Intent-Driven Input Device Arbitration for XR

Eric J. Gonzalez
Google
Seattle, WA, USA
ejgonz@google.com

Ishan Chatterjee
Google
Seattle, WA, USA
ishanc@google.com

Mar Gonzalez-Franco
Google
Seattle, WA, USA
margon@google.com

Andrea Colaço
Google
Mountain View, CA, USA
andreacolaco@google.com

Karan Ahuja*
Google
Seattle, WA, USA
karanahuja@google.com

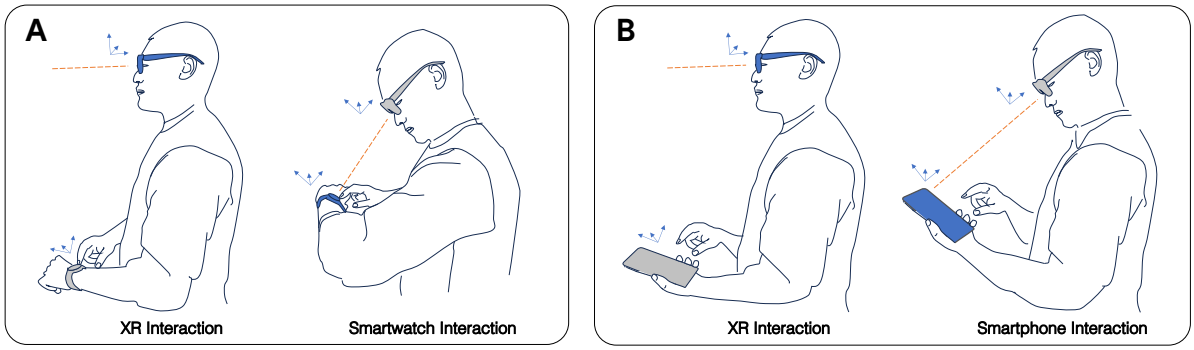

Figure 1: We present a method of device arbitration using inertial measurement units (IMUs) in a head-mounted display and a smartwatch/smartphone. For example, touch inputs to a smartwatch (A) or smartphone (B) can be used to navigate the UI of smartglasses (left), unless the user is looking at their device (right) in which case input is routed as normal to the watch/phone.

ABSTRACT

Interactions with Extended Reality Head-Mounted Displays (XR HMDs) require precise, intuitive, and efficient input methods. Current approaches either rely on power-intensive sensors, such as cameras for hand tracking, or specialized hardware such as controllers. Previous work has explored the use of familiar, available devices such as smartphones and smartwatches as more a more practical input alternative. However, this approach risks interaction overload – how can one determine whether the user's gestures on the watch or phone are directed toward control of the XR device or the mobile device itself? To this end, we propose a novel method for cross-device input arbitration based on the relative orientation between the HMD and target device as measured by on-device IMUs. In a validation study with 6 users, we demonstrate 93.7% accuracy in estimating the intended device of interaction. Our method offers a practical, energy-efficient way to leverage users' existing devices for input and enable seamless cross-device experiences in XR.

CCS CONCEPTS

• **Human-centered computing** → **Mixed / augmented reality**; *Interaction techniques.*

KEYWORDS

Mixed Reality, Cross-device, Multi-device, Input, Arbitration

ACM Reference Format:
Eric J. Gonzalez, Ishan Chatterjee, Mar Gonzalez-Franco, Andrea Colaço, and Karan Ahuja. 2024. Intent-Driven Input Device Arbitration for XR. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3613905.3650758

1 INTRODUCTION

Current extended reality (XR) input methods largely rely on power-intensive, camera-based hand tracking or require the use of specialized controllers, which add cost and reduce convenience. These limitations highlight the need for more practical and efficient input alternatives. Recognizing this, recent work has turned towards leveraging devices which are already an integral part of users' daily lives – specifically, the readily-available touchscreens of smartphones and smartwatches [6, 10, 13, 20]. This shift towards integrating ubiquitous user devices presents a novel opportunity for facilitating interaction within XR environments.

However, this approach introduces a critical challenge: discerning whether a user's gestures on a smartwatch or smartphone are

*Also with Northwestern University, Evanston, IL, USA.

CHI EA '24, May 11–16, 2024, Honolulu, HI, USA

ACM ISBN 979-8-4007-0331-7/24/05
https://doi.org/10.1145/3613905.3650758

meant to control the device itself or to interact with the XR application. Prior work has shown that a users' gaze directed at their smart device is a powerful indicator of their intent to interact [7, 12, 14, 19]. By determining if the user is actively looking at the mobile device or not, we can estimate the intended device of interaction. As a consequence, rather than requiring the user to lift their hand to tap and swipe the temple-mounted touchpad featured on a number of smartglasses [4, 5], the user can use their smartwatch for the same function in an eyes-free manner. If they intend for their interactions to control the smartwatch, rather than the smartglasses, they can simply look at their smartwatch (Figure 1). By integrating this cross-device input routing and device arbitration mechanism, we aim to transform user interactions within XR spaces.

To achieve this, we propose a novel framework that leverages the inertial measurement units (IMUs) present in users' everyday devices and XR head-mounted displays (HMDs). IMUs consume significantly less power than HMD vision systems, operating under a single milliamp even with continuous sampling [2] and requiring much less complex processing. Furthermore, IMUs are often always-on for these devices anyway – for HMDs it is necessary for head tracking and for phones/watches it is used for longitudinal activity recognition, like step counting [15, 17]. Finally, IMUs are privacy preserving and can work across all lighting conditions.

Our method hinges on analyzing the relative orientation between the XR headset and the mobile input device (either smartphone or smartwatch). We evaluate the efficacy of our system through a 6 person user study across a variety of contexts: sitting, standing, walking, and lounging. Our decision tree-based model has an arbitration accuracy of 93.7%. We further create three demonstrative applications to showcase the real-world utility of our approach.

2 RELATED WORK

2.1 Traditional XR Input

A wide variety of input and interaction methods have been proposed for XR HMDs. Most immersive virtual reality (VR) applications rely on vision-based, free-space hand-tracking, which can be tiring for extended use due to lack of mechanical grounding [11], or require specialized 6DOF controller hardware [1, 3].

In contrast, most augmented reality (AR) smartglasses do not have the power budget or headset real-estate for sensor instrumentation. Further, their on-the-go nature precludes carrying custom controllers. As a result, the use of a 1D or 2D touchpad on the side of the AR smartglasses [4, 5] has been explored as an input device. However, a major limitation is the ergonomics and social acceptability of lifting one's hand to interact with the touchpad. In fact, some studies in social XR have shown users using the lifted finger as a rude gesture [16]. Therefore, for both immersive HMDs and smartglasses, making use of existing ubiquitous devices (e.g., phones, watches) has been an active area of exploration to (1) expand interaction possibilities, (2) allow for greater comfort and social acceptability, and (3) allow for lower power operation without the needing additional hardware.

2.2 Cross-Device Interactions in XR

Given their ubiquity and precising sensing capabilities, there has been substantial research exploring how personal devices such as smartphones and smartwatches might be used to broaden the interactive canvas of XR.

With BiShare [20], Zhu & Grossman proposed a set of interaction techniques that leverage touch gestures and positioning of the phone to act as controller and display space for the headset, including the ability to capture input and transfer content. In the same vein, MultiFi [10] integrated various smart display form factors – watches, tablets, phones – by dynamically distributing interface components across each display. The always-available touch surface afforded by smartwatches has also been explored for XR input. Lang et al [13] used the touchscreen, bezel, and gesturing on a smartwatch to perform navigation and 3D modeling, while Ahn et al [6] explored the use of smartwatches for XR text entry.

While prior works make use of these mobile devices (phones and watches) for XR interactions, none address the fact that these devices still need to simultaneously function for their primary use-case in addition to being an accessory to the HMD. For this to occur, an arbitration scheme is needed to determine intended target of interaction. Our work seeks to address that need.

3 SYSTEM IMPLEMENTATION

Our system consists of a real-time sensor streaming framework and machine learning (ML) classifier which continuously estimates the user's device of intended interaction.

3.1 Sensor Streaming and Communication

We leverage the Cross-Device Toolkit (XDTK) [9] to stream sensor data from multiple Android devices to a central XR application built in Unity. The Unity application connects to the mobile devices on the same local network in a server-client architecture using UDP. Once connected, each mobile device starts transmitting its IMU (orientation, accelerometer, gyroscope and magnetometer) sensor data to the Unity server. The smartwatch and smartphone log and stream IMU data at 80 and 100 Hz respectively. Similarly, the headset samples its IMU-inferred orientation data at 60 Hz.

3.2 Intent-Driven Device Arbitration

The device targeted for interaction can either be the XR headset or one of the mobile devices – the smartwatch or the smartphone. We estimate this intended device engagement by modeling where the user is looking. If the user is looking at the mobile device, it is the intended device of interaction; else it is the XR headset. Note, without any spatial tracking information (akin to 6 DoF pose from controllers or vision based methods), our problem is inherently under-constrained in nature. That is, for any given orientation pairs of the headset and mobile device, multiple plausible solutions exist. Thus, our model needs to contend with this solution ambiguity and offer the best arbitration estimate. Furthermore, existing IMU-based algorithms such as tilt-to-wake (TTW) on the watch or watch are themselves insufficient to determine the locus of intention as the user may perform the same arm-lifting motion to provide input on the device screen in both touch device input-intent or XR input-intent scenarios. The additional information supplied by headset IMU helps avoid this common failure case.

To model what the user is actively looking at, we make use of the relative orientation between the mobile device and the XR headset,

captured via the IMU sensors of each device. In particular, we make use of the following features:

- 3D rotation of the mobile device represented as a quaternion
- Acceleration along X, Y and Z axis of the mobile device
- 3D rotation of the XR headset represented as a quaternion
- Difference in yaw, pitch and roll between the mobile device and the XR headset

Each instance of IMU data from the mobile device and the XR headset results in a length 14 feature vector, which serves as the input to a machine learning (ML) model. For our ML model, we make use of sklearn's Decision Tree Classifier (default parameters) [18] with a maximum depth of 3 to avoid overfitting. We train two Decision Tree Classifiers: one to disambiguate between phone and XR headset, and the other to disambiguate between watch and XR headset. Once trained, we export the rules of the decision tree from Python to Unity for real-time inference and integration into XR experiences. Our tree-based classifier also aided in easier debugging due to its interpretability.

4 EVALUATION

4.1 Data Collection

For our data collection, we utilized two types of mobile devices: a smartphone (Google Pixel 6 Pro) and a smartwatch (Google Pixel Watch 1st gen) in conjunction with a XR headset (Meta Quest 3). We collected data across six participants. Four of these participants wore the watch on the left hand, and two on the right hand. The data collection process was structured around three distinct conditions, each its own session: looking at the watch, looking at the phone, and interacting with the XR environment without looking at either device. To capture a broad range of user interactions, we conducted the data collection in four different contexts: sitting, standing, walking, and reclining on a chair (lounging).

During each session, the features described in Section 3 were logged from Unity at a rate of 60 Hz. Within each condition (look-at-watch, look-at-phone, XR interaction), participants performed the required interaction in each context (sitting, standing, walking, lounging) for approximately one minute each. Therefore, we recorded approximately four minutes of data for each condition per participant. Participants were given a 2 minute break between conditions. Data was logged continuously during each condition and all contexts were grouped together in subsequent analysis. Each datapoint was labeled according to its condition and treated as an independent sample. No additional data processing was performed.

Importantly, during data collection participants were instructed to behave naturally and look at the devices as they would in their daily lives, without any exaggerated or artificial movements. Participants were free to change the hand they were holding the smartphone in (left vs right) and were also free to change their arm positions as they saw fit. This approach was crucial to ensure that the data collected reflected realistic user behaviors and interactions, thereby enhancing the relevance and applicability of our model in actual XR settings.

Figure 2: A user exercising in a virtual nature trail (left) can view their fitness tracking data when looking at their smartwatch via dynamic passthrough (right).

4.2 Results

We evaluated the performance of our arbitration model on the user study data we captured. Specifically, we performed a leave one-user out cross validation. In this process, we train a Decision Tree Classifier on the 14 IMU features from five of our participants and test on the holdout sixth participant (all combinations, results averaged). The Decision Tree arbitration model between the smartwatch and the XR headset achieved a mean accuracy of 96.1% (SD = 3.4). The smartphone and XR headset arbitration achieved a slightly lower accuracy of 91.4% (SD = 4.8). This accuracy drop can be attributed to the higher degrees of freedom of the smartphone in a user's hand, as compared to a smartwatch, the position of which is limited by the arm's range of motion. We further tested other ML models such as Ensemble based classifiers (Random Forest) and Neural Network based methods, they achieved a comparable accuracy.

Our end-to-end pipeline runs at a frame rate of 60 Hz and has a mean latency of 32 ms from captured photo to output device arbitration prediction. This frame rate enables a variety of real-time XR applications as detailed in Section 5. Our arbitration classification modules for both watch and phone are very lightweight, running well above the sampling rate of the respective IMUs. Thus, the bulk of the latency comes from the rendering and display of the XR scene in Unity which runs at 60 Hz.

5 APPLICATIONS

Our method can determine whether the user is looking at their mobile device with 93.7% accuracy. This heuristic can be applied to a number of XR cross-device applications:

5.1 Dynamic Passthrough

Mixed reality HMDs can fully envelope the user in a virtual environment or allow for spatial awareness using passthrough. By utilizing our method, passthrough can be dynamically enabled when the user wishes to look at their mobile device. For example, if a notification arrives on the user's smartwatch or smartphone, the user simply needs to look at their device to activate passthrough, removing the need to take off their headset. Or if the user is in an immersive exercise experience, they can simply look at their watch to see their Fitbit activity metrics (see Fig. 2).

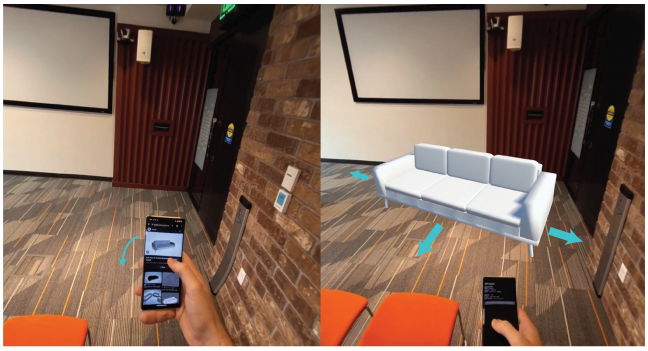

Figure 3: A user browsing furniture on their phone (left) can seamlessly instantiate a virtual sofa by lowering their phone attending to their environment. Their device automatically begins routing input to the HMD, allowing the user to position the sofa within their space using touch controls (right).

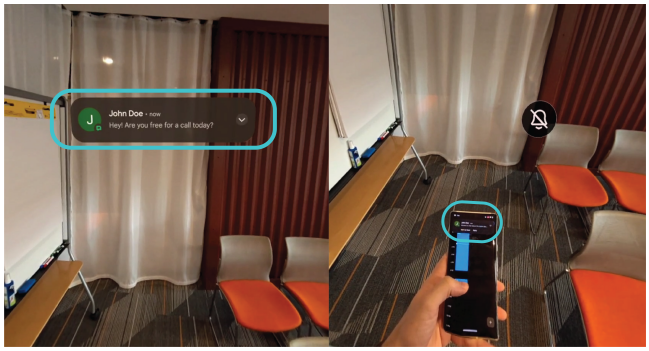

Figure 4: A user receives a notification in XR (left). While using their phone to respond, additional notifications are silenced on the HMD and routed directly to the phone (right).

5.2 Input Switching

The always-available touch surfaces on smartwatch and smartphone can provide grounded, precise interactions. In this mode, the touch device acts as an eyes-free, cross-device controller for the headset. For smartglasses, even the small touch area of a smartwatch can be used to enable UI navigation, avoiding the need to reach up to the touchpad on the glasses temple. If the user wishes to return to the primary function of their watch, they can look at it, returning the device to the main system experience.For immersive headsets, experiences can be enhanced with cross-device interaction. For example, the phone can be used as an interactive picker and manipulation surface for 3D objects. To switch between modes, the user may look at the phone to select the item they wish to view in AR, and then use the phone to manipulate the objects as they view it in AR via their headset (see Fig. 3).

5.3 Notification Routing

Our technique can be used to route alerts to actively attended devices, and suppress others. To stay informed on incoming messages, the user may choose to receive notifications via their smartglass display. As the user is notified within their glanceable field-of-view, they do not require their other devices to actively alert them to the incoming notification. Likewise, if the user is actively looking at their phone or smartwatch, a notification delivered on either device is sufficient to alert them, and the alert on the smartglass display can be suppressed (see Fig. 4).

6 LIMITATIONS AND FUTURE WORK

While the our system's accuracy is promising, there are several key limitations that will need to be overcome before it is ready for commercial adoption. First is the dataset. While we collected from six participants in lab settings, future work should extend this to a larger and more diverse participant pool encapsulating the variability of in-the-wild cross-device scenarios. More data would also help to improve the accuracy and robustness of our system.

Our current implementation is based on a snapshot of IMU data and does not account for any temporal consistencies. Prior work, such as IMUPoser [15], indicates that temporal inertial data can be instrumental in deducing spatial body pose, which in turn could reduce ambiguity and improve device arbitration accuracy. Furthermore, in addition to IMUs we can also explore other low-power sensors, such as UWB-based distance estimates between devices [8] to improve the robustness of our algorithm.

In this work we show a method to understand which device a user is looking at. Incorporating this utility into existing interaction technique frameworks such as BISHARE [20] or MutliFi [10] can help to make the concepts demonstrated in these works practically realizable, and expand the interactive potential of this method.

Lastly, in the future we would like to combine our approach with other heuristics that could aid in arbitration, such as screen state, UI content, tilt-to-wake motions, or gating touch gestures. These factors could offer additional context and enhance the ability to accurately discern user intentions. Incorporating these elements into future research could lead to a more context-aware approach for cross-device input management in XR environments.

7 CONCLUSION

We propose a framework for determining whether device-driven interactions in cross-device XR are intended for the HMD or the mobile device itself. For this we make use of the power-efficient IMUs on the respective devices, and model the relative orientation between them. Our user study, involving six participants, demonstrates the efficacy of our method, achieving an arbitration accuracy of 93.7% across participants, in a variety of contexts of use (sitting, standing, walking, and lounging). Our method offers a practical, energy-efficient way to leverage users' existing devices for input and better enable seamless cross-device experiences in XR.

REFERENCES

[1] [n. d.]. HTC VIVE. https://www.vive.com/us/ Accessed: 2024-01-25.
[2] [n. d.]. LSM6DSOX Inertial Module. https://www.st.com/en/mems-and-sensors/lsm6dsox.html Accessed: 2024-01-25.
[3] [n. d.]. Meta Quest. https://www.meta.com/quest/quest-3/ Accessed: 2024-01-25.
[4] [n. d.]. Vuzix Blade. https://www.vuzix.com/products/vuzix-blade-smart-glasses-upgraded Accessed: 2024-01-25.
[5] 2023. New Ray-Ban Meta Smart Glasses. https://about.fb.com/news/2023/09/new-ray-ban-meta-smart-glasses/ Accessed: 2024-01-25.
[6] Sunggeun Ahn, Seongkook Heo, and Geehyuk Lee. 2017. Typing on a smartwatch for smart glasses. In *Proceedings of the 2017 ACM International Conference on*

Interactive Surfaces and Spaces. 201–209.

[7] Rory M.S. Clifford, Nikita Mae B. Tuanquin, and Robert W. Lindeman. 2017. Jedi ForceExtension: Telekinesis as a Virtual Reality interaction metaphor. In *2017 IEEE Symposium on 3D User Interfaces (3DUI)*. 239–240. https://doi.org/10.1109/3DUI.2017.7893360

[8] Nathan DeVrio, Vimal Mollyn, and Chris Harrison. 2023. SmartPoser: Arm Pose Estimation with a Smartphone and Smartwatch Using UWB and IMU Data. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. 1–11.

[9] Eric J. Gonzalez, Khushman Patel, Karan Ahuja, and Mar Gonzalez-Franco. 2024. XDTK: A Cross-Device Toolkit for Input & Interaction in XR. In *2024 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*.

[10] Jens Grubert, Matthias Heinisch, Aaron Quigley, and Dieter Schmalstieg. 2015. 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*. 3933–3942.

[11] Ken Hinckley, Randy Pausch, John C. Goble, and Neal F. Kassell. 1994. A survey of design issues in spatial input. In *Proceedings of the 7th Annual ACM Symposium on User Interface Software and Technology* (Marina del Rey, California, USA) *(UIST '94)*. Association for Computing Machinery, New York, NY, USA, 213–222. https://doi.org/10.1145/192426.192501

[12] Mikko Kytö, Barrett Ens, Thammathip Piumsomboon, Gun A. Lee, and Mark Billinghurst. 2018. Pinpointing: Precise Head- and Eye-Based Target Selection for Augmented Reality. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (<conf-loc>, <city>Montreal QC</city>, <country>Canada</country>, </conf-loc>) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3173574.3173655

[13] Matěj Lang, Clemens Strobel, Felix Weckesser, Danielle Langlois, Enkelejda Kasneci, Barbora Kozlíková, and Michael Krone. 2023. A multimodal smartwatch-based interaction concept for immersive environments. *Computers & Graphics* 117 (2023), 85–95.

[14] Joseph J. LaViola, Ernst Kruijff, Ryan P. McMahan, Doug A. Bowman, and Ivan Poupyrev. 2017. *3D user interfaces: Theory and practice*. Addison-Wesley.

[15] Vimal Mollyn, Riku Arakawa, Mayank Goel, Chris Harrison, and Karan Ahuja. 2023. IMUPoser: Full-Body Pose Estimation using IMUs in Phones, Watches, and Earbuds. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–12.

[16] Fares Moustafa and Anthony Steed. 2018. A longitudinal study of small group interaction in social virtual reality. In *Proceedings of the 24th ACM Symposium on Virtual Reality Software and Technology* (Tokyo, Japan) *(VRST '18)*. Association for Computing Machinery, New York, NY, USA, Article 22, 10 pages. https://doi.org/10.1145/3281505.3281527

[17] Jun-geun Park, Ami Patel, Dorothy Curtis, Seth Teller, and Jonathan Ledlie. 2012. Online pose classification and walking speed estimation using handheld devices. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing* (Pittsburgh, Pennsylvania) *(UbiComp '12)*. Association for Computing Machinery, New York, NY, USA, 113–122. https://doi.org/10.1145/2370216.2370235

[18] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[19] Ken Pfeuffer, Benedikt Mayer, Diako Mardanbegi, and Hans Gellersen. 2017. Gaze + pinch interaction in virtual reality. In *Proceedings of the 5th Symposium on Spatial User Interaction* (Brighton, United Kingdom) *(SUI '17)*. Association for Computing Machinery, New York, NY, USA, 99–108. https://doi.org/10.1145/3131277.3132180

[20] Fengyuan Zhu and Tovi Grossman. 2020. Bishare: Exploring bidirectional interactions between smartphones and head-mounted augmented reality. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.