makes an interaction natural in the context of virtual spatial design, it is essential to examine the relationship between spatial cognition and user experience. Spatial cognition refers to the human ability to perceive, understand, and interpret the spaces, it also influences how we interpret abstract concepts such as time, relationships, and causality. For

example, the use of spatial terms to describe emotional states ("feeling distant") or organizational structures ("hierarchical levels") highlights how deeply spatial cognition informs human thought and communication.

In AR/VR environments, the cognitive framework can enable interactions that feel more intuitive by aligning virtual actions with the user's inherent understanding of space. Imagining a gesture-based system where users manipulate virtual objects by mimicking real-world actions, such as grabbing or pushing, by simulating the relationship of the objects, material, hands and friction force in it. These interactions might initially seem straightforward, and their alignment with users' innate expectations of spatial relationships, physical causality, and body schema makes it easier for users to understand.

4.3 Moving Beyond Learned Comfort

A critical distinction must be made between what users find comfortable and what is inherently natural. Current evaluation metrics for AR/VR interactions, such as efficiency or learning effort, often prioritize ease of use over a deeper investigation of naturalness. These metrics, while valuable, risk conflating familiarity with intuitiveness. An interaction learned through trial and error may become efficient over time, but this does not necessarily mean it is well-designed and aligns with the user's innate understanding.

To address this, it is necessary to develop new frameworks for evaluating natural interaction. These frameworks should:

1. Account for the cognitive and perceptual foundations of spatial understanding.
2. Incorporate both subjective and objective measures to evaluate users' sense of comfort, intentionality, and immersion.
3. Isolate the influence of learned behaviors from innate intuitions, ensuring that design decisions are grounded in fundamental human capabilities rather than acquired familiarity.

4.4 A Framework for Intuitive Interaction Design

The proposed framework begins by identifying core principles of spatial cognition that inform intuitive interaction. For example, research on intentionality the user's ability to predict outcomes based on their actions. In virtual environments, this could involve creating systems where

gestures or movements directly correspond to expected outcomes, minimizing the cognitive load required to translate intention into action.

Evaluation methods must also evolve to capture the nuances of natural interaction. Traditional questionnaires and task-based metrics, while useful, often fail to address the subjective and emotional dimensions of user experience. By integrating physiological data, such as heart rate variability or EEG signals, alongside qualitative feedback, researchers can gain a more comprehensive understanding of how users experience naturalness in virtual interactions. This multimodal approach allows for the triangulation of findings, reducing biases associated with self-reported data and highlighting discrepancies between perceived and actual intuitiveness.

5. Methodology

5.1 Wizard of Oz Experiments for Redefining Embodied Spatial Interaction

The WoZ methodology simulates an advanced AI system in VR to study how users naturally engage with generative 3D AI. By focusing on behaviors, interaction patterns, and user expectations, the study can inform the design of adaptive systems aligned with natural interaction principles.

Participants interact in VR prototypes, performing tasks like arranging 3D objects (trees, mountains, walls, etc), designing landscapes, or creating 3D models. These open-ended scenarios allow users to explore and express diverse behaviors. A researcher, acting as the wizard, controls AI responses in real-time, simulating actions such as generating 3D objects or modifying spatial elements.

Data collection combines behavioral observations, interaction logs, and user feedback from post-task interviews. Observations focus on gestures, spatial manipulation techniques, and verbal commands, while interviews capture user perceptions of interaction ease, alignment with intentions, and expectations for AI capabilities.

The wizard maintains consistency in responses while adapting to individual user styles using predefined templates, ensuring realistic and efficient simulations. Analysis identifies common interaction patterns, user assumptions about AI capabilities, and friction points in interaction flow, supplemented by qualitative insights into naturalness and intuitiveness.

5.2 Design-Based Experiment

This study involves iterative development and testing of AR/VR prototypes to explore embodied spatial interaction with AI assistance. Prototypes will integrate gestures, body movements, and speech for seamless interaction, supported by adaptive AI offering real-time assistance based on user preferences. Scenarios will include structured tasks for efficiency evaluation, exploratory tasks for creativity assessment, and collaborative design tasks to study user-AI interaction dynamics.

Participants, ranging from novices to experts, will interact with the prototypes in diverse tasks such as spatial arrangement and creative problem-solving. Data will include subjective

measures (e.g., immersion and workload via Presence Questionnaire and NASA-TLX), behavioral data (e.g., motion tracking and gesture patterns), and creative outputs (e.g., originality and task alignment). Semi-structured interviews will provide qualitative insights into user expectations and system usability.

Analysis will refine prototypes based on feedback, usability observations, and creative outcomes, aligning with principles of embodied interaction. This iterative process ensures the prototypes are applicable, intuitive, and effective in meeting user needs.

5.3 Quantitative Analysis

A comparative study will evaluate the developed embodied interaction prototypes against traditional screen-based systems to measure efficiency, creativity, and engagement. Participants will perform identical spatial design tasks across both conditions, enabling a consistent comparison of task success rates, error occurrences, and completion times. Physiological data (e.g., heart rate and skin conductance) will capture cognitive and emotional load, while behavioral engagement metrics will monitor interaction frequency and duration.

Data will be analyzed using statistical tests (e.g., ANOVA, t-tests) to identify significant differences between interaction methods. Results will also be interpreted through the lens of embodied cognition and HCI principles, validating the effectiveness of the prototypes and highlighting their advantages over traditional tools.

5.4 Overview

This research employs a mixed-methods approach combining elicitation studies, prototyping, Wizard of Oz experiments, and comparative analysis. Each method is chosen to address specific research questions, ensuring a robust exploration of embodied interaction and its role in spatial design.

6. Experiments

6.1 Experiment 1 : Breaking the Midas Spell [18]

This research focuses on how novice users use generative AI tools while they are designing spaces. We conducted a formal co-design session through Wizard-of-oz.

For the formal experiment, we will recruit 12 novice participants in spatial design, all of these subjects are between the ages of 18-35 years old, some of the candidates do not have any experience in spatial design at all, and some others have very limited experience in spatial design. Additionally, these non-professional participants will be required to have a basic interest in spatial design to avoid a mismatch of needs.

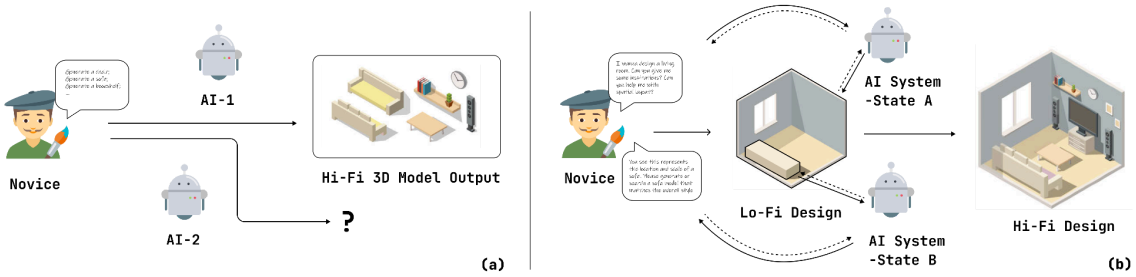

Figure 1: (a) the pipeline of the traditional AI-driven tools generates the output in one step, and these tools are scattered; (b) an envision of a progressive human-AI collaboration framework in Spatial Design, in which AI collaborates with a novice learner in the whole progress, understanding each embodied command, helping human develop from a vague idea in mind to a low-fidelity design, and finalizing with a high-fidelity spatial design scheme.

6.1.1 Apparatus

The experiment will be conducted in an online format, primarily using desktop computers. The participants will mainly use Figma to read the introduction and perform card-based tasks while using Spline to create 3D spatial design models through a split-screen setup. Figma is a user-friendly collaborative tool for discussion, note-taking, and design, while Spline is a 3D modeling tool designed for non-professionals, offering only basic modeling functionalities and pre-made model components.

For the staff, besides Figma and Spline, they will need to open various existing AI generation tools to perform tasks requested by the participants in the back end. These tools include ChatGPT (primarily for textual conversations and image explanations), MidJourney (for image generation), ComfyUI (used for specific, controllable image workflows like style transfer and generating refined images from sketches—these workflows can be downloaded from open-source repositories on GitHub), and Meshy (for fast generation of 3D models based on cloud platforms). Additionally, to accelerate the speed of providing suitable models, we provide a manual model retrieval workflow, using Sketchfab, the largest public model library available. Based on participants' needs, models are retrieved and imported into Spline.

6.1.2 Experiment Design

The experiment, which will be fully recorded and videotaped, is expected to last approximately 90 minutes. After the experimenter introduces the purpose, rules, cards, and design platforms, participants who agree to participate will sign a consent form and then proceed to watch three spatial design case studies. They will review the cards in Figma, read the descriptions and illustrations, and familiarize themselves with the basic operations and card functions in Spline. Once they feel prepared, participants will complete a 30-minute design task, followed by a 5-minute interference task, and then the second stage of its previous task with higher fidelity and detailed requirements, which also lasts for 30 minutes. The selection of questions is drawn from our predefined question bank, and the probability is randomized. This question bank covers various types of space design, ranging from virtual to real, and from indoor to outdoor environments, with a moderate level of difficulty.

6.1.3 Task 1: Low-Fidelity Spatial Design Task (30 min)

We will present participants with a typical scene design task. Participants will complete the design of a scene within 30 minutes, including aspects: layout, scale, and function. They will select design operation cards in Figma, including non-AI, AI-assisted, and custom options. These cards correspond to operations in the Spline platform. One researcher will monitor time and provide guidance, notifying participants when 5 minutes remain and checking if any design categories (positioning, color, function, material, lighting) are missing. another researcher acts as the AI, providing feedback based on participant requests. For example, if a participant uses a text-generation tool card, the researcher will input prompts into ChatGPT and paste the reply into Spline. If a participant selects a "Generate Model" card, the researcher will generate the

model and upload it to Spline. Participants will then adjust and refine their design within Spline based on the AI feedback.

6.1.4 Task 2: High-Fidelity and Reflective Spatial Design Task (30 min)

After a 5-minute break, we invited the participants to come back and they were encouraged to review their spatial design proposals, and then refine or modify specific details. Similarly, the work in this phase focuses not only on addressing layout, scale, and function issues that may have arisen during the first phase due to time constraints or operational errors, but also places greater emphasis on detailed elements such as lighting, color, materials, and decorations.

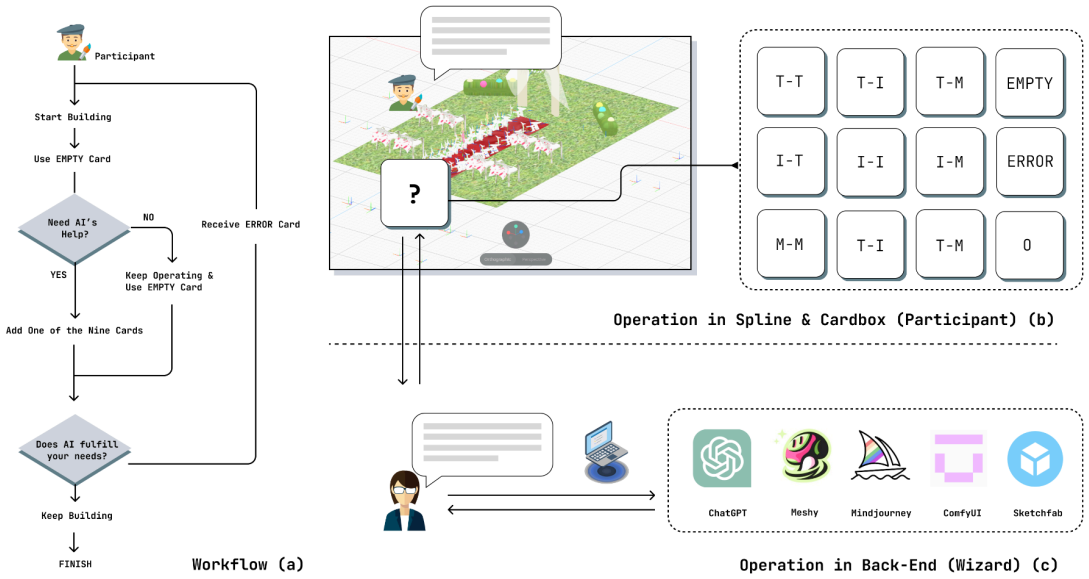

Figure 2: The formal Experiment Process

6.1.5 Common Workflow

The interaction process of the 12 participants was distilled into a common workflow, which most participants followed, as illustrated in Figure 4. The process consists of four key stages:

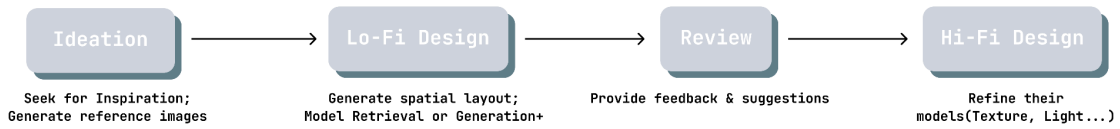

Figure 3: Common Workflow

6.1.5.1 Ideation Phase

In the initial phase, participants typically used T-T (Text to Text) to seek inspiration from the AI, asking for general design advice for the type of space they were tasked with creating. The AI-generated descriptions were then refined into more specific prompts. Participants would then use T-I (Text to Image) to generate reference images, which served as the foundational style guide, helping them form a concrete design concept.

6.1.5.2 Lo-Fi Design Phase

In this phase, participants manually built low-fidelity models of the design, establishing key elements such as the position, size, and shape of individual components using basic geometric forms. Alternatively, some participants ask for AI assistance, using T-I (Text to Image) to generate spatial layout plans, which were imported and overlaid on the workspace to guide the construction of the low-fidelity models. This phase also involved the use of M-M (Model to Model) to convert each low-fidelity model into a high-fidelity version. In most cases, users did not complete all models, leaving the design in a draft state where low- and high-fidelity models coexisted.

6.1.5.3 Review Phase

In this stage, participants used I-T (Image to Text) to take screenshots of their partially completed design from specific perspectives and submit them to the AI for review. The AI then provided feedback and suggestions for further refinement, offering directions for enhancing the design.

6.1.5.4 Hi-Fi Design Phase

Based on the AI's feedback, users proceeded to refine their models. They used T-I (Text to Image) to generate textures for objects or employed I-I (Image to Image) to have the AI generate lighting schemes directly onto the image.

Moreover, throughout both the Lo-Fi and Hi-Fi stages, users frequently sought AI assistance for operational tasks, such as helping to model a specific object or move an item to a particular position.

While the workflow outlined above represents the common process followed by most participants, there were also unique and unexpected workflows observed, as depicted in Figure 5. These alternative workflows were of particular interest, as they reflected the participants' varying levels of expertise, creative thinking, and personal approaches to design.

For example, some participants skipped certain phases of the process due to their prior experience or well-defined ideas. Participant P1, for instance, had some prior experience with spatial design and thus did not require AI inspiration during the ideation phase. Instead, she immediately used T-I (Text to Image) to generate reference images. Although she initially attempted to use AI to generate models, the results did not meet her expectations. However, she found that the AI-generated models had a distinct, concrete style that could be leveraged for the overall aesthetic of her design. This led her to repeatedly use the AI's model retrieval function, progressively adding concrete-style models to the scene and shaping the design accordingly.

Another interesting variation was observed in Participant P4, who approached the design task by separating the space and models. Designing a futuristic office scene, he treated the space as a model itself, first generating a textured "room box" for the environment and then incrementally generating individual models for doors, windows, chairs, and tables. This modular approach allowed him to focus on different elements of the design separately, creating a flexible and organized workflow.

Participant P7 employed a completely different approach by bypassing the need for a traditional floor plan. Instead, they instructed the AI to generate an isometric view, which simultaneously conveyed both the spatial layout and aesthetic style. The participant then extracted individual images from this isometric view and input them into the AI for further model generation, ensuring stylistic consistency across the generated models. Although this method posed some challenges due to current technological limitations, it showed great potential for maintaining coherence in design styles, and future advancements could further enhance this technique.

Finally, a particularly innovative approach was demonstrated by Participant P12, who utilized the I-I (Image to Image) function's style transfer feature—something no other participants explored.

By applying the distinctive style of the surrealist master Salvador Dalí to the generated models, P12 achieved remarkable results.

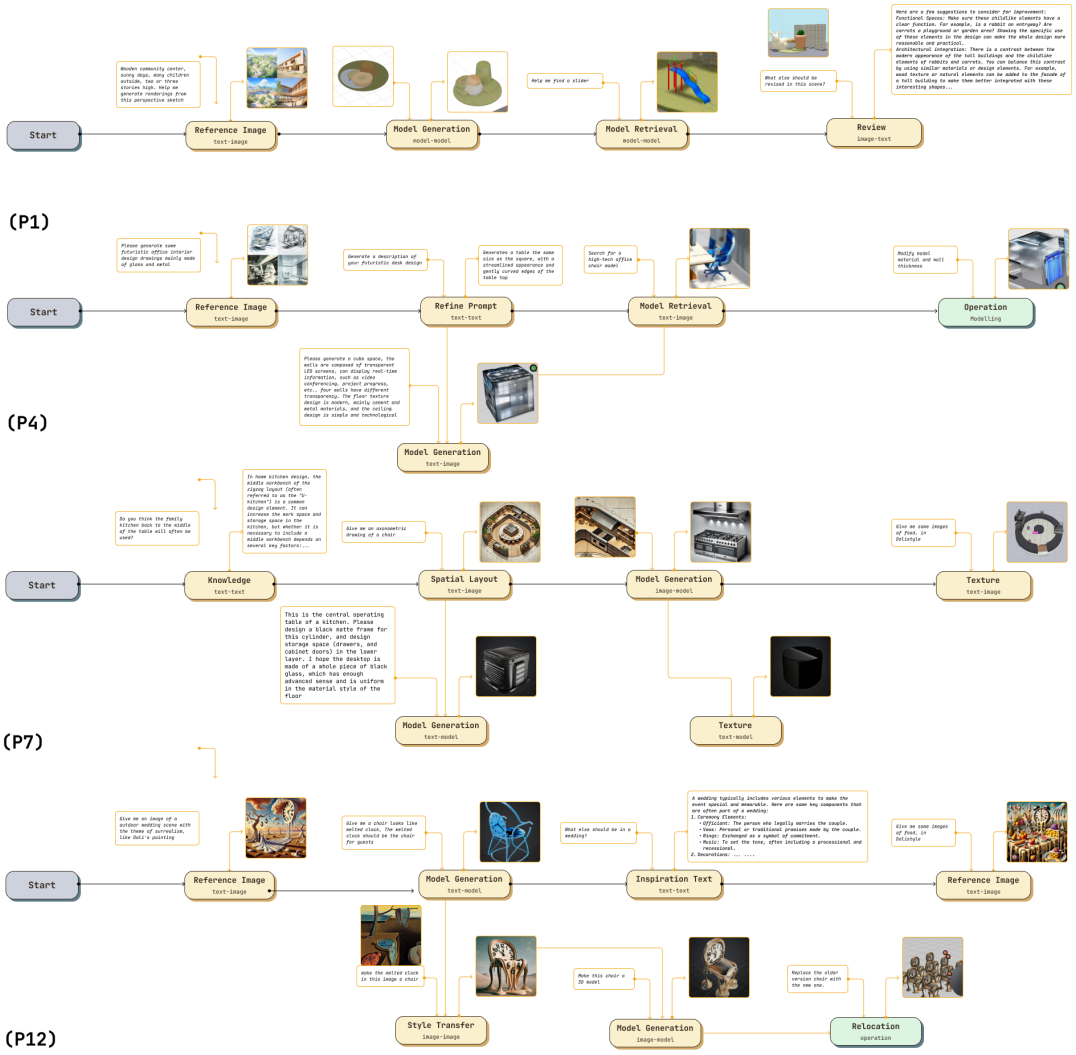

Figure 4: Users' action flow

6.1.6 Findings: Errors in Novice-AI Collaboration in Spatial Design

Our experiment revealed several critical areas where users' expectations of AI capabilities did not align with the system's limitations, alongside gaps in their understanding of spatial design

and modeling logic. These insights offer a deeper understanding of the challenges novice users face in AI-assisted spatial design.

6.1.6.1 Expectation vs. Capability Gaps for AI

One of the most prominent issues in the experiment was the users' inflated expectations of what AI could achieve. Many participants expected the AI to perform tasks far beyond its current abilities, particularly in generating complex or functional scenes.

a. Expectations for Full Scene Generation and Spatial Expansion

Some users assumed that the AI could autonomously generate entire functional scenes or expand an existing scene. For example, **p2** requested an "enterable" cube space but received a solid cube model instead. Similarly, **p11** expected the AI to generate a complete toilet scene based on minimal input, but the AI produced a chaotic result. **p12** tried to zoom out of their scene, expecting the AI to maintain spatial consistency, but it generated an entirely different environment.

Error Cases

> **p2**: The user expected the AI to generate a space they could enter but received a solid cube.
>
> **p11**: The user wanted the AI to generate an entire toilet scene in one step, resulting in a poorly structured model.
>
> **p12**: The user expected the AI to zoom out and expand their scene, but the output was disconnected from their original design.

Insight: Users often overestimated the AI’s ability to generate functional spaces and scenes in a single action, without recognizing the need for more incremental design steps. These expectation gaps reflect the users' unfamiliarity with the current limitations of AI technology in spatial design.

b. Expectations for AI’s Understanding of Modeling Context

Users frequently assumed that the AI could understand the context of their models and accurately generate objects that fit within their scenes. For instance, **p8** expected the AI to create a “cylinder table” within the existing scene, but the AI misinterpreted this and generated a generic operation desk. **p7** assumed the AI would correctly apply a texture to a curved wall, but the output was flawed due to issues with UV mapping and the wall not being unified.

Error Cases:

> **p8**: The AI generated a generic desk instead of the cylinder table the user requested for their scene.
>
> **p7**: The AI failed to apply the texture correctly to a curved wall due to misalignment and poor UV mapping.

Insight: Users expected the AI to accurately interpret their prompts in the context of the scene, but the system struggled with such tasks due to its limited object recognition and spatial awareness. This demonstrates the need for clearer user guidance on the capabilities of AI within specific modeling contexts.

6.1.6.2. Spatial Design and Modeling Logic

Another key challenge was the users' lack of spatial design knowledge and their difficulty in correctly modeling or describing their target objects. These gaps often contributed to the errors encountered during the experiment.

a. Lack of Spatial Design Knowledge

Some errors occurred because users did not fully understand fundamental spatial design principles. For instance, **p2** did not realize that creating an "enterable" space requires designing walls and floors separately. Similarly, **p11** attempted to create a multi-gender toilet in a single step, leading to confusion when the AI generated a mixed-gender design. In both cases, users either lacked the knowledge of how to properly structure their design or expected the AI to fill in these gaps for them.

Error Cases:

> **p2**: The user did not grasp that an enterable space requires separate walls and floors.

p11: The user tried to design a multi-gender toilet in one step, leading to AI confusion and a mixed output.

Insight: Users often lacked the foundational knowledge needed to structure spatial designs effectively, resulting in unrealistic expectations of what the AI could generate. This reflects a need for educational tools or guidance that teach users the basics of spatial design alongside AI usage.

b. Difficulty in Describing or Modeling Target Objects

In some cases, users struggled to accurately describe or model the objects they wanted the AI to generate. **p7** could not correctly apply a texture to a curved wall because they did not account for the complexities of UV mapping. **p8** had trouble conveying that they wanted a specific shape (a cylinder table) in their scene, leading to an incorrect object being generated.

Error Cases:

p7: The user attempted to apply a texture to a curved wall but did not understand the limitations of UV mapping.

p8: The user's request for a cylinder table resulted in a generic desk due to a lack of clarity in their prompt.

Insight: Users often struggled with accurately describing or modeling objects within their designs, leading to errors when the AI could not interpret vague or incorrect prompts. This highlights the importance of developing more intuitive input methods and better teaching users how to structure their requests for AI-assisted design.

6.1.6.3. Reinterpreting User Requirements and AI Functionality

In some cases, users selected the wrong AI function for the task at hand, resulting in errors. For instance, when **p12** tried to place trees in their scene, they called the model-generation AI, which could not handle the request. However, when the wizard switched to text-based AI and generated a guide (e.g., "place trees under the streetlights"), the user was able to quickly follow the instructions and improve their scene.

Insight: This suggests that, at times, user requirements may need reinterpretation. The system could offer suggestions or alternative AI functions based on the task, helping users select the most effective tool for their design process.

6.2 Experiment 2 VR Spatial Design Tools

This research focuses on how people think about natural and intuitive design processes and how they interact with gestural, bodily and verbal systems.

6.2.1 Co-design Session

In this empirical investigation, we recruited a cohort of participants to engage in a phenomenological thought experiment centered on spatial imagination and object creation. The experimental protocol required subjects to close their eyes and mentally conceptualize object generation within an imagined spatial context, with specific attention directed toward articulating the morphological characteristics and potential functional affordances of these conceptualized artifacts.

The experimental session was structured to span a duration of ten minutes, during which comprehensive observational data were systematically collected. The research methodology prioritized the documentation of participants' kinesthetic behaviors, gestural configurations, and linguistic expressions. The primary analytical objective was to examine and elucidate the emergent dynamics of embodied interaction throughout the cognitive process of imaginative object construction.

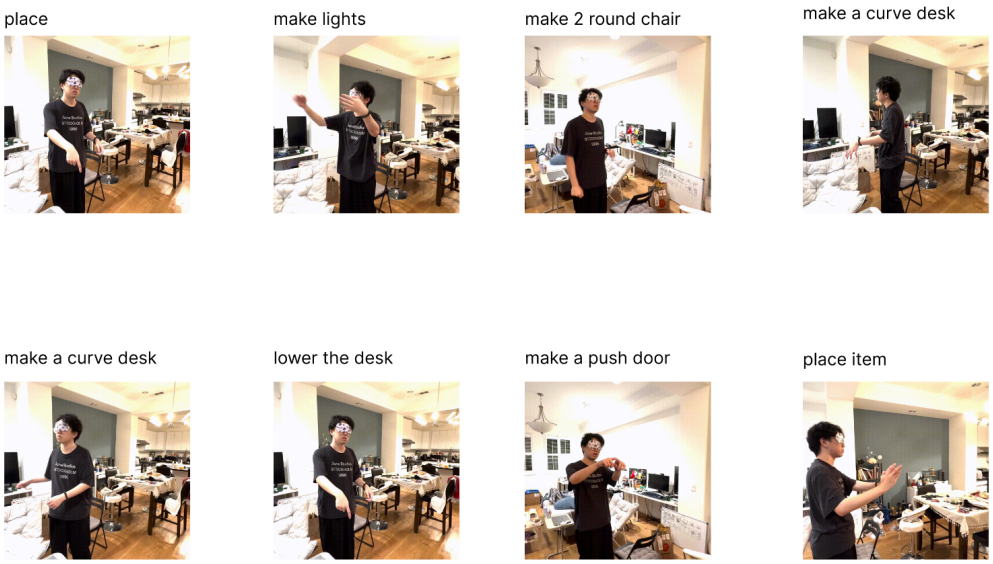

Figure 5: Users' action records

6.2.2 Finding and observation

In the course of this experimental investigation, we observed several intriguing patterns of embodied cognitive representation and interaction. Participants demonstrated distinct gestural strategies for conceptualizing geometric forms: specifically, they tended to employ a unilateral (single-hand) approach when describing circular geometries, while utilizing symmetrical bilateral hand movements to represent quadrilateral or rectangular shapes.

Moreover, the experimental protocol revealed a noteworthy cognitive progression wherein participants would not merely conceive of an object, but subsequently engage in an imaginative simulation of interactive potential. For instance, when a participant conceptually constructed a door within their mental spatial landscape, the subsequent gestural and narrative sequence frequently involved an imagined kinesthetic interaction—such as a pantomimed opening of the door followed by a metaphorical traversal of its threshold.

These observations provide compelling evidence for the intricate relationship between embodied cognition, spatial imagination, and gestural communication, suggesting that object conceptualization is fundamentally an active, dynamically enacted process rather than a static mental representation.

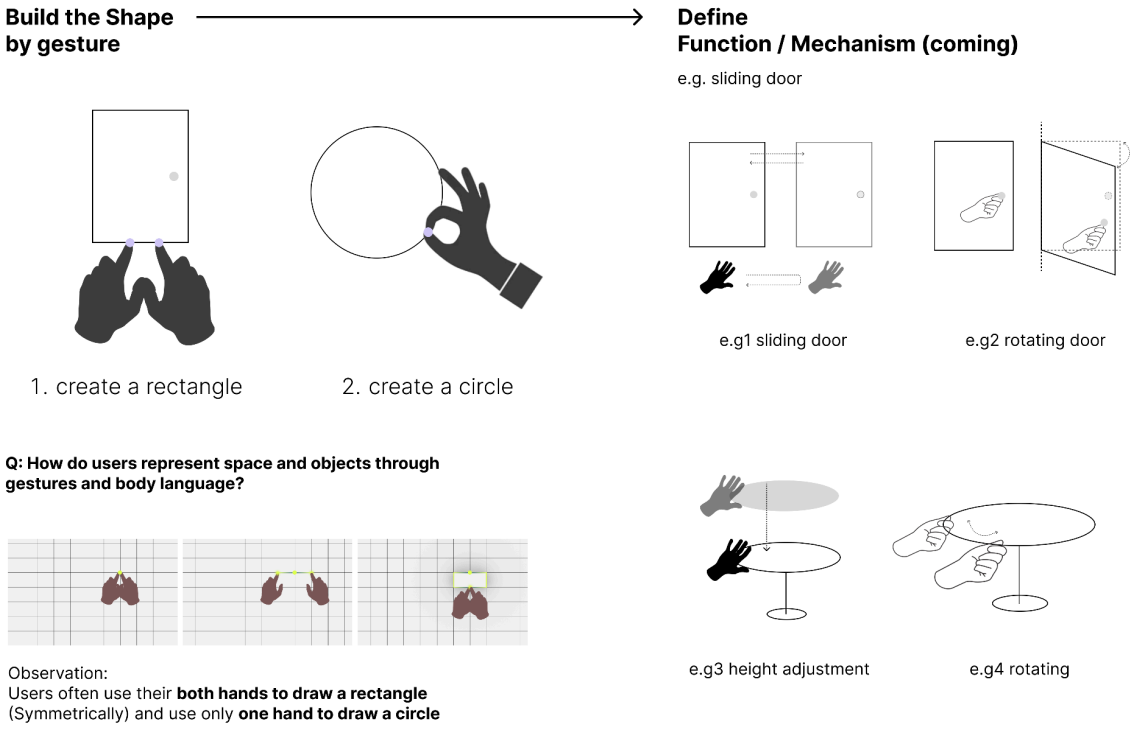

Figure 6: From shape to function

6.2.2 Prototyping

Based on these empirical findings regarding users' gestural patterns of spatial interaction, we proceeded with a novel prototyping approach for a Virtual Reality (VR) product. The primary objective of this design intervention was to enable users to translate their innate, embodied cognitive strategies of object conceptualization directly into the immersive VR environment.

Specifically, our prototype design sought to operationalize the observed gestural taxonomies—such as unilateral circular and symmetrical quadrilateral hand movements—as intuitive interaction mechanisms for spatial object generation. We conducted a subsequent experimental phase to validate the prototype's efficacy.

In this follow-up study, participants were positioned within the virtual spatial environment and tasked with a generative design challenge: to create visual representations that authentically

reflected their cognitive imaginations. The experimental protocol was structured to allow participants maximum creative autonomy, with minimal technological mediation, thus preserving the organic, embodied nature of their spatial ideation processes.

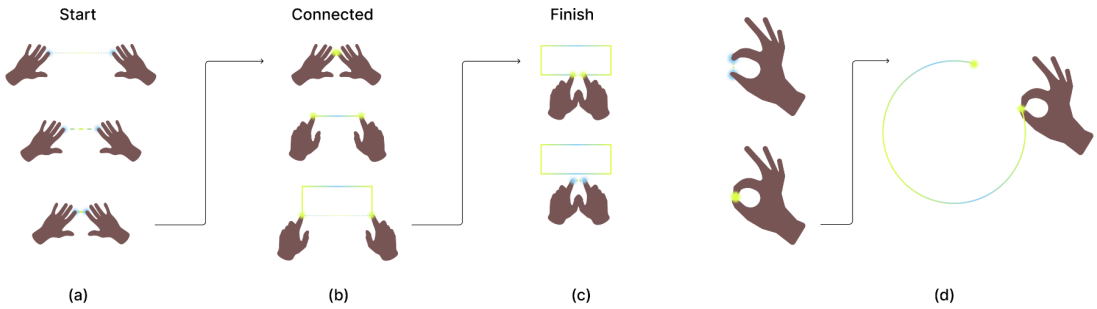

Figure 7: Gestural Interaction Design

Building upon our initial prototype, we developed Prototype 2, which introduced a sophisticated gesture-based object creation mechanism that significantly enhanced the user experience through advanced shape recognition technologies.

In this iteration, participants were instructed to generate objects through gestural drawing, with an intelligent shape recognition algorithm serving as the core translation mechanism for object placement and generation. The algorithmic approach enabled a mapping between users' gestural inputs and virtual object instantiation, with a feature of spatial projection that transcended individual body-scale limitations.

Comparative analysis between Prototype 1 and Prototype 2 revealed compelling insights. The new prototype demonstrated marked improvements in both user satisfaction and task completion metrics. Specifically, users reported a substantially more intuitive and engaging interaction paradigm compared to the previous design.

However, the system has its constraints. While the shape recognition and projection capabilities represented a significant technological advancement, the object generation remained fundamentally constrained by the predefined object library within the system.

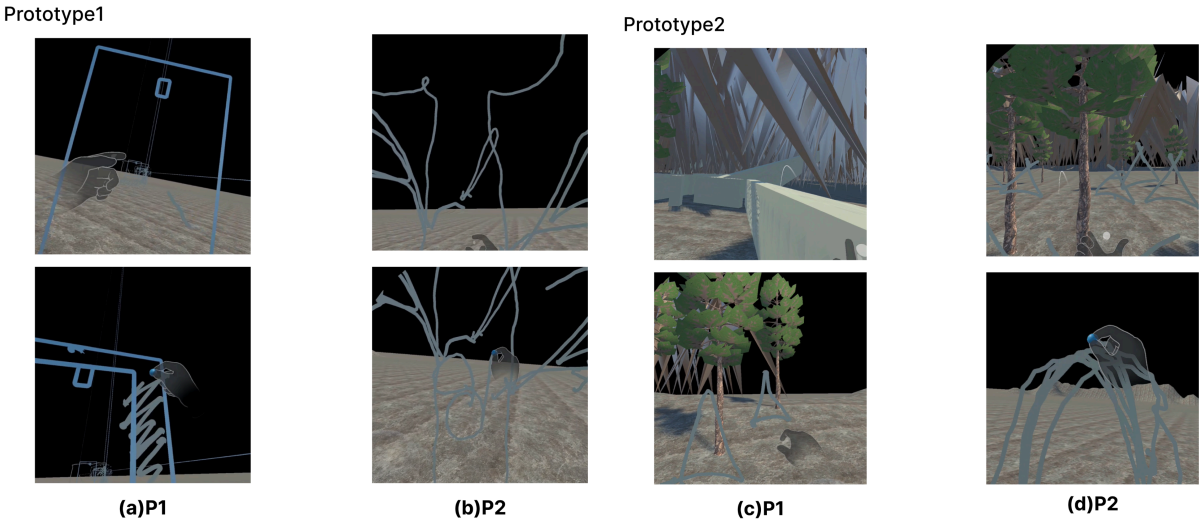

Figure 8: Screenshots of p1,p2 using prototype 1 & 2

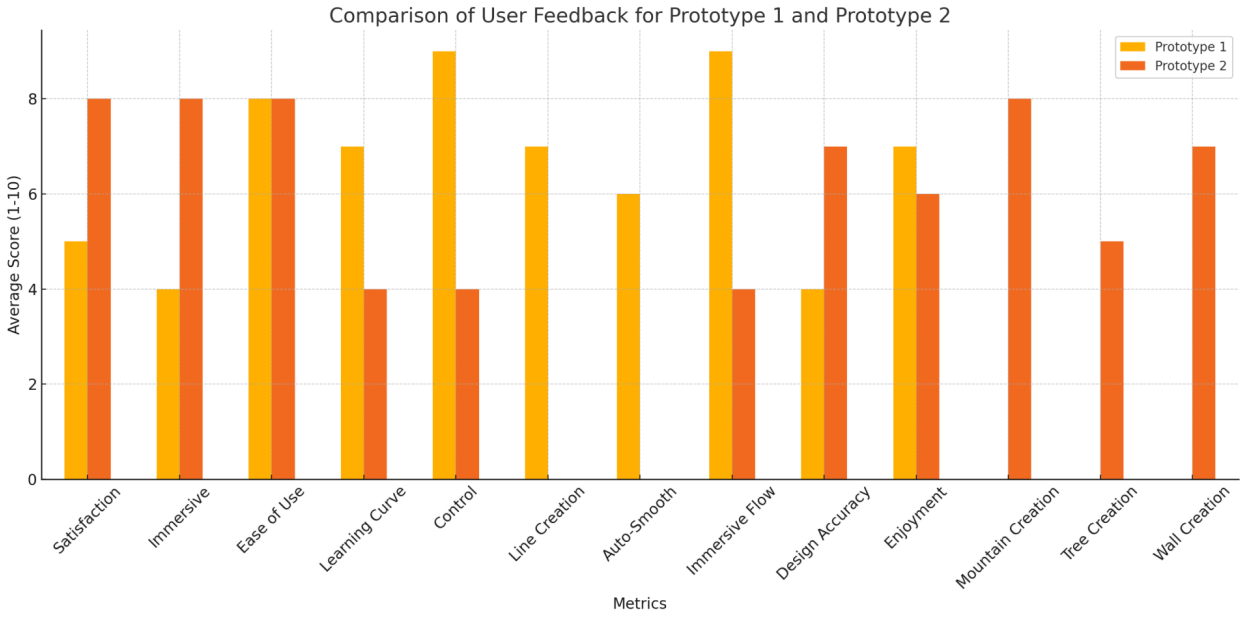

Figure 8: A survey for user experience of prototype 1 & 2

6.3 Insights

Through the systematic investigation of these two experimental prototypes, we have uncovered significant insights into the potential of embodied interaction as a transformative paradigm for user engagement and immersion in virtual environments.

Our research demonstrates that multimodal, multi-channel information input not only enhances user autonomy within interactive systems but also creates novel conceptual and technological spaces for artificial intelligence integration. Specifically, the experimental trajectory revealed critical limitations in existing AI models' approaches to object generation.

While traditional AI models predominantly focus on morphological generation, they frequently lack the nuanced capability to support controlled, incrementally-refined construction processes. This limitation stems from the insufficient granularity of user intent data—a challenge that demands more sophisticated methods of capturing and interpreting human creative expression.

The progression from Experiment 1 to Experiment 2 illuminates a promising trajectory: by synthesizing the gesture-based interaction mechanisms with advanced AI intent recognition, we can conceptualize a more naturalistic and intuitive object creation methodology. In this envisioned system, users would enjoy creative freedom, utilizing a comprehensive array of communicative modalities—including gestural, vocal, and corporeal expressions—to articulate their design intentions with remarkable precision.

The research suggests that future interaction design should prioritize systems that can holistically interpret and respond to the rich, nuanced communicative repertoire of human embodied expression, thus bridging the current phenomenological gap between human imagination and technological mediation.

7. Biography

[1] Alissa N. Antle, Greg Corness, Saskia Bakker, Milena Droumeva, Elise van den Hoven, and Allen Bevans. 2009. Designing to support reasoned imagination through embodied metaphor. Proceedings *of the seventh ACM conference on Creativity and cognition* (*C&C '09*), October 26, 2009. Association for Computing Machinery, New York, NY, USA, 275–284. https://doi.org/10.1145/1640233.1640275

[2] Suresh K. Bhavnani, Bonnie E. John, and Ulrich Flemming. 1999. The strategic use of CAD: an empirically inspired, theory-based course. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems* (*CHI '99*), May 01, 1999. Association for Computing Machinery, New York, NY, USA, 183–190. https://doi.org/10.1145/302979.303036

[3] Yulong Bian, Chao Zhou, Wei Gai, Juan Liu, and Chenglei Yang. 2023. The effect of embodied interaction designs on flow experience: examination in VR games. *Virtual Real.* 27, 2 (June 2023), 1549–1565. https://doi.org/10.1007/s10055-023-00758-3

[4] Paul Dourish. Embodied Interaction: Exploring the Foundations of a New Approach to HCI.

[5] Paul Dourish. *Where the Action Is: The Foundations of Embodied Interaction*. Retrieved November 23, 2024 from https://direct.mit.edu/books/monograph/3875/Where-the-Action-IsThe-Foundations-of-Embodied

[6] S.R. Ellis. 1994. What are virtual environments? *IEEE Comput. Graph. Appl.* 14, 1 (January 1994), 17–22. https://doi.org/10.1109/38.250914

[7] Martin Heidegger. 2010. *Being and Time*. SUNY Press.

[8] Eva Hornecker. 2011. The role of physicality in tangible and embodied interactions. *interactions* 18, 2 (March 2011), 19–23. https://doi.org/10.1145/1925820.1925826

[9] Eva Hornecker and Jacob Buur. 2006. Getting a grip on tangible interaction: a framework on physical space and social interaction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (*CHI '06*), April 22, 2006. Association for Computing Machinery, New York, NY, USA, 437–446. https://doi.org/10.1145/1124772.1124838

[10] Lian Loke and Toni Robertson. 2013. Moving and making strange: An embodied approach to movement-based interaction design. *ACM Trans Comput-Hum Interact* 20, 1 (April 2013), 7:1-7:25. https://doi.org/10.1145/2442106.2442113

[11] Elena Márquez Segura, Laia Turmo Vidal, Asreen Rostami, and Annika Waern. 2016. Embodied Sketching. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (*CHI '16*), May 07, 2016. Association for Computing Machinery, New York, NY, USA, 6014–6027. https://doi.org/10.1145/2858036.2858486

[12] G. Riva, M. T. Anguera, and B. K. Wiederhold. 2006. *From Communication to Presence: Cognition, Emotions and Culture Towards the Ultimate Communicative Experience*. IOS Press, Incorporated, Amsterdam, NETHERLANDS, THE. Retrieved November 26, 2024 from http://ebookcentral.proquest.com/lib/socal/detail.action?docID=280878

[13] Giuseppe Riva and Fabrizia Mantovani. 2012. From the body to the tools and back: A general framework for presence in mediated interactions☆. *Interact. Comput.* 24, 4 (July 2012), 203–210. https://doi.org/10.1016/j.intcom.2012.04.007

[14] John R. Searle. 1983. *Intentionality: An Essay in the Philosophy of Mind*. Cambridge University Press.

[15] Mel Slater and Sylvia Wilbur. 1997. A Framework for Immersive Virtual Environments (FIVE): Speculations on the Role of Presence in Virtual Environments. *Presence Teleoperators Virtual Environ.* 6, 6 (December 1997), 603–616. https://doi.org/10.1162/pres.1997.6.6.603

[16] Jakob Tholander and Martin Jonsson. 2023. Design Ideation with AI - Sketching, Thinking

and Talking with Generative Machine Learning Models. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference* (*DIS '23*), July 10, 2023. Association for Computing Machinery, New York, NY, USA, 1930–1940. https://doi.org/10.1145/3563657.3596014

[17] Mathias Peter Verheijden and Mathias Funk. 2023. Collaborative Diffusion: Boosting Designerly Co-Creation with Generative AI. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems* (*CHI EA '23*), April 19, 2023. Association for Computing Machinery, New York, NY, USA, 1–8. https://doi.org/10.1145/3544549.3585680

[18] Zijun Wan, Jiawei Tang, Linghang Cai, Xin Tong, and Can Liu. 2024. Breaking the Midas Spell:Understanding Progressive Novice-AI Collaboration in Spatial Design. https://doi.org/10.48550/arXiv.2410.20124

[19] Miao Wang, Lyu Xu-Quan, Li Yi-Jun, and Zhang Fang-Lue. 2020. VR content creation and exploration with deep learning: A survey. *Comput. Vis. Media* 6, 1 (March 2020), 3–28. https://doi.org/10.1007/s41095-020-0162-z

[20] Bob G. Witmer, Christian J. Jerome, and Michael J. Singer. 2005. The Factor Structure of the Presence Questionnaire. *Presence Teleoperators Virtual Environ.* 14, 3 (June 2005), 298–312. https://doi.org/10.1162/105474605323384654

[21] Zhuohao Wu, Danwen Ji, Kaiwen Yu, Xianxu Zeng, Dingming Wu, and Mohammad Shidujaman. 2021. AI Creativity and the Human-AI Co-creation Model. In *Human-Computer Interaction. Theory, Methods and Tools*, 2021. Springer International Publishing, Cham, 171–190. https://doi.org/10.1007/978-3-030-78462-1_13

[22] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding Conditional Control to Text-to-Image Diffusion Models. https://doi.org/10.48550/arXiv.2302.05543

[23] Reality3DSketch: Rapid 3D Modeling of Objects From Single Freehand Sketches | IEEE Journals & Magazine | IEEE Xplore. Retrieved November 27, 2024 from https://ieeexplore-ieee-org.libproxy1.usc.edu/document/10295995

[24] From presence to consciousness through virtual reality. - Document - Gale Academic OneFile. Retrieved November 26, 2024 from https://go-gale-com.libproxy2.usc.edu/ps/i.do?p=AONE&u=usocal_main&id=GALE%7CA188972927&v=2.1&it=r

[25] Explanation in Human-AI Systems: A Literature Meta-Review, Synopsis of Key Ideas and Publications, and Bibliography for Explainable AI - ProQuest. Retrieved November 27, 2024 from https://www.proquest.com/docview/2177026156?pq-origsite=primo&sourcetype=Working%20Papers

[26] From Geometry to Behavior. Retrieved November 28, 2024 from https://mitpress.mit.edu/9780262547116/from-geometry-to-behavior/