More Than Speed: User Agency and Social Comfort in Speech-based Interaction for Multiscale Medical AR Applications

Vasudev Agarwal* Texas A&M University
Regis Kopper† Iowa State University
Jong-in Lee‡ Texas A&M University

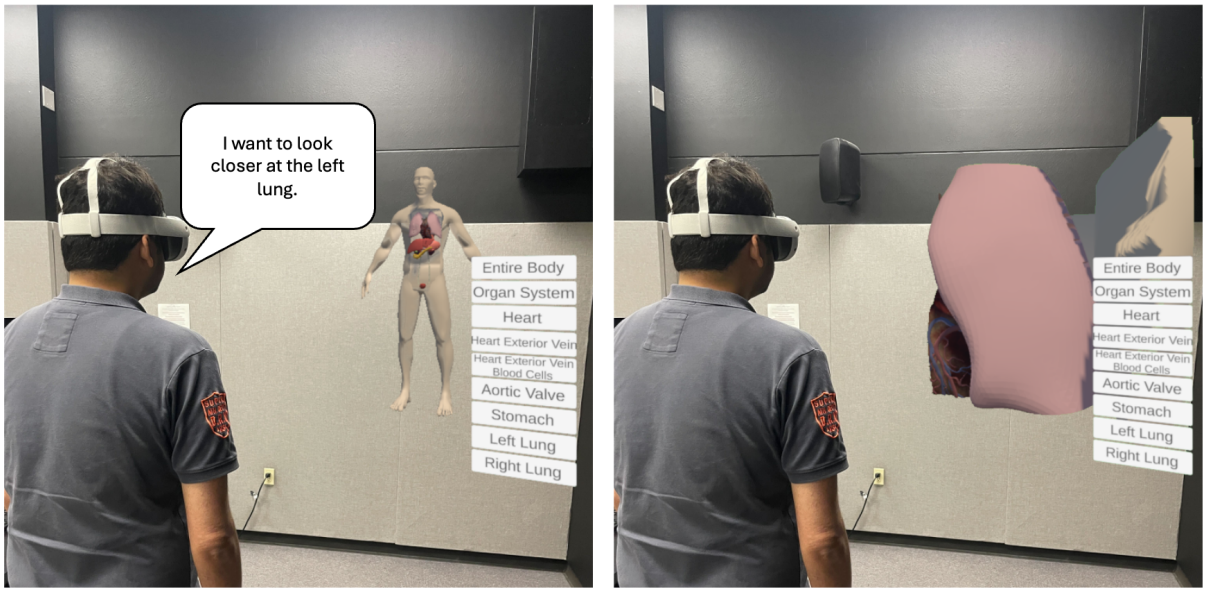

Figure 1: Our speech-based interaction technique enables intuitive exploration of 3D multiscale environments in Augmented Reality by interpreting conversational natural language, rather than requiring predefined commands. This flexible approach allows users to seamlessly transition across levels of scale through target-directed viewpoint control, supporting more natural interaction with complex spatial content.

Abstract

In augmented reality (AR), voice-based input offers a hands-free approach to interacting with 3D content across multiple domains, including medical training and surgical rehearsal environments. While existing 3D user interface research has explored voice-based interaction techniques, these have primarily focused on simple manipulation tasks using predefined commands. We present a natural speech- based technique for target-directed viewpoint control that enables users to navigate across different levels of scale in 3D multiscale data. Using an explanatory sequential mixed-methods study, we found that users tended to favor AI-assisted navigation for its autonomy, lack of social pressure, and sense of control through overall preference differences were not statistically significant. Notably, qualitative findings indicated that most users did not perceive AI-related response delays as more disruptive than those with human assistance. Our findings suggest that natural speech interfaces in AR may yield strong user acceptance, potentially driven more by psychological benefits—such as user agency and social comfort—than by absolute speed alone. These exploratory insights warrant further investigation in future research.

*e-mail: vasu14devagarwal@tamu.edu
†e-mail: kopper@iastate.edu
‡e-mail: jongin@tamu.edu

Index Terms: Human-centered computing Interaction techniques, Navigation, Augmented Reality, Voice User Interface

1 Introduction

The rapid advancement of augmented reality (AR) technologies has transformed how users interact with digital content across several domains, including healthcare, education, and training [24, 30, 7, 54, 6]. As AR systems become increasingly integrated into professional workflows, the need for more intuitive and efficient interaction methods continues to grow. Traditional AR interaction techniques—including controllers, hand gestures, and gaze-based methods—can be limiting, especially when users need their hands for primary tasks [40].

In medical AR applications, the need for hands-free interaction arises not only during active clinical procedures, where sterility and instrument handling are critical, but also throughout surgical rehearsal, simulation, and anatomy training environments [16, 21, 4, 8, 13, 2, 38, 15]. These settings increasingly aim to replicate real clinical conditions, including gloved hands, limited physical interaction, and sustained cognitive focus. In these contexts, interaction techniques that depend on manual input or frequent physical gestures can disrupt training flow and reduce realism [20, 36].

Voice-based interaction therefore offers a compelling solution for minimizing manual input in AR applications. While voice commands have been explored for navigation and manipulation, most implementations still rely on rigid, predefined command sets [55]. This restricts user agency and hinders fluid, conversational interaction. Alternative hands-free modalities have also been investi-

gated, such as gaze tracking, head-pose, or brain-computer interfaces. Each brings unique affordances but faces barriers for intuitive, multiscale navigation: gaze and head pose can lack the precision and context required for complex tasks, while brain-computer interfaces remain limited by current technology and user training demands [19, 26, 46]. None yet achieves the naturalness and flexibility required to seamlessly manage Level-of-Scale (LoS) transitions in complex 3D environments.

Navigating multiscale 3D content in AR—such as anatomical models, engineering assemblies, or architectural layouts—poses unique challenges. Users must fluidly transition between details and overviews while maintaining spatial awareness [43]. Current navigation techniques that enable such transitions often demand manual gestures or interruptive controller use, which can disrupt workflow.

To address these gaps, we present a natural speech-based technique for target-directed viewpoint control in multiscale AR environments as a key building block toward complete speech-based navigation interfaces. Leveraging advances in natural language processing, our system interprets users' conversational speech to enable intuitive transitions across levels of scale—without requiring memorized command sets [55]. This enables users to express navigation intentions in their own words, supporting more natural and context-aware interaction. While our current prototype employs push-to-talk activation for robust speech capture, the underlying natural language processing pipeline and interaction paradigm represent foundational components for future fully hands-free systems.

We report findings from an explanatory sequential mixed-methods user study [18, 56] comparing our AI-assisted speech interaction with human-assisted baseline conditions. Beyond assessing usability and efficiency, our study reveals key psychological factors shaping user acceptance: participants highly valued the autonomy, agency, and social comfort provided by AI-assisted interaction, with many preferring it even when response times were slightly slower than human assistance. These findings suggest that user acceptance of speech-based AR interfaces is driven more by experiential and psychological benefits than by absolute speed alone.

In the following sections, we detail our speech-based viewpoint control technique and discuss both technical and user-centered design implications based on our empirical evaluation.

2 Related Work

2.1 Hands-Free Interaction Techniques in AR

Prior work on hands-free interaction in AR has explored a wide range of modalities, including gesture-based, speech-based, gaze-based, and multimodal approaches. Multimodal interaction—especially the combination of gesture and speech—has been shown to provide more natural, efficient, and satisfying user experiences compared to unimodal techniques, supporting intuitive manipulation of virtual objects in 3D environments [65, 25, 35, 34, 68]. For instance, Williams et al. found that multimodal gesture-speech input enables natural-feeling interactions and reduces perceived workload, while user-defined gesture sets can further enhance intuitiveness [65, 68].

Comparative studies reveal that gesture controls are often preferred for their playfulness and directness, while voice commands are valued for efficiency and hands-free operation, especially in navigation tasks [28, 41]. Korkiakoski et al. reported no significant difference in overall user experience between gesture and voice controls, but noted individual variability in preferences, suggesting that context and user characteristics play a key role [28]. Gaze-based and gaze-assisted techniques, such as Gaze-Hand Alignment, have also been shown to outperform hands-only input for selection and manipulation tasks, offering rapid, precise, and less physically demanding interaction [37, 61].

Recent work has expanded the multimodal paradigm to include gaze, gesture, and speech in flexibly configurable systems, demonstrating that combining all three modalities can yield superior efficiency and user satisfaction [63, 62]. Other studies have explored the use of head gestures, eye tracking, and even wearable devices (e.g., smart rings) to enable subtle, private, or accessible hands-free interactions, further broadening the design space [61, 27, 45].

Systematic reviews and empirical studies consistently highlight that the optimal interaction technique depends on the specific task, user context, and individual preferences [41, 40, 42]. For instance, gaze and dwell are often preferred for selection tasks due to their speed and low workload, while multimodal approaches excel in complex manipulation or collaborative scenarios [62, 35, 40]. However, many prior systems rely on fixed command sets or require users to learn specific gesture-speech combinations, which can limit flexibility and expressiveness [65, 42].

In contrast, our approach enables users to navigate using natural, conversational speech interpreted by large language models, allowing for greater flexibility and expressiveness. Additionally, we uniquely address psychological factors like user agency and social comfort, which are often overlooked in previous systems. Recent literature also calls for more research into these social-psychological dimensions, as user preferences are shaped not only by technical performance but also by the desire for autonomy, comfort, and adaptability [28, 27, 41].

2.2 Speech-based Navigation in AR

Speech-based navigation has emerged as a promising hands-free interaction technique in augmented reality (AR), offering naturalness and accessibility across diverse application domains. Early systems demonstrated the feasibility of using speech interfaces for city exploration, enabling users to receive information about landmarks and request further details through voice commands, thus supporting intuitive and non-invasive navigation experiences [9]. More recent work has focused on integrating natural language understanding (NLU) into AR navigation interfaces, showing that NLU-powered systems improve command accuracy and ease the learning curve for new users compared to traditional voice interfaces without NLU [66].

Speech-based AR navigation has also been applied to support visually impaired users, where voice commands enable hands-free operation and synthesized speech provides real-time navigation instructions and object descriptions, enhancing mobility and independence [10, 59]. In medical and industrial contexts, voice user interfaces have been shown to facilitate effortless navigation and scene adjustment, especially when users' hands are occupied or sterility must be maintained, with studies reporting high user satisfaction and usability [22, 57].

Despite these advances, challenges remain in achieving robust real-time responsiveness, handling semantic ambiguities, and supporting truly conversational, flexible speech input. Furthermore, there is a persistent gap in addressing the psychological and social dimensions of speech-based AR navigation—such as user agency, comfort, and emotional expression—especially in real-time, dynamic environments [33, 58]. Our research addresses these gaps by developing a natural speech-based technique for target-directed viewpoint control and demonstrating that user acceptance depends more on psychological factors—particularly agency and social comfort—than on absolute performance speed. This work establishes both technical and user-centered foundations for future hands-free medical AR systems.

3 Speech-Based Target-Directed Viewpoint Control for Multiscale Medical AR

We developed a speech-based technique for target-directed viewpoint control that enables users to navigate between levels of scale

(LoS) in 3D multiscale data using natural speech prompts. To support natural speech prompts instead of predefined voice commands, the system employs a two-stage speech detection pipeline that combines real-time transcription and intent interpretation through a large language model that extracts the user's intended action and target from free-form speech. Together, these components allow the system to handle a wide variety of natural user inputs while maintaining consistent recognition accuracy during real-time interaction (Figure 2).

3.1 Speech Detection for Navigation

Our speech detection system consists of two core components: speech transcription and intent interpretation. Each component plays a critical role in enabling natural voice-based interaction. The system pipeline begins with detecting the user's utterance and transcribing it into text, then interpreting the transcribed text into a structured action-target pair that drives viewpoint transitions.

Speech Transcription. The first component converts recorded speech into text for downstream processing. When the user releases the right trigger, indicating the end of an utterance, the audio segment is transmitted via the Azure Speech-to-Text (STT) SDK. The API returns a text transcription of the user's speech. Azure STT was selected after testing due to its higher accuracy with non-native English speakers compared to open-source automatic speech recognition (ASR) models such as Whisper-Tiny [48]. While a larger model could offer more accuracy, it would require either cloud hosting with fine-tuning, which would introduce additional latency and complexity.

Intent Interpretation. The second component extracts the user's intended action and target from the transcribed text. This step uses Anthropic's Claude API [1], where we employ a large language model with a structured system prompt that constrains the model to produce a predefined output schema. (See Supplementary Appendix B for full prompt specifications and constraints.) Rather than performing a deterministic mapping, the system prompt provides domain-specific context and narrows the space of permissible outputs while preserving flexibility in natural language interpretation.

Initial baseline testing with BERT embeddings *all-MiniLM-L6-v2* [49] showed low accuracy for naturalistic phrasing, lacking domain-specific nuances. The LLM approach was selected for superior zero-shot reasoning, enabling immediate flexibility and high accuracy without training data. Future work may explore fine-tuned embeddings models to optimize latency. The LLM system prompt enforces two structured fields: *action* and *target*. The action field is restricted to a few, predefined set of navigation operations (e.g., zoom-in and zoom-out). The target field is constrained to a predefined anatomical hierarchy that enumerates all valid navigable body structures in the scene. In addition, the prompt includes illustrative examples to guide the model toward consistent structured generation.

For example, given the user utterance "Center the view on the heart exterior vein blood cells" produces JSON output: (action = zoom-in, target = Heart Exterior Vein Blood Cells, action confidence = 1.00, target confidence = 1.00). Another example, the utterance "to the stomach" is interpreted as (action = zoom-in, target = Stomach, action confidence = 0.70, target confidence = 1.00). The confidence values serve as interaction safeguards rather than evaluation metrics. When confidence falls below 0.70 for either field, the system requests user clarification instead of executing the command. A distinctive double haptic pulse via the controller alerts the user, while a visual prompt on the controller displays asks for reissue their command via push-to-talk. This multi-modal feedback approach reduces unintended responses while preserving natural, flexible interaction.

Push-to-Talk Activation. In addition to the core speech detection functionality, we implemented a push-to-talk feature to allow users to explicitly control the timing of speech input. Users press and hold the right controller trigger to initiate speech recording. A brief double haptic pulse signals that the system is ready to capture audio. Upon release of the trigger, a single haptic pulse confirms that the speech input has been received and is being processed.

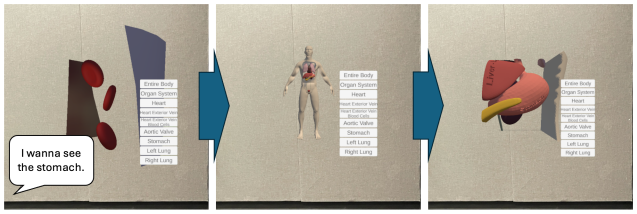

Figure 2: When the system detects the user's speech regarding target selection, it employs a two-phase viewpoint transition approach to maintain spatial context. First, it zooms out to the least common ancestor of the current and target LoS, then zooms in to the requested target, helping users preserve their spatial understanding throughout multiscale transitions.

We adopted this push-to-talk approach after encountering significant limitations with voice activity detection (VAD) during preliminary testing. VAD suffered from frequent false activations in noisy environments and failed to capture complete utterances when users paused mid-sentence, as the system would prematurely terminate recording during brief silences. The push-to-talk mechanism addresses these issues by giving users explicit control over speech boundaries, thereby ensuring robust and reliable speech capture throughout our user study. While this design choice means our current implementation does not achieve fully hands-free operation, it establishes the core natural language processing pipeline and interaction paradigm that serve as foundational components for future systems.

To achieve true hands-free operation, we propose experimenting with openWakeWord [44], an open-source framework that integrates Silero VAD. openWakeWord enables custom WakeWord recognition and offers edge-deployment solutions optimized for resource constrained XR devices. This approach will directly address our identified VAD failure modes by combining wake-word gating with integrated voice activity detection, eliminating push-to-talk requirements while maintaining robust speech capture. Similarly, frameworks like CognitiveEMS [64] demonstrate edge-centric architecture for hands-free operation without cloud dependency. Adapting similar continuous listening models would overcome current latencies and VAD errors, ensuring reliability in disconnected clinicial environment.

3.2 Viewpoint Transition between LoS

We designed seamless viewpoint transitions between LoS by dynamically adjusting the focus of the 3D multiscale data within its fixed-position clipping box, minimizing viewer disorientation through smooth viewpoint changes (Figure 2). The viewpoint transition employs a two-phase approach: first "zoom-out" then "zoom-in"—building upon techniques established in previous work [31, 32].

For these phases, the camera is not being transformed. Instead, we scale the model according to the LoS at the target viewpoint, while keeping that viewpoint centered inside a bounding box at a fixed distance from the camera. This bounding box is used to clip parts of the model that are not in focus at the current LoS so that they do not obstruct the real environment around the user.

When zooming out, the system transitions to the LoS of the least common ancestor (the body LoS in Figure 2) between the current

and target LoS. During this phase, the model is scaled down over time through linear interpolation to the LoS of the least common ancestor while keeping it centered within the bounding box. Once this phase is finished, the zoom-in phase begins, where the system transitions to the target viewpoint, and the model is scaled up over time through linear interpolation toward the LoS at the target viewpoint. This completes the viewpoint transition with the target viewpoint in focus, and parts of the model outside the bounding box clipped to minimize occlusion of the real physical space.

4 User Evaluation Overview

We adopted an explanatory sequential mixed methods approach to evaluate our natural speech-based AR scale control technique. First, we conducted a quantitative user study to assess usability and task performance using both **AI-assisted** and **Human-assisted** viewpoint control techniques. The **AI-assisted** technique represents our novel speech-based viewpoint control technique, while the **Human-assisted** viewpoint control technique serves as a baseline condition in which a human assistant controls the view of a 3D multiscale model based on the primary user's verbal commands within the AR application. **Human-assisted** viewpoint control as our baseline is motivated by its established use in high-stakes environments, such as AR-supported medical procedures, where an assistant operates digital interfaces and manages patient data by directly following the surgeon's verbal instructions [17]. In such contexts, the human assistant interprets commands, manipulates AR visualizations, and adapts to unexpected situations in real time, ensuring reliability, accuracy, and effective collaboration throughout the procedure. This paradigm establishes a high standard for responsiveness, adaptability, and user-centered control—making it an ideal baseline for benchmarking novel hands-free, AI-driven systems. Building upon the quantitative findings, we subsequently conducted qualitative interviews designed to explain user preferences, illuminate experiences with autonomy and social interaction, and explore contextual factors influencing user acceptance and technology adoption.

5 Quantitative Phase

In the first phase, we conducted a within-subjects user study in which participants performed viewpoint control tasks within a hierarchical 3D anatomy AR environment using both **AI-assisted** and **Human-assisted** techniques. We collected multiple forms of data to comprehensively evaluate each technique: (1) objective performance metrics, including response time and task accuracy; (2) cognitive workload measures using the NASA Task Load Index (NASA-TLX); and (3) subjective assessments through post-task questionnaires evaluating ease of use, perceived speed, perceived accuracy, and overall preference between the two methods.

5.1 Experimental Design

We employed a within-subjects design with constant target difficulty to isolate effects of AI vs. Human-assisted viewpoint control, independent of naming complexity Each session paired two participants: one as primary user, one as assistant. After the first participant completed both AI-assisted and Human-assisted conditions, they switched roles. We used counterbalancing (Latin Square) to control for order effects. Each participant pair completed four blocks: two as primary user, two as assistant, with interface order randomized Importantly, participants switched roles only once during each session, with the role switch occurring between the second and third condition across all orderings. This constraint was imposed to minimize cognitive disruption and repeated role transitions, while still allowing the interaction mode order to vary before and after the role switch.

In the **AI-assisted** condition, participants interacted directly with the speech-based viewpoint control system. They were instructed to navigate to a predefined list of anatomical targets using natural speech while holding the controller's right trigger to activate the push-to-talk function. Each utterance was processed by the speech pipeline, which generated an action-target pair and automatically triggered the corresponding camera movement in the XR environment. Participants could use any natural phrasing or refer to body parts in any order. In the **Human-assisted** condition, a human assistant monitored the participant's speech in real time and manually executed the corresponding navigation action by pressing the appropriate UI button. This setup mirrored the **AI-assisted** workflow but replaced automated speech interpretation with human mediation, providing a baseline for comparison in accuracy, responsiveness, and user experience.

5.2 Participants

A total of 16 participants (9 female, 7 male, ages 18-61) were recruited from the university and the local community. All of the participants had experience with speech technology, e.g., Amazon Alexa and Apple Siri; 14 participants had experience with XR technology.

5.3 Apparatus and Environment

The experiment was conducted using a Meta Quest 3S headset equipped with controllers for interaction. The headset provides a per-eye resolution of 2064 X 2208 pixels. Participants performed all tasks in a 3 m x 1.5 m controlled lab space while seated on chairs.

The virtual environment comprised a hierarchical, multiscale 3D human-anatomy model, rendered in Unity. A complete human body was included. Camera transitions between scales were implemented through smooth interpolation, enabling seamless navigation between macroscopic and microscopic regions.

Both **AI-assisted** and **Human-assisted** conditions used the same virtual environment and viewpoint targets. A heads-up display (HUD) presented the current navigation target, task instructions, and a directional cue guiding user movement. For the **Human-assisted** interface, the observer's desktop view mirrored the headset display and included UI buttons corresponding to each anatomical target.

5.4 Task

During each experimental session, participants performed a navigation task within a virtual 3D multiscale environment. While wearing the headset, participants viewed a list of body parts and were instructed to navigate to each target in any order, as quickly and accurately as possible, using natural speech. the target list was provided to standardize task difficulty across participants and was not intended to guide spatial navigation. A target was considered successfully reached when the virtual camera centered on the corresponding body part. Each target was visually distinct and spatially separated, requiring participants to deliberately adjust their viewpoint rather than continuously scan nearby areas.

Each navigation task consisted of selecting a single target from the list displayed as a floating panel on the left side of the user's view, which remained visible throughout the task. The participants navigated to the selected target by issuing a spoken command using natural language. For example, participants could say phrases such as "zoom into the heart" or "move to the stomach". After a spoken command was given, the system interpreted the intended target and initiated a camera transition to the corresponding target. Once the target was reached, it was visually marked as completed in the list before the participants went to the next target.

The targets were distributed across different anatomical regions and scales, requiring participants to reposition their viewpoint for each navigation task rather than rely on local scanning or mirror adjustments.

Participants completed navigation tasks under both **AI-assisted** and **Human-assisted** speech interface conditions. In the **AI-assisted** condition, the system automatically triggered the corresponding viewpoint transition upon completion of speech transcription and interpretation. In the **Human-assisted** condition, a human assistant listened to the participant's speech and manually initiated the same transition by pressing the appropriate GUI button.

A blind study design was considered to isolate user experience between AI-assisted and Human-assisted conditions. However, enforcing such separation would have created artificial interaction that don't reflect real-world use. Ecological validity was therefore a key consideration in our design choice, as in practical AR navigation scenarios users are aware of whether control is meditated by an AI or human assistant. Maintaining this awareness was necessary to preserve realistic user perceptions and experience.

5.5 Procedure

This study was approved by the Institutional Review Board (IRB) and upon giving informed consent, participants received a brief orientation explaining the study's purpose and procedures. They then completed a demographic and experience questionnaire assessing their personality traits, prior exposure to AR/VR systems, and speech-based interfaces. Next, participants performed a short practice session to familiarize themselves with the push-to-talk mechanism, the speech interface within the AR environment. The practice phase continued until participants demonstrated comfort and proficiency with the interaction methods. Following the practice session, participants completed the main navigation tasks under both **AI-assisted** and **Human-assisted** conditions, with the order counterbalanced as described previously. In each condition, the participants completed two blocks of tasks, with each block consisting of 7 navigation tasks targeting different anatomical structures. This resulted in a total of 14 navigation tasks per condition. Participants were allowed to select anatomical targets in any order and issued navigation commands using natural speech. After completing both blocks within each condition, participants assessed their cognitive workload by responding to the NASA TLX questionnaire. Upon completing all tasks under both conditions, participants completed post-task questionnaires evaluating usability, perceived speed and accuracy, satisfaction, sense of control, and overall interface preference. At the end of the session, participants were debriefed, thanked for their participation, and compensated for their time.

5.6 Results

We analyzed both objective performance measures and subjective user perceptions to evaluate the two viewpoint control techniques. One-way ANOVAs were performed for continuous, normally distributed data, and Aligned Rank Transform (ART) ANOVAs for ordinal or non-normally distributed data. All significance testing employed an alpha level of .05, and effect sizes are reported where applicable. The following subsections present findings for each measure.

5.6.1 Response Time

Response time served as a key metric for evaluating interaction efficiency across the **AI-assisted** and **Human-assisted** conditions. To ensure comparability between conditions, response time was defined as the duration between the start of the camera transition and end of user speech. These temporal markers were consistently logged throughout the user study to provide a precise, system-level measure of task execution latency. This metric reflects the total delay experienced from the moment the system receives the user's final input to the initiation of the corresponding visual response, thereby capturing the responsiveness of each assist type.

A one-way ANOVA was conducted to examine the effect of Assist Type on response time. The analysis revealed a significant main effect of Assist Type, F(1, 446) = 141.60, $p < .0001$. As illustrated in figure 3, participants took significantly longer to respond in the AI Assisted condition ($M = 2.00s$, $SD = 0.53$, 95% CI [1.90, 2.10]) compared to the Human Assisted condition ($M = 1.00s$, $SD = 1.12$, 95% CI [0.85, 1.15]). These findings indicate that AI assistance resulted in increased response latency relative to **Human-assisted** interactions, likely due to system processing time during the task.

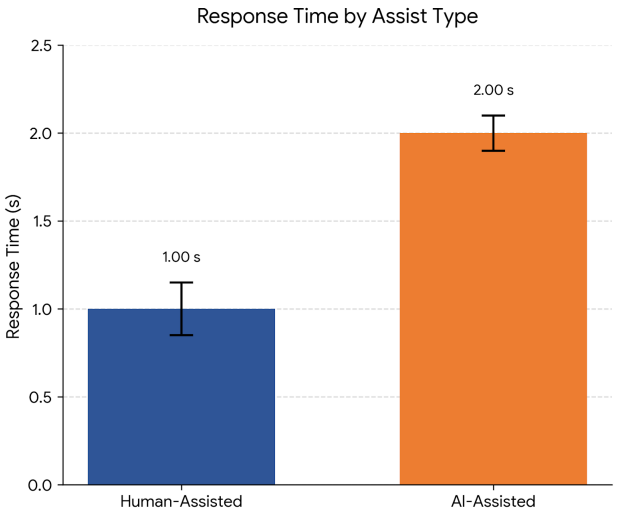

Figure 3: Response time by assist type. Error bars indicate the 95% confidence interval of the mean.

5.6.2 Task Accuracy

Task accuracy was assessed as an indicator of system reliability and user performance across both the **Human-assisted** and **AI-assisted** conditions. Accuracy was defined as the proportion of successful navigation attempts in which the user's intended target was correctly reached. This metric captures both human execution errors and system-level transcription providing a holistic measure of interaction precision.

In the **Human-assisted** condition, task accuracy was computed by cross-referencing transcribed user utterances with the corresponding UI selections made by the observer, as recorded in the log files. The results revealed a high level of precision, with an overall accuracy of 99.5%. Occasional discrepancies occurred in a small number of trials where the observer selected an incorrect button.

In the **AI-assisted** condition, each trial was analyzed to identify cases in which the AI system failed to accurately transcribe the user's utterance. The overall accuracy for this condition was 97.3%. The few observed failures primarily resulted due to incomplete speech capture and partial transcription errors.

5.6.3 NASA-TLX

The overall task load, as measured by the NASA-TLX questionnaire, didn't differ significantly between the **Human-assisted** and **AI-assisted**. A one-way Aligned Rank transform (ART) ANOVA revealed no main effect of assist type on overall workload ($F(1,30) = 0.25$, $p = 0.619$). As illustrated in Figure 4, the mean weighted TLX score was slightly higher for the **Human-assisted** condition ($M = 8.78$, $SD = 7.80$ 95% CI [4.63, 12.94]) compared to the **AI-assisted** condition ($M = 7.87$, $SD = 7.65$, 95% CI [3.79, 11.94]), but this difference was not statistically significant.

Further analysis of the six individual TLX dimensions showed no significant differences between assist types: Mental Demand ($F(1,30) = 1.34$, $p = .25$), Physical Demand ($F(1,30) = 0.157$,

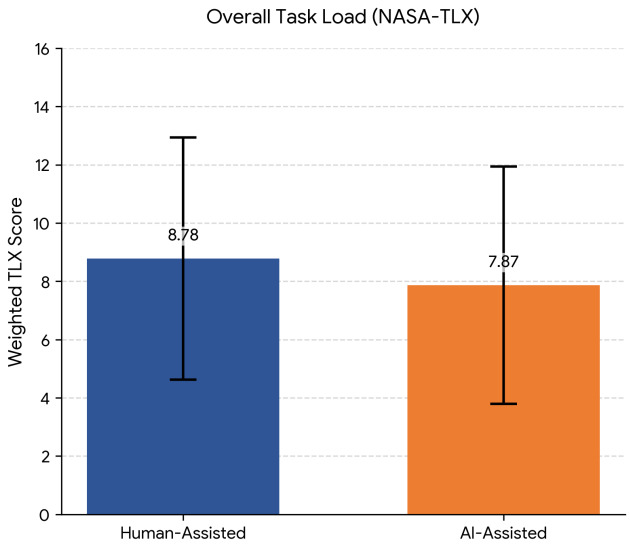

Figure 4: Overall task load (weighted NASA-TLX score) by assist type. Error bars indicate the 95% confidence interval of the mean.

$p = .69$), Temporal Demand ($F(1,30) = 0.36$, $p = .55$), Performance ($F(1,30) = 0.07$, $p = .79$), Effort ($F(1,30) = 0.23$, $p = .63$), and Frustration ($F(1,30) = 0.03$, $p = .86$)

Overall, participants reported comparable levels of perceived workload across both assist conditions. Descriptively, **Human-assisted** trials to induce slightly higher Mental and Temporal Demand, whereas **AI-assisted** trials elicited marginally greater Effort and Frustration ratings. however, these trends did not reach statistical significance.

5.6.4 Subjective Ratings

To evaluate participants' subjective experience with each viewpoint control mode, ratings of preference, ease of interaction, perceived speed, and perceived accuracy were analyzed using Aligned Rank Transform (ART) ANOVAs.

For Preference, there was no significant effect of assist types, $F(1,30) = 0.71$, $p = .406$. As illustrated in Figure 5, participants rated both Human- and **AI-assisted** technique favorably, with comparable mean scores **Human-assisted**: ($M = 5.63$, $SD = 1.09$, 95% CI [5.05, 6.20]); **AI-assisted**: ($M = 5.94$, $SD = 1.06$, 95% CI [5.37, 6.50]).

For Ease of Interaction, no significant difference was observed, ($F(1,30) = 0.14, p = .709$). Participants found both systems easy to use, with high and comparable ease ratings **Human-assisted**: ($M = 6.19$, $SD = 0.75$, 95% CI [5.79, 6.59]); **AI-assisted**: ($M = 5.94$, $SD = 1.12$, 95% CI [5.34, 6.54]).

For Perceived Speed, the effect approached but did not reach statistical significance, $F(1,30) = 3.14$, $p = .087$. Participants rated **Human-assisted** viewpoint control ($M = 5.88$, $SD = 1.20$, 95% CI [5.23, 6.52]) as somewhat faster than **AI-assisted** viewpoint control ($M = 5.13$, $SD = 1.31$, 95% CI [4.43, 5.82]).

For Perceived Accuracy, the effect again approached without statistical significance, $F(1,30) = 2.94$, $p = .097$. Participants rated both viewpoint control modes as highly accurate, though **Human-assisted** viewpoint control received slightly higher ratings ($M = 6.63$, $SD = 0.81$, 95% CI [6.20, 7.05]) compared to **AI-assisted** viewpoint control ($M = 6.31$, $SD = 0.70$, 95% CI [5.94, 6.69]).

Overall, participants' ratings indicated similarly positive impressions of both navigation systems, with only slight, non-significant differences in perceived speed and accuracy favoring the **Human-assisted**. Also, both assist types were positively received. **Human-assisted** technique tended to be seen as slightly faster, while preference, ease of interaction, and perceived accuracy and speed remained comparable across conditions.

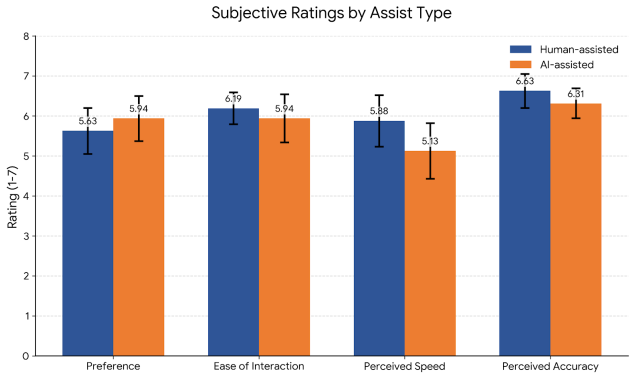

Figure 5: Subjective ratings for preference, ease of interaction, perceived speed, and perceived accuracy by assist type. Error bars indicate the 95% confidence interval of the mean.

5.7 Associations with Personality Traits

To explore the influence of individual differences, we conducted Pearson correlation analyses between core personality trait ratings and key survey outcomes (preference, trust, willingness to reuse). While none of the associations reached conventional statistical significance ($p < 0.05$), several marginal associations ($p < 0.10$) emerged that are suggestive and warrant further investigation:

- **Dependable, self-disciplined** was positively associated with both liking **AI-assisted** viewpoint control technique ($r = 0.49, p = 0.08$) and trust in the AI assistant ($r = 0.51, p = 0.06$). More dependable individuals tended to rate the AI system higher and trust it more.
- **Extraverted, enthusiastic** was negatively associated with liking **Human-assisted** viewpoint control ($r = -0.58, p = 0.05$), suggesting that less outgoing users may be more drawn to human assistance.
- **Sympathetic, warm** was positively associated with liking **Human-assisted** viewpoint control ($r = 0.52, p = 0.07$), indicating a greater appeal for social interaction among more sympathetic participants.
- **Critical, Quarrelsome** was negatively associated with willingness to continue using the AI assistant ($r = -0.54, p = 0.07$), indicating more critical users may be less likely to persist with AI systems.

These findings suggest that user psychological dispositions play a measurable, though not statistically definitive, role in shaping interface preference, trust, and technology acceptance in this context. Future research with larger samples should further investigate these individual differences.

5.8 Discussion

While **Human-assisted** viewpoint control was significantly faster in terms of response time, this speed advantage did not translate into significantly higher subjective ratings of preference, ease of use, accuracy, or satisfaction. Both viewpoint control modes were perceived as easy to use, and neither method was associated with excessive workload or frustration. The lack of significant

differences in NASA-TLX workload and subjective usability ratings underscores the value of both interfaces, particularly given the complex, attention-heavy nature of 3D multiscale navigation tasks. This aligns with prior research showing that well-designed AI and human-assisted systems can achieve comparable levels of usability and workload, provided they are transparent and compatible with user needs [51, 14].

Interestingly, despite the objective latency of the **AI-assisted** system, participants reported a slightly higher overall preference for **AI-assisted** viewpoint control. This suggests that factors other than raw performance metrics—such as agency, control, and the ability to operate independently—may play a substantial role in shaping user attitudes. Prior work highlights that perceived agency and autonomy in AI systems can significantly influence user trust and preference, sometimes outweighing efficiency or speed, as users value the sense of control and independence provided by AI assistance [14, 60]. The parity in perceived accuracy and preference further indicates that users were able to adapt to the AI system's slower response, possibly valuing its autonomy over sheer efficiency.

Notably, analysis of personality trait associations with interface responses revealed several marginal effects ($p < 0.10$). Participants who scored higher in dependability tended to express more enjoyment and trust toward the AI assistant. In contrast, those with higher levels of criticalness were less willing to continue using the AI, and less extraverted or more sympathetic participants tended to prefer **Human-assisted** viewpoint control. These observations are supported by a growing body of research demonstrating that personality traits—such as conscientiousness, agreeableness, extraversion, and locus of control—can meaningfully influence trust, enjoyment, and willingness to adopt or continue using AI systems [50, 29, 69, 12, 53, 52]. Although these associations did not achieve conventional significance ($p < 0.05$), they signal that dispositional traits may meaningfully—but subtly—influence technology acceptance patterns alongside objective performance measures. Larger studies may further clarify these relationships and their practical implications for XR interface personalization and adoption [50, 29, 69, 52].

Taken together, these findings highlight the complexity of user experience in spatial interface design. Optimal navigation systems should be evaluated not only on efficiency, but also on the psychological comfort, adaptability, and individual user characteristics that together enable technology acceptance in hands-free, cognitively demanding contexts [50, 60, 29, 69, 52].

6 Qualitative Phase

To further understand and explain the quantitative results, we conducted a series of semi-structured interviews with a subset of participants from the first phase. These interviews were designed to explore participants' navigation experiences in depth, probing the reasons behind their preferences, perceptions of autonomy and social dynamics, and experiences negotiating challenges such as delays or misunderstood commands. A thematic analysis revealed core motivators underlying participant choices, as well as the unique values each method provided.

6.1 Study Design

The qualitative phase followed an explanatory sequential mixed methods model, aimed at elaborating on user preferences and behavioral motivations observed in the survey study. Semi-structured interviews focused on interaction with the **AI-assisted** viewpoint control and **Human-assisted** baseline, as well as contextual and emotional factors shaping user experience.

6.2 Participants

Among all participants from section 5, 14 participants were recruited for follow-up interviews, purposefully sampled to represent both AI- and human-navigator preference groups. Participant diversity in AR/VR experience, speech technology use, and anatomy knowledge was maintained.

Theme	Description	Example Quote	Intercoder Agreement	Inter-rater Reliability (Cohen's kappa)
Autonomy/Agency	Valuing freedom to act independently and set one's pace	With AI, I was in control and could work at my own speed.	0.93	0.85
Social Comfort	Reduced social anxiety or pressure with AI vs. human	I didn't feel judged using the AI system.	0.9	0.81
Lag Perception	Experience and tolerance of system delay	I hardly noticed the delay; it was just part of the process.	0.88	0.8
Accuracy/Communication	Ease of resolving ambiguity and intent	The human could just clarify if they misunderstood me.	0.91	0.84
Feedback/Learning	Quality of feedback for growth or reassurance	The human gave me tips when I struggled with navigation.	0.88	0.82
Task Complexity	Preference shift depending on difficulty/context	I used the AI for easy parts, but needed the human for complex steps.	0.87	0.81

Table 1: Major themes from thematic analysis of user interviews, with representative quotes and intercoder reliability metrics

6.3 Interview Protocol

Each interview (20-40 minutes) used an evolving set of open-ended questions to probe experiences with autonomy ("doing it myself"), controllability, social dynamics (freedom from being judged or rushed), navigation errors, trust, and comfort with both methods. Participants described what they valued most and least about each approach, and how specific situations (such as accent/pronunciation or complex navigation) shaped their attitudes.

6.4 Procedure

All interviews were conducted within a week after the quantitative phase, audio-recorded, and fully transcribed. Questions and probes (Appendix A) were refined iteratively as new topics and themes emerged. Transcripts were anonymized for analysis.

6.5 Analysis

A thematic analysis approach [11] was used, combining inductive open coding with a systematic review of each transcript. Qualitative coding was conducted using ATLAS.ti software version 25.0 [5], which incorporates AI-based autocoding and thematic suggestion features. All AI-generated codes were reviewed and refined by the research team to ensure interpretive accuracy and analytic depth. Coding was performed iteratively, with two researchers—one author and an institute colleague—independently coding a subset of transcripts and discussing discrepancies to refine the codebook and enhance consistency. Coding consistency was assessed using both intercoder agreement (percent agreement) and inter-rater reliability (see Table 1), providing robust quantification of analytic rigor. An audit trail of codebook development and theme evolution was maintained throughout. Themes were refined via negative case analysis, ensuring that contradictory or outlier perspectives were examined and reported. Reflexivity was practiced through regular analytic meetings where coders reflected on assumptions and discussed potential biases. Thick descriptions and illustrative quotes were used to convey context and bolster credibility. Coding and theme comparison were performed across AI-preference and human-preference groups, with both shared and divergent experiences identified. Coding proceeded until thematic/code saturation was reached (no new themes emerged).

6.6 Results

Thematic analysis across all transcripts, including recent participants, revealed nuanced perspectives on navigation preferences and experiences.

- **Autonomy and Agency:** Participants frequently highlighted the value of autonomy provided by **AI-assisted** viewpoint control technique. Many appreciated being "in control" and

"able to go at [their] own pace," free from coordinating or relying on another person. This self-determination fostered a sense of mastery and comfort, particularly for those who valued independent exploration or learning.

- **Social Comfort and Reduced Pressure:** Consistently, participants described less social anxiety, self-consciousness, or fear of judgment when interacting with AI. Several recounted feeling "less rushed," "free to make mistakes," and generally more at ease focusing on the navigation task itself. These sentiments were pronounced among users who preferred solitude or felt nervous giving commands to another person.

- **Lag Perception and Response Time:** Interviewees' perceptions of system lag sometimes diverged from quantitative measurements. Many described little difference in speed, or even occasional advantages for AI—especially for simple, direct navigation commands. For a subset, delays felt "barely noticeable" or "not significant," while others admitted that only pronounced system errors really disrupted their workflow. **Human-assisted** viewpoint control was sometimes seen as slower for finding uncommon structures or when misunderstandings arose.

- **Accuracy and Communication: Human-assisted** viewpoint control was particularly valued for flexibility, rapid clarification of ambiguous or accented speech, and the ability to interpret intent seamlessly. Several participants described comfort in being "understood even without perfect words," and cited human feedback as vital during complex or uncertain navigation tasks.

- **Feedback and Learning:** Human assistants provided timely, encouraging feedback and could offer suggestions or error correction that AI systems currently lack. However, participants also appreciated the "less stressful, lower-stakes" environment created by AI for practice or solo exploration.

- **Task Complexity and Adaptability:** Across both groups, preference often shifted with task demands. **AI-assisted** viewpoint control was preferred for straightforward, well-defined transitions across scales, while human support became more valued as tasks increased in ambiguity, complexity, or required emotional reassurance.

Representative quotes include:

"I liked being able to just say what I wanted, not worry about being right or wrong."

"With the AI, it was my own pace—I didn't have to wonder if I was being too slow or asking too much."

"If I made a mistake, it wasn't awkward. I just tried again. With a human, I felt like I had to explain myself."

"Sometimes it took a moment for the AI to catch up, but usually it worked really well, and I didn't notice any real lag."

"The human just sort of knew what I meant, even if my words weren't exact."

7 Overall Discussion

This explanatory sequential mixed-methods study reveals important insights into user acceptance of speech-based viewpoint control in multiscale AR: participants valued autonomy and social comfort alongside—and often above—response speed and technical accuracy. Quantitative results show that while **Human-assisted** viewpoint control is rated as faster and slightly more accurate, participants reported both AI- and **Human-assisted** methods as easy to use and generally low in workload, consistent with prior findings that usability and satisfaction are often high across modalities, with preferences shaped by context and individual differences rather than performance alone [28, 62, 40, 47]. Despite these performance differences, **AI-assisted** viewpoint control earned a slightly higher overall preference and very strong trust and willingness-to-reuse ratings, echoing research that users may value autonomy and agency over sheer efficiency [69, 53, 39, 23].

A key element of our study is the integration of personality trait data with attitudes and preferences toward speech-based interaction. Understanding such individual differences is particularly relevant for XR training and rehearsal environments, where user with diverse backgrounds, experience levels and learning style interact with the same system. To analyze the influence of personality on user responses, we computed Pearson correlation coefficients between Likert ratings of core personality traits and outcomes such as preference, trust, and willingness to continue with the AI system. Moderate, statistically suggestive associations ($|r| >= 0.4, p < 0.1$) were interpreted as meaningful contributors to user attitudes. This is supported by research showing that personality traits such as conscientiousness, extraversion, and agreeableness significantly influence trust, enjoyment, and willingness to use AI or human-assisted systems [69, 53, 50, 29, 67, 12].

Importantly, this exploratory analysis revealed that more dependable participants were more likely to report higher enjoyment and trust in the AI assistant, while those reporting higher criticalness were less willing to continue using AI. Extraverted and sympathetic participants were more inclined to prefer and enjoy **Human-assisted** viewpoint control. Although these associations were moderate in strength, they suggest that individual psychological dispositions play a measurable role not only in interface preference but also in trust and technology acceptance, as highlighted in recent reviews and empirical studies [69, 53, 50, 29, 67, 12].

Crucially, user preferences and attitudes were shaped by more than system performance alone. While response time and recognition accuracy remain important dimensions of interactive systems, qualitative and quantitative findings indicate that participants valued autonomy, agency, and overall interaction comfort when using the **AI-assisted** viewpoint control. Users appreciated the ability to control viewpoint directly through natural speech and to progress at their own pace, and observed response delays or recognition errors did not undermine their overall trust in or willingness to use the system. Human assistants were most valued in situations requiring clarification adaptation, or additional support, consistent with prior findings that access to assistance influences user preference in XR and AI systems [39, 23].

While participants tolerated 2-second latency in the prototype, clinical deployment depends on task context: active intervention vs. cognitive support demand different thresholds. For high-stakes, intraoperative manipulation, the latency requirement is strict. Akasaka et al. [3] demonstrates that in telerobotic surgery, suboptimal network environments and resulting delays significantly degrade surgeon performance and increase fatigue by disrupting hand-eye coordination. In such contexts, our current 2-second delay would be unacceptable for controlling surgical tools or critical view adjustments during active dissection. For protocol based support, broader latency tolerance applies. Cognitive assistants in emergency medicine operates within 4-second windows [64], so our system is acceptable for advisory and pre-operative tasks. However, robust clinical systems must minimize cloud dependency [64]. Future work should prioritize edge deployment to achieve sub-second responsiveness bridging cognitive and active intervention tasks. We

mitigated LLM hallucinations through multi-stage validation: (1) strict JSON schema constraining outputs to predefined anatomical fields, (2) hierarchical validation rejecting invalid targets, and (3) confidence-based filtering (threshold = 0.70) requesting clarification when confidence drops. This conservative approach prioritizes safety and user control over full automation.

Technical improvements alone are insufficient for user acceptance of speech-based AR interfaces. Performance matters, but psychological factors like agency, comfort, control which ultimately drive adoption and user preferences reflect personal values rather than speed alone. Future XR systems should balance autonomous speech-driven control for independent users with adaptable human-in-the-loop options for those seeking collaborative support.[28, 69, 39, 23, 47].

8 LIMITATIONS AND FUTURE WORK

Our current implementation has several limitations that suggest directions for future improvement. Speech transcription quality and capture timing present one set of limitations. While Azure Speech-to-Text provides robust performance for non-native speakers, recognition accuracy could be further improved through domain-specific fine-tuning or more advanced ASR models. In addition, our system processes speech only after the user releases the trigger, introducing a brief delay between speaking and system response. Implementing real time streaming transcription could reduce this delay, but would introduce additional challenges, such as managing continuous buffering and handling partial or unstable transcription outputs in real time. Cloud-based APIs for speech and LLM processing incur network overhead, contributing to slower AI-assisted response times. Edge deployment could reduce latency but requires trade-offs in model size, memory footprint, and platform constraints The current push-to-talk design prevents true hands-free operation and limits applicability in hand-busy contexts like surgery. We adopted this approach to mitigate VAD limitations like false activations and incomplete capture during natural pauses.

our findings suggest that user acceptance may be driven more by psychological factors—particularly autonomy, agency, and social comfort—than by absolute response speed. Although quantitative difference in preference did not reach statistical significance, these exploratory findings suggest that design priorities for speech-based AR systems should extend beyond technical performance metrics to encompass experiential and psychological dimensions of user acceptance.

There are several key directions for future work. First, deploying the system with domain experts in real-world medical training and clinical rehearsal environments will help validate its effectiveness across different tasks and identify context-specific requirements. This feedback will guide refinement of the natural language understanding capabilities and viewpoint transition strategies to better serve authentic use cases.

Second, integrating robust voice activity detection (VAD) specifically, frameworks like openWakeWord with built-in Silero VAD represents a critical next step toward fully autonomous speech interaction. CognitiveEMS demonstrates that edge-deployable VAD can support real-time, hands-free operation on resource constrained devices without cloud dependency. By adopting this edge centric approach, our system could overcome current network latencies and VAD limitations, enabling reliable operation in clinical environments with limited or unavailable connectivity.

Finally, we plan to integrate this speech-based viewpoint control component into a complete hands-free Multimodal navigation interface for medical XR applications. By combining natural speech with complementary modalities such as gaze tracking and head gestures, we aim to create a comprehensive interaction system that adapts to varying environmental conditions, task demands, and user preferences. This integrated approach will support the full range of navigation tasks—including fine-grained spatial positioning, orientation control, and scale transitions—required for effective interaction with complex 3D medical data in AR environments.

A APPENDIX: INTERVIEW QUESTIONS AND PROBES

This section lists the semi-structured interview questions used for qualitative data collection. During each interview, follow-up *probes* were also used to clarify, elicit richer responses, or explore new themes as they emerged. Common probes included: “Can you tell me more about that?”, “Could you give an example?”, “What do you mean by that?”, “How did that make you feel?”, and “Can you explain further?”

1. **Experience Overview:** Describe your overall experience with both AI-assisted and Human-assisted navigation. What felt most intuitive, effective, or challenging? *Probes: Can you give a specific example? What made it feel intuitive or challenging?*

2. **Explaining Your Preference:** What specific reasons or moments led to your preference for [AI/Human]-assisted navigation? *Probes: Were there key events that changed your mind? What influenced your decision most?*

3. **Speed and Accuracy:** Which condition was faster? More accurate? *Probes: What led you to that conclusion? Did it feel the same throughout, or did it vary?*

4. **Navigating Complexity and Mistakes:** How did each technique support or hinder movement through scales and anatomical structures? Recall a situation when navigation was misunderstood or slow—which method felt more effective? *Probes: Can you walk me through that moment? How did you recover from mistakes?*

5. **Tolerances, Trade-offs, and Benefits:** (AI-preferring) What features outweighed AI drawbacks? (Human-preferring) What did human assistance provide that AI could not? What was most problematic with AI? *Probes: Did anything surprise you? Was there something you wish the system had done?*

6. **User Needs, Social, and Autonomous Aspects:** Did task type, setting (solo learning, group use, teaching, presenting), or social aspects shape your preference? How important were social reassurance, adaptive feedback, privacy, autonomy, or support? *Probes: Did the context affect your comfort or choice? Would you choose differently in another setting?*

7. **Background and Personal Factors:** How did your AR/VR experience, speech technology use, or anatomy knowledge impact navigation comfort? Did personality traits (patience, openness to tech, sociability) influence which system worked best? *Probes: Can you relate this to another technology you have used, or has your opinion changed over time?*

8. **Improvements and What-if Scenarios:** What changes would make your less-preferred method more appealing? If you could fix one limitation (lag, recognition, feedback style), would you reconsider? *Probes: What would be the single biggest improvement? How much would that change your use?*

9. **Open Reflection:** Anything else about your needs, tactics, or experience that would explain your preferences or challenges? *Probes: Is there anything we've missed? Any advice for designers of such systems?*

REFERENCES

[1] Anthropic claude api. https://www.anthropic.com/product/claude. Accessed: 2024-06-10. 3

[2] M. Abosheisha, R. Prabhu, M. Abdelglil, A. Swealem, M. Ali, Z. Al-Hamid, R. Tamanna, and M. Elhadidi. The role of augmented reality in surgical training: A narrative review. *Cureus*, 17, 2025. doi: 10.7759/cureus.95214 1

[3] H. Akasaka, K. Hakamada, H. Morohashi, T. Kanno, K. Kawashima, Y. Ebihara, E. Oki, S. Hirano, and M. Mori. Impact of the suboptimal communication network environment on telerobotic surgery performance and surgeon fatigue. *PLoS ONE*, 17(6):e0270039, 2022. 8

[4] C. Andrews, M. K. Southworth, J. Silva, and J. R. Silva. Extended reality in medical practice. *Current Treatment Options in Cardiovascular Medicine*, 21:1–12, 2019. doi: 10.1007/s11936-019-0722-7 1

[5] ATLAS.ti Scientific Software Development GmbH. Atlas.ti 25.0 [computer software]. https://atlasti.com/, 2024. 7

[6] Y. Baashar, G. Alkawsi, W. N. W. Ahmad, H. Alhussian, A. Alwadain, L. F. Capretz, A. Babiker, and A. Alghail. Effectiveness of using augmented reality for training in the medical professions: Meta-analysis. *JMIR Serious Games*, 10, 2021. doi: 10.2196/32715 1

[7] E. Barsom, M. Graafland, and M. Schijven. Systematic review on the effectiveness of augmented reality applications in medical training. *Surgical Endoscopy*, 30:4174 – 4183, 2016. doi: 10.1007/s00464-016-4800-6 1

[8] S. Barteit, L. Lanfermann, T. Bärnighausen, F. Neuhann, and C. Beiersmann. Augmented, mixed, and virtual reality-based head-mounted devices for medical education: Systematic review. *JMIR Serious Games*, 9, 2021. doi: 10.2196/29080 1

[9] P. Bartie and W. Mackaness. Development of a speech-based augmented reality system to support exploration of cityscape. *Transactions in GIS*, 10, 2006. doi: 10.1111/j.1467-9671.2006.00244.x 2

[10] M. M. J. Begum. Ai-powered visual navigation system. *International Journal for Research in Applied Science and Engineering Technology*, 2025. doi: 10.22214/ijraset.2025.69504 2

[11] V. Braun and V. Clarke. Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2):77–101, 2006. 7

[12] M. Böckle, K. Yeboah-Antwi, and I. Kouris. Can you trust the black box? the effect of personality traits on trust in ai-enabled user interfaces. pp. 3–20, 2021. doi: 10.1007/978-3-030-77772-2_1 7, 8

[13] A. Cangelosi, G. Riberi, P. Titolo, M. Salvi, F. Molinari, L. Ulrich, E. Vezzetti, M. Agus, and C. Calì. Augmented reality simulation framework for minimally invasive orthopedic surgery. *Computers in biology and medicine*, 189:109943, 2025. doi: 10.1016/j.compbiomed.2025.109943 1

[14] S. Cao and C.-M. Huang. Understanding user reliance on ai in assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 6:1 – 23, 2022. doi: 10.1145/3555572 7

[15] F. Chahartangi, N. Zarifsanaiey, M. Mehrabi, and B. Z. Ghoochani. Integrating augmented reality virtual patients into healthcare training: A scoping review of learning design and technical requirements. *PLOS One*, 20, 2025. doi: 10.1371/journal.pone.0324740 1

[16] T. E. Chemaly, C. A. Neves, F. Fu, B. A. Hargreaves, and N. Blevins. From microscope to head-mounted display: integrating hand tracking into microsurgical augmented reality. *International Journal of Computer Assisted Radiology and Surgery*, 19:2023 – 2029, 2024. doi: 10.1007/s11548-024-03224-w 1

[17] F. Cofano, G. Di Perna, M. Bozzaro, A. Longo, N. Marengo, F. Zenga, N. Zullo, M. Cavalieri, L. Damiani, D. J. Boges, et al. Augmented reality in medical practice: from spine surgery to remote assistance. *Frontiers in Surgery*, 8:657901, 2021. 4

[18] J. W. Creswell and V. L. P. Clark. *Designing and conducting mixed methods research*. Sage publications, 2017. 2

[19] A. T. Duchowski. Gaze-based interaction: A 30 year retrospective. *Computers & Graphics*, 73:59–69, 2018. 2

[20] N. Gasteiger, S. N. van der Veer, P. Wilson, and D. Dowding. How, for whom, and in which contexts or conditions augmented and virtual reality training works in upskilling health care workers: Realist synthesis. *JMIR Serious Games*, 10, 2021. doi: 10.2196/31644 1

[21] A. Grinshpoon, S. Sadri, G. J. Loeb, C. Elvezio, and S. K. Feiner. Hands-free interaction for augmented reality in vascular interventions. *2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 751–752, 2018. doi: 10.1109/vr.2018.8446259 1

[22] J. N. Hombeck, H. Voigt, and K. Lawonn. Voice user interfaces for effortless navigation in medical virtual reality environments. *Comput. Graph.*, 124:104069, 2024. doi: 10.1016/j.cag.2024.104069 2

[23] W. Hsu and M.-H. Lee. Semantic technology and anthropomorphism: Exploring the impacts of voice assistant personality on user trust, perceived risk, and attitude. *J. Glob. Inf. Manag.*, 31:1–21, 2023. doi: 10.4018/jgim.318661 8, 9

[24] A. Iqbal, A. Aamir, A. Hammad, H. Hafsa, A. Basit, M. O. Oduoye, M. W. Anis, S. Ahmed, M. I. Younus, and S. Jabeen. Immersive technologies in healthcare: An in-depth exploration of virtual reality and augmented reality in enhancing patient care, medical education, and training paradigms. *Journal of Primary Care Community Health*, 15, 2024. doi: 10.1177/21501319241293311 1

[25] A. Ismail, M. Billinghurst, M. S. Sunar, and C. S. Yusof. Designing an augmented reality multimodal interface for 6dof manipulation techniques - multimodal fusion using gesture and speech input for ar. pp. 309–322, 2018. doi: 10.1007/978-3-030-01054-6_22 2

[26] P. Jia, H. H. Hu, T. Lu, and K. Yuan. Head gesture recognition for hands-free control of an intelligent wheelchair. *Industrial Robot*, 34(1):60–68, 2007. 2

[27] T. Kojić, M. Vergari, M. Warsinke, D. Ali, S. Möller, and J.-N. Voigt-Antons. Multimodal user experience in extended reality: Exploring hand tracking, voice, and passthrough interactions. *Proceedings of the 17th International Workshop on IMmersive Mixed and Virtual Environment Systems*, 2025. doi: 10.1145/3712677.3720459 2

[28] M. Korkiakoski, P. Alavesa, and P. Kostakos. Preference in voice commands and gesture controls with hands-free augmented reality with novel users. *IEEE Pervasive Computing*, 23:18–26, 2024. doi: 10.1109/MPRV.2024.3364541 2, 8, 9

[29] A. Küper and N. C. Krämer. Psychological traits and appropriate reliance: Factors shaping trust in ai. *International Journal of Human–Computer Interaction*, 41:4115 – 4131, 2024. doi: 10.1080/10447318.2024.2348216 7, 8

[30] G. Lampropoulos, P. Fernández-Arias, A. del Bosque, and D. Vergara. Augmented reality in health education: Transforming nursing, healthcare, and medical education and training. *Nursing Reports*, 15, 2025. doi: 10.3390/nursrep15080289 1

[31] J.-I. Lee, P. Asente, and W. Stuerzlinger. A comparison of zoom-in transition methods for multiscale vr. In *ACM SIGGRAPH 2022 Posters*, pp. 1–2. 2022. 3

[32] J.-I. Lee, P. Asente, and W. Stuerzlinger. Designing viewpoint transition techniques in multiscale virtual environments. In *2023 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*, pp. 680–690. IEEE, 2023. 3

[33] L.-H. Lee, T. Braud, S. Hosio, and P. Hui. Towards augmented reality driven human-city interaction: Current research on mobile headsets and future challenges. *ACM Computing Surveys (CSUR)*, 54:1 – 38, 2020. doi: 10.1145/3467963 2

[34] M. Lee. Multimodal speech-gesture interaction with 3d objects in augmented reality environments. 2010. 2

[35] M. Lee, M. Billinghurst, W. Baek, R. Green, and W. Woo. A usability study of multimodal input in an augmented reality environment. *Virtual Reality*, 17:293–305, 2013. doi: 10.1007/s10055-013-0230-0 2

[36] D. Loeb, J. Shoemaker, A. Parsons, D. Schumacher, and M. W. Zackoff. How augmenting reality changes the reality of simulation: Ethnographic analysis. *JMIR Medical Education*, 9, 2023. doi: 10.2196/45538 1

[37] M. N. Lystbæk, P. Rosenberg, K. Pfeuffer, J. Grønbæk, and H.-W. Gellersen. Gaze-hand alignment. *Proceedings of the ACM on Human-Computer Interaction*, 6:1 – 18, 2022. doi: 10.1145/3530886 2

[38] I. A. C. Mangalote, O. Aboumarzouk, A. A. Al-Ansari, and S. Dakua. A comprehensive study to learn the impact of augmented reality and haptic interaction in ultrasound-guided percutaneous liver biopsy training and education. *Artificial Intelligence Review*, 57, 2024. doi: 10.1007/s10462-024-10791-6 1

[39] M. R. Miller, H. Jun, F. Herrera, J. Y. Villa, G. Welch, and J. Bailen-

son. Social interaction in augmented reality. *PLoS ONE*, 14, 2019. doi: 10.1371/journal.pone.0216290 8, 9

[40] P. Monteiro, H. Coelho, G. Gonçalves, M. Melo, and M. Bessa. Exploring the user experience of hands-free vr interaction methods during a fitts' task. *Computers & Graphics*, 117:1–12, 2023. 1, 2, 8

[41] P. Monteiro, G. Gonçalves, H. Coelho, M. Melo, and M. Bessa. Hands-free interaction in immersive virtual reality: A systematic review. *IEEE Transactions on Visualization and Computer Graphics*, 27:2702–2713, 2021. doi: 10.1109/TVCG.2021.3067687 2

[42] S. S. M. Nizam, R. Z. Abidin, N. C. Hashim, M. C. Lam, H. Arshad, and N. A. A. Majid. A review of multimodal interaction technique in augmented reality environment. *International Journal on Advanced Science, Engineering and Information Technology*, 2018. doi: 10.18517/IJASEIT.8.4-2.6824 2

[43] E. Olshannikova, A. Ometov, Y. Koucheryavy, and T. Olsson. Visualizing big data with augmented and virtual reality: challenges and research agenda. *Journal of Big Data*, 2:1–27, 2015. 2

[44] openWakeWord Contributors. openwakeword: Open-source wakeword detection framework, 2024. MIT License. Accessed: 2025-01-15. 3

[45] K.-B. Park, S. Choi, J. Y. Lee, Y. Ghasemi, M. Mohammed, and H. Jeong. Hands-free human–robot interaction using multimodal gestures and deep learning in wearable mixed reality. *IEEE Access*, 9:55448–55464, 2021. doi: 10.1109/ACCESS.2021.3071364 2

[46] F. Putze, A. Vourvopoulos, A. Lécuyer, D. Krusienski, S. Bermudez i Badia, T. Mullen, and C. Herff. Brain-computer interfaces and augmented/virtual reality. *Frontiers in Human Neuroscience*, 14:144, 2020. 2

[47] R. C. Quesada and Y. Demiris. Multi-dimensional evaluation of an augmented reality head-mounted display user interface for controlling legged manipulators. *ACM Transactions on Human-Robot Interaction*, 13:1 – 37, 2024. doi: 10.1145/3660649 8, 9

[48] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever. Whisper tiny model, 2025. 3

[49] N. Reimers and I. Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1410 3

[50] R. Riedl. Is trust in artificial intelligence systems related to user personality? review of empirical evidence and future research directions. *Electronic Markets*, 32:2021 – 2051, 2022. doi: 10.1007/s12525-022-00594-4 7, 8

[51] H. Saeidi and Y. Wang. Incorporating trust and self-confidence analysis in the guidance and control of (semi)autonomous mobile robotic systems. *IEEE Robotics and Automation Letters*, 4:239–246, 2019. doi: 10.1109/LRA.2018.2886406 7

[52] A. Schepman and P. Rodway. The general attitudes towards artificial intelligence scale (gaais): Confirmatory validation and associations with personality, corporate distrust, and general trust. *International Journal of Human–Computer Interaction*, 39:2724 – 2741, 2022. doi: 10.1080/10447318.2022.2085400 7

[53] N. N. Sharan and D. Romano. The effects of personality and locus of control on trust in humans versus artificial intelligence. *Heliyon*, 6, 2020. doi: 10.1016/j.heliyon.2020.e04572 7, 8

[54] M. F. Siddiqui, S. Jabeen, A. Alwazzan, S. Vacca, L. Dalal, B. Al-Haddad, A. Jaber, F. F. Ballout, H. K. A. Zeid, J. Haydamous, R. E. H. Chehade, and R. Kalmatov. Integration of augmented reality, virtual reality, and extended reality in healthcare and medical education: A glimpse into the emerging horizon in lmics—a systematic review. *Journal of Medical Education and Curricular Development*, 12, 2025. doi: 10.1177/23821205251342315 1

[55] A. W. Stedmon, H. Patel, S. C. Sharples, and J. R. Wilson. Developing speech input for virtual reality applications: A reality based interaction approach. *International journal of human-computer studies*, 69(1-2):3–8, 2011. 1, 2

[56] J. P. Takona. Research design: qualitative, quantitative, and mixed methods approaches. *Quality & Quantity*, 58(1):1011–1013, 2024. 2

[57] J.-L. Tseng. Intelligent augmented reality system based on speech recognition. *International Journal of Circuits, Systems and Signal Processing*, 2021. doi: 10.46300/9106.2021.15.20 2

[58] S. D. Ubur and D. Gracanin. Narrative review of emotional expression support in xr: Psychophysiology of speech-to-text interfaces. 2024. 2

[59] A. L. Valvo, D. Croce, D. Garlisi, F. Giuliano, L. Giarré, and I. Tinnirello. A navigation and augmented reality system for visually impaired people †. *Sensors (Basel, Switzerland)*, 21, 2021. doi: 10.3390/s21093061 2

[60] B. S. Vanneste and P. Puranam. Artificial intelligence, trust, and perceptions of agency. *Academy of Management Review*, 2024. doi: 10.5465/amr.2022.0041 7

[61] Z. Wang, J. Sun, M. Hu, M. Rao, W. Song, and F. Lu. Gazering: Enhancing hand-eye coordination with pressure ring in augmented reality. *2024 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 534–543, 2024. doi: 10.1109/ISMAR62088.2024.00068 2

[62] Z. Wang, H. Wang, H. Yu, and F. Lu. Interaction with gaze, gesture, and speech in a flexibly configurable augmented reality system. *IEEE Transactions on Human-Machine Systems*, 51:524–534, 2021. doi: 10.1109/thms.2021.3097973 2, 8

[63] Z. Wang, H. Yu, H. Wang, Z. Wang, and F. Lu. Comparing single-modal and multimodal interaction in an augmented reality system. *2020 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, pp. 165–166, 2020. doi: 10.1109/ISMAR-Adjunct51615.2020.00052 2

[64] K. Weerasinghe, S. Kim, S. Iyer, S. Janapati, J. A. Stankovic, X. Ge, and H. Alemzadeh. Real-time multimodal cognitive assistant for emergency medical services. *arXiv preprint arXiv:2403.06734*, 2024. 3, 8

[65] A. S. Williams, J. Garcia, and F. Ortega. Understanding multimodal user gesture and speech behavior for object manipulation in augmented reality using elicitation. *IEEE Transactions on Visualization and Computer Graphics*, 26:3479–3489, 2020. doi: 10.1109/TVCG.2020.3023566 2

[66] J. Zhao, C. J. Parry, R. K. D. Anjos, C. Anslow, and T. Rhee. Voice interaction for augmented reality navigation interfaces with natural language understanding. *2020 35th International Conference on Image and Vision Computing New Zealand (IVCNZ)*, pp. 1–6, 2020. doi: 10.1109/IVCNZ51579.2020.9290643 2

[67] J. Zhou, S. Luo, and F. Chen. Effects of personality traits on user trust in human–machine collaborations. *Journal on Multimodal User Interfaces*, 14:387 – 400, 2020. doi: 10.1007/s12193-020-00329-9 8

[68] X. Zhou, A. S. Williams, and F. Ortega. Eliciting multimodal gesture+speech interactions in a multi-object augmented reality environment. *Proceedings of the 28th ACM Symposium on Virtual Reality Software and Technology*, 2022. doi: 10.1145/3562939.3565637 2

[69] Y. Zhu, G. Hua, X. Liu, C. Wang, and M. Tang. Trust in machines: how personality trait shapes static and dynamic trust across different human–machine interaction modalities. *Frontiers in Psychology*, 16, 2025. doi: 10.3389/fpsyg.2025.1539054 7, 8, 9