# Practical Input Considerations for Wearable AI Assistants in XR

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

## ABSTRACT

Recent advances in extended reality (XR) and multimodal AI have enabled new opportunities for wearable assistants. This paper discusses practical input considerations for designing such systems, emphasizing the need for intuitive, context-aware interactions that do not sacrifice convenience, comfort, or social acceptability. Specifically, we emphasize the value of multimodality (e.g., voice, gaze, gesture), leveraging users' existing smart devices (e.g., smartphone, smartwatch), making use of real-world surfaces, and enabling discreetness for wearable AI inputs. This work is intended to facilitate discussion and guide future research toward more effective interactions with wearable AI assistants in XR.

## CCS CONCEPTS

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

## KEYWORDS

Extended Reality, Input, Wearables, Assistants

**ACM Reference Format:**
Eric J. Gonzalez. 2025. Practical Input Considerations for Wearable AI Assistants in XR. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems, April 26 – May 2 2025, Yokohama, Japan.* ACM, New York, NY, USA, 3 pages.

## 1 INTRODUCTION

The rise of real-time, interactive AI is fundamentally changing how we interact with computers. Multi-modal large language models (LLMs) are now able to engage in search-grounded Q&A, write functioning code, analyze hundreds of documents, reason about images and videos, operate traditional computing devices, and more – all in response to simple queries from the user.

In parallel with these capabilities, computing platforms themselves have also been evolving. Over the last several years, extended reality (XR) systems have advanced to the point where seamless, context-aware blending of digital content into the real world has become feasible. While XR has not yet experienced widespread adoption on the scale of smartphones, improvements in display resolution, spatial tracking, and natural input have made it uniquely well-positioned to leverage these powerful AI models in ways not possible through a laptop or phone [24, 43].

Nearly all XR devices sense the world around the user. When coupled with real-time multi-modal AI, the are many opportunities for meaningful day-to-day impact: contextual information about anything in front of you, live audio descriptions, memory aids, etc. This is without considering the potential of visual outputs optimized for XR: in-world navigation, contextual overlays, semantic highlighting, environment augmentation, and more. We have already seen the first steps towards XR-AI systems in the form of industry research prototypes (e.g., Google's Project Astra [1], Meta's Orion [4], Snap Spectacles [3]) as well as consumer products (e.g., Ray-Ban Stories [2]).

Most of these early spatially-aware AI interfaces can be viewed as "assistants", available to provide on-demand information, help with tasks, and act as external memory (e.g., "Where did I leave my keys?"). The focus of such interfaces is on parsing multi-modal input from the user and their environment to provide simple but helpful responses and/or digital actions. While there is clearly interest in such systems within both industry and academia [9, 13, 29, 42], it remains to be seen if and how they will be adopted by users.

One important factor influencing technological adoption is how effectively users can communicate their intent to the system [14, 19]. Just as the mouse helped unlock the potential of PCs by enabling simple, precise, and direct interactions, the inputs for AI-powered XR (XR-AI) must also enable interactions that feel intuitive and effective in order to get the most out of wearable assistants. This raises the following important question:

> How can we enable users to **precisely** convey intent to wearable AI assistants in a **convenient**, **comfortable**, and **socially acceptable** way?

As a step towards addressing this question, this work presents practical considerations for designing XR-AI inputs inspired by recent research. Here we take a more holistic view of Human-AI Interaction [6], rather than focusing on technical details such as tracking requirements, scene understanding, or model architecture. While not exhaustive, this work aims to provide researchers value for future discussions on XR-AI inputs.

## 2 CONSIDERATIONS

### 2.1 Support Broad Multimodality

The first commercially available wearable AI assistants are likely to focus on voice interactions: world-facing cameras facing provide the assistant with context, while user input is provided through explicit voice queries. While this approach is certainly powerful (and requires little additional instrumentation), it does not capture many of the rich nuances with which people convey information when discussing the world around them. Eye gaze, head gaze, and pointing gestures, for example, provide the foundations for establishing *joint attention* [31] – the shared state in which two or more

*CHI Workshops '25, April 26 – May 2 2025, Yokohama, Japan*

individuals are intentionally focused on the same thing. This is critical for resolving ambiguities in queries about the world (e.g., "What is *that*?") as well as commands ("Put-that-there" [7]). Without such spatial disambiguation capabilities, the user would likely need to be overly verbose in their queries, which can be frustrating and impractical [32]. Recently, Lee et al. demonstrated how gaze and pointing can be integrated with XR voice assistant interactions in *GazePointAR* [29]. Moreover, with precise eye-tracking, AI assistants in XR can gain additional context about what the user is reading in order to better answer their questions [8]. Additional cues such as posture, facial expressions, and more can further enrich interactions by conveying emotion or emphasis that voice alone might miss. As previous research highlights, the key challenge in leveraging these modalities is how to efficiently map them to inputs a multimodal LLM can robustly interpret.

## 2.2 Leverage Existing Devices

AI-enabled XR devices are entering into an existing ecosystem of personal devices (e.g., smartphones, smartwatches). Rather than rely entirely on the sensing and computation capabilities of the XR devices themselves (which may be limited for certain form-factors), an alternative approach is to make use of the powerful devices we already have. For example, smartphones offers a readily available high-precision multi-touch surface, precise rotational input via IMU, and even 6DOF tracking (e.g., via ARCore[1]). Combined with simple pointing and selection mechanisms, such capabilities may enable a good portion of the spatial interactions enabled by hand tracking (e.g., raycasting) while requiring significantly less power. As we move towards lower-profile, sleeker, all-day wearables, this trade-off may be acceptable to the user. Several recent projects have explored the use of personal devices for input in XR [21, 22], both to extend the capabilities of traditional controllers (e.g., with multi-touch gestures [44]) and to improve ergonomics (e.g., limited space [28]). Specifically for wearable AI assistants, smartphones can be used to more precisely indicate areas of interest (e.g., by controlling a cursor), provide additional vantage points via camera, enable familiar swipe and multi-touch gestures, and allow for simple and subtle text communication with the world-grounded agent.

## 2.3 Repurpose Objects & Surfaces

With AI-powered XR headsets, everyday objects and surfaces can possibly be repurposed as interactive canvases and contextual anchors for wearable assistants. Tabletops and walls are ubiquitous and provide wide areas of passive haptic feedback for touch interactions, as well as ergonomic support that enables greater precision [10]. Unique object affordances can be leveraged to generate ad-hoc user interfaces [15], further enriching the interaction experience. As the real-time environmental and spatial awareness of AI assistants improves, it is reasonable to envision a system that overlays contextual information onto a wall or table while simultaneously capturing user touches or gestures on that surface. Even solely as an input, a surface-anchored touch interface may offer more convenience than voice input and be less fatiguing than mid-air gestures [10]. Recent research has shown significant progress in capturing touch interactions on everyday surfaces via XR [16, 37], though considerable work remains to achieve the precision and robustness users expect. By incorporating sensing data from existing wearables like smartwatches [30], it may be possible to further enhance on-surface gesture recognition.

## 2.4 Enable Discreetness

For users to benefit from wearable assistants, they must be comfortable interacting with them. It is well-known that users are reluctant to interact with voice assistants in public due to privacy and social acceptability concerns [17, 36]. Similarly, mid-air gestures can be seen as obtrusive [25, 38], in addition to being fatiguing [23, 35]. While the alternative input methods described above (e.g., existing devices) can offer discretion and precision, it is worth exploring how speech and gestures may be used in more acceptable ways. Enabling speech is particularly desirable both for accessibility reasons and because it is a high-bandwidth form of communication. One explored approach has been the enabling of "silent speech" recognition [11, 20], allowing users to speak in a whisper or lower volume while still being interpreted by the system. Most silent speech interfaces have required cumbersome sensors (e.g., electromagnetic articulography [18], ultrasound [12], custom microphones [33]), though in recent years advances in machine learning have led to more practical approaches such as detecting silent speech via earbuds ([26, 41]) or lip-reading via computer vision [5]. Further research is needed, but such approaches may be more compatible with XR headsets or glasses, providing a path towards more discreet voice interactions in XR. Combining this with microgestures [27] and touch interactions enabled by existing devices, it may be possible to create a unified interaction vocabulary for wearable XR-AI assistants that is effective, discreet, and comfortable to users.

# 3 CONCLUSION & FUTURE DIRECTIONS

In this work, we present a number of input considerations for wearable AI assistants in XR, aimed at conveying intent while maintaining user convenience and comfort. We highlight that supporting broadly multimodal input (such as voice + gaze + touch) better mirrors principles of human-to-human communication (e.g., joint attention). We discuss how existing devices like smartphones can be leveraged as precision tools for interacting with XR-AI to compliment natural inputs (e.g., voice). Furthermore, we highlight that surfaces should similarly be considered as canvases for input, given their ubiquity and ergonomic benefits. Finally, we argue that input systems for XR-AI should support discreet interactions for improved social acceptability.

To achieve these goals, of course, significant research is needed to further optimize multimodal agents, improve XR perception systems and semantic understanding, and develop more efficient architectures which can better support the computation required for many of the above considerations. Critically, privacy and ethical concerns must also be addressed [34, 40] (e.g., bystander privacy, transparency, excessive data collection, superrealism [39]). Overall, it is exciting to see the potential of XR-AI and wearable assistants come into clearer view over the past few years. We hope this work sparks meaningful discussion amongst researchers on effective input design for XR-AI.

[1] ARCore: https://developers.google.com/ar

# REFERENCES

[1] [n. d.]. Project Astra. https://deepmind.google/technologies/project-astra/

[2] [n. d.]. Ray-Ban Stories. https://www.ray-ban.com/usa/discover-ray-ban-meta-smart-glasses/clp

[3] [n. d.]. Snap Spectacles. https://www.spectacles.com/

[4] 2024. Orion: True AR Glasses Have Arrived. https://www.meta.com/blog/orion-ar-glasses-augmented-reality/

[5] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. 2018. Deep lip reading: a comparison of models and an online application. *arXiv preprint arXiv:1806.06053* (2018).

[6] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. 2019. Guidelines for human-AI interaction. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–13.

[7] Richard A Bolt. 1980. "Put-that-there" Voice and gesture at the graphics interface. In *Proceedings of the 7th annual conference on Computer graphics and interactive techniques*. 262–270.

[8] Riccardo Bovo, Steven Abreu, Karan Ahuja, Eric J Gonzalez, Li-Te Cheng, and Mar Gonzalez-Franco. 2024. Embardiment: an embodied ai agent for productivity in xr. *arXiv preprint arXiv:2408.08158* (2024).

[9] Sonia Castelo, Joao Rulff, Erin McGowan, Bea Steers, Guande Wu, Shaoyu Chen, Iran Roman, Roque Lopez, Ethan Brewer, Chen Zhao, et al. 2023. Argus: Visualization of ai-assisted task guidance in ar. *IEEE Transactions on Visualization and Computer Graphics* 30, 1 (2023), 1313–1323.

[10] Yi Fei Cheng, Tiffany Luong, Andreas Rene Fender, Paul Streli, and Christian Holz. 2022. ComforTable user interfaces: Surfaces reduce input error, time, and exertion for tabletop and mid-air user interfaces. In *2022 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, 150–159.

[11] Bruce Denby, Tanja Schultz, Kiyoshi Honda, Thomas Hueber, Jim M Gilbert, and Jonathan S Brumberg. 2010. Silent speech interfaces. *Speech Communication* 52, 4 (2010), 270–287.

[12] Bruce Denby and Maureen Stone. 2004. Speech synthesis from real time ultrasound images of the tongue. In *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 1. IEEE, I–685.

[13] Mustafa Doga Dogan, Eric J Gonzalez, Karan Ahuja, Ruofei Du, Andrea Colaço, Johnny Lee, Mar Gonzalez-Franco, and David Kim. 2024. Augmented Object Intelligence with XR-Objects. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*. 1–15.

[14] Paul Dourish. 2001. *Where the action is: the foundations of embodied interaction*. MIT press.

[15] Ruofei Du, Alex Olwal, Mathieu Le Goc, Shengzhi Wu, Danhang Tang, Yinda Zhang, Jun Zhang, David Joseph Tan, Federico Tombari, and David Kim. 2022. Opportunistic interfaces for augmented reality: Transforming everyday objects into tangible 6dof interfaces using ad hoc ui. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. 1–4.

[16] Camille Dupré, Caroline Appert, Stéphanie Rey, Houssem Saidi, and Emmanuel Pietriga. 2024. TriPad: Touch Input in AR on Ordinary Surfaces with Hand Tracking Only. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–18.

[17] Aarthi Easwara Moorthy and Kim-Phuong L Vu. 2014. Voice activated personal assistant: Acceptability of use in the public space. In *International conference on human Interface and the Management of Information*. Springer, 324–334.

[18] Michael J Fagan, Stephen R Ell, James M Gilbert, E Sarrazin, and Peter M Chapman. 2008. Development of a (silent) speech recognition system for patients following laryngectomy. *Medical engineering & physics* 30, 4 (2008), 419–425.

[19] Shabnam FakhrHosseini, Kathryn Chan, Chaiwoo Lee, Myounghoon Jeon, Heesuk Son, John Rudnik, and Joseph Coughlin. 2024. User adoption of intelligent environments: A review of technology adoption models, challenges, and prospects. *International Journal of Human–Computer Interaction* 40, 4 (2024), 986–998.

[20] Masaaki Fukumoto. 2018. Silentvoice: Unnoticeable voice input by ingressive speech. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*. 237–246.

[21] 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*. 1–5.

[22] 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)*.

[23] Jeffrey T Hansberger, Chao Peng, Shannon L Mathis, Vaidyanath Areyur Shanthakumar, Sarah C Meacham, Lizhou Cao, and Victoria R Blakely. 2017. Dispelling the gorilla arm syndrome: the viability of prolonged gesture interactions. In *Virtual, Augmented and Mixed Reality: 9th International Conference, VAMR 2017, Held as Part of HCI International 2017, Vancouver, BC, Canada, July 9-14, 2017, Proceedings 9*. Springer, 505–520.

[24] Teresa Hirzle, Florian Müller, Fiona Draxler, Martin Schmitz, Pascal Knierim, and Kasper Hornbæk. 2023. When xr and ai meet-a scoping review on extended reality and artificial intelligence. In *Proceedings of the 2023 CHI conference on human factors in computing systems*. 1–45.

[25] Yi-Ta Hsieh, Antti Jylhä, Valeria Orso, Luciano Gamberini, and Giulio Jacucci. 2016. Designing a willing-to-use-in-public hand gestural interaction technique for smart glasses. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 4203–4215.

[26] Yincheng Jin, Yang Gao, Xuhai Xu, Seokmin Choi, Jiyang Li, Feng Liu, Zhengxiong Li, and Zhanpeng Jin. 2022. EarCommand: " Hearing" Your Silent Speech Commands In Ear. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 2 (2022), 1–28.

[27] Chirag Kandoi, Changsoo Jung, Sheikh Mannan, Hannah VanderHoeven, Quincy Meisman, Nikhil Krishnaswamy, and Nathaniel Blanchard. 2023. Intentional microgesture recognition for extended human-computer interaction. In *International Conference on Human-Computer Interaction*. Springer, 499–518.

[28] Mohamed Kari and Christian Holz. 2023. Handycast: Phone-based bimanual input for virtual reality in mobile and space-constrained settings via pose-and-touch transfer. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–15.

[29] Jaewook Lee, Jun Wang, Elizabeth Brown, Liam Chu, Sebastian S Rodriguez, and Jon E Froehlich. 2024. GazePointAR: A context-aware multimodal voice assistant for pronoun disambiguation in wearable augmented reality. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–20.

[30] Manuel Meier, Paul Streli, Andreas Fender, and Christian Holz. 2021. TapID: Rapid touch interaction in virtual reality using wearable sensing. In *2021 IEEE Virtual Reality and 3D User Interfaces (VR)*. IEEE, 519–528.

[31] Chris Moore, Philip J Dunham, and Phil Dunham. 2014. *Joint attention: Its origins and role in development*. Psychology Press.

[32] Chelsea Myers, Anushay Furqan, Jessica Nebolsky, Karina Caro, and Jichen Zhu. 2018. Patterns for how users overcome obstacles in voice user interfaces. In *Proceedings of the 2018 CHI conference on human factors in computing systems*.

[33] Yoshitaka Nakajima, Hideki Kashioka, Kiyohiro Shikano, and Nick Campbell. 2003. Non-audible murmur recognition input interface using stethoscopic microphone attached to the skin. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, Vol. 5. IEEE, V–708.

[34] Joseph O'Hagan, Pejman Saeghe, Jan Gugenheimer, Daniel Medeiros, Karola Marky, Mohamed Khamis, and Mark McGill. 2023. Privacy-enhancing technology and everyday augmented reality: Understanding bystanders' varying needs for awareness and consent. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 4 (2023), 1–35.

[35] Eduardo GQ Palmeira, Alexandre Campos, Ígor A Moraes, Alexandre G de Siqueira, and Marcelo GG Ferreira. 2023. Quantifying the 'Gorilla Arm' Effect in a Virtual Reality Text Entry Task via Ray-Casting: A Preliminary Single-Subject Study. In *Proceedings of the 25th Symposium on Virtual and Augmented Reality*. 274–278.

[36] Laxmi Pandey, Khalad Hasan, and Ahmed Sabbir Arif. 2021. Acceptability of speech and silent speech input methods in private and public. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–13.

[37] Mark Richardson, Fadi Botros, Yangyang Shi, Pinhao Guo, Bradford J Snow, Linguang Zhang, Jingming Dong, Keith Vertanen, Shugao Ma, and Robert Wang. 2024. StegoType: Surface Typing from Egocentric Cameras. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*. 1–14.

[38] Marcos Serrano, Barrett M Ens, and Pourang P Irani. 2014. Exploring the use of hand-to-face input for interacting with head-worn displays. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 3181–3190.

[39] Mel Slater, Cristina Gonzalez-Liencres, Patrick Haggard, Charlotte Vinkers, Rebecca Gregory-Clarke, Steve Jelley, Zillah Watson, Graham Breen, Raz Schwarz, William Steptoe, et al. 2020. The ethics of realism in virtual and augmented reality. *Frontiers in Virtual Reality* 1 (2020), 1.

[40] Lorenzo Stacchio, Roberto Pierdicca, Marina Paolanti, Primo Zingaretti, Emanuele Frontoni, Benedetta Giovanola, and Simona Tiribelli. 2024. XRAI-Ethics: Towards a Robust Ethical Analysis Framework for Extended Artificial Intelligence. In *2024 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*. IEEE, 214–219.

[41] Xue Sun, Jie Xiong, Chao Feng, Haoyu Li, Yuli Wu, Dingyi Fang, and Xiaojiang Chen. 2024. Earssr: Silent speech recognition via earphones. *IEEE Transactions on Mobile Computing* 23, 8 (2024), 8493–8507.

[42] Xin Wang, Taein Kwon, Mahdi Rad, Bowen Pan, Ishani Chakraborty, Sean Andrist, Dan Bohus, Ashley Feniello, Bugra Tekin, Felipe Vieira Frujeri, et al. 2023. Holoassist: an egocentric human interaction dataset for interactive ai assistants in the real world. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 20270–20281.

[43] Carolin Wienrich and Marc Erich Latoschik. 2021. extended artificial intelligence: New prospects of human-ai interaction research. *Frontiers in Virtual Reality* 2 (2021), 686783.

[44] Li Zhang, Weiping He, Huidong Bai, Qianyuan Zou, Shuxia Wang, and Mark Billinghurst. 2023. A hybrid 2D–3D tangible interface combining a smartphone and controller for virtual reality. *Virtual Reality* 27, 2 (2023), 1273–1291.