PilotAR: Streamlining Pilot Studies with OHMDs from Concept to Insight

NUWAN JANAKA, National University of Singapore, Singapore
RUNZE CAI, National University of Singapore, Singapore
ASHWIN RAM*, National University of Singapore, Singapore
LIN ZHU*, Tsinghua University, China
SHENGDONG ZHAO†, City University of Hong Kong, China
YONG KAI QI, National University of Singapore, Singapore

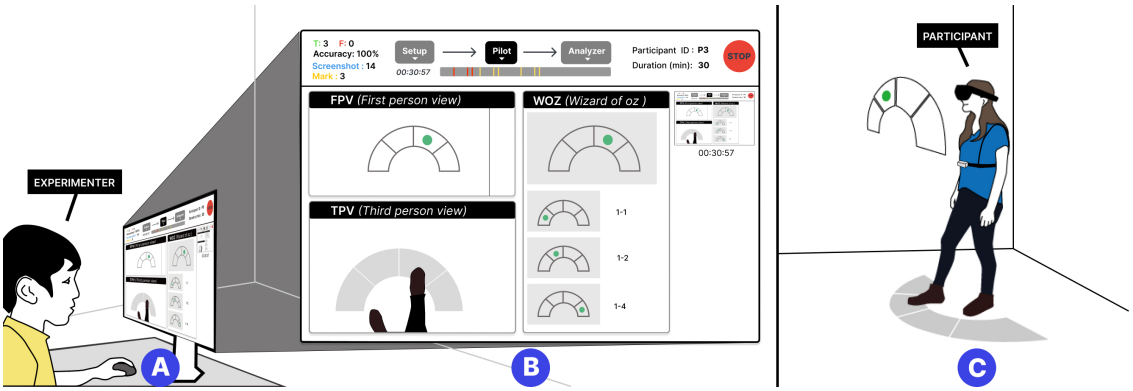


Fig. 1. (A) The experimenter employs *PilotAR*, a desktop-based experimenter tool, for Optical See-Through Head-Mounted Displays (OHMD) based pilot studies. (B) *PilotAR* facilitates real-time monitoring of participants' experiences from both first-person and third-person perspectives, enabling experimenters to track ongoing studies dynamically. In addition, the tool's annotation features allow for the precise marking and capture of significant moments in a photo or video format. Quickly logging quantitative metrics, such as event time, can be done using shortcut keys. Furthermore, a real-time summary of the observed moments and recorded data, available for post-study interviews, promotes in-depth discussions, insights, and support for collaborative review and interpretation. (C) In a separate room, the participant interacts with the simulated AR system, maintaining communication with the experimenter.

Pilot studies in HCI research serve as a cost-effective approach to validate potential ideas and identify impactful findings before extensive studies. Yet, the additional requirements of AR/MR, such as multi-view observations and increased multitasking, make it challenging to conduct pilot studies effectively, hindering innovations in this field. Based on interviews with 12 AR/MR researchers, we identified the key challenges associated with conducting AR/MR pilot studies with Optical See-Through Head-Mounted Displays (OST-HMDs, OHMDs), including the inability to observe and record in-context user interactions, increased task load, and difficulties with in-context data analysis and discussion. To tackle these challenges, we introduce *PilotAR*, a desktop-based tool designed iteratively to enhance OHMD-based AR/MR pilot studies. *PilotAR* facilitates data collection via live first-person and third-person views, multi-modal

*Authors contributed equally to this research.
†Corresponding Author.

Manuscript submitted to ACM

annotations, and flexible wizarding interfaces. It also accommodates multi-experimenter settings, streamlines the study process with configurable workflows and shortcuts, records annotated data, and eases results sharing. Formative testing, conducted using three case studies, has highlighted the significant benefits of *PilotAR*, as well as its potential for further development and refinement.

CCS Concepts: • **Human-centered computing** → **Ubiquitous and mobile computing systems and tools**; **User interface toolkits**; **Mixed / augmented reality**.

Additional Key Words and Phrases: toolkit, tool, pilot, heads-up computing, augmented reality, OST-HMD, smart glasses, evaluation, interaction

ACM Reference Format:

Nuwan Janaka, Runze Cai, Ashwin Ram, Lin Zhu, Shengdong Zhao, and Yong Kai Qi. 2024. *PilotAR*: Streamlining Pilot Studies with OHMDs from Concept to Insight. In . ACM, New York, NY, USA, 35 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Thomas Alva Edison's journey to perfect the light bulb involved conducting thousands of experiments with various potential solutions [30]. This highlights a common pattern in scientific discoveries and technological innovations: significant breakthroughs often emerge from thorough explorations of various hypotheses and potential solutions [7, 60].

In the field of human-computer interaction (HCI), exploring alternative hypotheses and potential solutions is closely linked to conducting pilot studies. Traditionally, pilot studies are defined as small-scale preliminary studies that evaluate the feasibility, duration, cost, and possible adverse events, aiming to refine the study design before a full-scale research project is undertaken [41, 49, 61]. However, within the HCI context, the term "pilot study" does not only refer to scaled-down versions of larger studies but also encompasses formative testing of various prototypes, such as early samples, models, or product releases of interactive solutions [39]. The underlying principle remains consistent: both scientific discovery and technological innovation involve venturing into the unknown, where conducting a full-scale investigation without preliminary exploration can be risky; thus, it is wise to send out low-cost probes to gather more information, and based on the results, decide on how to proceed to the next steps.

The objective of pilot studies is to gather as much insight into the unknown as possible while minimizing the costs of the investigation. This principle is widely practiced in HCI, where cost-effective methods and tools are employed extensively. Techniques such as low-fidelity prototyping and the Wizard of Oz (WOz) testing allow researchers and designers to test and refine potential solutions without significant effort [18, 22, 23].

However, the scenario changes when it comes to conducting pilot studies in Augmented Reality (AR) and Mixed Reality (MR) using Optical See-Through Head-Mounted Displays (OST-HMDs, OHMDs). As an emerging field, OHMD-based AR/MR is garnering considerable attention due to its potential to realize concepts such as the metaverse and heads-up computing [72]. OHMDs have the capability to enable users to interact seamlessly with a blended environment of physical and digital elements, regardless of their location and time [3, 31]. This capability aligns closely with the vision of ubiquitous computing (UbiComp), which advocates for technology that integrates effortlessly into everyday life, making digital interactions as natural as those in the physical world [69].

Despite their potential, pilot studies in AR/MR using OHMDs face unique challenges. Insights from interviews with twelve AR/MR researchers highlight the complexities involved in these studies. They point out difficulties such as setting up complex testing environments and procedures, monitoring virtual content and real-world interactions at the same time, managing various tasks, including observation, facilitation, and real-time manual adjustments during experiments,

and the need for quick analysis and sharing of results with collaborators. These complexities not only increase the costs associated with conducting pilot studies but also hinder the researchers' ability to gather meaningful insights efficiently. As a result, pilot studies in AR and MR using OHMDs are less straightforward compared to traditional HCI methods.

We developed *PilotAR* (Figure 1), a desktop-based tool specifically designed for AR/MR research using OHMDs, targeting WOz pilot studies. *PilotAR* features a guided workflow to simplify setup and use. The tool offers manual and automatic event tagging and shortcut annotation interactions to reduce experimentation costs. It facilitates the easier conduct of pilot studies by streamlining task distribution between the during- and post-pilot phases. To generate more and better insights, *PilotAR* supports rich data capture through integrated first-person and third-person video streaming, enhancing real-time understanding of user interactions. The previously introduced tagging and annotation mechanisms are designed to simplify post-study analysis, thereby increasing the potential to generate deeper insights.

A preliminary usability evaluation of *PilotAR*, conducted with three AR/MR research teams using OHMDs, suggests that it effectively reduces the costs of conducting studies and enhances insight generation. This positions *PilotAR* as a promising tool to accelerate research and innovation in OHMD-based AR/MR.

The contribution of this paper is threefold: 1) It provides empirical knowledge into AR/MR pilot study practices and challenges; 2) It introduces *PilotAR*, an open-source tool tailored for these studies; 3) It demonstrates the tool's comprehensive data collection capabilities, reduced experimenter workload, and enhanced support for in-context discussions through empirical validation.

2 RELATED WORK

In this section, we review the literature on pilot studies in HCI, particularly for developing AR/MR technology. We then explore the challenges AR/MR researchers face during experimentation and review existing AR/MR tools, highlighting their advantages and limitations.

2.1 Pilot Studies

A *pilot studies* traditionally refers to small-scale preliminary investigations conducted before the main study to test hypotheses, validate experimental procedures/designs, and identify possible errors or issues [41, Ch 5][49, Ch 55][28, 59, 61, 63, 64, 67]. In the context of HCI, the term "pilot study" can also refer to informal or small-scale evaluations of various prototypical solutions [39, 41, 61]. These evaluations are crucial for assessing the feasibility of a concept and identifying any usability issues early in the development process. This approach allows researchers and designers to make necessary adjustments before further development and larger-scale testing [41, 59, 61, 63].

Pilot studies are designed to be low-cost and small-scale, typically involving a limited number of participants and less rigorous testing procedures [28, 41, 59, 61, 63]. Their primary goal is to refine hypotheses and solutions, preparing the groundwork for more comprehensive investigations. This approach helps ensure that resources are used efficiently and effectively in the early stages of research.

Pilot studies serve an essential but often less noticeable role in human-computer interaction (HCI) research. While formal studies, such as large-scale controlled experiments, are featured in scientific publications due to their rigorous methodologies and larger sample sizes, pilot studies frequently go unmentioned. Researchers might report only those pilot studies that support the narrative of their papers, often omitting many other attempts that did not yield the desired outcomes, primarily due to space constraints in publications. Despite their lower visibility, pilot studies are crucial for shaping these comprehensive investigations. Without the preliminary insights they provide, the structured and formal

studies commonly seen in academic papers would not be as well-founded or effectively designed [28, 41, 49, 61, 63, 64]. Below, we further elaborate on a few specific reasons why pilot studies are indispensable for HCI research.

Firstly, conducting comprehensive formal studies is both labor-intensive and costly. A single oversight could invalidate months of work. Therefore, seasoned researchers rely on pilot studies to test procedures and designs and to identify potential errors early in the process. Secondly, even well-executed formal studies can sometimes yield uninspiring results or fail to demonstrate improvements over existing approaches. Pilot studies provide a lower-cost avenue to assess the potential for significant findings, increasing the likelihood that the formal studies will lead to meaningful breakthroughs. Lastly, the complexity of HCI research often requires consideration of multiple variables. It's impractical to address all these in one formal study. Pilot studies help in refining the focus by eliminating less relevant factors, thereby narrowing the scope of the study. In conclusion, pilot studies are indispensable methods/tools that facilitate quicker advancements in knowledge and innovation in the field of HCI [41, 49, 61]. Effective pilot studies aim to achieve maximum learning/insights with minimum effort.

2.2 Challenges in OHMD-based AR/MR Research and Pilot Studies

The importance of conducting pilot studies is amplified when the cost of running the formal study increases. This is particularly significant in the context of HCI studies related to OHMD-based Augmented Reality (AR) or Mixed Reality (MR) technologies. Unlike traditional UI design (e.g., 2D interfaces), OHMD-based AR/MR research encompasses both the virtual and physical worlds and their interconnections [23, 56], which entails higher costs due to the inherent complexities and challenges associated with setting up and executing such studies [2, 34, 47, 57]. Furthermore, as an emerging field, researchers encounter more challenges during the design, development, and testing phases of OHMD-based AR/MR research due to a lack of authoring tools that require minimal technical competencies yet still provide the desired functionalities [34, 47], and a deficiency of experimentation tools supporting data capture and analysis [6, 13].

Raffaillac and Huot emphasize that the research studying the requirements of HCI researchers is surprisingly sparse compared to the array of toolkits designed for them [51]. While the challenges and needs concerning AR/MR design and development aspects have been examined [2, 34, 47, 57], there remains a lack of information about the needs of experimenters regarding AR/MR testing and evaluation of research (e.g., [52]). Carter et al. [13] investigated experimenter needs in the domain of ubicomp experiments, but given the unique characteristics of AR/MR experiments (e.g., context, interface, relations [70]), certain needs (e.g., different views to understand the relationships) were overlooked in their study. Thus, we conducted a formative interview study (Sec 3) to build upon previous research, providing new empirical insights into the experimentation process of AR/MR researchers with pilot studies.

2.3 Tools for AR/MR Pilot Studies

Our formative study (Sec 3) identified that AR/MR researchers using OHMDs require support across all phases of the pilot study, including **pre-pilot (e.g., setup)**, **during the pilot (e.g., experimentation)**, and **post-pilot (e.g., analysis and summarization)**. In reviewing related work, most previous studies fall into one of two categories. The first category consists of tools that support all study phases but are designed for formal studies. These tools often require significant effort to set up and use, which makes them unsuitable for the rapid iteration needed in pilot studies. The second category includes tools that are lightweight and easy to use, but they only cater to one stage of the pilot study lifecycle and do not support the entire process.

2.3.1 Tools for Formal Studies. MRAT [48] is designed to support high-fidelity MR studies involving all phases. While being useful, it demands advanced skills in developing MR scenes using Unity3D, which is both time-consuming and requires specific technical skills, making it less effective as a pilot study tool.

2.3.2 Tools Supporting Specific Phases of the Pilot Study Lifecycle.

Tools for Setup. Existing AR/MR tools primarily focus on the initial setup phase [20, 26, 48], emphasizing rapid prototyping and content creation. This includes content authoring tools (reviewed in [46, 47]) and rapid prototyping tools (reviewed in [23]), as well as gesture interaction tools (e.g., [68, 71]), which are mainly used for the pre-pilot phase. This gap has motivated us to develop a tool that supports the entire pilot study lifecycle.

Wizard-of-Oz in AR/MR Pilot Studies. During initial study phases, including pilots, experimenters often use low-fidelity prototypes and basic applications to accelerate iterations and reduce setup costs [13, 18, 19, 23] (Sec 3.2). The wizard-of-oz (WOz) protocol [17, 22], where experimenters simulate the expected application behavior using (low- to high-fidelity) prototypes, is commonly used in AR/MR studies to reduce setup and simulation costs [5, 18, 21, 22]. Our tool, *PilotAR*, supports this approach by facilitating the setup process with a range of interfaces from low-fidelity (e.g., paper [14]) to high-fidelity (e.g., Unity3D [1] [10]), enhancing flexibility[2] and reducing the resources[3] needed for the pre-pilot setup.

Tools for Experimentation. During the pilot phase, which includes observation, data collection, and task management (Sec 3.4), *PilotAR* offers functionalities similar to AR tools like the Immersive eXperimenter Control Interface (IXCI) [53, 54] and the Designer's Augmented Reality Toolkit (DART) [24, 40]. However, unlike these platforms that often require high-fidelity implementation skills (i.e., higher setup cost), *PilotAR* provides more accessible multi-view observations and in-situ data annotations, supporting even low-fidelity prototypes (i.e., lower setup cost).

Tools for Analysis. In the post-pilot phase, which involves data analysis and summarization (Sec 3.5), *PilotAR*, while not as specialized as tools like ReLive [29] or MIRIA [12] (or others[4] [16, 50, 58]), provides essential functionalities for immediate retrospective observations and efficient note-taking enabling higher insight capturing, which are crucial for the iterative design and refinement of pilot studies.

In summary, *PilotAR* distinguishes itself by streamlining every phase of the OHMD-based AR/MR pilot study lifecycle, effectively the need for rapid iterative exploration by optimizing both costs and insight generation at various stages. For a comprehensive feature comparison, please refer to Appendix A.

3 STUDY 1: UNDERSTANDING THE CHALLENGES FACED BY RESEARCHERS DURING THE EARLY STAGES OF AR/MR STUDIES

To understand the experimenters' challenges during AR/MR pilot studies which are underrepresented in literature (sec 2.2), we conducted semi-structured interviews with 12 AR/MR researchers (R1-R12), all of whom have experience with OHMD-based experiments ranging from 2 to 10 years (see Appendix B.1-Table 2 for details). We employed the

[1] https://unity.com/

[2] enabling to use existing prototyping tools, serving to generate content or function as wizarding interfaces such as remote paper prototypes [14], 360 experiences with paper [45], in-situ 3D sketches in video prototypes [38], spatial prototypes incorporating real-world motion [44], and cross-reality prototypes [25]

[3] lowers the technical skill barriers for high-fidelity prototyping (e.g., IXCI [53, 54], Welicit [5], UXF [10])

[4] non-MR tools like Noldus's Observer XT[5] [74], ANVIL [33], EXCITE [42], EagleView [11]

critical incident technique [55] to discern the design requirements for our tool. Thus, we asked the researchers about their past AR/MR research projects, the tools or methods they used, the team collaboration, the stages of the projects, how they progressed, challenges faced during the early stages of the projects, and their mitigating strategies. The interviews, each lasting approximately 45 minutes, were transcribed and subsequently thematically analyzed following Braun and Clarke [8] (see Appendix B.2 for details). The insights from this study offer a more nuanced picture of the challenges experimenters face during different phases of OHMD-based AR/MR pilot studies and subsequently helped us articulate design goals for the *PilotAR*.

3.1 Purposes of Pilot Studies in AR/MR

As our interviewees detailed, pilot studies play a crucial role in the early stages of AR/MR research projects, serving three primary functions: 1) guiding design space exploration to identify potential research avenues (10/12). —"*I often use pilots to see how conditions change and if it looks promising ... how it affects user behaviors... they help narrow down on things to test and their practicality* (R2)", 2) comparing multiple interfaces, interactions, or systems to discern their pros and cons (9/12) quickly. —"*I compared our system with others [during piloting] to see whether the formal study would work* (R1)", 3) identifying usability concerns to improve them (12/12). —"*Pilot studies helped me identify usability issues to refine our proposed interface and layout.* (R1)"

3.2 Pilot Study Process

All researchers conducted multiple iterative pilot studies, integrating findings from each study into the next, leading to either a formal study or project discontinuation. They employed prototyping [18] at varying fidelity levels, either alone (5/12) or in combination with the wizard-of-oz technique (7/12) [18, 22], for quick and systematic design testing and validation (5th-6th column of Table 2).

Similar to the formal experimental lifecycle (i.e., setup, experimentation, analysis, summarizing) [5, 13], pilot studies encompass multiple steps, which we categorized into *pre-pilot*, *during-pilot*, and *post-pilot* phases. *Pre-pilot*, experimenters create and set up testing environments, like AR/MR content and OHMDs. *During-pilot*, they observe behaviors, take notes, and collect data to evaluate their designs, preliminary research questions, and hypotheses. Finally, in *post-pilot*, experimenters interview participants to resolve any ambiguities and gain deeper insights. They analyze data, summarize the evidence supporting or opposing the research questions, and validate the hypotheses. This phase ends in discussions with collaborators to plan future/next steps (e.g., iteration).

The following sections describe the challenges researchers faced in each phase. Due to the iterative nature of pilot studies, some challenges spanned multiple phases. Additionally, certain challenges are not unique to AR/MR pilots with OHMDs but also relevant to other HCI studies, like ubicomp experiments [13].

3.3 Challenges in the *Pre-pilot* Phase

Aligned with previous research, researchers faced challenges in content preparation [2, 34] and tool usage [23, 24, 34, 47], particularly due to the absence of quick authoring tools for AR/MR experiences. To address these issues, researchers employed low-fidelity prototyping methods, such as PowerPoint/Google Slides, paper, and video (Appendix B.1-Table 2). They also utilized wizard-of-oz techniques with digital tools like Figma, Miro, and Google Slides via platforms such as Zoom or Google Meet, in addition to mirroring 2D content to AR/MR devices by connecting devices like Nreal/Xreal Light glasses to a tablet. However, because there are no standardized procedures or preparation guides, this process often becomes ad hoc and tedious and consumes a significant amount of time and energy to set up.

A prominent but less-known challenge is the necessity of testing AR/MR content on OHMDs in realistic settings, as opposed to other platforms like mobile phones or video see-through HMDs (mentioned by 10/12 researchers). This need stems primarily from color blending issues on transparent displays [31]. As R2 highlighted, "*I create video content on the computer, which looks great. However, when transferred to smart glasses, it differs significantly, especially outdoors or in motion. The content that looks good on one pair of glasses may need recreation for another.*"

3.4 Challenges in the *During-pilot* Phase

During the pilot study, researchers faced challenges in study execution and data collection. Generic challenges included managing multiple tasks simultaneously (8/12), such as observing, note-taking, and wizarding, leading to task overload and fatigue [5], and unfamiliarity with the pilot steps (3/12), resulting in inconsistent experiments. This multitasking often hindered detailed observation note-taking (9/12), hindering subsequent in-depth post-interview questioning.

Although these challenges could be partly mitigated by enlisting additional experimenters, over half of the researchers (7/12) conducted pilot studies alone due to resource constraints, such as limited trained personnel.

3.4.1 Challenges with Data Collection. AR/MR studies pose unique challenges compared to traditional usability lab tests, especially in observing participants (11/12) and the system (10/12).

In traditional settings, such as desktop (2D) environments, researchers observe participants' behaviors and digital interactions from a third-person view (TPV). However, in AR/MR environments, TPV alone is insufficient to capture interactions with virtual content, which remains invisible from this perspective. Access to the first-person view (FPV), which includes egocentric views with virtual content, is often restricted by experimenters' lack of specialized skills or knowledge. This limitation is particularly problematic in wizard-of-oz (WOz) methodologies [1, 22, 23, 37], leading researchers to "*guess*" participants' intentions. Such constraints often result in the need for repeated trials when errors are recognized post-trial without access to real-time monitoring capabilities.

When experimenters have access to both TPV and FPV, managing multiple viewpoints increases multitasking demands and complicates detailed note-taking, especially when they must simultaneously operate multiple devices or tools (e.g., wizarding).

To address these challenges, researchers employed semi-automatic recording methods such as audio, video, and system logs [54], alongside tools from OHMD vendors for virtual content observation, including Windows Device Portal[6] for HoloLens 2 and Meta Quest Developer Hub. They also implemented think-aloud protocols and pre-tested systems before trials. However, similar to the challenges encountered in the *pre-pilot* phase, the absence of standardized procedures or preparation guides often renders this process ad hoc, tedious, and highly time-consuming.

3.5 Challenges in the *Post-pilot* Phase

In the *post-pilot* phase of AR/MR studies, researchers commonly face two primary challenges: data analysis and results sharing. These challenges, while typical in HCI research [13], are particularly pronounced in AR/MR settings due to their context-dependent nature [6], especially when contextual information is lacking during analysis and results sharing.

3.5.1 Challenges with Data Analysis. Researchers (7/12) encountered difficulties during interviews, primarily due to insufficient contextual information and a lack of processed quantitative data. On the one hand, the absence of contextual

[6]https://learn.microsoft.com/en-us/windows/mixed-reality/develop/advanced-concepts/using-the-windows-device-portal

information sometimes made it difficult to ask relevant questions or recall specific user experiences without detailed notes. On the other hand, participants often forgot their earlier behaviors, which hampered the research process. Additionally, without preliminary quantitative analysis, researchers found it challenging to validate hypotheses or discern between the effectiveness of different techniques, such as technique A versus technique B. This limitation made it difficult to pose meaningful questions during post-pilot interviews that could provide deeper insights into specific issues, which typically only emerged after a detailed quantitative analysis of the results.

To address these issues, researchers took detailed notes and recorded key moments. Some attempted "live" analysis, which involves consolidating measures or processing data in real time, to better prepare for interviews. Although a few researchers (3/12) tried using video analysis to aid their work, they struggled with efficiently navigating to relevant segments due to limited search capabilities. Conversely, another group (3/12) opted to avoid detailed recordings altogether to save time on extensive post-pilot analysis.

3.5.2 Challenges with Results Sharing. Sharing AR/MR findings with collaborators who hadn't experienced the application firsthand was particularly challenging (6/12). Traditional text notes often failed to adequately convey the nuanced, situated experiences of participants, making it difficult for collaborators to grasp the behaviors observed and suggest effective design improvements fully.

To tackle these challenges, researchers employed detailed note-taking and video recording of crucial moments. Additionally, some researchers (3/12) allowed on-site collaborators to directly "try out" the pilot study, which enabled them to experience the participants' perspectives firsthand and contribute to more informed discussions and feedback. However, this direct involvement isn't always feasible. In cases where first-hand observation isn't possible, it becomes challenging for collaborators not actively involved in the study to understand the outcomes and provide valuable input fully.

3.6 Summary and Design Goals

Based on our interviews and related literature, we formulated design requirements for a tool that addresses common challenges in OHMD-based AR/MR pilot studies.

Ease of Setup in Conducting Pilot Studies: To significantly reduce the costs associated with setting up pilot studies (Sec 3.3), it is essential for the tool to offer a guided setup procedure, pre-configured templates, and user-friendly interfaces. These features should be designed to accommodate users with varying levels of technical expertise, streamlining the setup process and minimizing the time and resources required. This approach enhances efficiency and makes the pilot study process more cost-effective.

Support for Familiar Wizarding/Simulation Interfaces: The tool should be equipped to handle a range of familiar wizarding and simulation interfaces, facilitating quick iterations across diverse experimental setups (Sec 3.2). This feature is instrumental in the rapid testing of ideas and ensures compatibility with existing presentation and simulation technologies (Sec 2.3.2), thereby reducing both technical and financial barriers for researchers and experimenters.

Support for Observations in Situated Contexts: To achieve comprehensive data collection, the tool needs to support monitoring user interactions from both first-person and third-person perspectives (Sec 3.4, [56]) in their natural environments. This multi-perspective approach is vital for accurately identifying system errors and unexpected user behaviors, enriching the depth of insights derived from the study.

Reduce Task Load of Experimenters: Automating processes such as recording, note-taking, and measuring can significantly alleviate experimenters' cognitive and physical load. Moreover, enabling collaborative use by multiple experimenters (Sec 3.4) helps evenly distribute the workload. This ensures continuous observation and maintains data accuracy while reducing the overall effort required to manage the study.

Expedite Data Recording, Analysis, and Generation of Creative Insights: The tool should offer immediate access to data recordings and support easy annotation during and after experiments. These features facilitate faster turnaround times in data analysis and the generation of creative insights (Sec 3.1–3.2). Additionally, functionalities, like annotated video recordings [33, 74] and straightforward navigation to specific instances during post-experiment reviews, are crucial for quickly pinpointing relevant data, thus enhancing both the speed and quality of insight generation.

4 *PILOTAR* TOOL

In this section, we delineate the functionalities of the tool that meet the design goals specified in Sec 3.6 and describe a typical usage scenario of *PilotAR* (Figure 2). For details on iterative tool design and its role in verifying the design goals and elucidating detailed requirements, refer to Appendix C.

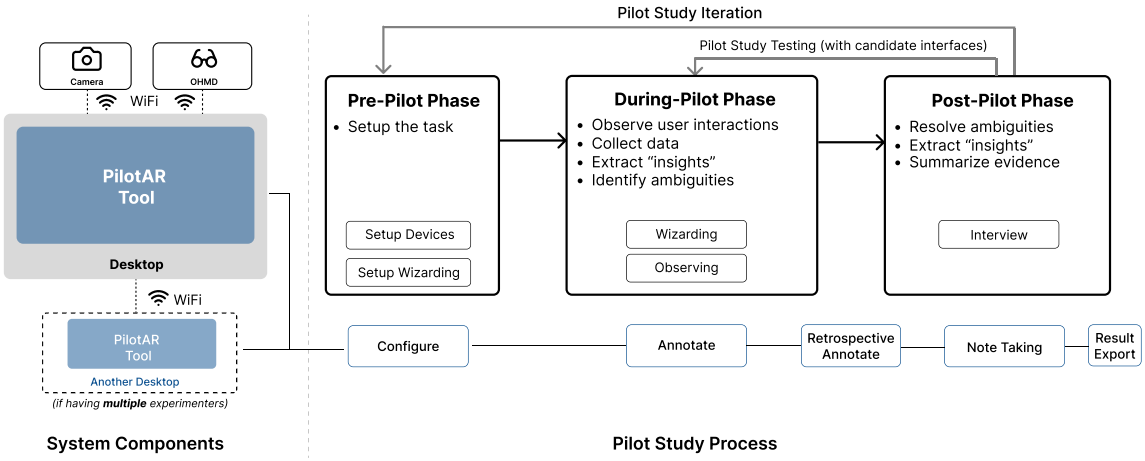


Fig. 2. Overview of the system components and workflow with *PilotAR*.

4.1 Major Functions

4.1.1 FPV and TPV Live Streaming. Although relatively straightforward in design, as an essential feature, we enabled experimenters to observe participants wearing OHMD in situated contexts through the live first-person view with grids (FPV)[7] and third-person view (TPV). These video streams are simultaneously recorded for subsequent analysis. Specifically, FPV streams the overlay of digital content and the realistic environment rendered by the OHMD. TPV streams video from a user-attached camera or one positioned by experimenters.

[7]The current implementation accommodates multiple camera/video streams, with successful testing for up to three streams. As determined through iterative design (Appendix C), we have implemented grids on the video stream to assist experimenters in locating and positioning virtual content.

4.1.2 Annotations with Function Shortcuts. To facilitate documentation during pilot study observations, we enable a variety of annotations. These encompass *Screenshot* (to capture the current screen, optionally with a colored block highlighting a specific Region of Interest (ROI)), *Focus* capturing only a selected screen region), *Correct* and *Incorrect* (for accuracy calculations), and *Counter* (for tracking interaction attempts). The communication between experimenters and participants is recorded and transcribed to *Voice Annotation* in text format. During pilot studies, experimenters can use customized keyboard shortcuts to activate *Annotation* functions. These shortcuts can be mapped to UI, user, or experimenter actions for automatic annotations. Additionally, each *Annotation*'s color can be customized for easy identification, and all annotations are time-stamped for later review.

4.1.3 Multi-experimenter Support. To reduce task load during pilot studies, we support multi-experimenter scenarios alongside single-experimenter setups. In a single-experimenter scenario, the experimenter concurrently manipulates the wizarding interface, conducts observations, and makes annotations. In the multi-experimenter configuration, one experimenter can act as the wizard, adjusting the interface based on users' actions observed via FPV and TPV, and another experimenter can focus solely on observation and annotations. After the pilot, annotations from both experimenters are seamlessly synchronized[8].

4.1.4 Analyzer. To allow experimenters to get a real-time summary of the collected data, we implemented the *Analyzer* view. By reviewing the annotation index on the recording's timeline, experimenters can identify key moments and use video playback to assist participants in recalling their experiences. Experimenters can adjust annotations recorded during the pilot session (e.g., change timestamp, modify manipulation correctness, modify notes), add new notes, and take screenshots. The analyzer also provides a quick summary of accuracy and the time duration between two indices of *Annotation* and corresponding events.

4.1.5 Summary Review. To facilitate information sharing among collaborators, a comprehensive review of the pilot results can be exported from the analyzer, including overall descriptive statistics, selected annotation timestamps, notes, and screenshot images. Raw data (e.g., video) can be shared for subsequent analyses.

4.2 *PilotAR* Usage Scenario

Experimenters might adopt various strategies with *PilotAR*. Here, we outline a basic approach for conducting a pilot study using *PilotAR*, with the replication of 'Mind the Tap' [43] as an example to highlight its usage.

> *Mary, an AR researcher, conceives a novel idea employing foot-tapping as an input interaction for OHMDs [43] (**Figure 1**). She identifies two potential interactions:* direct *(i.e., the menu appears on the floor within leg's reach) and* indirect *(i.e., the menu displays in front of the eyes, requiring users to use proprioception to associate it with their foot, Figure 1C). She aims to discern the strengths and limitations of each foot-tap interaction. Choosing a within-subject design for an initial comparison, Mary opts to employ the wizard-of-oz technique to minimize developmental efforts in a tangible system (e.g., Unity development with optical tracking) and to persuade colleagues to explore this concept further.*

4.3 System Components

To support the scenario described above, as shown in **Figure 2** and **Figure 1**, we utilize additional hardware and software components besides the *PilotAR*. These include an OHMD, specifically the HoloLens2[9], and a TPV camera

[8]Note: Additional tags have been added for annotations made by the other experimenter.
[9]https://www.microsoft.com/en-us/hololens/hardware

stream, which can be provided by devices such as a phone, tablet, laptop camera, USB camera, or IP camera (e.g., DroidCam[10] mobile app). On the software side, we utilize a wizarding interface to display and manipulate OHMD content based on user reactions. This interface can range from low-fidelity solutions like slides (e.g., Google Slides[11]) and whiteboards (e.g., Miro[12], Figma[13]) with communication software (e.g., Google Meet[14], Zoom[15], MS Teams[16]), to high-fidelity prototypes such as Unity3D[17] or Unreal Engine applications with holographic remoting capabilities (e.g., Holographic Remoting Player[18]).

4.4 Interface and Workflow

The main workflow using *PilotAR* is divided into three phases: *pre-pilot*, *during-pilot*, and *post-pilot*. This section demonstrates how Mary can utilize *PilotAR*'s interfaces throughout these phases.

4.4.1 Pre-pilot Phase. As shown in Figure 3, the experimenter set up system components and configures *PilotAR*, which involved role selection, device configuration, checklist creation, and shortcut key customization for *Annotations*.

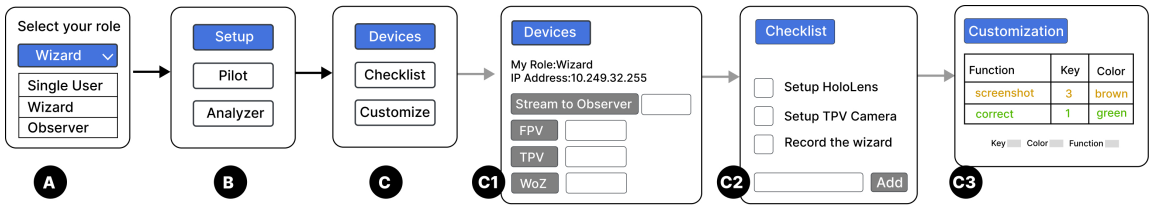


Fig. 3. Workflow of Setup UI. Upon starting the tool, the experimenter is prompted to select the role (A), including single- and multi-experimenter (wizard/observer). Then, menu (B) indicates the three major steps of conducting a pilot study: Setup, Pilot, and Analyzer. In Setup (C), there are three sub-steps, including device configurations (C1), checklist configuration (C2), and annotation customization (C3).

Mary quickly crafts a wizarding interface using Google Slides with a 2x4 menu, where the target location randomizes on subsequent slides. She mirrors these slides to the HoloLens 2 (HL2) via Google Meet on a browser. She uses a phone camera as the TPV by linking it to Google Meet. For direct *interactions, the mirrored WOz interface is fixed on the floor. Conversely, for* indirect *interactions, it's positioned in front of the users' eyes.*

Role Selection (Figure 3A). Upon launching the tool, the experimenter is prompted to select their role: *single-user* for single-experimenter pilots, or *wizard/observer* for multi-experimenter pilots.

Device Configuration (Figure 3C1). This task allows the experimenter to input essential information such as FPV and TPV connections (e.g., IP address, credentials), *Wizarding Interface* (e.g., Google Slides URL link or python file path), and screen recording inputs (e.g., video and audio source), making them all displayed on the monitor.

[10] https://play.google.com/store/apps/details?id=com.dev47apps.droidcam&hl=en&gl=US
[11] https://docs.google.com/presentation
[12] https://miro.com/
[13] https://www.figma.com/
[14] https://meet.google.com/
[15] https://zoom.us/
[16] https://www.microsoft.com/en/microsoft-teams/group-chat-software
[17] https://unity.com/
[18] https://learn.microsoft.com/en-us/windows/mixed-reality/develop/native/holographic-remoting-player

Checklist Creation (Figure 3C2). The checklist aids in remembering crucial steps during the pilot study, such as confirming OHMD, TPV camera, and recording. Customizable items can be added by typing in the provided space at the bottom.

Shortcut Key Customization (Figure 3C3). Experimenters can manage which *Annotations* are displayed during the pilot session (known as Pinned *Annotation*) and customize aspects like color, name, and shortcut key.

Mary initiates the PilotAR, selects 'Single User' (Figure 3A), and sets up the devices (Figure 3B) with the HL2 IP address for FPV, a Google Meet link for TPV, and Google Slides for the Wizarding Interface (Figure 3C1). She then adds a "Check foot visibility" checklist item (Figure 3C2) to verify the FPV setup is accurate before each pilot session. To ascertain accuracy and usability, she enables (Figure 3C3) Correct, Incorrect, Counter, and Screenshot annotations.

4.4.2 During-pilot Phase. After setting up and confirming the checklist, experimenters can enter the anticipated duration[19] and participant and session ID and initiate the "Pilot" phase by clicking the "Start/Stop" button on the top bar (Figure 4A).

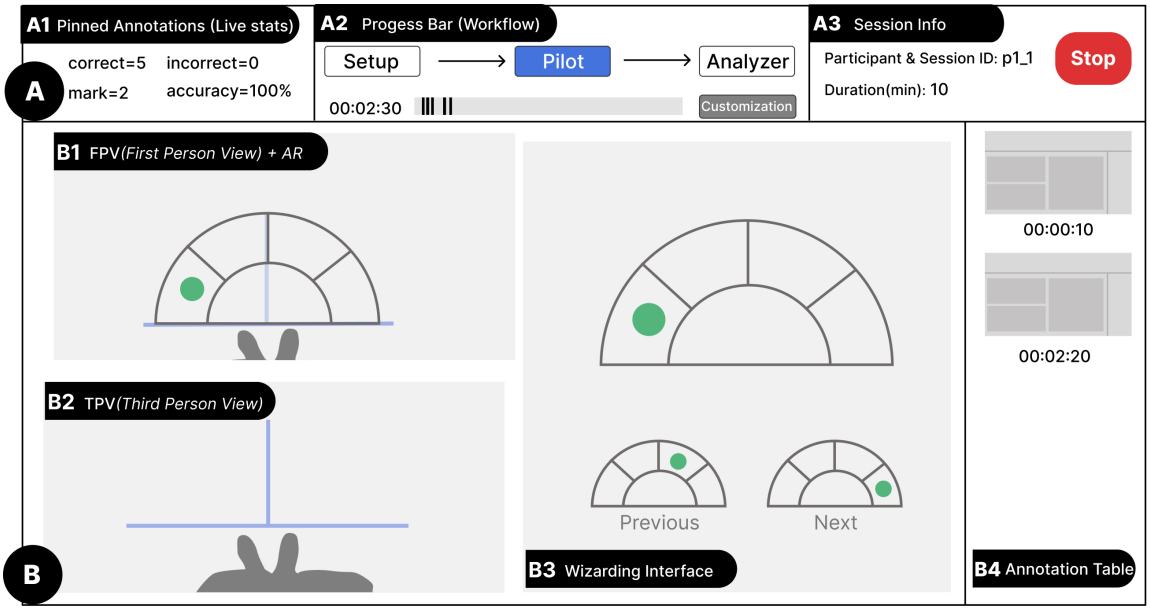


Fig. 4. Pilot interface, which includes two major areas. Area (A) is the Top Bar showing (pinned) *Annotations*' live statistics (A1), the session progress (A2), and session information (A3). Area (B) presents the main working panel housing the FPV (B1, which shows the digital interface and user's feet from their FPV), TPV (B2), *Wizarding Interface* (B3), and a sidebar for the annotation table (B4).

Top Bar (Figure 4A). The top bar displays session-related metadata, including live statistics of measures (e.g., count of Annotations, Figure 4A1), session progress (e.g., duration and timeline, Figure 4A2), and session information (e.g., participant info, anticipated duration, Figure 4A3). Experimenters will receive a notification when the anticipated time has elapsed and can stop the session by clicking the "Stop" button located at the right corner of the top bar.

[19]The experimenter can estimate the session's duration; exactness is not required.

Main Working Panel (Figure 4B). The working panel displays FPV (Figure 4B1), TPV (Figure 4B2), and *Wizarding Interfaces* (Figure 4B3), with a layout that can be customized according to the experimenter's preferences. In the right corner of the working panel, the captured *Screenshot* and *Focus* annotations using keyboard shortcut keys (e.g., "3" key key) are shown as images with timestamps in the Annotation Table (see Figure 4B4). Clicking on these images opens a pop-up window, allowing the experimenter to add notes to the annotations.

[Piloting with the First Interface] Mary then invites a friend to participate in the pilot, affixing the TPV phone to their chest to monitor foot interactions (Figure 4B1). After the briefing and training, the pilot starts with the direct *interface (Figure 1C). Adjusting the target location on the Wizarding Interface (Figure 4B3), she annotates accuracy across ten trials, taking screenshots of any unusual or interesting behaviors (Figure 4B4). Mary also monitors the trial count and accuracy via the live statistics dashboard (Figure 4A1).*

4.4.3 Post-pilot Phase. The final step involves a *post-pilot* analysis. Upon completion of the pilot session, the *Analyzer* window appears (Figure 5), displaying the video panel on the left (Figure 5A) and *Annotations* panel on the right (Figure 5B).

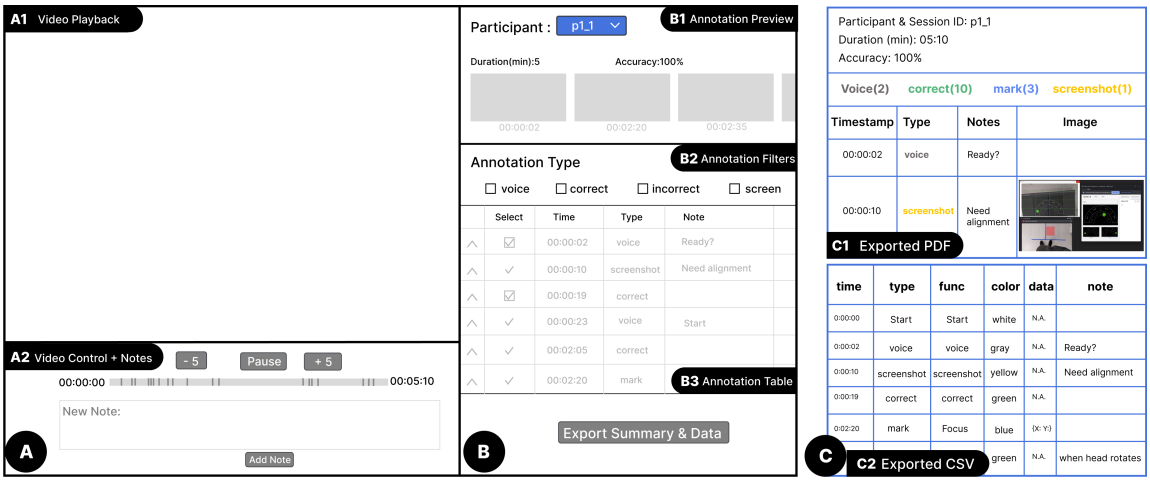


Fig. 5. The Analyzer interface comprises two main panels: the video panel (A) and the annotation panel (B). The video panel includes video playbacks of the pilot (A1), video controls, and a new note panel (A2). The annotation panel features an annotation preview (B1), annotation filtering options (B2), an annotation table (B3), and an exporting button. The Analyzer supports exporting the annotations (C) in PDF format (C1) and CSV format (C2).

Video Panel (Figure 5A). The video panel can play[20] the recorded video (Figure 5A1) and navigate to any timestamp by clicking the timeline (Figure 5A2) or using three buttons to rewind, pause, and fast-forward. Experimenters can create new *Annotations* with notes in the "New Note" area below the video timeline (Figure 5A2).

Annotation Panel (Figure 5B). The annotation panel features an annotation preview (Figure 5B1), annotation filtering options (Figure 5B2), an annotation table (Figure 5B3), and an exporting button. The annotation preview (Figure 5B1)

[20] at a 0.5x, 1x, 2x; the default is 1x

provides an overview of the pilot, including its duration, manipulation accuracy, and collected screenshots. Experimenters can click on these screenshots to pinpoint annotated moments in the recorded video.

Within the Annotation table (Figure 5B3), experimenters have the capability to view and adjust annotation details by double-clicking on a cell. Additionally, specific Annotations can be highlighted by clicking the corresponding icon in the first column or applying the filters available (Figure 5B2). The tool also facilitates the export of summaries and selected *Annotations* in both PDF and CSV formats (Figure 5C).

[Analysis] Upon finishing the session, the Analyzer activates, presenting screenshots, accuracy data, and annotations (Figure 5). Before the interview, Mary reviews these annotations and accuracy (Figure 5B1-B3), devising questions for further inquiry. For clarity on specific screenshots, she replays footage from 5 seconds prior (Figure 5A1-A2). She then conducts the interview, discussing the participant's experiences and challenges, and incorporates their feedback into the annotation notes (Figure 5B3).

Experimenters can return to the "Pilot" session for subsequent pilot studies and initiate new recordings. All interactions in the *Analyzer* are stored, enabling experimenters to switch between different pilot recordings using the drop-down menu in Figure 5B1.

[Piloting with the Alternative Interface] After assessing the direct *interface, Mary tests the* indirect *interface in the same approach.*

[Overall Analysis] After piloting both interfaces, Mary invites the participant for an overall interview, utilizing the Analyzer to toggle between pilot recording sessions or view them simultaneously (Figure 5B1). This comparison offers insights into "rough" accuracy and usability variations, which are noted in Analyzer (e.g., direct one is slightly more accurate while causing neck pain for long usage, (Figure 5B3).

[Repeating] Mary replicates this process with three more participants, counterbalancing the interface. Mary exports participant data summaries in PDF (Figure 5C1) and shares them with colleagues to convince the differences between direct *and* indirect *interfaces. She cites participant feedback and replays specific recordings for context when queried for details.*

[Further Exploration: Multi-experimenter] Seeing the team's interest, Mary broadens their exploration to assess how interaction accuracy and speed vary between two interfaces as menu size changes. She trains a colleague to act as the wizard, thus reducing the wizarding workload and focusing more on observations. After creating additional slides for varied menu sizes (e.g., 1x2, 2x4, 3x6), they conduct pilot tests with four participants using a between-subjects design. To calculate the speed of interactions, they combine Correct/Incorrect annotations with custom annotations that automatically mark target changes (linked to slides' changes). After each pilot session, data is exported to CSV (Figure 5C2) for graph generation in Excel, which facilitates comparing relationships among speed, accuracy, and menu size. Convinced that their pilot study has uncovered a notable trend, the team decides to transition to a formal study.

[Summary] Employing the wizard-of-oz methodology with PilotAR, the team expedites (e.g., less than one week as opposed to a full-fledged motion tracking application, which can take several weeks to months) the identification of viable research directions. Using PilotAR, experimenters can overcome challenges in rapidly evaluating diverse concepts, gathering preliminary quantitative measures for comparison, and convincing colleagues, significantly shortening the knowledge discovery phase.

4.5 Implementations

We used Python (3.9) as our primary programming language due to its cross-platform compatibility (e.g., Windows, MacOS). To achieve the tool's functionalities, we incorporated several third-party packages. The user interface (UI) was developed using Tkinter[21] and related theme packages, such as CustomTkinter[22]. The *PilotAR* utilizes Pynput[23] to monitor user inputs and FFmpeg[24] to handle screen recording. For video playback, we used Python-VLC[25] and audio transcription we used Whisper[26]. FFmpeg and websocket were incorporated to enable video and data streaming between the wizard and the observer in multi-experimenter settings. Detailed information about the **open-source** implementation can be found in https://github.com/Synteraction-Lab/PilotAR.

5 STUDY 2: CASE STUDY EVALUATION AND EXPERT REVIEW

To assess the usage of *PilotAR*, we adopt the usability study approach outlined by Ledo et al. [36]. We observed three research teams using *PilotAR* for their initial investigations to understand whether and how *PilotAR* can facilitate AR/MR pilots. In addition, we presented *PilotAR* to two renowned senior AR/MR research experts and sought their input. None of the volunteers had participated in previous studies or received any compensation[27].

5.1 Observation Study

To evaluate the usage of *PilotAR* in realistic settings, we partnered with a local research institution specializing in smart systems related to HCI, design, XR, AI, and robotics for both academia and industry, and performed three case studies with three teams (T1, T2, T3), as detailed in **Figure 6**. Two teams used a single experimenter setting, while one team used a multi-experimenter setting.

We tracked tool usage during the pilots, conducted post-pilot interviews with experimenters, and gathered questionnaire data on *PilotAR*.

5.2 Expert Review

We extended invitations to two renowned senior researchers (E1, E2) with more than 15 years of experience in the AR/VR field. E1 is a pioneer in AR development, AR display technology, and 3D rendering. E2 is an expert in wearable devices, including smart eyewear, attention-aware computing, and embedded systems. They provided feedback on its usage after a walkthrough demonstration of the functionalities of *PilotAR* and using it in a wizard-of-oz study.

Throughout the usability study, volunteers engaged in a messaging task while multitasking. The participants replied to OHMD text messages sent by the wizard (i.e., experimenter) using speaking or typing (***Wizarding Interface*: Python program**). Using the *PilotAR*, E1 and E2 acted as the experimenters who had to identify the usability issues with the OHMD texting prototype and observe how participants' texting behavior changed with multitasking complexity.

5.3 Findings

We categorized our observation notes and user interview feedback based on *PilotAR*'s design requirements and other emerging themes, using the same analysis method as in *Study 1* (see Appendix B.2).

[21] https://docs.python.org/3/library/tkinter.html
[22] https://github.com/TomSchimansky/CustomTkinter
[23] https://pypi.org/project/pynput
[24] https://ffmpeg.org
[25] https://pypi.org/project/python-vlc/
[26] https://openai.com/blog/whisper/
[27] All participants were given free access to use the *PilotAR* in their future studies.

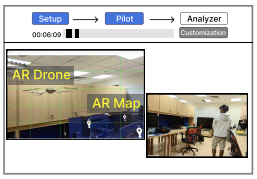
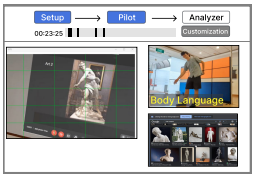
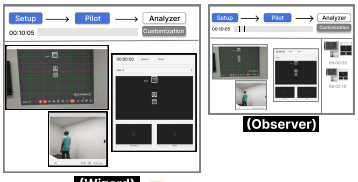
	Case Study 1: Drone-based Inspection	**Case Study 2:** Multimodal Search	**Case Study 3:** Head-Gestures for Menu Selection
Goal	Investigate how AR can assist in inspecting indoor building spaces and make the process more accessible to novice users	Explore novel search methods using verbal and non-verbal inputs (gestures, poses) in MR	Evaluate the proposed head-gesture menu selection technique (*Head-Pitch*), comparing it against prevalent methods such as the *Dwell* technique during walking
Team	Team **T1**, **Single-Experimenter** 2 PhD students 1-2 years of OHMD experience	Team **T2**, **Single-Experimenter** 1 Postdoc, 2 Undergraduates 0.25 - 4 years of OHMD experience	Team **T3**, **Multi-Experimenter** 1 Postdoc, 2 Master's students 1-3 years of OHMD experience
System	 unity + zoom • High-fidelity Unity3D app as the wizarding interface. • Holographic Remoting to mirror the wizarding view on OHMD. • Zoom on the phone for TPV	 Google Images + Google Meet • Low-fidelity Google Images web page as the wizarding interface. • Google Meet to mirror the wizarding view on OHMD. • Google Meet on the phone as TPV	 Google Slides + Google Meet • Low-fidelity Google Slides as the wizarding interface. • Google Meet to mirror the wizarding view on OHMD. • DroidCam app on phone as TPV
Task	**Objective** • Explore users' path-planning behaviors and associated usability issues **Participant** • Plan the drone's flight path in the real environment by placing virtual waypoints within a miniature virtual environment mode	**Objective** • Investigate how users formulate queries using multimodal input. **Participant** • Search for various common (e.g., items for a birthday party) and obscure (e.g., a statue in a specific pose) objects using multimodal input.	**Objective** • Compare the *HeadPitch* with the *Dwell* approach in AR menu selection **Participant** • Select the designated menu item in 10 random trials for each technique
Sessions	• 2 sessions (2 participants per session, 15 minutes per participant)	• 1 indoor session (4 participants, 15 minutes per participant). • 1 outdoor session (1 participant, 45 minutes)	• 1 single-experimenter session (1 participant, 10 minutes). • 1 multi-experimenter session (4 participants, 10 minutes per participant).
During-pilot phase	**Experimenter** • Guided participants through system operation using verbal instructions • Modified system's status for correct interaction. • Took Screenshot annotations	**Experimenter** • Observed participants' interaction with the system using multimodal inputs. • Mentally noted how participants formulated search queries. • Did not take any annotations during observations.	**Experimenter: Wizard** • Emulated menu selections based on the user's head gestures. • Used FPV and TPV to recognize the gestures. • Marked head-gesture start time with a custom Counter. **Experimenter: Observer** • Assessed interactions as correct or incorrect from FPV. • Marked the correctness using annotations. • Added notes on usability issues after each trial.
Post-pilot phase	**Experimenter** • Conducted interviews using *Analyzer* • Utilized pre-captured annotations to ask questions • Added new notes • Exported PDF summary for later use	**Experimenter** • Marked the remembered interesting moments on *Analyzer* • Fast-forwarded through the video for new annotations and notes • Conducted interviews navigating through annotations • Updated notes based on participant responses • Showed recordings to participants for clarification • Exported results for team sharing	**Experimenter: Observer** • Verified and corrected annotations on *Analyzer* • Exported data in CSV format for analysis in Excel • Conducted interviews, noting insights directly into Analyzer • Exported data as PDF for documentation and collaboration

Fig. 6. Details of the three case studies, including team, system, task, sessions, and experimenter activities in the *during-pilot* and *post-pilot* phases.

All experimenters (T1-T3) and experts (E1-E2) expressed a positive outlook on the tool's design and features. They also proposed suggestions for the tool's future improvement. We observed that *PilotAR*'s all-in-one support capabilities, such as centralized views, recording, annotating, note-taking, and exporting, enable experimenters to conduct OHMD-based pilot studies more efficiently and gather quick insights for further exploration. Moreover, the system usability score, *SUS* [9] of $M = 76$, $SD = 3$ (T1: 73, T2: 78, T3: 78) indicates that *PilotAR* has 'Good' [4] usability supporting both single and multi-experimenter settings with *familiar mixed-fidelity wizarding/simulation interfaces*. Figure 7 shows the subjective ratings for *PilotAR*'s use in pilot studies on the selected wizarding interface, demonstrating its effectiveness in *simplifying the piloting process and reducing associated costs* such as setup time, analysis time/effort, results sharing efforts, and human resources.

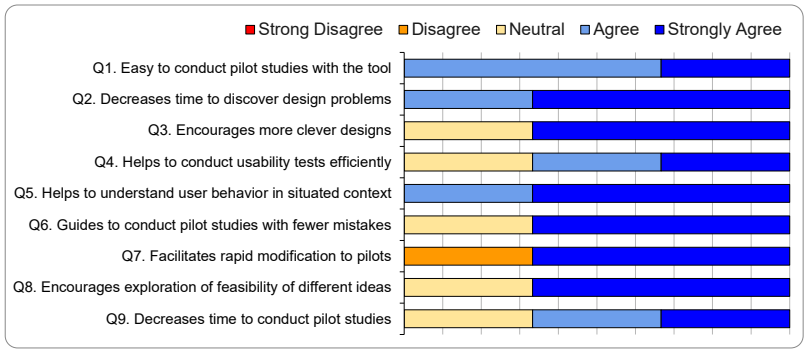


Fig. 7. Results: 100% stacked bar chart of three teams' ratings on "How much do you disagree or agree with the following statements about *PilotAR*?" (Q1-Q9). The questionnaire is derived from previous work on tool usage [27, 38]. Note: The disagree and neutral ratings were from T1 who used a high-fidelity wizarding application (Unity3D) during piloting, limiting them from quick modifications to pilot iterations.

5.3.1 Support for Observations in Situated Contexts. As expected, the combined FPV and TPV were essential and complementary —"*First-person view helps me to see how a user observes a virtual environment ... while the third-person view helps me to see the user's full-body movements in the physical world.* (T1)"

Notably, we identified a previously unnoticed trend: experimenters' usages for the two views are influenced by the experiment's configuration and task load. In a single-experimenter setting, the FPV became the primary focus during the piloting since it typically conveyed users' intentions and actions within the context of the displayed content and surroundings. This allowed for immediate responses to user behaviors. —"*I mainly rely on the first-person view to understand how the user navigates the virtual environment and interacts with it, such as pointing and manipulating digital entities.* (T1)" —"*The first-person view shows the interactions between users and the environment, and this is enough for my wizarding requirements.* (T3)"

On the other hand, the TPV was predominantly utilized during the post-pilot analysis. Experimenters opined that it presented a "*fresh perspective*" on the study, an angle that was typically less noticed during the actual study due to the pressing demands of multitasking —"*I already know everything going on from that [FPV] point of view. I would like to re-observe the whole event unfolding from another point of view [TPV] where I could potentially pick up more things. ... small moments where he could do subconscious actions, like gestures.* (T2)"

During the piloting with the multiple-experimenter setting, the observer's attention was mostly directed at the FPV, while the wizard balanced their focus between both FPV and TPV. This prioritization of the FPV was essential in aiding

experimenters in approximating interaction durations. During analysis, the observer's attention was primarily on the FPV, extracting insights on user interactions (e.g., menu selection for T3). The TPV subsequently provided an auxiliary perspective to better grasp usability issues and recollect participants' actions.

These different usages of different views resonate with our initial goal of facilitating simultaneous observation of virtual content via FPV and real-world interactions via TPV, fostering a holistic comprehension of the user-system interplay.

5.3.2 Reduce Task Load of Experimenters. Given experimenters' multiple responsibilities during a pilot—including wizarding, observing, and recording—facilitating multitasking emerged as an important goal for *PilotAR*. Our findings underscore that *PilotAR* has largely achieved this objective. Experimenters found it "*easier to conduct pilots*" and "*reduced time pressure*" owing to features like annotation shortcuts, automatic recording, and the convenience of revisiting content subsequently when necessary —"*Thanks to this [integrated view], I don't need to juggle multiple devices.* (T1)" —"*It's reassuring to know that every spoken word and every action is logged. This guarantees I can always revisit and assess them when needed.* (T2)"

Various teams employed distinct methodologies to enable simultaneous wizarding and observing. T1 utilized custom annotations to monitor system state changes and user inputs, and applied manual annotations to document unexpected behaviors. This approach was facilitated by the high-fidelity prototype, which enabled semi-automatic wizarding.

However, due to task load, it is not always possible to record in-situ real-time/instantaneous annotations. T2, for example, use an alternative strategy —"*[During the study, my focus is on asking questions [wizarding] and observing. I don't engage in immediate analysis. If an event stands out, I don't capture it right away; instead, I recall and note it during the review [analysis] phase.* (T2)"

Certainly, this limitation can be alleviated by adding more experimenters.T3, the team with multiple experimenters, didn't express concerns about task load since they delegated tasks among members to gather necessary observations and measures. Additionally, they leveraged custom annotations to automatically register the wizard's actions based on keyboard presses to measure time.

While *PilotAR* supports both instantaneous and retrospective annotations, the choice of annotation strategy depends on the session length. T1 and T2 indicated that for longer pilot sessions, exceeding 30 minutes, they prefer instantaneous annotation to avoid the time-consuming process of recalling all relevant events and reviewing lengthy recordings for retrospective annotations.

5.3.3 Expedite Data Recording, Analysis, and Generation of Creative Insights. The utility of the *Analyzer* was evident across all teams, as they all turned to it immediately after the study. Their feedback indicated that the *Analyzer* substantially accelerated data analysis and enhanced subsequent interviews by enabling them to identify, filter, annotate, and add user feedback to "*interesting*" moments and access quantitative results (e.g., the accuracy of interactions by T3) while the memory is fresh.

As T3 articulated the value of immediate analysis, "*[Observer] I appreciate having access to accuracy and incorrect instances immediately after the pilot. It allowed me to pinpoint areas of concern precisely [where the user performed the head gesture incorrectly] and seek clarification [on the reasons for inaccuracy]. This is invaluable for understanding the strengths and weaknesses of our technique, even before a full-fledged study [begins].*"

Furthermore, the *Annotations* displayed within the *Analyzer* proved essential in providing context during interviews, sparking more in-depth discussions. As noted by T2, during discrepancies between experimenters' observations and participants' perceptions, "*I can now show participants, along with the video and audio, what they did while saying this*

and ask why that was the case [to clarify the discrepancy]". However, experimenters used playback selectively, mainly to refresh participants' memories or highlight subconscious behaviors, aware that it might alter participants' subjective perceptions, which can differ from their objective behaviors.

This suite of features consistently led experimenters to discover new insights, allowing them to identify observations they might have previously overlooked. The testimony of T2 illustrates this: "*Before using Analyzer, I could only recall this gesture [posture of the statue] as noteworthy. That was the only interesting thing I noticed [as a novel way of using full-body posture for multimodal input]. After using Analyzer, I found two more interesting moments, such as the mental image he [the user] used to describe the plant artwork [imagining the toilet seat cover in his home with a similar plant]... and another instance when he abandoned the gesture and simply used one word to describe his search [believing the gesture was insufficient, even though it was adequate].*"

In summary, the annotated recordings played a crucial role in the success of *PilotAR*. They not only facilitated a more insightful pre-interview analysis but also enabled experimenters to focus their attention more effectively during the study —"*I could concentrate on conducting the experiments rather than immediately noting [down] interesting findings, knowing that I could do so [note interesting moments] later, as everything is recorded.* (T2)"

5.3.4 Usage of Exported Data Summaries. All experimenters recognized the value of *PilotAR*'s exporting capabilities, agreeing that it enhanced the efficient communication of findings to collaborators. This consensus stemmed from the tool's ability to share context-rich screenshots highlighting key study moments, which could guide future pilot study designs. As noted by T1, "*I found the visualization [PDF, e.g., Figure 5C1] very useful. It serves as a reference for comparison with upcoming pilot iterations.*"

Furthermore, T3 praised the tool's ability to export data in analysis-friendly formats such as CSV. This feature facilitated interviews that could delve into higher-level insights beyond just raw data. For example, *PilotAR* presents the calculation of average interaction durations of a user across trials and makes them immediately accessible for the experimenter after the pilot study. This allows experimenters to tailor interview questions more effectively based on user performance, which proved particularly useful for comparing interaction speeds between techniques and exploring the reasons behind observed differences.

Remote collaborators, such as those from T2, highlighted the benefits of exporting documentation and recordings that include rich context, such as FPV and TPV screenshots. This comprehensive perspective is particularly useful for remote team members who did not participate directly, as it fosters more insightful discussions by vividly re-creating the situated user interactions.

5.3.5 Collaborative Use of PilotAR. One of the notable findings from our interview data was the symbiotic relationship between the use of *PilotAR* and the engagement of additional experimenters. This synergy either allowed for a reduction in personnel without compromising the quality of observations or leveraged extra hands to enhance the depth of observations.

For T1, *PilotAR* could significantly reduce the need for an additional person. They highlighted the tool's ability to automate the quantification of specific user interactions, such as zooming in/out of a miniature view, adding new waypoints for a drone path, or testing the drone path, through custom annotations. However, T2 elucidated the broader capabilities of *PilotAR*, emphasizing that its potential was not only substitutional but also collaborative; as stated by T2, "*It is even more helpful than an additional experimenter. I don't believe another person can replace the functionality of this application [due to retrospective observation and annotation support]... With two skilled experimenters, one can provide instructions, and another can focus on taking [in-situ] screenshots and notes.*" Additionally, T3 highlighted the benefits

of including an additional experimenter for quantitative measures, as this helped to calculate interaction durations[28], reducing the reliance on high-fidelity prototypes during pilot sessions.

5.3.6 Expert Feedback. Both experts deemed *PilotAR* as "*very useful*" and expressed their desire to utilize it in their studies because it could "*generate insights faster*". E1 suggested the TPV could be enhanced by using an on-body 360 camera or a drone for mobile or varying view settings. E2 stated that *PilotAR* could "*undoubtedly make the current pilot studies much more informative, smooth, and interactive*". E2 elaborated that *PilotAR*'s remote monitoring capabilities could mitigate experimenter biases, such as the Hawthorne effect[29], by observing users remotely in real environments and measuring their responses using integrated views. This would create a "*new form*" of pilots, where users are not confined by strict study protocols intended to minimize biases/confounding factors. Instead, these biases could be measured as variables by observing them both in real-time [using TPV and FPV in *PilotAR*] and post-analysis [using *Analyzer* of *PilotAR*] across participants, leading to more informed decisions on whether user responses that are "*beyond the scope of the study procedure*" are valid.

6 GENERAL DISCUSSION

PilotAR has demonstrated its effectiveness by enabling experimenters to conduct pilots efficiently and rapidly gain insights, as evidenced by three case studies. This efficiency stems from its integrated observation views, multi-modal annotations, and support for swift pilot studies across prototypes of varying fidelity.

6.1 Situated Annotations with Pre-Interview Analysis Enhance Insight Generation Process

As expected, making annotations during pilots facilitated the filtering and selection of significant moments for post-analysis and interviews. Additionally, automated annotations linked to experimenter or user reactions alleviated the burden of manual annotation. Retrospective observations of under-observed viewpoints before interviews gave experimenters an auxiliary perspective on user reactions (Sec 5.3.1), enabling them to observe previously unnoticed behaviors more deeply with additional annotations. Such situated annotations, coupled with post-analysis before user interviews, enabled experimenters to pose contextually relevant questions to users and meticulously document their responses, thereby enhancing the understanding of user behaviors and interactions.

6.2 Integrating Situated Live and Retrospective Observations Improves Workload Distribution

Experimenters utilized the tool's immediate replay capabilities for situated observations and analysis, effectively reducing their instantaneous workload (Sec 5.3.1-5.3.2). This workload reduction was achieved through the spatial distribution of observed content, focusing on one view at a time, and the temporal distribution, which entailed shifting attention to less-observed pilot views during analysis. While prior work has demonstrated that spatial distribution helps reduce workload by allowing a focus on primary tasks [32], the insights from using *PilotAR* reveal that the temporal distribution of tasks further enables experimenters to prioritize critical tasks (e.g., wizarding) and allocate more cognitive resources (e.g., attention) to these tasks. This is done with the understanding that less immediate analyses can be performed retrospectively. Combined, these two strategies improve task management, reduce experimenter fatigue, and facilitate insight generation, though they require additional time for analysis.

[28] using *Annotations* time difference to the nearest second
[29] The phenomenon where participants in lab-based experiments may alter their behavior due to the awareness of being observed [35, Ch 2.5]

6.3 Trade-offs of Single vs. Multi-Experimenter Setup with *PilotAR*

The choice between a single and multi-experimenter setup with *PilotAR* depends on the specific demands of the study. A multi-experimenter setup is preferable for complex studies requiring simultaneous wizarding and detailed observations as it benefits from a division of labor, enabling comprehensive analysis with the expense of additional resources (e.g., manpower). Such a setup becomes indispensable when quick and frequent content manipulation/wizarding is required, precise timing measurements are essential, or the experimenter faces constraints on time for retrospective observations. However, a single-experimenter setup may suffice for studies with lesser workloads, especially given *PilotAR*'s capabilities for automating data capture and annotation, thereby reducing the need for additional personnel. Appropriate scenarios for a single-experimenter approach include less intensive wizarding tasks, employment of high-fidelity prototypes for automated content manipulation, or situations with limited trained manpower.

6.4 Cost-Benefit Trade-off of Using *PilotAR* in Pilot Studies

Although pilot studies are frequently associated with the "quick-and-dirty" approach—suggesting that both setup and results are expedited through less rigorous methods—this does not imply that the outcomes lack valuable insights. As emphasized in Sec 2.1, pilot studies must balance the resources (e.g., time, development effort) with the benefits of early insight. *PilotAR* supports this by facilitating the recording of detailed data and performing rigorous analyses to quickly gain early insights while minimizing effort (e.g., recording, filtering). These insights, while not directly usable in final reports due to the less rigorous methods employed, can indicate whether the main studies are likely to yield significant results or success [61]. Contrary to traditional approaches that may depend on quick-and-dirty analyses, *PilotAR* enables detailed analysis with quick-and-dirty setups, maximizing insight gains with reduced effort.

Inspired by Edward Tufte's concept of the data-ink ratio [62], we introduce the *insight-to-cost ratio* as a valuable concept for evaluating tools that support pilot studies. While a detailed quantification of costs and insights has not been precisely defined, they can be roughly assessed using subjective metrics[30]. Costs are quantified by the effort, time, and human resources required to conduct pilot studies and collect preliminary data, including setup and development costs. Insights are quantified by the information gathered to address research questions or test hypotheses, encompassing both holistic and specific data. By optimizing the *insight-to-cost ratio*, we can design and develop more effective tools for pilot studies.

PilotAR is designed to improve the *insight-to-cost ratio* by: 1) reducing the *costs* of setting up pilot studies through a guided process; 2) decreasing the *costs* of simulating AR/MR experiences by enabling seamless integration with existing presentation and simulation tools; 3) lowering experimentation *costs* through support for automation, shortcuts, and multi-experimenter collaboration; 4) enhancing *insight* generation by supporting detailed, multi-perspective monitoring of both the study process and outcomes; 5) improving *knowledge* discovery via quick data analysis and sharing capabilities. Our case studies have shown significant progress toward these goals, as evidenced by the qualitative feedback we have received.

6.5 What Kind of Studies Is *PilotAR* Best Suited For?

As mentioned in Sec 4.3, fully leveraging *PilotAR* in pilot studies, such as Wizard-of-Oz studies (see Sec 5), requires integrating it with a video feed (e.g., FPV, TPV) and a wizarding interface connected to an OHMD. This integration necessitates additional effort in setting up pilot studies and becoming acquainted with the *PilotAR* workflow. For

[30] Inspired by Chewar et al. [15] in defining Interruption Cost

instance, in simplistic AR/MR pilot studies where an FPV with an AR view is unnecessary or in complex pilot studies requiring precise objective measurements (e.g., parameter studies close to formal studies), *PilotAR* may not be the ideal choice due to either underutilization of its capabilities or insufficient functionality for the required analysis.

However, *PilotAR*'s utility extends beyond OHMD-based AR/MR pilot studies, encompassing extended observations and analyses in non-OHMD-based research, as demonstrated in three **three additional scenarios we observed**. In the first scenario, T2 employed *PilotAR* to observe a participant in real-world environments, such as unmanned retail spaces and parks, for 45 minutes. The participant's interactions were recorded from both first- and third-person views, with one experimenter annotating behaviors and another managing the camera. In the second scenario, two authors utilized *PilotAR* for capturing screenshots, annotating observations, and documenting interactions during pilot study sessions (see Section 5, with each session lasting 20-60 minutes) over several weeks. The *Analyzer* tool facilitated interviews by exploring unexpected events, showcasing the value of real-time annotations in longer studies (>20 minutes), in contrast to shorter studies where annotations were primarily made *post-pilot* phase (sec 5). This approach enabled deeper analytical insights, especially when no wizarding tasks were involved.

The third scenario involved T2 using *PilotAR* in pilot studies with high-fidelity prototypes, excluding live recording or real-time annotations. *PilotAR* was pivotal for the post-analysis of museum study recordings with six participants and hour-long sessions. It streamlined the analysis, synthesis, and dissemination of knowledge, comparable to established video analysis tools [33, 74]. *PilotAR* enhanced the interview process by facilitating questioning during specific frames/annotations review, marking important frames, and generating PDF tables for sharing insights in weekly remote team meetings. Furthermore, based on its technical implementation, *PilotAR* can be adapted for use with other AR/MR devices, such as video see-through (VST) displays (e.g., VST-HMD, smartphones, tablets), when streaming the AR/MR view as a video feed.

In summary, *PilotAR* serves multifaceted roles in research, functioning as an OHMD-based WOz study facilitator, a real-time observation support tool, and a standalone video analysis platform.

6.6 Supporting Study Replication and Fostering Creative Exploration

A crucial milestone that we hope *PilotAR* can help the research community achieve is facilitating study replication by enabling experimenters to preserve their study configurations and data, including video recordings and annotations. Other researchers, equipped with *PilotAR*, can leverage this archived data to replicate the study with new participants or to review the data for verification of results. This capability can enhance the replication and transparency of research [66]. Additionally, by integrating all phases of a pilot study—ranging from workflows and configurations to checklists—within a single tool, *PilotAR* ensures consistent quality in observation and analysis. Its support for varying fidelity levels in wizarding interfaces, collaborative experimentation, and sharing of contextual findings further promotes innovative exploration across multiple pilot study iterations.

6.7 Areas of Improvements

Although *PilotAR* received primarily positive feedback, several areas remain for enhancement: post-analysis, multi-setup, and measures.

6.7.1 Enhancing Virtual Content Display. In a particular session, a lag in the FPV relative to the TPV caused the experimenter to rely more on the TPV. This was due to network issues requiring a high-performance WiFi router to mitigate. Developing a dedicated OHMD application with reduced latency streaming can address these issues and

ensure compatibility with other OHMDs (e.g., Nreal Light, Magic Leap), and minimize potential data privacy concerns related to third-party tool usage (e.g., Google).

6.7.2 Enhancing Post-analysis. One team recommended checklists not only for the pre-pilot but also for the post-pilot phase to ensure consistent post-analysis. All teams noted that comparing various sessions can yield new "*insights*" and prompt further questions for participants. Integrating post-questionnaires into *PilotAR* and allowing the export of user responses alongside annotation notes can simplify subsequent statistical analysis. Another proposal involves audio recording interviews and utilizing AI tools, such as ChatGPT[31] to summarize them. Instead of exporting actions as static images, using short video snippets or animated images can foster better sharing and understanding among collaborators.

6.7.3 Accommodating Varied Setups. This encompasses support for mobile configurations via adaptable TPVs (e.g., drones, 360 cameras, multiple TPVs, body-attached cameras) and more compact devices like tablets[32]. While *PilotAR* currently supports two experimenters with one wizard and an observer, it should be expanded to include multiple observers. Enhancing remote monitoring capabilities to facilitate remote studies, as in [52], can address challenges like expert user recruitment and conducting studies when in-person interactions are challenging.

6.7.4 Enhancing Measuring Capabilities. The present limitations of *PilotAR* restrict its use in formal or pilot studies requiring precise quantitative recordings [37], such as sub-second-level time measurements. Such capabilities are currently tied to the wizarding interface (Sec 2.3). *PilotAR* offers a few quantitative metrics (e.g., time to the nearest second, time gap, accuracy, count) to aid experimenters in formulating interview questions, planning subsequent iterations, or identifying potential statistically significant outcomes in formal studies. One method to support precise measurements involves expanding *PilotAR* to incorporate more *Annotations* programmatically.

7 CONCLUSION

While tools exist to support studies, many current options do not adequately support observations and recordings in pilot studies. AR/MR experimenters find it especially challenging to filter out important moments for post-pilot discussions, as they must observe multiple viewpoints and manage extensive data. As OHMD-based AR/MR technology is poised to shape the future immersive world, including the metaverse, facilitating interactions between digital and physical entities becomes paramount. This underscores the importance of tools tailored for refining these interactions through pilot studies. As an initial step, we introduce *PilotAR*, an open-source tool (https://github.com/Synteraction-Lab/PilotAR) designed to support such studies. It enables real-time and retrospective multi-viewpoint observations, notes, and filters of crucial observations, thereby facilitating comprehensive discussions with participants and researchers to discover insights effectively. Additionally, it has the ability to share the pilot study process, data, and insights with the larger research community (e.g., OSF[33]). This capability can enhance the replication and transparency of research, but it requires community adoption. These enhancements can streamline the research process, promoting efficient data collection and analysis, and advancing OHMD-based AR/MR technologies. We believe integrating Artificial Intelligence (e.g., a virtual experimenter) can further enhance this tool, but such integration should be approached with care to address potential privacy and research integrity concerns. Such an upgrade would help pinpoint critical observations, summarize data, manage workloads, and enable researchers to focus more effectively on observation and analysis.

[31] https://chat.openai.com/
[32] While the current *PilotAR* supports Windows tablets, such as the Microsoft Surface Pro, it requires enhancements for touch interactions.
[33] https://osf.io/

ACKNOWLEDGMENTS

We would like to express our gratitude to the volunteers who participated in our studies (e.g., Interviews, Paperthon) and the Synteraction (formerly NUS-HCI) Lab members for their constructive feedback. We would also like to thank Tan Si Yan and Siddanth Ratan Umralkar for developing specific system components. Additionally, we wish to thank the anonymous reviewers for their valuable time and insightful comments, which helped improve this paper.

This research is supported by the National Research Foundation, Singapore, under its AI Singapore Programme (AISG Award No: AISG2-RP-2020-016). The CityU Start-up Grant 9610677 also provides partial support. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not reflect the views of the National Research Foundation, Singapore.

REFERENCES

[1] Günter Alce, Mattias Wallergård, and Klas Hermodsson. 2015. WozARd: a wizard of oz method for wearable augmented reality interaction—a pilot study. *Advances in Human-Computer Interaction* 2015 (June 2015). https://doi.org/10.1155/2015/271231

[2] Narges Ashtari, Andrea Bunt, Joanna McGrenere, Michael Nebeling, and Parmit K. Chilana. 2020. Creating Augmented and Virtual Reality Applications: Current Practices, Challenges, and Opportunities. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–13. https://doi.org/10.1145/3313831.3376722

[3] Ronald T. Azuma. [n. d.]. The road to ubiquitous consumer augmented reality systems. 1, 1 ([n. d.]), 26–32. https://doi.org/10.1002/hbe2.113

[4] Aaron Bangor, Philip T. Kortum, and James T. Miller. 2008. An Empirical Evaluation of the System Usability Scale. *International Journal of Human–Computer Interaction* 24, 6 (July 2008), 574–594. https://doi.org/10.1080/10447310802205776

[5] Andrea Bellucci, Telmo Zarraonandia, Paloma Díaz, and Ignacio Aedo. 2021. Welicit: A Wizard of Oz Tool for VR Elicitation Studies. In *Human-Computer Interaction – INTERACT 2021 (Lecture Notes in Computer Science)*, Carmelo Ardito, Rosa Lanzilotti, Alessio Malizia, Helen Petrie, Antonio Piccinno, Giuseppe Desolda, and Kori Inkpen (Eds.). Springer International Publishing, Cham, 82–91. https://doi.org/10.1007/978-3-030-85607-6_6

[6] Steve Benford, Andy Crabtree, Martin Flintham, Adam Drozd, Rob Anastasi, Mark Paxton, Nick Tandavanitj, Matt Adams, and Ju Row-Farr. 2006. Can you see me now? *ACM Transactions on Computer-Human Interaction* 13, 1 (March 2006), 100–133. https://doi.org/10.1145/1143518.1143522

[7] W. I. B. Beveridge. 2020. *The art of scientific investigation.*

[8] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (Jan. 2006), 77–101. https://doi.org/10.1191/1478088706qp063oa

[9] John Brooke. 1996. SUS - A quick and dirty usability scale. *Usability evaluation in industry* 189, 194 (1996), 7.

[10] Jack Brookes, Matthew Warburton, Mshari Alghadier, Mark Mon-Williams, and Faisal Mushtaq. 2020. Studying human behavior with virtual reality: The Unity Experiment Framework. *Behavior Research Methods* 52, 2 (April 2020), 455–463. https://doi.org/10.3758/s13428-019-01242-0

[11] Frederik Brudy, Suppachai Suwanwatcharachat, Wenyu Zhang, Steven Houben, and Nicolai Marquardt. 2018. EagleView: A Video Analysis Tool for Visualising and Querying Spatial Interactions of People and Devices. In *Proceedings of the 2018 ACM International Conference on Interactive Surfaces and Spaces (ISS '18)*. Association for Computing Machinery, New York, NY, USA, 61–72. https://doi.org/10.1145/3279778.3279795

[12] Wolfgang Büschel, Anke Lehmann, and Raimund Dachselt. 2021. MIRIA: A Mixed Reality Toolkit for the In-Situ Visualization and Analysis of Spatio-Temporal Interaction Data. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3411764.3445651

[13] Scott Carter, Jennifer Mankoff, and Jeffrey Heer. 2007. Momento: support for situated ubicomp experimentation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '07)*. Association for Computing Machinery, New York, NY, USA, 125–134. https://doi.org/10.1145/1240624.1240644

[14] Kevin Chen and Haoqi Zhang. 2015. Remote Paper Prototype Testing. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. Association for Computing Machinery, New York, NY, USA, 77–80. https://doi.org/10.1145/2702123.2702423

[15] C. M. Chewar, D. Scott McCrickard, and Alistair G. Sutcliffe. 2004. Unpacking critical parameters for interface design: evaluating notification systems with the IRC framework. In *Proceedings of the 5th conference on Designing interactive systems: processes, practices, methods, and techniques (DIS '04)*. Association for Computing Machinery, New York, NY, USA, 279–288. https://doi.org/10.1145/1013115.1013155

[16] Hyunsung Cho, Matthew L. Komar, and David Lindlbauer. 2023. RealityReplay: Detecting and Replaying Temporal Changes In Situ Using Mixed Reality. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 7, 3 (Sept. 2023), 90:1–90:25. https://doi.org/10.1145/3610888

[17] N. Dahlbäck, A. Jönsson, and L. Ahrenberg. 1993. Wizard of Oz studies — why and how. *Knowledge-Based Systems* 6, 4 (Dec. 1993), 258–266. https://doi.org/10.1016/0950-7051(93)90017-N

[18] Marco de Sá and Elizabeth Churchill. 2012. Mobile augmented reality: exploring design and prototyping techniques. In *Proceedings of the 14th international conference on Human-computer interaction with mobile devices and services (MobileHCI '12)*. Association for Computing Machinery,

New York, NY, USA, 221–230. https://doi.org/10.1145/2371574.2371608

[19] Saul Delabrida, Thiago D'Angelo, and Ricardo A. Rabelo Oliveira. 2015. Fast prototyping of an AR HUD based on google cardboard API. In *Adjunct Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2015 ACM International Symposium on Wearable Computers (UbiComp/ISWC'15 Adjunct)*. Association for Computing Machinery, New York, NY, USA, 1303–1306. https://doi.org/10.1145/2800835.2807928

[20] Arindam Dey, Mark Billinghurst, Robert W. Lindeman, and J. Edward Swan. 2018. A Systematic Review of 10 Years of Augmented Reality Usability Studies: 2005 to 2014. *Frontiers in Robotics and AI* 5 (2018). https://doi.org/10.3389/frobt.2018.00037

[21] Steven Dow, Jaemin Lee, Christopher Oezbek, Blair MacIntyre, Jay David Bolter, and Maribeth Gandy. 2005. Wizard of Oz interfaces for mixed reality applications. In *CHI '05 Extended Abstracts on Human Factors in Computing Systems (CHI EA '05)*. Association for Computing Machinery, New York, NY, USA, 1339–1342. https://doi.org/10.1145/1056808.1056911

[22] S. Dow, B. MacIntyre, J. Lee, C. Oezbek, J.D. Bolter, and M. Gandy. 2005. Wizard of Oz support throughout an iterative design process. *IEEE Pervasive Computing* 4, 4 (Oct. 2005), 18–26. https://doi.org/10.1109/MPRV.2005.93 Conference Name: IEEE Pervasive Computing.

[23] Gabriel Freitas, Marcio Sarroglia Pinho, Milene Selbach Silveira, and Frank Maurer. 2020. A Systematic Review of Rapid Prototyping Tools for Augmented Reality. In *2020 22nd Symposium on Virtual and Augmented Reality (SVR)*. 199–209. https://doi.org/10.1109/SVR51698.2020.00041

[24] Maribeth Gandy and Blair MacIntyre. 2014. Designer's augmented reality toolkit, ten years later: implications for new media authoring tools. In *Proceedings of the 27th annual ACM symposium on User interface software and technology (UIST '14)*. Association for Computing Machinery, New York, NY, USA, 627–636. https://doi.org/10.1145/2642918.2647369

[25] Uwe Gruenefeld, Jonas Auda, Florian Mathis, Stefan Schneegass, Mohamed Khamis, Jan Gugenheimer, and Sven Mayer. 2022. VRception: Rapid Prototyping of Cross-Reality Systems in Virtual Reality. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3491102.3501821

[26] Anhong Guo, Ilter Canberk, Hannah Murphy, Andrés Monroy-Hernández, and Rajan Vaish. 2019. Blocks: Collaborative and Persistent Augmented Reality Experiences. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 3 (Sept. 2019), 83:1–83:24. https://doi.org/10.1145/3351241

[27] Björn Hartmann, Leith Abdulla, Manas Mittal, and Scott R. Klemmer. 2007. Authoring sensor-based interactions by demonstration with direct manipulation and pattern recognition. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '07)*. Association for Computing Machinery, New York, NY, USA, 145–154. https://doi.org/10.1145/1240624.1240646

[28] Melody A. Hertzog. 2008. Considerations in determining sample size for pilot studies. *Research in Nursing & Health* 31, 2 (2008), 180–191. https://doi.org/10.1002/nur.20247

[29] Sebastian Hubenschmid, Jonathan Wieland, Daniel Immanuel Fink, Andrea Batch, Johannes Zagermann, Niklas Elmqvist, and Harald Reiterer. 2022. ReLive: Bridging In-Situ and Ex-Situ Visual Analytics for Analyzing Mixed Reality User Studies. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. Association for Computing Machinery, New York, NY, USA, 1–20. https://doi.org/10.1145/3491102.3517550

[30] Paul Israel. 2000. *Edison: A Life of Invention* (first edition ed.). John Wiley & Sons, New York, NY.

[31] Yuta Itoh, Tobias Langlotz, Jonathan Sutton, and Alexander Plopski. 2021. Towards Indistinguishable Augmented Reality: A Survey on Optical See-through Head-mounted Displays. *Comput. Surveys* 54, 6 (July 2021), 120:1–120:36. https://doi.org/10.1145/3453157

[32] Youn-ah Kang and John Stasko. 2008. Lightweight task/application performance using single versus multiple monitors: a comparative study. In *Proceedings of Graphics Interface 2008 (GI '08)*. Canadian Information Processing Society, CAN, 17–24. https://dl.acm.org/doi/10.5555/1375714.1375718

[33] Michael Kipp. 2014. ANVIL: The Video Annotation Research Tool. In *The Oxford Handbook of Corpus Phonology*, Jacques Durand, Ulrike Gut, and Gjert Kristoffersen (Eds.). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199571932.013.024

[34] Veronika Krauß, Alexander Boden, Leif Oppermann, and René Reiners. 2021. Current Practices, Challenges, and Design Implications for Collaborative AR/VR Application Development. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3411764.3445335

[35] Jonathan Lazar. 2017. *Research methods in human computer interaction* (2nd edition ed.). Elsevier, Cambridge, MA.

[36] David Ledo, Steven Houben, Jo Vermeulen, Nicolai Marquardt, Lora Oehlberg, and Saul Greenberg. 2018. Evaluation Strategies for HCI Toolkit Research. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–17. https://doi.org/10.1145/3173574.3173610

[37] Minkyung Lee and Mark Billinghurst. 2008. A Wizard of Oz study for an AR multimodal interface. In *Proceedings of the 10th international conference on Multimodal interfaces (ICMI '08)*. Association for Computing Machinery, New York, NY, USA, 249–256. https://doi.org/10.1145/1452392.1452444

[38] Germán Leiva, Cuong Nguyen, Rubaiat Habib Kazi, and Paul Asente. 2020. Pronto: Rapid Augmented Reality Video Prototyping Using Sketches and Enaction. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376160

[39] Youn-Kyung Lim, Erik Stolterman, and Josh Tenenberg. 2008. The anatomy of prototypes: Prototypes as filters, prototypes as manifestations of design ideas. *ACM Transactions on Computer-Human Interaction* 15, 2 (July 2008), 7:1–7:27. https://doi.org/10.1145/1375761.1375762

[40] Blair MacIntyre, Maribeth Gandy, Steven Dow, and Jay David Bolter. 2004. DART: a toolkit for rapid design exploration of augmented reality experiences. In *Proceedings of the 17th annual ACM symposium on User interface software and technology (UIST '04)*. Association for Computing Machinery, New York, NY, USA, 197–206. https://doi.org/10.1145/1029632.1029669

[41] I. Scott MacKenzie. 2013. *Human-computer interaction: an empirical research perspective* (first edition ed.). Morgan Kaufmann is an imprint of Elsevier, Amsterdam.

[42] Nicolai Marquardt, Frederico Schardong, and Anthony Tang. 2015. EXCITE: EXploring Collaborative Interaction in Tracked Environments. In *Human-Computer Interaction – INTERACT 2015 (Lecture Notes in Computer Science)*, Julio Abascal, Simone Barbosa, Mirko Fetter, Tom Gross, Philippe Palanque, and Marco Winckler (Eds.). Springer International Publishing, Cham, 89–97. https://doi.org/10.1007/978-3-319-22668-2_8

[43] Florian Müller, Joshua McManus, Sebastian Günther, Martin Schmitz, Max Mühlhäuser, and Markus Funk. 2019. Mind the Tap: Assessing Foot-Taps for Interacting with Head-Mounted Displays. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300707

[44] Leon Müller, Ken Pfeuffer, Jan Gugenheimer, Bastian Pfleging, Sarah Prange, and Florian Alt. 2021. SpatialProto: Exploring Real-World Motion Captures for Rapid Prototyping of Interactive Mixed Reality. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3411764.3445560

[45] Michael Nebeling and Katy Madier. 2019. 360proto: Making Interactive Virtual Reality & Augmented Reality Prototypes from Paper. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300826

[46] Michael Nebeling, Shwetha Rajaram, Liwei Wu, Yifei Cheng, and Jaylin Herskovitz. 2021. XRStudio: A Virtual Production and Live Streaming System for Immersive Instructional Experiences. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3411764.3445323

[47] Michael Nebeling and Maximilian Speicher. 2018. The Trouble with Augmented Reality/Virtual Reality Authoring Tools. In *2018 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*. 333–337. https://doi.org/10.1109/ISMAR-Adjunct.2018.00098

[48] Michael Nebeling, Maximilian Speicher, Xizi Wang, Shwetha Rajaram, Brian D. Hall, Zijian Xie, Alexander R. E. Raistrick, Michelle Aebersold, Edward G. Happ, Jiayin Wang, Yanan Sun, Lotus Zhang, Leah E. Ramsier, and Rhea Kulkarni. 2020. MRAT: The Mixed Reality Analytics Toolkit. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376330

[49] A. Ant Ozok. 2012. Survey Design and Implementation in HCI. In *Human Computer Interaction Handbook* (3 ed.). CRC Press. https://doi.org/10.1201/b11963

[50] Michael Prilla and Lisa M. Rühmann. 2018. An Analysis Tool for Cooperative Mixed Reality Scenarios. In *2018 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*. 31–35. https://doi.org/10.1109/ISMAR-Adjunct.2018.00026

[51] Thibault Raffaillac and Stéphane Huot. 2022. What do Researchers Need when Implementing Novel Interaction Techniques? *Proceedings of the ACM on Human-Computer Interaction* 6, EICS (June 2022), 159:1–159:30. https://doi.org/10.1145/3532209

[52] Jack Ratcliffe, Francesco Soave, Nick Bryan-Kinns, Laurissa Tokarchuk, and Ildar Farkhatdinov. 2021. Extended Reality (XR) Remote Research: a Survey of Drawbacks and Opportunities. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3411764.3445170

[53] Alejandro Rey, Andrea Bellucci, Paloma Diaz, and Ignacio Aedo. 2021. A Tool for Monitoring and Controlling Standalone Immersive HCI Experiments. In *The Adjunct Publication of the 34th Annual ACM Symposium on User Interface Software and Technology (UIST '21)*. Association for Computing Machinery, New York, NY, USA, 20–22. https://doi.org/10.1145/3474349.3480217

[54] Alejandro Rey Lopez, Andrea Bellucci, Paloma Diaz Perez, and Ignacio Aedo Cuevas. 2022. IXCI:The Immersive eXperimenter Control Interface. In *Proceedings of the 2022 International Conference on Advanced Visual Interfaces (AVI 2022)*. Association for Computing Machinery, New York, NY, USA, 1–3. https://doi.org/10.1145/3531073.3534489

[55] Maria Rosala. 2020. The Critical Incident Technique in UX. https://www.nngroup.com/articles/critical-incident-technique/

[56] Maximilian Speicher, Brian D. Hall, and Michael Nebeling. 2019. What is Mixed Reality?. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3290605.3300767

[57] Maximilian Speicher, Brian D. Hall, Ao Yu, Bowen Zhang, Haihua Zhang, Janet Nebeling, and Michael Nebeling. 2018. XD-AR: Challenges and Opportunities in Cross-Device Augmented Reality Application Development. *Proceedings of the ACM on Human-Computer Interaction* 2, EICS (June 2018), 7:1–7:24. https://doi.org/10.1145/3229089

[58] Yuki Sugita, Keita Higuchi, Ryo Yonetani, Rie Kamikubo, and Yoichi Sato. 2018. Browsing Group First-Person Videos with 3D Visualization. In *Proceedings of the 2018 ACM International Conference on Interactive Surfaces and Spaces (ISS '18)*. Association for Computing Machinery, New York, NY, USA, 55–60. https://doi.org/10.1145/3279778.3279783

[59] Lehana Thabane, Jinhui Ma, Rong Chu, Ji Cheng, Afisi Ismaila, Lorena P. Rios, Reid Robson, Marroon Thabane, Lora Giangregorio, and Charles H. Goldsmith. 2010. A tutorial on pilot studies: the what, why and how. *BMC Medical Research Methodology* 10, 1 (Jan. 2010), 1. https://doi.org/10.1186/1471-2288-10-1

[60] Stefan H. Thomke. 2003. *Experimentation Matters: Unlocking the Potential of New Technologies for Innovation.* Harvard Business Review Press, Boston, Mass.

[61] Khai Truong. 2017. Pilot Studies: When and how to conduct them when conducting user studies. *GetMobile: Mobile Computing and Communications* 20, 4 (April 2017), 8–11. https://doi.org/10.1145/3081016.3081020

[62] Edward R. Tufte. 2013. *The Visual Display of Quantitative Information.* Cheshire, Conn.

[63] Edwin R. van Teijlingen and Vanora Hundley. 2001. The importance of pilot studies. (2001). https://aura.abdn.ac.uk/handle/2164/157

[64] Edwin R. Van Teijlingen, Anne-Marie Rennie, Vanora Hundley, and Wendy Graham. 2001. The importance of conducting and reporting pilot studies: the example of the Scottish Births Survey. *Journal of Advanced Nursing* 34, 3 (2001), 289–295. https://doi.org/10.1046/j.1365-2648.2001.01757.x

[65] Christian David Vazquez, Afika Ayanda Nyati, Alexander Luh, Megan Fu, Takako Aikawa, and Pattie Maes. 2017. Serendipitous Language Learning in Mixed Reality. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems - CHI EA '17*. ACM Press, Denver, Colorado, USA, 2172–2179. https://doi.org/10.1145/3027063.3053098 00000.

[66] Chat Wacharamanotham, Lukas Eisenring, Steve Haroz, and Florian Echtler. 2020. Transparency of CHI Research Artifacts: Results of a Self-Reported Survey. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–14. https://doi.org/10.1145/3313831.3376448

[67] Maurice Waite. 2002. *Concise oxford thesaurus in clair A-Z from.* Oxford University Press. https://pks.pua.edu.eg/social_sciences_books/2320

[68] Tianyi Wang, Xun Qian, Fengming He, Xiyun Hu, Yuanzhi Cao, and Karthik Ramani. 2021. GesturAR: An Authoring System for Creating Freehand Interactive Augmented Reality Applications. In *The 34th Annual ACM Symposium on User Interface Software and Technology (UIST '21)*. Association for Computing Machinery, New York, NY, USA, 552–567. https://doi.org/10.1145/3472749.3474769

[69] Mark Weiser. 1991. The Computer for the 21 st Century. *Scientific American* 265, 3 (1991), 13. https://doi.org/10.1145/329124.329126

[70] Xuhai Xu, Anna Yu, Tanya R. Jonker, Kashyap Todi, Feiyu Lu, Xun Qian, João Marcelo Evangelista Belo, Tianyi Wang, Michelle Li, Aran Mun, Te-Yen Wu, Junxiao Shen, Ting Zhang, Narine Kokhlikyan, Fulton Wang, Paul Sorenson, Sophie Kim, and Hrvoje Benko. 2023. XAIR: A Framework of Explainable AI in Augmented Reality. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–30. https://doi.org/10.1145/3544548.3581500

[71] Hui Ye and Hongbo Fu. 2022. ProGesAR: Mobile AR Prototyping for Proxemic and Gestural Interactions with Real-world IoT Enhanced Spaces. In *CHI Conference on Human Factors in Computing Systems (CHI '22)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3491102.3517689

[72] Shengdong Zhao, Felicia Tan, and Katherine Fennedy. 2023. Heads-Up Computing Moving Beyond the Device-Centered Paradigm. *Commun. ACM* 66, 9 (Aug. 2023), 56–63. https://doi.org/10.1145/3571722

[73] John Zimmerman, Jodi Forlizzi, and Shelley Evenson. 2007. Research through design as a method for interaction design research in HCI. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '07)*. Association for Computing Machinery, New York, NY, USA, 493–502. https://doi.org/10.1145/1240624.1240704

[74] Patrick H. Zimmerman, J. Elizabeth Bolhuis, Albert Willemsen, Erik S. Meyer, and Lucas P. J. J. Noldus. 2009. The Observer XT: A tool for the integration and synchronization of multimodal signals. *Behavior Research Methods* 41, 3 (Aug. 2009), 731–735. https://doi.org/10.3758/BRM.41.3.731

A TOOL COMPARISON

Table 1 compares the high-level **features** of *PilotAR* with other AR/MR tools.

B *STUDY 1*

B.1 Demographics of Participants

Table 2 shows the background of the researchers in *study 1*.

B.2 Analysis

One coauthor undertook the thematic analysis of the interview transcripts and observation notes, adhering to the guidelines established by Braun and Clarke [8]. This analytical process was multi-faceted and consisted of several stages.

Initially, the coauthor familiarized themselves with a section of the data (comprising four transcription files and corresponding observation notes), from which they derived preliminary codes encapsulating the key concepts. These initial codes were then reviewed and discussed with a second coauthor to resolve any discrepancies or conflicts before applying them to the remainder of the data (an additional eight transcription files with observation notes).

Following this, the coauthor grouped these codes into common themes, using their content as the basis for categorization. To guarantee the validity of the analysis, the two coauthors worked together to discuss, interpret, and rectify any discrepancies or conflicts during the theme-grouping process.

Table 1. Summary of the feature comparisons between the tools for conducting AR-related studies. Here, FPV = first-person view, TPV = third-person view, AR = virtual content view. Note: This list is not exhaustive. Although DART [24, 40] is meant for authoring AR/MR content, we have added it here for comparison as it supports various functions that could also be used for conducting AR/MR experiments.

Tool/Toolkit	Lee et al. [37]	Rey et al. [53, 54] (IXCI)	MacIntyre et al. [24, 40] (DART)	Nebeling et al. [48] (MRAT)	Proposed tool (***PilotAR***)
Purpose	Identify multimodal inputs for AR manipulation tasks and how AR display conditions affect them	Support research by streamlining immersive user studies	An authoring tool enabling rapid prototyping of AR applications by designers/non-technologists	An experimenter support tool for analyzing MR experiences	An experimenter support tool for conducting AR/MR **pilots**, data collection, and analysis
Target studies	WOz studies	Unity3D-based studies	AR studies	Unity3D-based studies	Pilot studies in AR/MR, including WOz
Prototype fidelity	High	High	Low-High	High	Low-High
Multiple experiment support	Single	Multiple	Multiple	Multiple	Multiple
Observation support	FPV, TPV	AR	FPV, AR, TPV	Interaction datapoints	FPV with AR, TPV
Recording support	✓	✗	✓	Processed spatial-temporal interaction data points	✓
Note taking	✗	✗	✗	✗	✓
Post-analysis	✗	✗	Not applicable	✓	✓
Summarizing and exporting	✗	✗	Not applicable	✓	✓

The final stage involved a thorough review of the transcripts and audio recordings. Specific quotes relevant to each identified theme were extracted to provide more context and enrich the analysis.

C ITERATIVE DESIGN OF THE *PILOTAR*

While the core concepts such as enabling multiple views, screen recording, annotations, and summarizing persisted throughout each iteration of the tool, each feature continued to be refined and extended as we progressed in each iteration—as shown in Figure 8 and Table 3—while addressing the design goals detailed in Sec 3.6. Here, we describe our research-through-(tool)-design process [73]. Four AR researchers with over two years of experience in AR/MR research involving OHMDs were selected as experimenters.

The initial *PilotAR* prototype (Figure 8a) was crafted using readily available tools. Following formative testing, we further refined this to an enhanced version (Figure 8b). This version underwent further testing and refinement, culminating in the final design (Figure 3-5) detailed earlier in Sec 4.

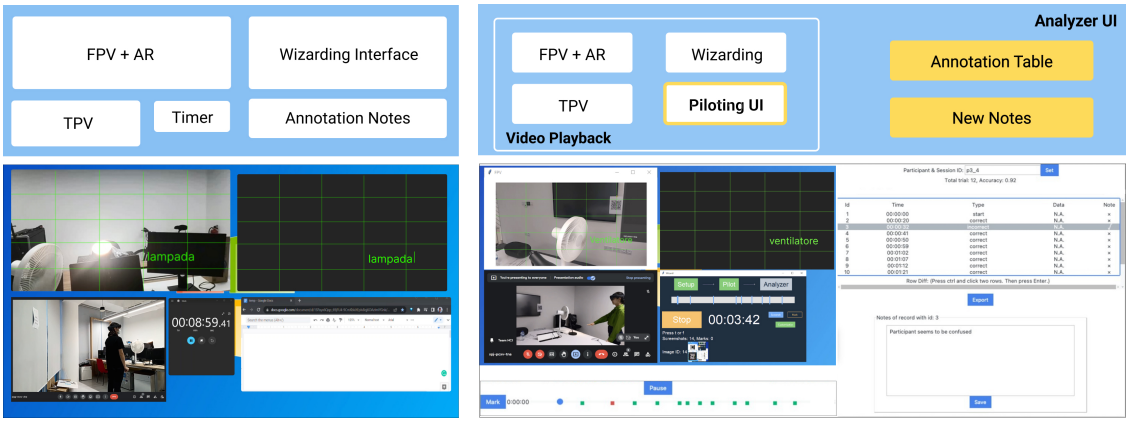
Task and Procedure. We chose a pilot study task that required design space exploration and usability issue recognition, common in AR/MR pilot studies (Sec 3.1). We employed a wizard-of-oz approach, typically used in early pilot studies,

Table 2. The background of the AR/MR researchers interviewed in *study 1*. Note: ⋆ The fidelity of the prototyping tools varied depending on familiarity and the project stage. For instance, early pilot studies of R1 often employed low-fidelity tools like Google Slides, whereas later stages used high-fidelity tools such as Unity3D.

ID	Occupation	Experience (years)	AR/MR Research projects	AR/MR platforms	Prototyping Tools⋆
R1	Professor	10	Perception (Dementia eyes), Sports spectating, Learning, Navigation	HMD (Magic Leap, Vive Pro, Google Cardboard), Phone	Unity3D, Unreal, Figma, Google Slides, Miro, Paper
R2	Postdoc	4	Video learning, Video adaptation, Mental health (Mindfulness), Gesture interactions	HMD (Nreal Light, BT-300, Vuzix Blade, HoloLens2)	iMovie, Adobe Premier, Keynote, HTML+JS
R3	Postdoc	2.5	Idea generation, Writing, Text presentation	HMD (Nreal Light)	Google Doc, Miro, Zoom
R4	Postdoc	3.5	Memory aids, Mental health (relaxation), Decluttering	HMD (Magic Leap, HoloLens, HoloLens 2, Epson), Phone	Unity3D, Miro, Android
R5	Postdoc	2.5	Text editing, Measurement, Voice-based AR assistant	HMD (Vuzix Blade, BT-300), Phone	Android, HTML+JS, Paper
R6	Industry researcher	3	Assembly guidance, AI assistant	HMD (Nreal Light, HoloLens2, BT-300), Tablet (iPad), Phone	Unity3D
R7	PhD student (5yr.)	4	Assembly guidance, Augmenting TV	HMD (HoloLens2, BT-300), Tablet (iPad), Phone	PowerPoint, HTML+JS, Android
R8	PhD student (4yr.)	2.5	Display news, Building architecture	HMD (HoloLens2), Phone	Fologram, Rhino 3D, Figma, Paper
R9	PhD student (1yr.)	2.5	IoT manipulation, Fire disaster management, AI-based text editing	HMD (Nreal Light, HoloLens2)	Unity3D, Protopie, Figma, Google Meet
R10	PhD student (2yr.)	3.5	Drone control, Multi-modal searching, Dynamic text displays, Gaze interactions	HMD (HoloLens2, Nreal Light), Phone	Unity3D, Google Slides, Zoom, Photoshop, Pygame, Android, Paper
R11	Research engineer	2	Mental health (mindfulness)	HMD (Nreal Light)	iMovie, Keynote, HTML+JS
R12	Master student	2	Text presentation, Multitasking	HMD (HoloLens2, Nreal Light)	Unity3D, PowerPoint, Python, Paper

to emulate AR systems (Sec 3.3). Therefore, we designed a contextual language task inspired by Serendipitous Learning [65]. Here, the AR/MR experimenter (the target participant) acted as the wizard to simulate the AR system, while a user functioned as the language learner. This task, as shown in Figure 9, enabled experimenters to not only identify the optimal modality for user object selection and feedback (i.e., interaction design space exploration) but also to evaluate the advantages or limitations of using OHMDs for serendipitous learning tasks (i.e., usability issue recognition).

Experimenter actions were recorded, and subsequent interviews were conducted to understand tool usage, design challenges, and potential enhancements.



(a) First iteration of *PilotAR*. (b) Second iteration of *PilotAR*.

Fig. 8. The first and second iterations of *PilotAR*, where the top figures represent the layout of UI elements while the bottom figures represent the actual UI implementations. (a) The first iteration of UI for *PilotAR* using third-party commodity software (e.g., MS Timer, Google Doc, Google Meet, Miro) representing the Piloting UI (i.e., the UI of *PilotAR* during the pilot study). (b) The second iteration of *PilotAR* represents the Analyzer UI (i.e., UI of *PilotAR* during the analysis and post-interview phase) implemented using Python and commodity third-party software. The Piloting UI is shown inside the video playback of the Analyzer UI.

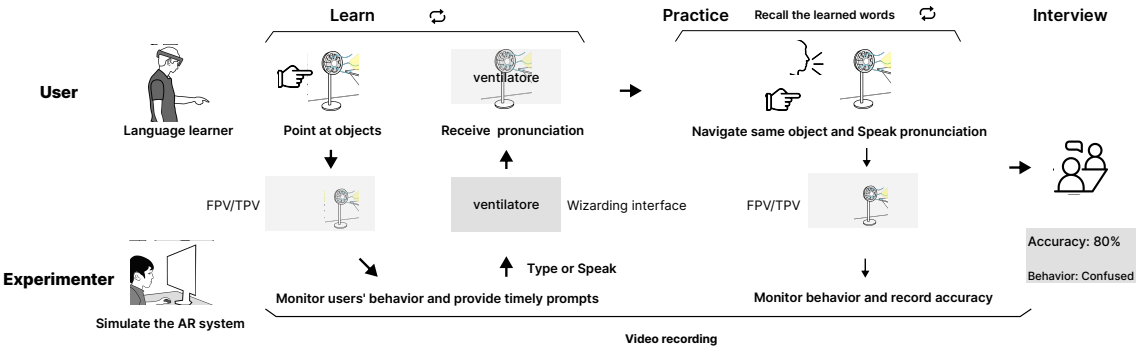


Fig. 9. The user and experimenter tasks during tool iterations. The user learns foreign vocabulary by pointing at laboratory objects, receiving their foreign language pronunciation, and later recalling the learned words while interacting with the same objects. As wizards, the experimenters monitor user behavior, provide timely prompts (e.g., pronunciation of pointed objects), and identify usability issues.

C.1 First iteration: Findings

Although the use of off-the-shelf software provided some assistance during the pilot, the formative study participants (i.e., experimenters) expressed several usability concerns regarding this approach, detailed in Table 4, emphasizing the necessity of a dedicated tool.

The disconnection between multiple UIs was solved by streamlining them into a workflow enabling *Ease of Setup in Conducting Pilot Studies* (***D0***). Unsynchronized views, crucial to *Support Observations in Situated Contexts* (***D1***), were mitigated by dedicated video players. To *Reduce Task Load of Experimenters* (***D2***), the disconnection between observations (e.g., screenshots) and notes was rectified by linking them. Live analysis (e.g., accuracy) via shortcuts

Table 3. Iterations of the *PilotAR*, covering the implementation, new features, and findings from formative testing. Here, ***D0***, ***D1***, ***D2***, and ***D3*** represent *Ease of Setup in Conducting Pilot Studies* (***D0***), *Support Observations in Situated Contexts* (***D1***), *Reduce Task Load of Experimenters* (***D2***), and *Expedite Data Recording, Analysis, and Generation of Creative Insights* (***D3***), respectively.

	First (Figure 8a)	**Second** (Figure 8b)	**Final** (Figure 3-5)
		Implementation	
D0	• Utilized third-party tools • Single device setup	• Employed Python along with third-party components • Added workflow support	• Employed Python along with third-party components • Developed a unified GUI • Improved system feedback
		New Features	
D1	• Enabled TPV, FPV, AR	• Incorporated a dedicated player	• Added device configurations
D2	• Provided a customizable layout • Added screenshot functionality • Added note-taking functionality • Added timer support	• Enabled live analysis with shortcuts • Linked notes with screenshots • Enable targeted snapshots • Previewed screenshots	• Added support for multiple experimenters • Provided customizable annotations
D3	• Enabled tool screen recording • Enabled observation note editing	• Linked notes and screenshots with recordings • Facilitated additional note-taking during interviews • Enabled exporting of summary notes	• Implemented audio transcription functionality • Enabled filtering and highlighting of annotations • Allowed export of selected annotations in both PDF and CSV formats
		Formative Testing and Findings	
	• 4 AR researchers	• 4 AR researchers	• Sec 5.1- 5.2
D1	• Latency in FPV causes unsynchronized view	• More configurations are needed to manage third-party components	• Sec 5.3, Sec 6.7
D2	• Separated notes and screenshots increase post-analysis time • Lack of highlighting makes it hard to identify the interested observation quickly • Difficulty in recording the accuracy of user responses	• When wizarding, high task load makes adding fine-grain observation details challenging • Annotations are hard to customize and modify	• Sec 5.3, Sec 6.7
D3	• Difficulty in navigating recorded video due to manual timestamp searching • Additional effort is required to take screenshots with notes and summarize them	• Insufficient indicators of communication between users and experimenters • Insufficient support for easy navigation and filtering through annotated moments during analysis	• Sec 5.3, Sec 6.7

simplified the recording of user responses. Similarly, to *Expedite Data Recording, Analysis, and Generation of Creative Insights* (***D3***), difficulties in navigating the recorded video were mitigated by linking annotations (e.g., screenshots, notes) to the video, allowing direct navigation through timestamps. Additionally, a summary was automatically generated to help experimenters to better focus during interviews.

Table 4. Concerns and solutions for the **first** iteration (Figure 8a). Here, ***D0***, ***D1***, ***D2***, and ***D3*** represent *Ease of Setup in Conducting Pilot Studies* (***D0***), *Support Observations in Situated Contexts* (***D1***), *Reduce Task Load of Experimenters* (***D2***), and *Expedite Data Recording, Analysis, and Generation of Creative Insights* (***D3***), respectively.

	Issue Description	Design Solution & Features
General ***D0***	*Difficulty in setting up separate components* - Since individual UI components are not interconnected, each one needs to be operated separately (e.g., multiple software applications need to be opened), leading to the possibility of forgetting to enable certain functions (e.g., screen recording) due to the numerous operations involved.	The tool consolidates all UI components into a workflow that guides users through each key step (such as setting up, conducting a pilot, and analyzing observations). Each UI component can be accessed from the workflow control panel.
D1	*Latency in FPV* - The more than 2-second delay in accessing FPV via WDP causes an unsynchronized FPV and TPV, making it challenging to infer the user's intentions during wizarding.	The tool's player integration of FPV using a streaming API reduces the latency to less than 1 second.
D2	*Disconnection between notes and screenshots* - Experimenters had to manually link notes with screenshots because they were not automatically connected, making post-pilot analysis time-consuming.	The tool enables notes to be directly attached to screenshots, ensuring their linkage.
	Inability to highlight specific parts of screenshots - Although full-screen screenshots were useful, experimenters found it challenging to identify which part to focus on during an interview without additional location indications.	The tool allows for screenshotting a selected screen part and highlighting the area of interest.
	Absence of screenshot indications - Although audio feedback when taking screenshots was useful, experimenters needed a way to view the screenshot while simultaneously observing the participants.	The tool enables a preview of the screenshots taken.
	Difficulty in recording the accuracy of user responses - Experimenters found it challenging to record and consolidate users' recalled foreign language accuracy during the evaluation phase.	The tool enables live analysis, calculates accuracy (using correct/incorrect annotations), and displays statistics.

Table 4 continued from previous page

D3	*Difficulty in navigating recorded video* - Manually searching through the video based on timestamps from screenshots was time-consuming and distracted experimenters from focusing on the interviews.	The tool links screenshots and notes with the recorded video and enables direct navigation to corresponding timestamps by clicking on screenshots.
	Additional effort required to take screenshots with notes and summarize them - Experimenters found it demanding to take additional screenshots from the recorded video during analysis and copying them manually to the note documents was burdensome when summarizing the observation notes.	- The tool enables taking additional notes and screenshots quickly during the post-study interview. - It also auto-generates a summary view based on recorded screenshots and notes.

C.2 Second Iteration: Findings

Despite participants appreciating the integrated interface, they expressed new concerns, outlined in Table 5, which were addressed in the final iteration (Sec 4) as follows:

The cluttered and inconsistent UI was addressed by redesigning it for uniformity and integrating all GUIs into one, enabling *Ease of Setup in Conducting Pilot Studies* (***D0***). To *Support Observations in Situated Contexts* (***D1***), difficulties in managing third-party components were mitigated by seamless integration with the device configuration feature. To *Reduce Task Load of Experimenters* (***D2***), we enabled annotation customization, modification, and multi-experimenter support to reduce task loads. To *Expedite Data Recording, Analysis, and Generation of Creative Insights* (***D3***), the introduction of audio transcription and corresponding annotations eased the difficulties in identifying critical feedback or instruction during analysis. To simplify the selection of necessary annotations for interviews and sharing results, we supported annotation highlighting and filtering, allowing tabular data to export in both PDF and CSV formats for easier viewing and analysis.

Table 5. Concerns and solutions for the **second** iteration (Figure 8b). Here, ***D0***, ***D1***, ***D2***, and ***D3*** represent *Ease of Setup in Conducting Pilot Studies* (***D0***), *Support Observations in Situated Contexts* (***D1***), *Reduce Task Load of Experimenters* (***D2***), and *Expedite Data Recording, Analysis, and Generation of Creative Insights* (***D3***), respectively.

	Issue Description	Design Solution & Features
General ***D0***	*UI is cluttered* - Difficulty in recognizing what UI to focus on during the study. - Lack of consistent look throughout the interfaces.	- Use a single uniformed GUI and add others as sub-GUIs. - Redesign the UI to make it more consistent.

Table 5 continued from previous page

	Lack of proper system feedback - Lack of confirmation to stop the pilot session. - Lack of feedback when the pilot session exceeds the anticipated duration.	*Enhance the system feedback to users* - Prompt for confirmation for stopping the pilot. - Play an alert when the anticipated duration is over.
D1	*More configurations are needed to manage third-party components* - Difficulty in setting up third-party components (e.g., wizarding interface) as they are not linked to the tool. - Difficulty in selecting correct video and audio sources for recording as the tool may record incorrect data due to multiple sources. - Difficulty in positioning and locating virtual content.	- Extend the device configuration feature (e.g., configure TPV's IP, *Wizarding Interface*'s address) - Allow configuration of video and audio recording sources. - Shows customizable grids (e.g., 4x4) on the FPV stream.
D2	*Annotations are hard to customize and modify* - Difficulty in taking annotations using familiar shortcut keys. - Difficulty in modifying the annotations during the post-study analysis if errors were made when recording.	Enable customization/modification of annotation types and their properties (e.g., timestamp) before the study and during the analysis.
	High task load makes adding fine-grain observation details difficult when wizarding - Difficulty in highlighting areas of interest when wizarding (i.e., typing the pronunciation.) - Difficulty in adding notes to screenshots when wizarding. - Forgetting to annotate interesting observations when too focused on wizarding.	Support multi-experimenters to delegate work between wizarding and observing.
D3	*Insufficient indicators of communication between users and experimenters.* - Certain moments and content of communication, such as the provision of feedback or instruction, were identified by experimenters as crucial for post-study analysis. However, it required considerable manual effort to navigate and pinpoint these instances in the recordings when the user or experimenter was providing feedback or instruction.	- Enable transcription of the audio from the recording, and present these transcriptions as voice annotations with corresponding timestamps in the *Analyzer*. - Use different colors or icons to distinguish between types of annotations.

Table 5 continued from previous page

Insufficient support for easy navigation through captured screenshots during analysis.	Integrate a photo gallery that is linked to the recording for use during analysis.
Filtering and analyzing annotations require extra effort.	Allow filtering and highlighting of annotations.
Not all annotations need to be discussed.	Allow exporting only the selected annotations.
Further analysis requires the use of familiar software (e.g., Excel).	Provide the option to export annotations in both PDF and CSV formats.

Received Nov 2023; revised May 2024; accepted Sept 2024