Beyond the Phone: Exploring Phone-XR Integration through Multi-View Transitions for Real-World Applications

Category: Research

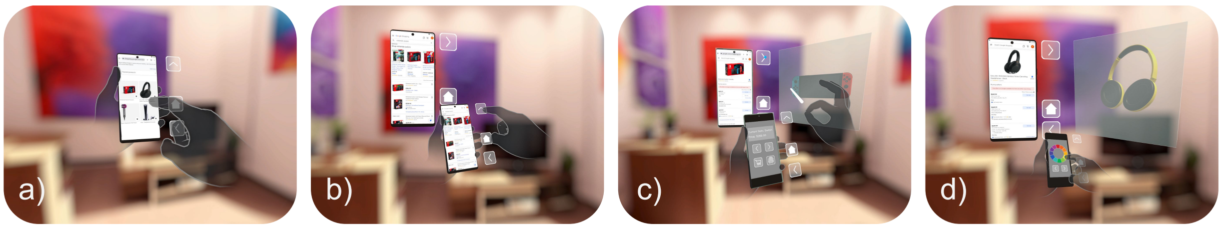


Figure 1: An example user journey with *Beyond the Phone*, enabling context-aware controls and spatial augmentation across multiple states: a) The user starts by reviewing shopping items in a mirrored phone view within XR; b) transitions to a magnified view for better readability; c) expands into an augmented view to examine the item in 3D; d) uses the tailored view on the phone to control the 3D model in the augmented view.

Abstract

Despite the growing prevalence of Extended Reality (XR) headsets, their integration with mobile phones remains limited. Existing approaches primarily replicate the phone's interface in XR or use the phone solely as a 6DOF controller. This paper introduces a novel framework for seamless transitions among mirrored, magnified, and augmented views with a tailored phone interface that dynamically adapts to the content and state of mobile applications in XR, enhancing the user experience. To achieve this, we establish a design space through literature reviews and expert workshops, outline user journeys with common real-world applications, and develop a prototype system that automatically analyzes UI layouts to provide enhanced controls and spatial augmentation. We validate our prototype system with a user study to assess its adaptability to a broad spectrum of applications at runtime. The findings offer valuable insights into the effectiveness of our approach, highlighting its potential to advance a cross-device ecosystem between mobile phones and XR.

Index Terms: Cross-Device Interaction, Phone-XR Intergration.

1 Introduction

The rapid evolution of Extended Reality (XR) headsets is ushering in a new era of immersive computing. As XR technology increasingly integrates into daily life, understanding how head-mounted displays (HMDs) can seamlessly coexist with our broader ecosystem of devices, especially mobile phones, becomes crucial [7]. Given the ubiquitous presence of mobile phones in modern society, it is imperative to explore how they can complement and enhance the XR experience, not merely as peripheral tools but as integral components of a cohesive interactive system.

Recent research has explored integrating mobile phones with XR devices in various ways: employing phones as 6DOF controllers [44, 3, 10, 22]; mirroring phone screens within XR environments [44, 12, 28, 3, 11, 46]; and extending phone interfaces with additional 2D panels [28] or 3D elements [44]. However, existing work often focuses on isolated modalities, either using controllers or enhancing display functionalities alone. In real-world applications, interactions are frequently multimodal and evolve throughout different stages of a task, necessitating fluid transitions between different modalities.

For example, a user might begin by using phone controllers for input and later rely on augmented visuals during the preview phase. Such a modality switching remains largely unexplored in current research. Furthermore, much of the prior work concentrates on mock-up applications with interaction modes designed specifically for single-state scenarios. This approach limits generality and fails to provide scalable solutions applicable to broader contexts beyond these prototype examples.

Addressing these limitations, we propose *Beyond the Phone*, a comprehensive framework that enables seamless multimodal interaction between mobile phones and XR head-mounted displays. Our framework integrates the strengths of both devices and supports fluid transitions between various interaction modalities, catering to the dynamic nature of real-world applications. Our design is guided by three key principles:

1. **Generality**: The design should be applicable to a wide range of applications, ideally supporting automation without requiring extensive re-engineering efforts.
2. **Bi-directionality**: It should support both phone-centric and HMD-centric tasks, enabling seamless interaction and data exchange between the two devices.
3. **Adaptability**: The framework should permit dynamic switching and smooth transitions between different modes of interaction involving the phone and XR user interfaces.

A user journey, depicted in Figure 1, illustrates our framework above. The user initiates interaction with a shopping app mirrored within the XR environment(Figure 1a), viewing the phone interface virtually while using the physical phone for input. Leveraging the unlimited display space of XR, the user transitions to a magnified view for enhanced readability (Figure 1b). Upon searching for a specific item, the user expands into an augmented view to examine the product in 3D (Figure 1c). Simultaneously, the phone interface in the user's hand transforms into a tailored control panel (Figure 1d), facilitating direct interaction with the augmented content. For instance, if the item offers color variants, the phone interface dynamically updates to present a color palette for selection. This dynamic modality switching, adapting the phone's role according to the application's stage, enables seamless transitions between phone-centric and HMD-centric interactions.

To achieve this framework, we conducted an expert workshop and literature review to develop a design space accommodating both generalizability and dynamic transitions in phone-XR integration. Based on this design space, we built a prototype system capable of automatically analyzing app interfaces and enabling semi-automatic content adaptation aligned with application states.

To demonstrate our approach, we augmented six real-world applications downloaded from the App Store and conducted a user study involving 12 experienced XR professionals to test the feasibility and applicability of our system. Our evaluation examined whether the proposed dynamic transition approach is universally optimal or if some applications benefit more from a single-view paradigm, providing insights into our framework and informing future research directions.

Beyond the Phone is therefore a combination of theory and practice in which we:

1. Based on an expert workshop and existing literature, depict a design space and a set of user journeys for a diverse range of mobile applications, transitioning among different views that demonstrate both bi-directionality and adaptability.
2. Develop and implement a phone-in-XR prototype system that supports real-world applications, featuring a semi-automatic mechanism to switch between different views based on the content and the application's state. This approach highlights the generality of our system, allowing it to adapt to various applications without requiring extensive customization.
3. Validate the prototype system in a user study with 12 experienced XR professionals, in which we show the potential of our framework to improve phone-in-XR experience with real-world applications.

2 Related Work

Our work is inspired by prior research in cross-device interaction between phone and XR, as well as contextual UI understanding.

2.1 Integrating Smartphones with XR Environments

In recent years, considerable research has explored the integration of smartphones within XR contexts. One primary focus has been enabling users to view and interact with their phones while engaged in immersive XR experiences. Techniques such as screen mirroring [3] and passthrough windows [11] allow users to seamlessly attend to notifications or respond to text messages without leaving the immersive environment.

However, integrating smartphones into XR presents significant technical challenges. Accurate interaction with a virtual representation of the phone requires precise touch calibration [45, 23], which depends on highly accurate reference frame alignment [21, 17, 26] and fingertip estimation [45, 23]. Additionally, estimating hand dexterity is crucial for reliable interaction [40].

Leveraging smartphones as spatial controllers within XR environments has been another significant research direction. By incorporating 6DOF tracking and utilizing the phone's inherent haptic feedback, researchers have explored various interaction techniques. These include manipulating 3D widgets [25, 26, 23], text entry methods [12, 3], menu selections [44, 22], two-factor authentication [47], and data visualization [10]. Additionally, mid-air gestures using smartphones have been investigated as an input method for XR applications [6, 24].

Extending phone-centric content beyond the physical screen through XR has also been a focus. Techniques involving extended displays [28, 14] enable new interactions with 3D content [25, 34] and bring static content to life [8]. Recent efforts have shifted towards creating cohesive design spaces for integrating smartphones within XR [44] and developing hybrid input that combine phone and controller interactions [42].

Overall, these projects have unveiled essential techniques, but often concentrates on isolated interaction modalities, focusing either on using smartphones as controllers or enhancing display functionalities independently. This compartmentalized approach overlooks the fluid nature of real-world interactions, where users frequently need to transition seamlessly between different input methods as tasks evolve. Moreover, prior studies often limit themselves to prototype applications designed for single-state scenarios, restricting their applicability to broader, more dynamic contexts.

In this work, we address these limitations by (1) highlighting the importance of seamless transitions between interaction modes driven by user intent and contextual adaptation, and (2) adopting an application-centric perspective to identify which functionalities are most relevant for diverse use cases.

2.2 Contextual UI Understanding in XR

To enable seamless transitions based on application content, it is essential to develop a contextual understanding of user interfaces within XR environments. This involves analyzing the application's state and interpreting the user's intent. Recognizing this need, Grubert et al. [13] envisioned a pervasive and context-aware XR, highlighting the unique potential of these devices not only in rendering spatial content but also in perceiving and interpreting the user's environment and context. They provided a foundational taxonomy at a time when the necessary technology was still emerging.

With advancements in the perception capabilities of XR devices, researchers have leveraged these improvements to dynamically adapt content—such as spatial user interfaces—based on environmental context [36, 5, 29, 32]. For instance, Lindlbauer et al. [20] integrated user awareness with spatial insights to determine both the placement and content of spatial widgets, underscoring the significance of context in enhancing user interactions.

Building on these developments, subsequent research has further advanced UI understanding in XR. Li et al.'s *HoloDoc*[18], for example, explored a document-aware XR workspace that adapts to user activities. Similarly, Cheng et al.'s *SemanticAdapt*[9] proposed an optimization-based approach to generate XR layouts by leveraging virtual-physical semantic connections.

Insights from cross-device interaction research have also informed XR development, particularly in the digital augmentation of devices and objects using smartphones. Examples include the creation of extended display devices in spatial environment [33, 12], visualizing content from physical devices [10, 1], and utilizing the real environment as a canvas for interaction [2]. These approaches demonstrate how AR methods can enhance XR experiences, as seen in the transfer of accessibility features from AR on phones to HMDs to augment reading experiences [4].

Despite these advancements, determining how and what to augment based on real-world mobile applications and user interactions remains a complex challenge. In our work, we address this gap by performing contextual analysis of both the smartphone app UI and the user's behavior. By interpreting the application's state and the monitoring where the user's attention is directed, we dynamically decide on the most appropriate augmentation methods, managing different views and interaction states within the XR environment. This approach enables seamless transitions between interaction modes, enhancing the user experience by making interactions more intuitive and context-aware.

3 Beyond the Phone Framework

To explore and expand smartphone-XR integration, we conducted an expert workshop to inform a design space of display enhancements, phone interface, and Phone-XR interactions.

3.1 Expert Workshop

In line with our goal to create a generalizable framework applicable to a broad range of real-world applications beyond mock-ups, we conducted an expert workshop to explore how daily smartphone applications can be utilized within XR environments and to identify the opportunities presented by such cross-device collaboration. The workshop involved nine professional researchers and software engineers from <an anonymous institution>, each possessing extensive experience in developing immersive VR and mobile applications, thereby providing valuable insights for our analysis.

Application	Example	Pain Points on Phones	Display Enhancement	Tailored Interface
Web Browsing	Wikipedia	• Text readability • Screen size	• Expanded displays • Focused reading and summarization modes	• Touchpad (tap & multitouch) • Context dependent menus
Collaborative Work	Google Docs	• Inconsistent layouts • Editing challenges	• Multiple panels for drafting, editing, revising, etc.	• Keyboard (typing) • Touchscreen for markups
Photo Browsing	Google Photos	• Screen size	• Immersive gallery • Panoramic views	• Spatial (raycast) • Editing palette
Video Watching	YouTube	• Screen size • Environment distractions	• Immersive cinema view • Extended device screen	• Intuitive video scrubbing
Shopping	Amazon	• Lack of 3D and in-situ visualizations	• 3D in-home gallery view • Color/style palette	• Spatial (placement & manipulation)
Communication	FaceTime	• Limited embodiment • Transcription & augmentation	• 3D avatar views • Summarization views	• Keyboard (typing)
Navigation	Google Maps	• Lack of 3D visualization • Screen size	• 3D Overlays • Earth view	• Touchpad (tap & multitouch) • Context dependent menus
Social Media	Twitter	• Text readability • Screen size	• Embodied content • Immersive videos	• Keyboard (typing)

Figure 2: Results from the expert workshop. We have outlined the pain points of daily smartphone applications, proposed XR display enhancements, and potential phone-integrated input improvements.

During the workshop, we began by reviewing existing literature that discusses phone usage in XR, to ensure our team understand the essential aspects of this domain. After a thorough introductory discussion, we initiated a brainstorming session, engaging them in evaluating daily applications by considering the following aspects:

1. Identifying the pain points of the application when accessed solely through a physical phone.
2. Exploring potential display enhancements that could be applied to this application within an XR environment.
3. Considering alternative interfaces that could transition from the phone to enhance input when the application is expanded into an XR format.

The results of our workshop, illustrated in Figure 2, revealed a common need for enhanced display capabilities—such as larger screen sizes, multiple views to showcase diverse content, and the ability to view 3D content at life-size scale. Through their discussions, participants also concluded that while leveraging the tangible nature of smartphones for input methods is desirable, implementing such inputs requires them to be triggered based on the application's different states and contexts. This necessitates a close alignment with the phone's content and context.

Furthermore, during the workshop, participants discussed that beyond focusing solely on interfaces attached directly to the tracked phone as previous works [44, 3, 22, 28] did, exploring the use of magnified views floating in the air with multiple view setups holds significant potential. This approach can accommodate different requirements for augmented content and phone modalities while keeping the initial phone interface accessible in the same context.

This finding underscores the practicality of alternating between mirrored and tailored phone interfaces as needed. These insights have laid the groundwork for proposing a design space for enhanced applications that combines phone and XR together.

3.2 Design Space

As previously highlighted in the workshop, existing research has focused on interfaces that are solely attached to the physical phone, typically applying only a static modality—either as a controller or as extended views—without transitioning between modalities across different application states.

To bridge the gap between these limitations and the need for dynamic modality transitions, we propose a solution that encompasses different modality views, including mirrored, magnified, and augmented views, utilizing the *Magnified View* as an **intermediary state for transition**. Users can simultaneously utilize both the mirrored and magnified views or employ the magnified view as a preliminary step to initiate augmented views when needed. When the augmented views are presented, the Magnified View retains the original application content, while the phone interface transitions to a tailored interface, incorporating appropriate input widgets that correlate with the augmented content currently in use, such as buttons. The capability to dynamically alternate between display enhancements fosters a cohesive transition experience, effectively unifying what were once isolated components.

To effectively guide the design process and align it with the points proposed above, we introduce a design space depicted in Figure 3. This design space outlines different states for both display enhancements and phone interfaces, allowing transitions between them. It includes considerations for display enhancements in the XR setting and context-responsive adjustments to the phone interface. Moreover, our proposed interactions integrate touch inputs with spatial dynamics, encompassing the movement and spatial relationship between the phone and the XR system.

We emphasize that our approach does not entail the creation of each component in isolation. Similar to the approach applied by previous works [44, 10], our work primarily involves the synthesis and integration of existing research knowledge to effectively construct a design space tailored to specific needs. In the following section, we provide a detailed explanation of how each element functions within this design space and discuss related works that exhibit similar concepts.

3.2.1 Display Enhancements Aligned with App States

Our system implements three types of display enhancements that are closely aligned with application states and allow for **seamless transitions between each other** based on the application's context. We describe each enhancement below.

Mirrored View [3, 16]: This enhancement creates an exact digital replica of the smartphone's interface within the XR

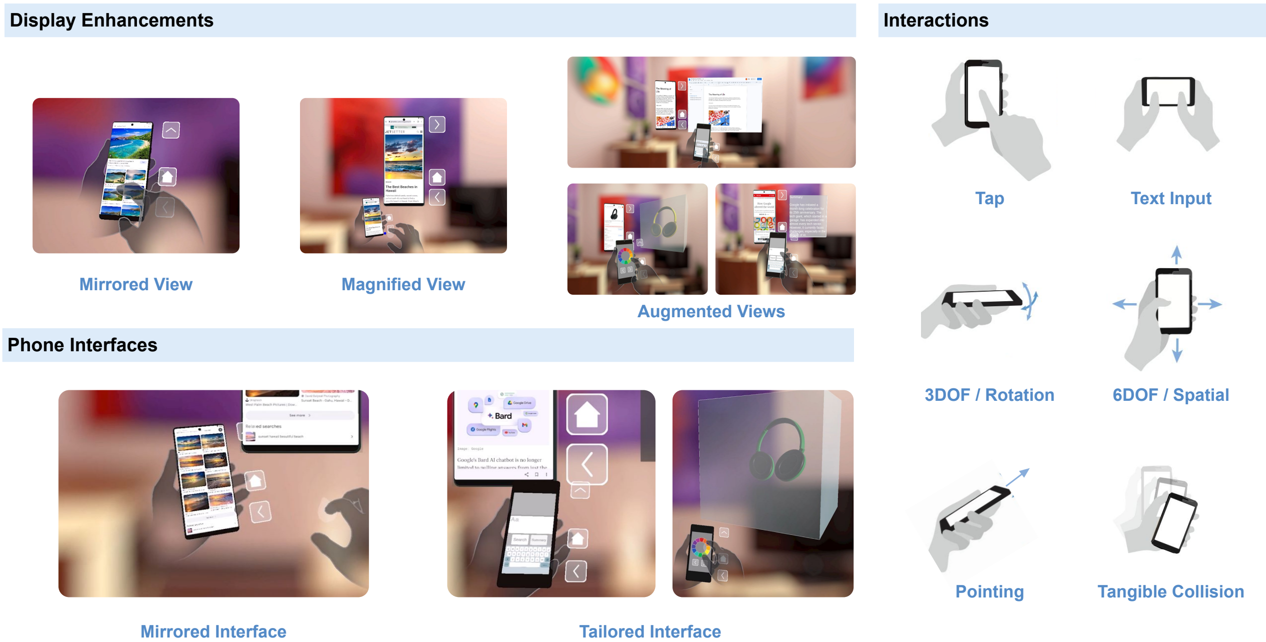


Figure 3: The *Beyond the Phone* design space we explore with, including: (1) Display Enhancement methods of varying immersion, (2) Phone Interfaces that applied based on different context, and (3) Interactions.

environment, aligned with the physical phone's spatial orientation. In application states where direct interaction with the phone's native interface is most appropriate, users can interact with their phone just as they would in the real world with this setup.

Magnified View [12, 28, 23, 46]: When the application requires enhanced readability or a larger display area, the system transitions to the magnified view. This enhancement projects the smartphone's interface onto a larger virtual canvas within the XR environment, overcoming the limitations of the phone's physical screen size. Users can interact with both the physical touchscreen via the mirrored view and the enlarged content of the magnified view using mid-air gestures. The magnified view acts as an intermediary state, allowing seamless transitions between direct touch and spatial interactions, aligned with the application's needs.

Augmented View [44, 12, 23]: For application states that benefit from additional content or 3D representations, the system enhances the display by overlaying augmented content onto the XR environment. This may include alternate user interfaces, realistic 3D representations for object previews (e.g., for shopping), or supplementary 2D content (e.g., news summaries). The physical phone interface adapts to become a customized controller, incorporating input widgets that correlate with the augmented content relevant to the current application state.

3.2.2 Seamless Transitions Between Enhancements

Our approach allows for dynamic transitions between these display enhancements based on the application's state and user actions. The magnified view serves as a bridge between the mirrored and augmented views. Users can activate the magnified view from the mirrored view when they need a larger display, and from there, they can initiate the augmented view to access enhanced content. Throughout these transitions, the original application content remains accessible, and the input modalities adapt to provide the most intuitive interaction methods.

By aligning the display enhancements with application states and enabling smooth transitions between them, our system fosters a cohesive and adaptable user experience. This design ensures that users can naturally progress through different phases of interaction without disruption, effectively unifying what were once isolated. Which also guaranteed the bi-directionality during the transition [44].

3.2.3 Phone Interfaces

The *phone interface* in our framework refers to the interface that is superimposed onto the physical device. We utilize two distinct modalities:

The **Mirrored Interface** [3, 16, 11, 46] refers to the exact replication of the phone's interface within the virtual environment. This modality serves a dual purpose, offering both a **Mirrored View** and control over content within the **Magnified View**.

The **Tailored Interface** refers to the alternative interface that transform the phone into a controller to facilitate enhanced interaction when the **Augmented View** is activated. This typically involves using the smartphone as a tactile controller [22, 26, 23, 37, 12, 44], equipped with widgets that are contextually relevant to the current content displayed.

3.2.4 Phone Interactions

To fully leverage the potential of phone interactions in XR, which can enhance context understanding and improve usability, we categorize phone interactions into three distinct levels:

Touch Input: The phone acts as a touch input device, accommodating familiar interactions such as tapping and text input, addressing the precise input needs that often lack with XR devices.

Spatial Input: The phone's spatial features can also be leveraged as an input. Users can utilize its 3DOF orientation or its full spatial movement capabilities in 6DOF to provide additional input modalities.

Spatial Relations: In this level, interactions are based on the spatial relationship between the phone and augmented elements within the XR environment. This includes pointing at augmented objects to select or interact with them, or initiating tangible collisions with virtual elements to simulate physical interactions.

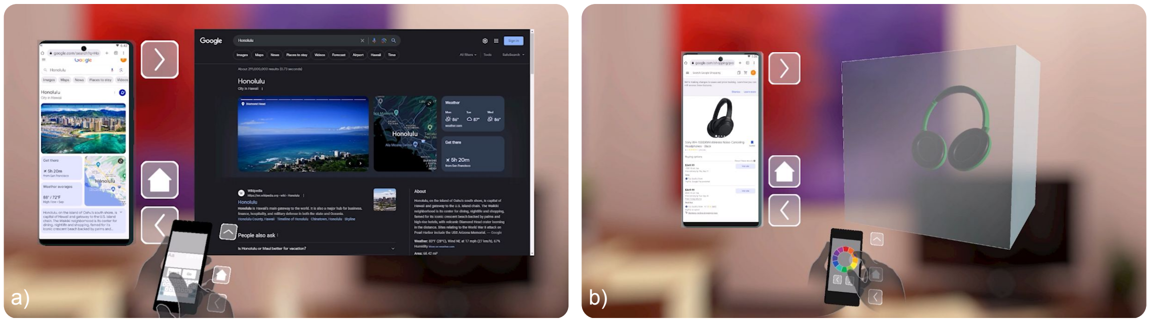


Figure 4: Example of how different applications are augmented uniquely in *Beyond the Phone*: a) a desktop webpage view and text-input view to support web-search application; b) a 3D product preview and customizable color palette for shopping application.

3.3 Application-Adapted Views

Our workshop revealed that application-centric information can be effectively leveraged as a context source [13] for augmented smartphone interactions in XR. This perspective differs from much of the prior research on Phone + XR, which has primarily adopted a device-centric approach [44].

Drawing inspiration from methodologies presented by Lindlbauer et al. [20, 9, 38], we propose a Smartphone UI Understanding approach that applies augmented content across different applications. Our approach utilizes the content and potential controls within on-phone applications to enhance interactions in XR environments. For example, Figure 4 demonstrates two scenarios: in a web-search application, the augmented view is transformed into a desktop-like interface for improved visualization, with the phone's tailored interface becoming a virtual keyboard for easier text entry. In a shopping application, the augmented view displays 3D product previews, while the phone interface adapts by incorporating a color palette, allowing for style selection.

Our expertise workshop results serve as a valuable reference for implementing these augmented interactions, supporting their generality and adaptability within our frameworks. Further technical details on how to implement this approach will be discussed in the following section.

4 Implementation

To gain a hands-on understanding of the design decisions in our framework and to assess it with actual users, we developed an interactive prototype integrated with UI understanding, with the system architecture shown in Figure 5.

Previous studies have developed various methods for phone tracking and user interface mirroring in virtual reality environments. Most research utilizes an external tracking system to track the physical phone [44, 22, 3]. For the interface attached to the physical phone, existing works either employ direct video streaming [3, 16] or construct a simulated interface in their demonstration applications, which does not accurately mirror the actual content visible on the phone [44, 22]. These methods often do not separately process input and output related to the smartphone interface during deployment.

For our system, *Beyond the Phone*, we aim to enable the phone to support multiple views, particularly as the device transitions between phone and controller modalities. This requires a separate processing approach where the phone's interface is treated as output, and input is fused with both spatial input and touch, independent of the phone's touch surface. This builds upon the foundational work of SAPIENS-in-XR [30] and XRStudio [27], which explore XR system architectures and interaction frameworks, which leading us into our system design that achieves the following:

1. The interface that represents the phone content is a real mirror of the physical phone, rather than a mock-up. This applies to both the **Mirrored View** and **Magnified View**.
2. Input from the physical screen does not directly influence the original phone interface shown in VR. Instead, the VR system controls the input, combining touch and spatial gestures as a unified input source.
3. The updates of the **Mirrored View** and **Magnified View** are independent of physical touch from the phone, instead controlled by combined inputs from touch, spatial input, and commands/events from the VR system.
4. The **Augmented View** is updated with UI analysis results from the context of the current phone application, rather than preset mock-ups.

Figure 5 shows how we achieve these requirements. In general, the phone interface and touch events are separately processed via our protocol. The UI analysis and input system are regenerated through our system. Drawing inspiration from prior research such as Bai et al. [3], we leveraged a customized WebRTC approach to relay the screen texture from the phone to our backend. Since the virtual phone's input and output do not strictly map to the physical characteristics of the real phone due to tracking errors, we employed an separation process of phone interface that processed via Android Emulator, in addition to a physical phone capturing touch input and 6DOF spatial tracking for prototype development. This emulator runs applications that could be downloaded from app stores like Google Play, processes inputs from the physical phone and XR, and transmits the screen texture to XR.

In the following sections, we will introduce two techniques from our exploration. The Hybrid Input technique is designed to eliminate the influence of tracking errors, while the UI analysis process provides a general solution for content understanding.

4.1 Hybrid Input

The Hybrid Input technique, based on the work by Zhu et al. [45], is established to eliminate the inevitable tracking errors between hands and phones. In terms of input for a multi-device system, we can identify three unique "Pointer Events" within our system:

- **Virtual Touch**: This event is detected when the virtual fingertip interacts with the virtual phone.
- **Physical Touch**: This event is triggered when the user's real finger comes into contact with the physical phone.
- **Spatial Input**: These are spatial gestures, such as raycasting + pinch to interact with the Magnified Screen.

If tracking was perfectly accurate, both the Virtual Touch and Physical Touch events would coincide without any disparity. Yet, achieving accurate tracking of both the phone and hands in XR

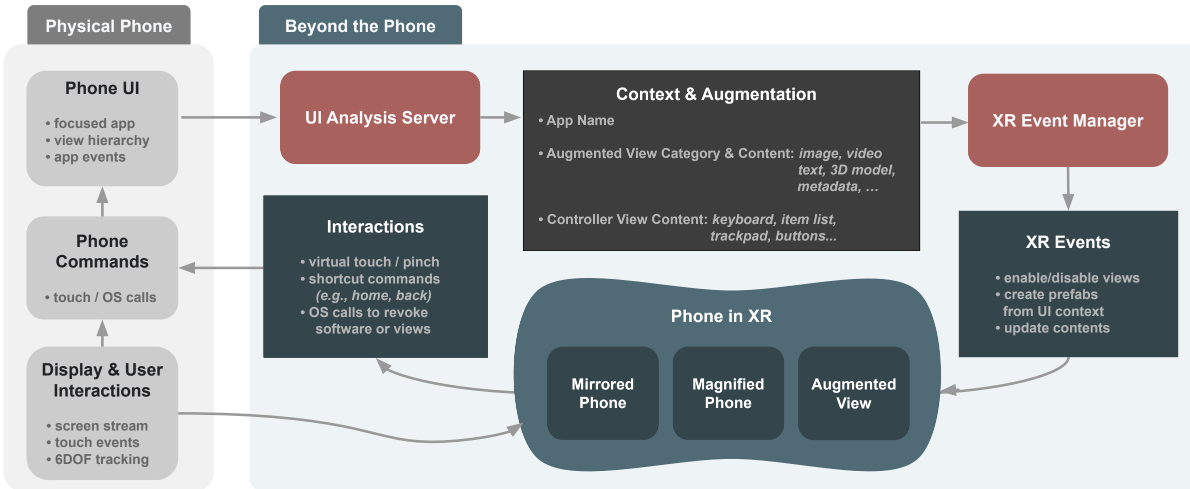


Figure 5: *Beyond the Phone* prototype architecture: On the left is the physical phone, with input and output conceptually divided for clarity. The right side maps out the process, including a UI Analysis Server, a XR event manager, and proper data flow towards different phone views in XR. The XR input will also be injected into the physical phone, either as pointer events or shortcut commands.

is rarely the case in real scenarios. Such deviations can lead to mismatches between these two events, in which the touch visually perceived by the user is different to what the touchscreen registered, akin to the "fat-finger problem" in touchscreen input.

To overcome this problem, and drawing from the insights of Zhu et al. [45], we implement a **Hybrid Touch** strategy. This approach utilizes the Virtual Touch to determine the 2D touch point on the virtual phone, while the Physical Touch on the real screen is used to confirm the touch upon physical contact with the surface.

In order to capture the spatial position and touch events from the phone, we designed a companion application that captures and transmits all touchscreen interactions to the XR + phone experience. This ensures precise implementation of hybrid touch and spatial input, while also enabling the phone to function as a custom spatial controller.

For quick access functions, such as 'home' and 'back', we embedded several UI widgets surrounding the virtual phone model (See Figure 1 and Figure 5). These widgets are tied to specific `adb shell` commands, pragmatically triggering the desired inputs. This design not only facilitates intuitive operation in the XR environment but also minimizes false inputs typically associated with global gesture controls, as these widgets are spatially distinct and float in the virtual space.

4.2 Screen Streaming

Unlike previous works that directly synchronize the phone screen with the system [3, 22], our approach, which integrates Hybrid Touch, Spatial Input, and System Shortcuts, updates the Android view only after processing these inputs collectively. This is achieved through system-level input injection into the Android emulator, involving necessary adjustments to events and visual presentations. Once updated, this view is then sent to the UI Analysis Server for further processing.

Beyond transmitting the screen image, our system leverages application states for content augmentation. Therefore, we also send metadata, such as the view hierarchy, through the same channel. The details of this process will be further discussed in the section §4.3.

4.3 UI understanding

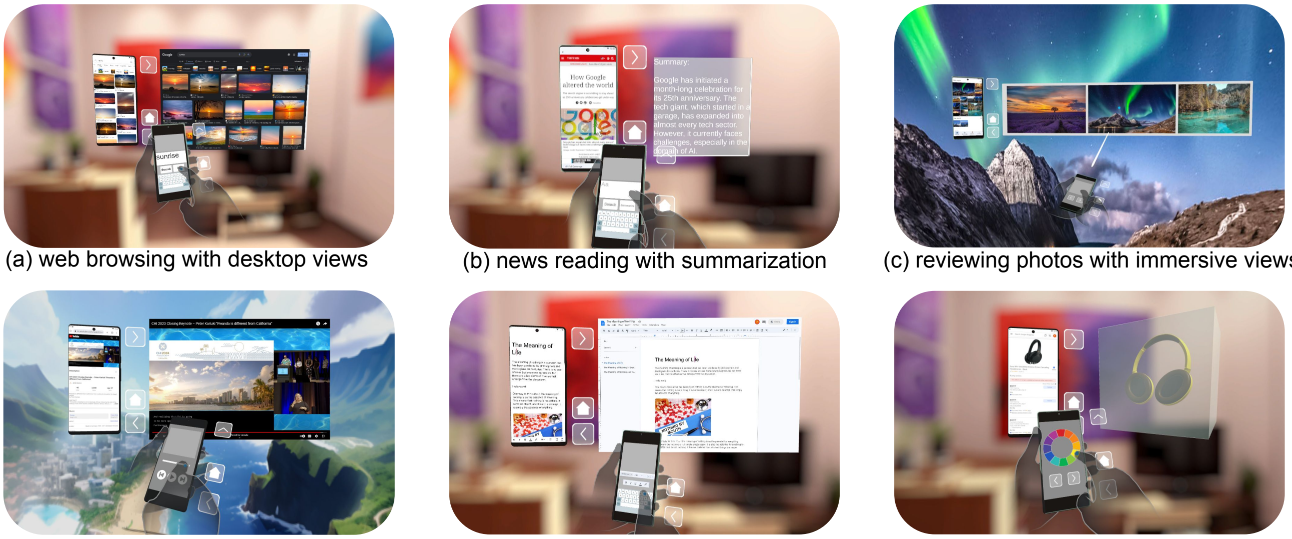
For our prototype, it is crucial to identify the currently focused application and content. This foundation enables us to discern what immersive content should be projected spatially and how to optimize the phone's functionality as a controller.

As indicated by prior research, the optimal method involves UI analysis utilizing the *UIAutomation* tools that Android natively offers. Drawing on the procedures adopted by previous work [39, 15, 43], we implemented a streamlined UI Analysis Server to obtain context-sensitive metadata. This server acquires screenshot, the view hierarchy in XML (see an example in the Appendix), and event log from the phone. Subsequently deducing metadata such as the **Application Name, Focused Content, and Main Activity Bounding Box.** We then attempt to match those data with the outcomes from our design workshop, which contains a set of different widgets for tailored views and content automation that could be applied for augmented views. As for our current design, the potential content that could be augmented includes 3D objects preview for certain types of objects, gallery views for multiple photos, digest of article, video player, online text editor and web browser based content. The potential tailored views includes widgets like keyboard, video player controller, color palette, list selection elements, and ray based pointer. Those widgets could be organized with multiple objects in the same view. For unlisted applications, the bounding box of the Focused Content is returned for direct content expansion (see Figure 5).

After successful matching, the server sends the relevant details back to the XR application, including the content meant for augmentation and the anticipated category at the time of the request. The XR application uses this information to update the XR assets appropriately while also toggling enable/disable states based on user inputs.

4.4 Setup and Apparatus

Our prototype uses a Pixel 6 Pro phone paired with a Meta Quest Pro headset. The software was developed using Unity 2021.3.4f. The mobile companion app was directly built in Android and relays touch events and tracking details to the XR application through UDP, with screen broadcasting via a specialized WebRTC approach. UI automation procedures were initiated through `adb shell` commands from the server end. Oculus SDK supplied the hand tracking information. To track the phone's position and orientation in the XR environment, the companion app runs ARCore's 6DOF tracker in the background. Meanwhile, the hand tracker is used to calibrate the phone's coordinate alignment with



(a) web browsing with desktop views (b) news reading with summarization (c) reviewing photos with immersive views

(d) video watching with media controls (e) document editing with format controls (f) shopping with color palette & 3D views

Figure 6: Example applications augmented with *Beyond the Phone* for the evaluation: (a) browsing image-extensive webpages with desktop views, (b) using text-heavy apps with summarizing views, (c) exploring photos with gallery and immersive views, (d) using YouTube with media controls, (e) editing online documents with phone keyboards and format controls, and (f) online shopping with color palette in phone and 3D augmentation spatially.

the XR space when held in a specified manner.

Specifically, since the system encompasses multiple communication channels with varying layers of information exchange between the XR system and the smartphone, necessitating attention to different levels of latency. During testing, the mirroring interface and synchronized tracking took approximately 100ms for a complete cycle, while the UI automation procedures via ADB required around 3 seconds for server-side analysis. Therefore, in our final development, we set tracking and screen mirroring to update continuously for smooth interactions. In contrast, UI analysis via UI automation is triggered as needed, as illustrated in Figure 5.

Following the insights gained from our workshop, we selected six applications for the evaluation (see Figure 6). All the applications are directly downloaded from Google App Store without any modification.

The applications include a 3D object viewer for an enhanced shopping experience, a 2D photo gallery extension compatible with Google Photos, a concise news summary feature, a document editing tool utilizing the phone's keyboard, a video player with on-phone controls, and a spatial web browser.

The selection of these applications was based on their varied configurations and the diversity of the content they present. Each application represents a distinct use case, ranging from 3D object interaction to 2D media browsing, document editing, and video playback, offering a comprehensive spectrum of interaction modalities. These applications will serve as key components in our user study, allowing us to evaluate the system's performance across different interaction contexts and validate its versatility.

5 User Study

To validate our proposed system and make an assessment of multi-view transitions and their effectiveness in real-world applications, we conducted a user study to gather user feedback on these features.

5.1 Study Setup

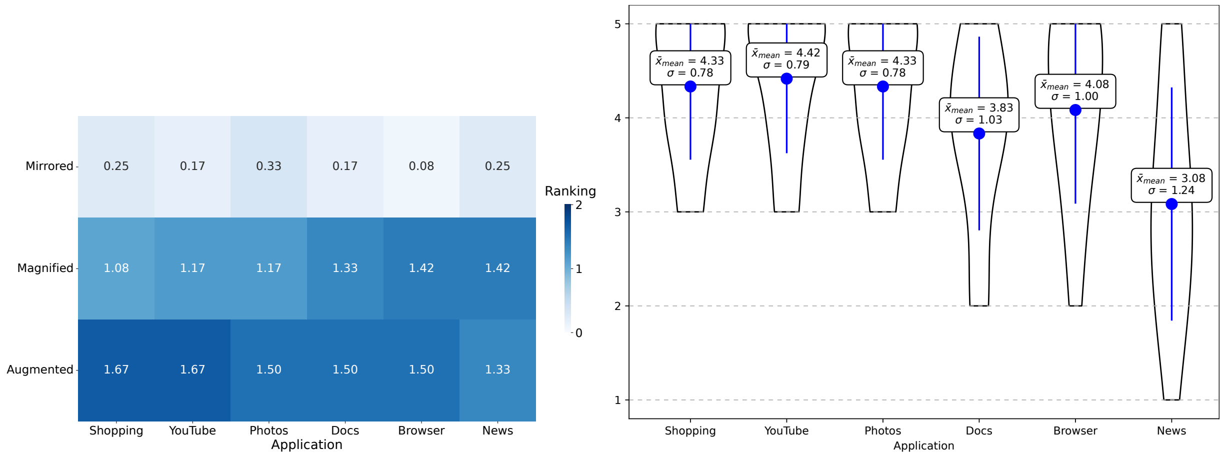
We conducted an empirical study of *Beyond the Phone* with 12 participants (3 female, aged M=33.7, SD=9.9). The participants in this study were all professional UI designers, researchers, or software engineers employed at <an anonymous institution>. Each participant had substantial experience in creating immersive experiences, providing a knowledgeable and relevant user base for our evaluation.

The study aimed to evaluate participants' preferences for different views of each application, as well as assess the perceived coherence and consistency across these views. As shown in Figure 1, participants assessed various transitions between views, including phone mirroring, spatial magnification, multi-view setups, and augmented configurations.

Each session lasted 45–60 minutes and consisted of four stages: (1) a tutorial; (2) an exploration of *Beyond the Phone* with the six applications mentioned above; (3) a questionnaire that captured preferences for different views and perceived coherence and consistency during transitions; and (4) a follow-up interview.

Tutorial and Exploration. Before starting the study, participants first signed a consent form and answered a demographic questionnaire. Each participant was then introduced to our system through a series of introductory slides and tutorial videos. This familiarization process included instructions on starting the application, calibrating the tracked phone, and launching features like the Magnified Phone View and Augmented View. Subsequently, participants engaged with the six applications depicted in Figure 6. Participants began with a standardized application journey. After successfully completing the controlled tasks, they were allowed to freely explore content that suited their needs, such as different YouTube videos or news articles. The order of applications was randomized for each participant to ensure unbiased feedback. This exploration allowed participants to interact with diverse functionalities, including document editing, web browsing, online shopping, news summarizing, video watching, and photo browsing.

Questionnaire and Interviews. Upon completing the exploration phase, participants undertook a thorough review of each application scenario. This was achieved through a detailed questionnaire, where they were asked to rank their personal preferences regarding phone representation for each application. Meanwhile, since all the applications combined the use of three



(a) Preference rankings of Mirrored, Magnified, and Augmented Views across each application.

(b) Perceived level of coherence and consistency across different views for each application.

Figure 7: Results from the quantitative studies.

views in transition, we also asked the participants about their perceived coherence and rationality level of the whole transition process in this immersive environment using a 5-point Likert scale.

After completing the questionnaire with individual applications, participants were asked about their overall mental load and physical fatigue compared to traditional VR applications. They were also requested to comment the overall usability of the system using text.

We then conducted interviews to gather participants' rationale and suggestions for improving the prototype. This structured approach allowed us to collect both quantitative and qualitative insights into user preferences and perceptions of the various functionalities offered by *Beyond the Phone*.

5.2 Results

All participants found our system intuitive and easy to use. No one reported perceived fatigue in either mental or physical load compared to standard VR applications.

For the quantitative analysis of preferences, participants ranked their preferences for Screen Mirror, Magnified, or Augmented Views for each application. A scale of (0, 1, 2) was used, where 0 represents the least preferred view and 2 represents the most preferred. The values presented in Figure 7a represent the average preference scores based on participants' rankings.

The results reveal two key insights:

1. Participants consistently favored views that went beyond simple mirroring, with Mirrored Views receiving the lowest preference scores across all applications.

2. Preferences for Augmented and Magnified Views varied depending on the application. Applications featuring media such as 3D shopping previews, videos, and photos saw a clear preference for Augmented Views. In contrast, for applications designed primarily for 2D content, the preferences for Magnified and Augmented Views were more balanced, with minimal differences in preference.

We conducted a Friedman chi-square test to analyze the differences in preference scores across the three views (Mirrored, Magnified, Augmented) for each application. Significant differences were observed for all applications. For example, Browser ($\chi^2 = 15.17, p < .01$) and Shopping ($\chi^2 = 16.55, p < .01$) both showed highly significant results. Pairwise Wilcoxon tests with Bonferroni correction revealed significant differences between Mirrored views and the other two views (Magnified and Augmented) across most applications.

Further analysis focused on the perceived coherence and rationality for each application, assessed on a 5-point Likert scale (1 = strongly disagree, 5 = strongly agree). As shown in Figure 7b, the results are visualized through violin plots, displaying the distribution of scores.

Most applications received high scores, suggesting that participants generally found the combination of views and transition process to be coherent and rational. However, for the news reading application, lower scores were observed. Participants indicated that simpler displays, such as extended or magnified views, were more appropriate for this context, as confirmed during follow-up interviews.

5.3 Interview Findings

In general, feedback from the interviews was overwhelmingly positive. All participants found our interface easy to follow. One participant remarked: *"When I saw the phone in my hand, it was very intuitive, because I'm really used to like, picking up my phone and doing things."* (P4). Many participants were excited by *Beyond the Phone*, with comments such as: *"You guys are the future"* (P1), *"I can view my object in a more immersive way, compared to being constrained in a small display device."* (P2), and *"I am very impressed by having users interact with whatever apps or experiences through a pop-up dedicated UI."* (P5).

Enhanced Legibility and Usability: Participants particularly appreciated the **enhanced legibility** of our system, noting significant improvements in reading and feeling more immersed. As one participant said: *"The ability to read small text is greatly improved with the larger views."* (P6). Another shared: *"I feel more immersed when I browse the web."* (P9). The **enhanced usability** was also widely commended: *"I like the idea of using the phone as a pointer."* (P11), and *"I appreciate the concept of keyboard typing on the phone."* (P9). These responses highlight the effectiveness of our prototype in leveraging the proposed framework and user journeys, particularly in creating a more immersive and interactive experience.

Application-Specific Preferences: As previously reported, we observed different preferences for Augmented and Magnified views across different applications. One potential reason for this is the overwhelming amount of information presented in Augmented View setups when reading is required. As noted by participants: *"I keep looking up and down, which may not be ideal."* (P5) and *"Switching views from the magnified phone to the smaller one felt disorienting."* (P8). This suggests a **desire for more explicit view options**. Some participants also provided application-specific feedback: *"If I'm watching YouTube Shorts, the large phone would be preferred. Traditional videos would need the extended view, as they are in landscape format."* (P3), and *"For Google Docs, I always prefer reading in the Augmented Views."* (P6).

Additionally, participants expressed a desire for configurable, personal choices in view settings. For example, P11 noted: *"For the news scenario, the physical phone seems redundant, but it might be indispensable for a shopping preview."* This prospect warrants further exploration in future studies.

Spatially Enriched Content: The content in the augmented view corresponding to the application was another highlight, with users emphasizing the impressiveness of viewing 3D models towards the time-saving benefits of concise text summaries: *"Viewing 3D headphones is really impressive."* (P10), *"The text summary saved my time."* (P2), highlighting that our framework was intuitive throughout the transitions between mirrored and augmented views.

While the content is generated in real time based on the application's context, an additional concern raised by users was regarding the trust and robustness of Augmented Views. Although most participants found the feature to be *"nice work"* (P3) and *"really cool"* (P4), one participant expressed concerns about the reliability of the results: *"I prefer to read articles in full; I would doubt the authenticity of the summary."* (P6).

In Conclusion, *Beyond the Phone* successfully enhances legibility, usability, and immersion in Phone-XR integration, with participants praising its flexibility across applications. Discussions highlighted opportunities to further optimize view setups and enrich content presentation in Augmented Views and summaries. These insights reinforce *Beyond the Phone*'s strong potential and will guide its continued development in the future.

6 Discussion and Future Works

With *Beyond the Phone*, our goal was to create an enhanced Phone-XR experience that goes beyond simple screen mirroring or a one-size-fits-all controller by enabling dynamic modality switching and adaptability across real-world applications. We are pleased to report that our deployment results demonstrated the system's effectiveness. Feedback from the user study was overwhelmingly positive, with participants adeptly navigating through various views and finding the transitions not only useful but also well-suited for real-time interactions. This outcome aligns with our initial vision of elevating the role of the phone from a mere controller or display mirror to an integral component of the XR experience.

Compared to existing projects [44, 28], our work introduces two key innovations: the use of real-world applications instead of mock-ups, and a seamless transition mechanism between multiple views that aligns with both application states and content. Our evaluations revealed that users had distinct preferences for different views based on the application, offering valuable insights for future development of phone applications within XR environments.

Our system, which focuses on integrating smartphones into XR, fits within a broader multi-device landscape that includes desktops, tablets, and wearables such as smartwatches. The strategies we developed—particularly for transitioning between views and augmenting content within the Phone-XR interface—could be adapted for these devices, opening new avenues for future research to extend our approach to a wider range of platforms.

Additionally, our method for implementing augmented views on phone interfaces could serve as a foundation for enhancing existing 2D applications within XR environments. By analyzing existing UIs and applying augmented content across multiple views, our framework provides a roadmap for seamlessly transitioning these applications into XR settings.

In our preliminary prototype, we implemented a universal solution with one augmented view per application. Future iterations could feature multiple views per application, customized to better understand the application context. There is also potential for user-configurable view settings based on personal preferences. This variability presents an exciting challenge: determining the optimal view setup for each application, which we identify as a key area for future exploration.

In our preliminary prototype, we implemented a universal solution with one augmented view per application. Future iterations could introduce multiple augmented views by further integrating with the application's content and context. There is also potential for user-configurable view settings based on personal preferences. This variability presents an exciting challenge: determining the optimal view setup for each application, which remains a key area for future exploration.

One limitation of our study is the small number of experienced XR professionals in our user study. Due to the difficulty of recruiting such experts, the sample size was relatively small, which may have impacted the generalizability of the results.

Another limitation of our proof-of-concept system is that *Beyond the Phone* 's UI Analysis Server is currently tailored to a specific set of applications, and its 3D asset generation is limited to predefined references. However, we are optimistic about future improvements, as emerging technologies like text-to-3D conversion [31, 19] have the potential to significantly enhance the XR experience by enabling the creation of more dynamic 3D content.[1]

We also anticipate that advancements in AI, particularly in scene understanding and segmentation [35], will further enrich UI context understanding in XR environments. Future work could leverage OS-level metadata to enable seamless recognition and augmentation of phone apps in immersive settings.

7 Conclusion

In this paper, we introduced *Beyond the Phone*, a novel framework that enables seamless integration of smartphones into XR environments through dynamic multi-view transitions for real-world applications. Unlike traditional methods that limit smartphones to screen mirroring or basic controller functionality, *Beyond the Phone* enhances the XR experience by enabling fluid transitions between mirrored, magnified, and augmented views, dynamically adapting to the content and the current state of the application. By leveraging user-centered design and real-time adaptability, our system creates a more immersive and interactive Phone-XR experience.

Beyond the Phone distinguishes itself by offering an adaptable, multi-view interaction model that breaks away from the conventional binary view of smartphones as either simple controllers or mirrored displays in XR. This flexibility opens up new possibilities for richer, more interactive mobile experiences in XR environments, moving beyond the limitations of existing methods.

While *Beyond the Phone* represents a meaningful step forward in integrating smartphones within XR environments, we recognize it as an early stage in addressing this complex challenge. We are excited about future developments that will build on this foundation, contributing to continued innovation in this exciting domain.

References

[1] K. Ahuja, S. Pareddy, R. Xiao, M. Goel, and C. Harrison. Lightanchors: Appropriating point lights for spatially-anchored augmented reality interfaces. In *Proceedings of the 32nd Annual ACM Symposium*

[1] 3D model searches could be improved via advanced Google Search, and generative models [41] may eventually bridge this gap.

on User Interface Software and Technology, UIST '19, p. 189–196, 2019. doi: 10.1145/3332165.3347884 2

[2] R. Arora, R. Habib Kazi, T. Grossman, G. Fitzmaurice, and K. Singh. SymbiosisSketch: Combining 2D & 3D Sketching for Designing Detailed 3D Objects in Situ. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pp. 1–15, 2018. doi: 10.1145/3173574.3173759 2

[3] H. Bai, L. Zhang, J. Yang, and M. Billinghurst. Bringing Full-Featured Mobile Phone Interaction Into Virtual Reality. *Comput. Graph*, 97:42–53, 2021. doi: 10.1016/j.cag.2021.04.004 1, 2, 3, 4, 5, 6

[4] S. Bang and W. Woo. Enhancing the Reading Experience on AR HMDs by Using Smartphones As Assistive Displays. In *2023 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*, pp. 378–386, 2023. doi: 10.1145/3234695.3236361 2

[5] J. M. E. Belo, M. N. Lystbæk, A. M. Feit, K. Pfeuffer, P. Kán, A. Oulasvirta, and K. Grønbæk. AUIT – the Adaptive User Interfaces Toolkit for Designing XR Applications. *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, 2022. doi: 10.1145/3526113.3545651 2

[6] E. Brasier, E. Pietriga, and C. Appert. AR-enhanced Widgets for Smartphone-centric Interaction. In *Proceedings of the 23rd International Conference on Mobile Human-Computer Interaction*, MobileHCI '21, pp. 1–12, 2021. doi: 10.1145/3447526.3472019 2

[7] F. Brudy, C. Holz, R. Rädle, C.-J. Wu, S. Houben, C. N. Klokmose, and N. Marquardt. Cross-device taxonomy: Survey, opportunities and challenges of interactions spanning across multiple devices. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, p. 1–28, 2019. doi: 10.1145/3290605.3300792 1

[8] W. Büschel, K. Krug, K. Klamka, and R. Dachselt. Demonstrating CleAR Sight: Transparent Interaction Panels for Augmented Reality. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI EA '23, pp. 1–5, 2023. doi: 10.1145/3544549.3583891 2

[9] Y. Cheng, Y. Yan, X. Yi, Y. Shi, and D. Lindlbauer. SemanticAdapt: Optimization-Based Adaptation of Mixed Reality Layouts Leveraging Virtual-Physical Semantic Connections. *The 34th Annual ACM Symposium on User Interface Software and Technology*, 2021. doi: 10.1145/3472749.3474750 2, 5

[10] N. Chulpongsatorn, W. Willett, and R. Suzuki. Holotouch: Interacting with mixed reality visualizations through smartphone proxies. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI EA '23, 2023. doi: 10.1145/3544549.3585738 1, 2, 3

[11] A. P. Desai, L. Peña-Castillo, and O. Meruvia-Pastor. A Window to Your Smartphone: Exploring Interaction and Communication in Immersive VR With Augmented Virtuality. In *2017 14th Conference on Computer and Robot Vision (CRV)*, pp. 217–224, 2017. doi: 10.1109/CRV.2017.16 1, 2, 4

[12] J. Grubert, M. Heinisch, A. Quigley, and D. Schmalstieg. MultiFi: Multi Fidelity Interaction with Displays On and Around the Body. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, pp. 3933–3942, 2015. doi: 10.1145/2702123.2702331 1, 2, 4

[13] J. Grubert, T. Langlotz, S. Zollmann, and H. Regenbrecht. Towards Pervasive Augmented Reality: Context-Awareness in Augmented Reality. *IEEE Transactions on Visualization and Computer Graphics*, 23(6):1706–1724, 2016. doi: 10.1109/TVCG.2016.2543720. 2, 5

[14] S. Hubenschmid, J. Zagermann, D. Leicht, H. Reiterer, and T. Feuchtner. ARound the Smartphone: Investigating the Effects of Virtually-Extended Display Size on Spatial Memory. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, pp. 1–15, 2023. doi: 10.1145/3544548.3581438 2

[15] I. Kim, H. Goh, N. Narziev, Y. Noh, and U. Lee. Understanding User Contexts and Coping Strategies for Context-Aware Phone Distraction Management System Design. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol*, 4(4), dec 2020. doi: 10.1145/3432213 6

[16] W. Kim, K. T. W. Choo, Y. Lee, A. Misra, and R. K. Balan. Empath-D: VR-Based Empathetic App Design for Accessibility. In *Proceedings of the 16th Annual International Conference on Mobile Systems, Applications, and Services*, pp. 123–135, 2018. doi: 10.1145/3210240.3211108 3, 4, 5

[17] S. Kyian and R. Teather. Selection performance using a smartphone in vr with redirected input. In *Proceedings of the 2021 ACM Symposium on Spatial User Interaction*, SUI '21, 2021. doi: 10.1145/3485279.3485292 2

[18] Z. Li, M. Annett, K. Hinckley, K. Singh, and D. Wigdor. Holodoc: Enabling mixed reality workspaces that harness physical and digital content. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, p. 1–14, 2019. doi: 10.1145/3290605.3300917 2

[19] C.-H. Lin, J. Gao, L. Tang, T. Takikawa, X. Zeng, X. Huang, K. Kreis, S. Fidler, M.-Y. Liu, and T.-Y. Lin. Magic3d: High-resolution text-to-3d content creation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 300–309, June 2023. 9

[20] D. Lindlbauer, A. M. Feit, and O. Hilliges. Context-aware online adaptation of mixed reality interfaces. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology*, UIST '19, p. 147–160, 2019. doi: 10.1145/3332165.3347945 2, 5

[21] A. Makhsadov, D. Degraen, A. Zenner, F. Kosmalla, K. Mushkina, and A. Krüger. Vrysmart: a framework for embedding smart devices in virtual reality. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI EA '22, 2022. doi: 10.1145/3491101.3519717 2

[22] F. Matulic, A. Ganeshan, H. Fujiwara, and D. Vogel. Phonetroller: Visual Representations of Fingers for Precise Touch Input With Mobile Phones in VR. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, pp. 1–13, 2021. doi: 10.1145/3411764.3445583 1, 2, 3, 4, 5, 6

[23] F. Matulic, T. Kashima, D. Beker, D. Suzuo, H. Fujiwara, and D. Vogel. Above-screen fingertip tracking with a phone in virtual reality. CHI EA '23, 2023. doi: 10.1145/3544549.3585728 2, 4

[24] F. Matulic and D. Vogel. Pen+Touch+Midair: Cross-Space Hybrid Bimanual Interaction on Horizontal Surfaces in Virtual Reality. In *Graphics Interface 2023*, 2023. 2

[25] A. Millette and M. J. McGuffin. DualCAD: Integrating Augmented Reality With a Desktop GUI and Smartphone Interaction. In *2016 IEEE International Symposium on Mixed and Augmented Reality (ISMAR-Adjunct)*, 2016. doi: 10.1109/ISMAR-Adjunct.2016.0030 2

[26] P. Mohr, M. Tatzgern, T. Langlotz, A. Lang, D. Schmalstieg, and D. Kalkofen. TrackCap: Enabling Smartphones for 3D Interaction on Mobile Head-Mounted Displays. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, pp. 1–11, 2019. doi: 10.1145/3290605.3300815 2, 4

[27] M. Nebeling, S. Rajaram, L. Wu, Y. Cheng, and J. Herskovitz. 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, 2021. doi: 10.1145/3411764.3445323 5

[28] E. Normand and M. J. McGuffin. Enlarging a Smartphone With AR to Create a Handheld VESAD (Virtually Extended Screen-Aligned Display). In *2018 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 123–133, 2018. doi: 10.1145/3544548.3581438 1, 2, 3, 4, 9

[29] B. Nuernberger, E. Ofek, H. Benko, and A. D. Wilson. SnapToReality: Aligning Augmented Reality to the Real World. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 2016. doi: 10.1145/2858036.2858250 2

[30] C. Pamparău, O.-A. Schipor, A. Dancu, and R.-D. Vatavu. SAPIENS in XR: operationalizing interaction-attention in extended reality. *Virtual Reality*, 27(3):1765–1781, Sept. 2023. doi: 10.1007/s10055-023-00776-1 5

[31] B. Poole, A. Jain, J. T. Barron, and B. Mildenhall. DreamFusion: Text-to-3D Using 2D Diffusion. *ArXiv Preprint ArXiv:2209.14988*, 2022. doi: 10.48550/arXiv.2209.14988 9

[32] X. Qian, F. He, X. Hu, T. Wang, A. Ipsita, and K. Ramani. Scalar: Authoring semantically adaptive augmented reality experiences in virtual reality. In *Proceedings of the 2022 CHI Conference on Human*

Factors in Computing Systems, CHI '22, 2022. doi: 10.1145/3491102.3517665 2

[33] M. Serrano, B. Ens, X.-D. Yang, and P. Irani. Gluey: Developing a Head-Worn Display Interface to Unify the Interaction Experience in Distributed Display Environments. In *Proceedings of the 17th International Conference on Human-Computer Interaction With Mobile Devices and Services*, pp. 161–171, 2015. doi: 10.1145/2785830.2785838 2

[34] A. Steed and S. Julier. Design and implementation of an immersive virtual reality system based on a smartphone platform. In *2013 IEEE Symposium on 3D User Interfaces (3DUI)*, pp. 43–46, 2013. doi: 10.1109/3DUI.2013.6550195 2

[35] J. Tian, L. Aggarwal, A. Colaco, Z. Kira, and M. Gonzalez-Franco. Diffuse, Attend, and Segment: Unsupervised Zero-Shot Segmentation Using Stable Diffusion. *ArXiv Preprint ArXiv:2308.12469*, 2023. 9

[36] Y. Tian, C.-W. Fu, S. Zhao, R. Li, X. Tang, X. Hu, and P.-A. Heng. Enhancing Augmented VR Interaction via Egocentric Scene Analysis. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol*, 3(3), sep 2019. doi: 10.1145/3351263 2

[37] A. E. Unlu and R. Xiao. PAIR: Phone as an Augmented Immersive Reality Controller. In *Proceedings of the 27th ACM Symposium on Virtual Reality Software and Technology*, VRST '21, pp. 1–6, 2021. doi: 10.1145/3489849.3489878 4

[38] B. Wang, G. Li, and Y. Li. Enabling Conversational Interaction With Mobile UI Using Large Language Models. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, 2023. doi: 10.1145/3544548.3580895 5

[39] B. Wang, G. Li, X. Zhou, Z. Chen, T. Grossman, and Y. Li. Screen2Words: Automatic Mobile UI Summarization With Multimodal Learning. In *The 34th Annual ACM Symposium on User Interface Software and Technology*, UIST '21, p. 498–510, 2021. doi: 10.1145/3472749.3474765 6

[40] Y.-T. Yeh, F. Matulic, and D. Vogel. Phone Sleight of Hand: Finger-Based Dexterous Gestures for Physical Interaction With Mobile Phones. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, pp. 1–19, 2023. doi: 10.1145/3544548.3581121 2

[41] T. Yi, J. Fang, G. Wu, L. Xie, X. Zhang, W. Liu, Q. Tian, and X. Wang. Gaussiandreamer: Fast generation from text to 3d gaussian splatting with point cloud priors. *arXiv preprint arXiv:2310.08529*, 2023. 9

[42] L. Zhang, W. He, H. Bai, Q. Zou, S. Wang, and M. Billinghurst. A hybrid 2d–3d tangible interface combining a smartphone and controller for virtual reality. *Virtual Reality*, 27(2):1273–1291, 2023. doi: 10.1007/s10055-022-00735-2 2

[43] X. Zhang, L. de Greef, A. Swearngin, S. White, K. Murray, L. Yu, Q. Shan, J. Nichols, J. Wu, C. Fleizach, A. Everitt, and J. P. Bigham. Screen Recognition: Creating Accessibility Metadata for Mobile Applications From Pixels. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, 2021. doi: 10.1145/3411764.3445186 6

[44] F. Zhu and T. Grossman. BISHARE: Exploring Bidirectional Interactions Between Smartphones and Head-Mounted Augmented Reality. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, pp. 1–14, 2020. doi: 10.1145/3313831.3376233 1, 2, 3, 4, 5, 9

[45] F. Zhu, Z. Lyu, M. Sousa, and T. Grossman. Touching the droid: Understanding and improving touch precision with mobile devices in virtual reality. In *2022 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 807–816. IEEE, 2022. doi: 10.1109/ISMAR55827.2022.00099 2, 5, 6

[46] F. Zhu, M. Sousa, L. Sidenmark, and T. Grossman. Phoneinvr: An evaluation of spatial anchoring and interaction techniques for smartphone usage in virtual reality. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '24. Association for Computing Machinery, New York, NY, USA, 2024. doi: 10.1145/3613904.3642582 1, 4

[47] H. Zhu, W. Jin, M. Xiao, S. Murali, and M. Li. BlinKey: A Two-Factor User Authentication Method for Virtual Reality Devices. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 4(4), dec 2020. doi: 10.1145/3432217 2