

VueBuds: Visual Intelligence with Wireless Earbuds

Maruchi Kim

Paul G. Allen School,
University of Washington, WA, USA
mkimhj@cs.washington.edu

Rasya Fawwaz

Electrical & Computer Engineering,
University of Washington, WA, USA
fawwa001@uw.edu

Zhi Yang Lim

Paul G. Allen School,
University of Washington, WA, USA
zylim@cs.washington.edu

Brinda Moudgalya

Electrical & Computer Engineering,
University of Washington, WA, USA
brindam@uw.edu

Hexi Wang

Electrical & Computer Engineering,
University of Washington, WA, USA
whx0627@uw.edu

Yuanhao Zeng

Electrical & Computer Engineering,
University of Washington, WA, USA
yhzeng9@uw.edu

Shyamnath Gollakota

Paul G. Allen School,
University of Washington, WA, USA
gshyam@cs.washington.edu

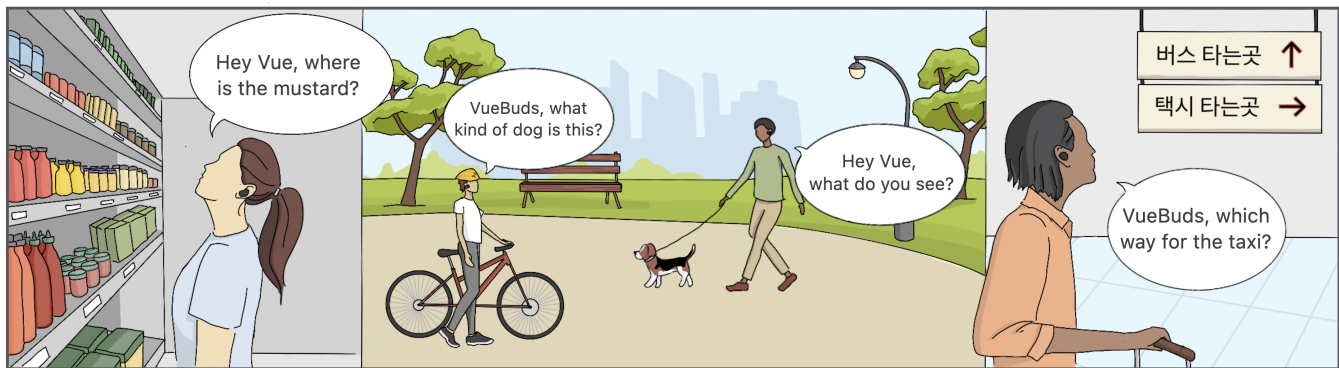


Figure 1: Applications of VueBuds. Our camera-integrated wireless earbuds enable natural language queries for everyday visual tasks such as locating items in a store, identifying objects, obtaining scene-level descriptions, and interpreting foreign text for navigation.

Abstract

Despite their ubiquity, wireless earbuds remain audio-centric due to size and power constraints. We present VueBuds, the first camera-integrated wireless earbuds for egocentric vision, capable of operating within stringent power and form-factor limits. Each VueBud embeds a camera into a Sony WF-1000XM3 to stream visual data over Bluetooth to a host device for on-device vision language model (VLM) processing. We show analytically and empirically that while each camera's field of view is partially occluded by the face, the combined binocular perspective provides comprehensive forward coverage. By integrating VueBuds with VLMs, we build an end-to-end system for real-time scene understanding, translation, visual reasoning, and text reading; all from low-resolution monochrome cameras drawing under 5mW through on-demand activation. Through online and in-person user studies with 90 participants, we compare VueBuds against smart glasses across 17 visual

question-answering tasks, and show that our system achieves response quality on par with Ray-Ban Meta. Our work establishes low-power camera-equipped earbuds as a compelling platform for visual intelligence, bringing rapidly advancing VLM capabilities to one of the most ubiquitous wearable form factors. Video demos at: <https://vuebuds.cs.washington.edu/>.

CCS Concepts

• Human-centered computing → Ubiquitous and mobile computing; • Computing methodologies → Artificial intelligence.

Keywords

Visual Computing, Earables, Multimodal Interaction

ACM Reference Format:

Maruchi Kim, Rasya Fawwaz, Zhi Yang Lim, Brinda Moudgalya, Hexi Wang, Yuanhao Zeng, and Shyamnath Gollakota. 2026. VueBuds: Visual Intelligence with Wireless Earbuds. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26)*, April 13–17, 2026, Barcelona, Spain. ACM, New York, NY, USA, 21 pages. <https://doi.org/10.1145/3772318.3791322>



This work is licensed under a Creative Commons Attribution 4.0 International License. *CHI '26, Barcelona, Spain*

© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2278-3/2026/04
<https://doi.org/10.1145/3772318.3791322>

1 Introduction

The emergence of large language models has transformed human-computer interaction, enabling natural language conversations with intelligent systems across diverse applications [20, 52]. This has accelerated further with the integration of visual intelligence, giving rise to vision language models (VLMs) that can comprehend and reason about images alongside text [8, 66]. These advances have found rapid adoption in consumer devices, with smartphones leveraging on-device visual intelligence for enhanced photography and accessibility [89], while smart glasses like Ray-Ban Meta demonstrate the potential of wearable visual computing in everyday life [71].

Despite advances in integrating visual intelligence into mobile and wearable devices, wireless earbuds remain largely limited to audio-centric functionality. Current designs incorporate low-power peripherals such as microphones, inertial measurement units, biometric sensors, and speakers [16, 73]. The absence of visual capabilities in this ubiquitous form factor represents a notable gap in the wearable ecosystem, particularly since wireless earbuds have orders-of-magnitude greater commercial adoption, with an estimated user base 150–200x larger than that of smart glasses [69, 83].

We introduce VueBuds, the first wireless earbud system integrating low-power cameras with vision language model interaction. VueBuds allow users to capture visual context from their environment and engage with vision language models through a familiar, everyday wearable platform, without requiring specialized eyewear. Our binaural system integrates dual forward-facing cameras, leveraging binocular vision to overcome facial occlusions and capture the wearer’s egocentric view.

Achieving this requires addressing three core research questions:

- *RQ1: Can truly wireless earbuds support camera hardware within strict size, weight, and power (SWaP) constraints?* Camera sensors and visual processing consume far more power than typical earbud components, raising questions about whether earbuds can support cameras while preserving acceptable battery life and form factor. Furthermore, camera data demands significantly higher bandwidth than audio, making it unclear whether the wireless protocols used in earbuds can effectively transmit a binocular visual stream.
- *RQ2: To what extent can cameras positioned at ear level provide robust egocentric views for visual perception and interaction?* Unlike smart glasses, where cameras are unobstructed and align closely with the user’s eyes, ear-level cameras have a posterior-lateral offset. This placement introduces potential facial occlusion, raising an unresolved question of whether such a vantage point can support effective egocentric vision.
- *RQ3: Can a fully wireless, Bluetooth-based pipeline support real-time multimodal interaction with vision–language models?* Answering user queries (e.g., “Where are my keys?”) requires capturing contextually relevant imagery, streaming it via low-bandwidth Bluetooth,¹ performing multimodal inference using an on-device vision–language model, and synthesizing an audio response. Meeting real-time latency constraints across this end-to-end pipeline remains an open systems challenge.

¹Bluetooth consumes far less power than Wi-Fi but offers significantly lower bandwidth. Smart glasses can support Wi-Fi due to their larger batteries and form factor.



Figure 2: VueBuds integrated with Sony wireless earbuds. The custom camera module (left) is powered directly from the earbud battery, with 3D-printed enclosures (middle) enabling forward-facing capture. VueBuds charge via the original case (right).

We address these questions through four key contributions:

- **Camera-integrated wireless earbud hardware.** We develop the first dual-camera earbud prototype, maintaining practical size, weight, and power expected in this form factor (Fig. 2). The custom camera module attaches to commodity Sony WF-1000XM3 earbuds and operates at under 5 mW (§4.3.1), adding only 11–14% battery overhead even under extreme usage of 60 visual queries per hour.
- **Binocular vision for facial occlusion.** We introduce an ear-level egocentric capture system that leverages dual viewpoints to resolve unilateral obstructions. Analytical modeling and empirical validation demonstrate that this approach significantly reduces blind spots, maintaining occlusion depths well below the Harmon distance threshold, the practical limit for supporting effective egocentric interaction (§3.2.2).
- **End-to-end system optimizations for real-time operation and VLM integration.** In §3.1, we detail optimizations that maximize wireless throughput and minimize latency for concurrent binocular video streaming. We demonstrate that VLMs such as Qwen2.5-VL can support robust scene understanding, translation, and text reading despite the earbuds’ low-resolution, monochrome imagery. Comparing input strategies, we find that stitching L/R images outperforms separate processing by eliminating visual redundancy, improving time-to-first-token (TTFT) by 46%. VueBuds achieves an end-to-end latency of under 3 seconds, with identified optimizations (§4.3.2) capable of reducing this to 1.14s.
- **User studies and comparative evaluation with smart glasses.** Across two user studies ($n = 90$), we assess real-world feasibility and user acceptance (§4.2). An online study ($n = 74$) shows that earbuds are far more widely adopted than regular glasses (93.3% vs. 62.7%) and that VueBuds+Qwen2.5-VL delivers visual question answering (VQA) performance across 17 tasks comparable to Ray-Ban Meta (MOS: 3.33 vs. 3.32). An in-person study ($n = 16$) demonstrates strong real-time performance in object recognition (82.5%), optical character recognition (OCR) (94.3%), and translation (83.8%). Participants also reported comfort similar to everyday earbuds and perceived broad applicability.

Key findings. Our results demonstrate that: (1) integrating cameras into earbuds is feasible within strict SWaP constraints; (2) binocular ear-level capture effectively mitigates facial occlusion to provide egocentric views; (3) our fully-wireless pipeline with modern VLMs can operate effectively on low-resolution, monochrome images in real time. Further, VueBuds is perceived as highly accessible for multimodal interaction, achieving utility competitive with Ray-Ban Meta glasses. Together, these findings establish earbuds as a promising platform for egocentric visual intelligence applications.

Table 1: Comparison of egocentric visual wearable systems. XR headsets (e.g., Vision Pro) offer eye-level alignment with rich deictic cues (gaze + hand tracking) but are bulky and isolating. Smart glasses (e.g., Ray-Ban Meta) resemble ordinary eyewear, but lack gaze tracking. Body-worn cameras (e.g. AI Pin, GoPro) provide torso-level viewpoints with optional projection feedback. VueBuds utilize ear-level cameras in a compact, familiar form factor. Beyond these technical trade-offs, device adoption is shaped by social factors: earbuds represent the most widely adopted wearable category, with over 340 million units shipped annually compared to approximately 2 million smart glasses [45], suggesting that ear-worn devices may face lower barriers for everyday use.

Form Factor	Viewpoint Alignment	Deictic Cues	Privacy Signaling	Feedback Modality
XR Headsets	Eye-level	Gaze + Hand + Voice	High	Visual + Audio
Smart Glasses	Eye-Level	Hand + Voice	Low-Medium	Visual + Audio
Body-worn Cameras	Torso-level	Hand + Voice	Medium-High	Projector + Audio
VueBuds (Ours)	Ear-Level	Hand + Voice	Low	Audio

2 Related work

We position VueBuds within three interconnected research areas. First, we examine existing wearable visual intelligence systems, from smart glasses to AR headsets, and their approaches to situated interaction and context-aware assistance. Second, we review the evolution of ear-worn sensing platforms, identifying the absence of visual capabilities as a key gap. Finally, we address the cross-cutting challenges of power consumption and privacy that any wearable camera system must navigate.

Head-mounted visual wearables. Existing visual wearable systems have primarily focused on glasses and head-mounted cameras [21, 30]. For example, Pupil Invisible glasses from Pupil Labs provide calibration-free gaze tracking [61]. CAPturAR [101] uses a customized AR head-mounted device to author context-aware applications by referencing users’ previous activities. GazePointAR [63] combines eye gaze, pointing gestures, and conversation history to disambiguate speech queries in augmented reality. In the consumer market, products such as Ray-Ban Meta [71] and Google Glass [37] feature forward-facing cameras for photo/video capture and AI-based visual question answering (VQA), though Google Glass faced negative press due to privacy concerns [40]. Other commercial offerings like the XREAL Aura [103] focus on display functionality rather than visual intelligence, serving as a wired secondary monitors for productivity. In parallel with these commercial systems, the advent of large language models has sparked recent research in exploring multimodal interaction techniques that combine visual perception with user intent. GesPrompt [43] demonstrates how co-speech gestures enable more precise object selection in vision language model interactions, allowing users to specify a specific object while pointing. In contrast to these prior works that focus on face-mounted visual wearables, ear-level cameras for general visual intelligence remain commercially unavailable and underexplored in the research community, despite the widespread adoption of wireless earbuds as a wearable platform.

Ear-worn sensor platforms and hearables. Modern hearables have evolved to incorporate a variety of sensors, including microphones, inertial measurement units (IMUs), and photoplethysmography (PPG) sensors, enabling health monitoring and motion tracking [34, 36, 42, 73, 92]. Prior work has explored diverse activity recognition and health-tracking applications using these sensor platforms [22, 23, 68, 79, 102]. Recently, several earable platforms have been developed that integrate multiple of these sensors [25, 47, 55, 84, 85],

but none have incorporated cameras. These works demonstrate the feasibility of compact, ear-worn devices for rich physiological and facial movement sensing, yet they generally lack egocentric vision capabilities. The most closely related work [72] uses an inward-facing camera to capture ear images for authentication, rather than for general visual intelligence, and relies on wired hardware with an evaluation board that does not meet practical power or size constraints. Similarly, [27] employs cameras to reconstruct facial expressions, not for LLM integration, and, like [72], is not wireless. Though rumored to be in development in tech media [80], no camera-integrated wireless earbuds have been publicly released.

Low-power camera systems. Prior work has explored several approaches to reducing the power consumption of vision systems, including the design of more power-efficient camera sensors [17, 50, 75], mixed-signal vision integrated circuits [104], and specialized processors [86, 91]. Low-power wireless video has been demonstrated in various contexts, for example, Bluetooth-based video streaming for robots and sensor systems [48, 99], as well as backscatter techniques and low-power machine learning algorithms [54, 62, 76, 88]. IRIS [57] further extends this space with a vision-enabled ring integrated with object detection. In contrast to these efforts, our work designs and demonstrates the first vision-enabled wireless earbud system for real-time interaction with VLMs.

3 System Design

We design VueBuds to integrate visual intelligence into wireless earbuds while preserving the familiar form factor that makes them widely adopted. This section describes our hardware platform, binocular vision approach for capturing the user’s forward-facing perspective, and end-to-end system integration with on-device vision language models.

3.1 VueBuds Hardware

3.1.1 Hardware Overview. Our custom hardware design integrates an ultra-low-power Himax HM01B0 CMOS image sensor (1/11" format, 324x324 pixel array), an Analog Devices MAX77650 power management integrated circuit, and a Nordic nRF52840 Bluetooth Low Energy system-on-chip. The design includes status LEDs and a single-pole single-throw switch for debugging and development purposes.

The custom companion PCB is integrated into Sony WF-1000XM3 earbuds by interfacing with the existing battery system, preserving

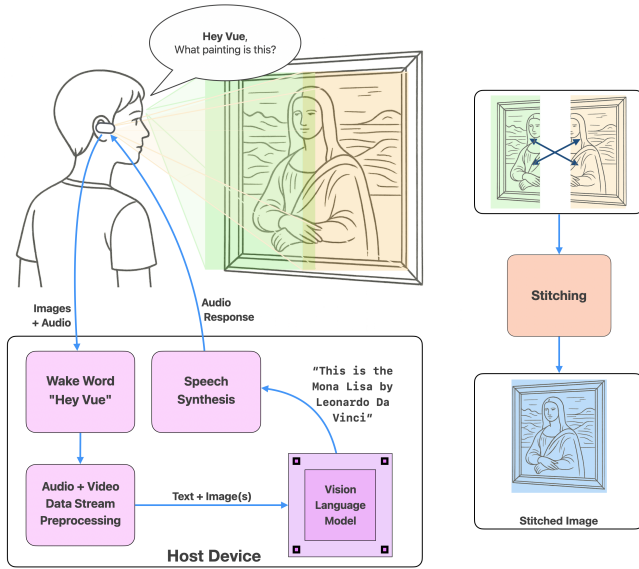


Figure 3: VueBuds system overview. VueBuds utilizes a vision language model to process the user’s voice and binocular camera data for real-time multimodal interaction.

all audio functionality while adding visual capabilities (Fig. 4). We selected the Sony WF-1000XM3 platform based on accessibility for disassembly and battery integration, enabling a maintainable research scope without requiring a complete earbud redesign from the ground up. Modern wireless earbuds incorporate sophisticated components, including digital signal processors, multiple microphones, in-ear sensors, class D amplifiers, and custom application-specific integrated circuits (ASICs). By building upon an existing commercial platform, we demonstrate that camera integration is feasible within the constraints of commercial earbud designs rather than simplified prototypes.

Proceeding hardware integration, we designed custom 3D-printed enclosures that seal each modified earbud while positioning the camera sensor to achieve forward-facing capture aligned with the user’s field of view. The enclosure design maintains the overall form factor expectations of modern earbuds while accommodating the additional camera hardware. Camera positioning optimization and angular field of view considerations are detailed in §3.2.2.

3.1.2 Camera and Bluetooth Chip Integration. Since the nRF52840 Bluetooth Low Energy SoC lacks a dedicated camera interface, we configure the HM01B0 CMOS sensor to operate in single-wire data transfer mode over its Display Video Port (DVP) interface. In this configuration, the HM01B0 effectively functions as an SPI controller, transmitting pixel data to the nRF52840 through its SPI peripheral (SPIS) port. Image transfer is coordinated through three key signals: Frame Valid (FVLD) indicates frame start with a rising edge, while Pixel Clock Out (PCLKO) and Data (D0) deliver synchronized pixel data. These signals map to the standard SPI protocol as chip select (CS), serial clock (SCLK), and controller-out-peripheral-in (MOSI).

Clock synchronization between the camera sensor and microcontroller is achieved by generating an 8 MHz master clock (MCLK) signal from the nRF52840 to drive the HM01B0, ensuring the pixel

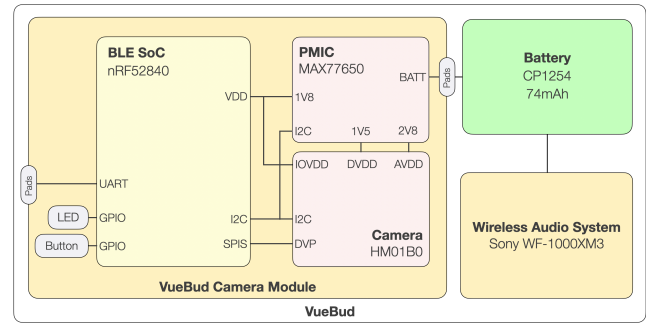


Figure 4: Hardware block diagram. Each VueBud integrates a Himax CMOS Imaging Camera and a Power Management Integrated Circuit (PMIC). Power is delivered through the onboard battery inside a Sony WF-1000XM3 wireless earbud.

clock output remains within the nRF52840’s maximum SPI frequency specification of 8MHz. This approach maintains tight timing synchronization between both devices throughout the image capture process. The interface requires signal polarity adaptation since the nRF52840 SPI peripheral expects an active-low chip select signal, while the HM01B0’s FVLD output is active-high. We address this by configuring FVLD as a GPIO interrupt source that triggers an external loopback connection, generating the required active-low chip select signal for the SPI interface. This approach could be optimized in future revisions by incorporating an inverter gate between FVLD and the SPI chip select line [49, 57]. Finally, the nRF52840’s SPI DMA controller imposes a 64 kB maximum transfer size, while a full 324×239 pixel image requires 77.4 kB of data transfer. We overcome this limitation by splitting each image capture into two sequential SPI transactions, using the external GPIO loopback to trigger the second transaction mid-frame.

3.1.3 Low-Power Optimizations. The VueBuds camera module implements a three-state power management architecture to minimize energy consumption: OFF, IDLE, and ACTIVE. In the OFF state, the camera module is completely powered down, corresponding to scenarios where the earbuds are removed from the user’s ears (detected via in-ear proximity sensors) or placed in the charging case. The IDLE state represents the standby mode, where VueBuds hardware is powered on with an interrupt from the earbud’s primary SoC. In this state, the nRF52840 maintains an active Bluetooth Low Energy connection to the host device, while the HM01B0 camera remains configured via I2C but clock-gated to minimize power consumption. The system transitions from IDLE to ACTIVE state upon wake-word detection (e.g., "Hey Vue" or "VueBuds"), enabling immediate camera activation without configuration delays. This wake-word paradigm mirrors existing earbud voice interactions such as "Hey Siri" or "Ok Google," facilitating seamless adoption within established user interaction patterns.

During ACTIVE operation, VueBuds enables full camera functionality and streams visual data to the connected host device. To balance responsiveness with power efficiency, the system automatically returns to IDLE state after 3 seconds of streaming, ensuring minimal impact on overall earbud battery life while maintaining user interaction capabilities. This strategy enables VueBuds to operate with minimal impact on the existing earbud power budget.

Table 2: Design space for dual ear-level camera orientation.

Angular Orientation	Forward Blind Spot Length	Added Field of View	Binocular Field of View	Overlap at Harmon Distance
0°	14.1cm	0°	88°	64%
5°	18.6cm	10°	98°	46%
10°	24.7cm	20°	108°	28%
15°	34.0cm	30°	118°	14%
20°	50.7cm	40°	128°	N/A

3.1.4 Frame Rate and Wireless Latency. To maximize frame rate, we configure the system with the shortest connection interval permitted by the BLE specification (7.5ms) while transmitting 5 packets of 247 bytes per interval. VueBuds operates at the maximum supported Bluetooth Low Energy data rate of 2 Mbps using LE 2M PHY [90]. This configuration achieves approximately 992 kbps of effective throughput [77].

We evaluate two camera configurations from the HM01B0 sensor: QQVGA (162×119 pixels) and QVGA (324×239 pixels), generating 19.3 kB and 77.4 kB per frame, respectively. Given the maximum available BLE throughput, the theoretical wireless transmission limits are 6.4 fps for QQVGA and 1.6 fps for QVGA operation. However, image acquisition introduces additional latency through the 8MHz SPI interface, requiring 19.3ms for QQVGA capture and 77.4ms for QVGA capture. Accounting for both acquisition and transmission overhead, the system achieves effective frame rates of 5.7 fps and 1.4 fps for QQVGA and QVGA configurations, respectively (see Table. 3). These rates could be improved through pipelining that overlaps image acquisition with wireless transmission [60].

Table 3: Resolution against frame rate and latency.

Camera Configuration	Frame Rate	Frame Latency
QQVGA (162×119)	5.7 fps	175 ms
QVGA (324×239)	1.4 fps	714 ms

3.1.5 Fabrication. We integrate custom 3D-printed enclosures with commercial Sony WF-1000XM3 earbuds to house the camera hardware. The enclosures were designed in OnShape CAD and manufactured on a Bambu Labs A1 Mini 3D printer. VueBuds printed circuit boards were designed with the open-source KiCad eCAD tool and fabricated as 2-layer PCBs through PCBWay, with component assembly completed by a local contractor. The enclosure design ensures seamless integration with the original earbud form factor while providing precise camera positioning. The HM01B0 camera sensor is mounted behind a window in the enclosure lid, oriented forward to capture the user’s natural field of view. This positioning aligns the camera with the user’s perspective when the earbuds are worn, facilitating intuitive visual queries by allowing users to simply look toward objects of interest.

3.2 Binocular Vision for Ear-Level Cameras

3.2.1 Camera Field-of-View and Facial Occlusion. We first quantify the extent of facial occlusion introduced by ear-level positioning and demonstrate how sensor windowing reduces obstruction while introducing coverage trade-offs. The HM01B0 camera sensor provides an 87-degree horizontal field of view, which poses significant challenges when positioned at ear level on wireless earbuds. Unlike temple-mounted cameras on smart glasses, which benefit from forward positioning relative to the user’s face, earbud-mounted cameras suffer from substantial facial obstruction, with the user’s

face blocking a considerable portion of each camera’s field of view, specifically in the inward-facing regions.

Facial obstruction severely impacts the ability to capture egocentric viewpoints and can cause issues with vision language model interpretation and panoramic stitching correspondence. To mitigate this obstruction, we implement the HM01B0’s windowed readout mode, which crops the sensor’s active area to reduce the effective field of view from 87° to 65°. This windowing approach eliminates the facial obstruction zone while (1) maintaining forward-facing capture capability, and (2) reducing unnecessary data transmission over the bandwidth-constrained Bluetooth Low Energy link.

This 25% reduction in captured visual data creates a fundamental trade-off: while windowed readout helps resolve facial obstruction, it significantly reduces peripheral visual coverage, particularly limiting the left and right outskirts of the user’s natural field of view. This necessitates alternative strategies to recover comprehensive environmental coverage through dual-camera coordination.

3.2.2 Camera Angular Orientation and Blind Spot Characterization.

To compensate for the visual context lost through windowed readout, we systematically evaluate camera angular orientations at 0°, 5°, 10°, 15°, and 20° outward from the forward-facing position. This angular adjustment serves the dual purpose of reducing potential facial obstructions and expanding the combined field of view.

In the completely forward-facing orientation (0°), windowed readout enables VueBuds to capture approximately 88° of the user’s forward perspective per camera. However, outward camera angling introduces competing design constraints that must be carefully balanced. Angling cameras outward conflicts with our core design principle of capturing the user’s natural forward-facing perspective and creates an expanding “blind spot” directly in front of the user, where objects held at close distances fall outside both cameras’ fields of view.

We address this trade-off by incorporating the Harmon distance, the average comfortable reading distance of 36.8cm established across 233 individuals [19]. Objects held closer than this distance typically fall outside normal interaction patterns for reading and detailed visual examination. By designing around this ergonomic constraint, we establish a tolerable blind spot that minimally impacts practical usage scenarios.

To characterize the forward-facing blind spot region, we model the binocular camera field of view on a person’s head at 5-degree increments until the forward blind spot length exceeds the Harmon distance. We define this blind spot region as the distance from a person’s eye to where a 5 cm wide object must be present in each camera’s field of view to ensure sufficient correspondence for image stitching algorithms and provide conservative redundancy between left and right images. Our analysis in Table. 2 demonstrates that camera orientations of 5° and 10° maintain forward blind spot

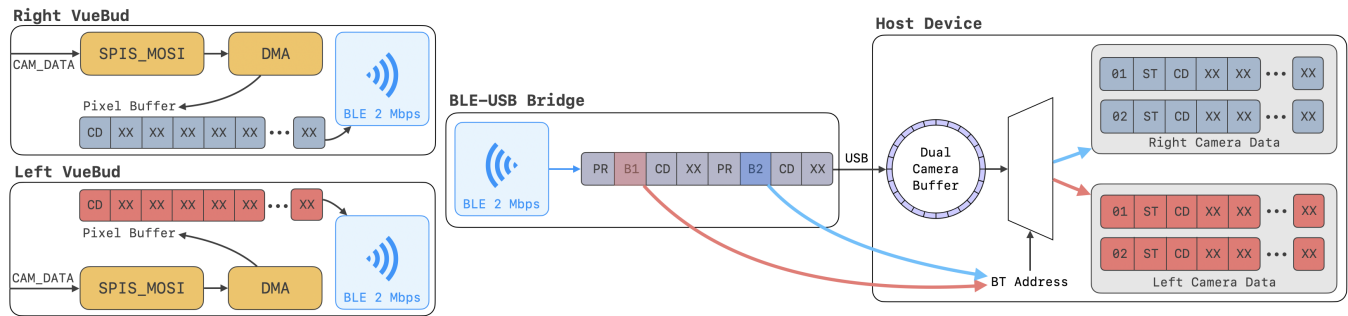


Figure 5: VueBuds to host device wireless data streaming. VueBuds stream dual camera data to the BLE-USB bridge, which multiplexes the binocular image feed for transmission to the host device, where they are ultimately demultiplexed into left and right images.

lengths well below the Harmon distance (18.6 cm and 24.7 cm, respectively), while adding 10-20° of additional field of view. At 15°, the blind spot (34.0 cm) approaches the Harmon distance threshold, representing a practical limit for maintaining usable interaction patterns. Beyond 15° of outward angling, the blind spot expansion significantly degrades user experience, potentially requiring users to step backward or extend objects to uncomfortable distances when looking directly at an object. We verify our blind spot simulations using our hardware in §4.1.1.

3.2.3 Image Stitching for Redundancy Reduction. While the simple approach is to feed each camera image from the left and right earbuds into the vision language model, we also explore image stitching to reduce redundant visual information and improve vision language model inference runtime.

Overlaps exist between left and right image pairs captured by VueBuds, particularly in the far-field regions. At the Harmon distance, we calculate 64%, 46%, and 28% overlap between images at 0°, 5°, and 10° camera orientations, respectively. To reduce effective input tokens and improve end-to-end latency, we implement light-weight stitching using ORB (Oriented FAST and Rotated BRIEF) feature detection, selected for computational efficiency [87]. At a high level, these techniques work by detecting repeatable key-points in both images, converting them to compact binary feature vectors, and then matching corresponding features to estimate geometric transformations for alignment and stitching. Our stitching pipeline operates without post-processing operations such as image trimming to preserve maximum visual information for the vision language model and minimize processing latency.

3.3 End-to-End System Integration

Shown in Fig. 5, we demonstrate wireless data flow from VueBuds hardware to vision language model processing.

3.3.1 BLE-USB Bridge. The system employs an nRF52840 development board as a BLE-USB bridge that converts Bluetooth Low Energy communication to USB connectivity for the host processing device. The receiver establishes connections with both VueBuds devices and manages both camera data streams simultaneously.

Upon receiving packets from each earbud, the receiver prepends a fixed preamble and the originating device’s Bluetooth MAC address to enable stream demultiplexing. This packetized data is transmitted via USB serial to the host device, where our Python-based parsing module processes the incoming stream. The parser uses

the preamble for packet boundary detection, the MAC address to route data into separate left and right image buffers, and a frame indicator to detect image boundaries for frame reconstruction.

3.3.2 Vision Language Model Processing. We implement on-device vision language model processing using Ollama, evaluating five models for optimal performance characteristics: Qwen2.5VL [12], Moondream [74], MiniCPM-V [105], LLaVA [67], and Gemma3 [94]. Model selection is based on systematic evaluation across accuracy and latency metrics, detailed in §4.1.2. Our implementation incorporates an adaptive system prompt that adjusts based on image input type, providing context-aware prompts for stitched images versus dual independent camera feeds. While the system supports both 162×119 and 324×239 resolutions, we select the higher resolution configuration based on its superior text recognition performance observed during evaluation.

3.3.3 Audio Processing Pipeline. Our audio pipeline uses the Bluetooth Hands-Free Profile for audio data transfer, and incorporates automatic speech recognition (ASR) and synthesis (TTS) capabilities. Our system integrates TinyWhisper for real-time speech-to-text conversion, chosen for its computational efficiency and accuracy in wake-word detection.

Our wake-word detection implementation uses a fuzzy matching algorithm to handle variations in pronunciation and ambient noise levels. The system recognizes wake phrases including "VueBuds", "Hey Vue", and phonetic variations, triggering camera activation and visual query processing. The system incorporates confidence thresholding to minimize false-positive wake-word activations while maintaining responsiveness. Speech synthesis utilizes Apple’s built-in text-to-speech engine to deliver VLM responses through the earbuds’ integrated speakers, enabling private audio playback. The entire audio pipeline operates in tandem with the visual pipeline, using a multi-process architecture to maximize system performance.

4 Evaluation

We present the system’s vision capabilities along with accuracy and latency results across several vision language models. Then, we select the best-performing VLM for our system and share the results from our online and in-person user studies. Finally, we assess system performance in terms of power and end-to-end latency.

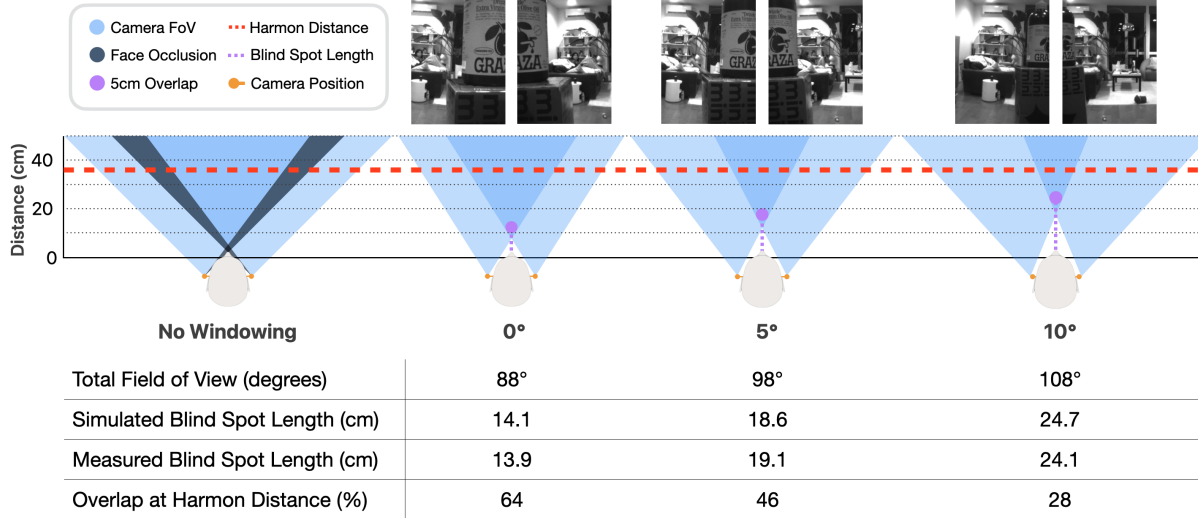


Figure 6: Blind spot evaluation. Geometric modeling of dual camera field of view (blue) showing facial occlusion and blind spot regions at various angular orientations. The red line indicates the Harmon distance. Empirical measurements (top) closely align with the model.

4.1 Vision Capability Benchmarks

Here, we evaluate VueBuds’ tolerance to blind spots within our camera angle design space, and then present VLM accuracy and latency results using images from our camera to explain why we chose Qwen2.5VL 7B among 5 selected vision language models.

4.1.1 Blind Spot Evaluation. As discussed in §3.2, the usability of camera-integrated earbuds depends critically on capturing the user’s forward field of view while preserving natural interaction patterns. Users should not need to hold objects at awkward angles or step backward to accommodate system limitations.

Methodology. In §3.2.2, we developed a geometric model to project the dual camera fields of view from ear-mounted positions. Using a mannequin wearing VueBuds, we measured a lateral camera offset of 2.3 cm. Our model projects the windowed 65° field of view from each camera position at angular orientations of 0°, 5°, 10°, 15°, and 20° outward from the forward-facing direction. The blind spot length was defined as the distance from the user’s eye center to the point where a 5 cm object would be captured in its entirety by both cameras’ fields of view.

To validate our theoretical values, we constructed an adjustable test rig using L-brackets that allowed precise angular positioning of the VueBuds cameras. For each angular configuration, we empirically measured the blind spot using a circular bottle with a text label as our test object. We placed the bottle at the mannequin’s eye center and gradually moved it backward until visual commonality was confirmed across both left and right camera views. This methodology directly measured the practical blind spot distance with our actual hardware system. As shown in Fig. 6, our empirical measurements closely aligned with the geometric model across all tested configurations. This strong correspondence (within 3% error) validated our geometric modeling approach for system design.

Design Tradeoffs. The total effective field of view, measured from the leftmost edge of the left camera to the rightmost edge of the right camera, expands from 88° at 0° orientation to 128° at 20°. As discussed in §3.2.2, this expanded coverage comes at the expense of increasing the blind spot length. This adversely affects usability,

as orientations beyond 15° potentially require users to hold objects farther than normal interaction distances. At 15°, the blind spot length (34.3 cm) approaches the Harmon distance threshold of 36.8 cm, while the 20° configuration creates a 49.5 cm blind spot that would severely degrade natural interaction patterns. Based on this analysis, we designed VueBuds with angular orientations between 5° and 10°, balancing expanded field of view coverage while preserving intuitive interaction where users can simply look toward objects of interest without adjusting their behavior.

4.1.2 VLM Accuracy and Latency Performance. Selecting an appropriate vision language model for VueBuds requires balancing accuracy, latency, and on-device deployment constraints.

Methodology. We evaluated five models under 8B parameters: Qwen2.5-VL (7B), Moondream (1.8B), MiniCPM-V (8B), LLaVA (7B), and Gemma3 (4B); all supported through Ollama. Each model processed 20 scenes captured by VueBuds at both 160px and 320px resolutions across three task categories: (1) Object/Scene Recognition: identifying common objects and describing scenes, (2) Optical Character Recognition: reading text from signs, labels, and documents, and (3) Translation/Reasoning: answering questions requiring visual understanding and logical inference. All benchmarks were performed on a Mac Mini M4 Pro in its base configuration [10], with human evaluators verifying response accuracy.

Resolution Impact. Fig. 7 shows that increasing resolution from 160px to 320px (a 4x increase in total pixels) yields substantial accuracy improvements across all models. Overall accuracy increased by 35% for both Qwen2.5-VL and Moondream, 61% for MiniCPM-V, and 96% for LLaVA. The impact was most pronounced for text reading tasks, where Qwen2.5-VL and Moondream achieved perfect accuracy at 320px within our dataset, representing improvements of 170% and 84% respectively from their 160px baselines. This suggests OCR capabilities degrade rapidly once character boundaries become indistinguishable at lower resolutions. For queries like "What does this sign say?" or "Can you translate this for me?", the higher resolution proves essential. This validated our decision to use HM01B0’s 324x239 resolution despite the increased latency.

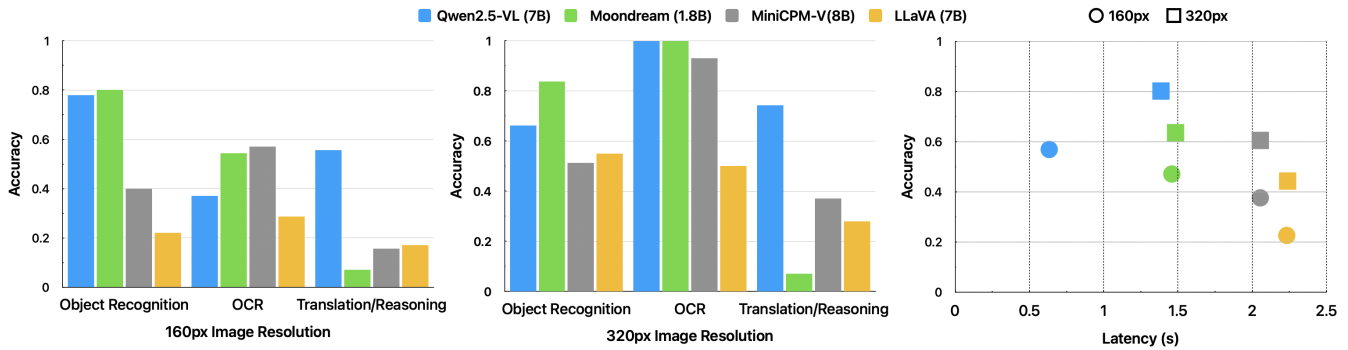


Figure 7: VLM comparison. Task-specific accuracy at 160px (left) and 320px (center), and accuracy against latency (right).

Accuracy-Latency Tradeoffs. Response latency directly impacts user experience in conversational interfaces. Fig. 7 illustrates this design space, with time-to-first-token ranging from 0.6 seconds (Qwen2.5-VL at 160px) to 2.24 seconds (LLaVA at 320px). Notably, Gemma3 exhibited severe latency issues, averaging over 12 seconds to first token, a known KV cache quantization bug [78] that rendered it challenging for real-time application use despite its compact 4B parameter size. Qwen2.5-VL achieved the optimal balance, delivering the highest overall accuracy (80.1%) while maintaining the lowest latency (1.39s at 320px). Its architectural optimizations, including window attention in the vision transformer [13], enable efficient processing along with its dynamic resolution adaptation. While Qwen2.5-VL shows the largest relative latency increase when scaling from 160px to 320px, this corresponds to its dynamic resolution processing that adjusts the number of visual tokens based on input complexity. Crucially, even at 320px, Qwen2.5-VL remains the fastest model in our evaluation.

Model Selection Rationale. While matching Qwen2.5-VL on object recognition and OCR tasks, Moondream failed on reasoning queries (7.1% accuracy), often returning empty responses for complex questions [31, 70]. This limitation disqualifies it for queries requiring inference, such as "How many calories are in this?" and similar questions. MiniCPM-V exhibited overfitting to our system prompts, generating templated responses that failed to adapt to varied user inputs. LLaVA produced inconsistent outputs when processing our low-resolution monochrome images. These evaluations confirm Qwen2.5-VL as a better model for VueBuds, providing reliable performance across diverse visual queries while maintaining sub-1.5-second response times.

4.2 User Studies

To comprehensively evaluate VueBuds, we conducted user studies with 90 participants examining three components: platform accessibility, response quality compared to Ray-Ban Meta, and real-world performance across standardized visual question answering (VQA) tasks. The first two components were administered as a two-part online survey through Google Forms to efficiently gather comparative assessments at scale, while the third was conducted through in-person sessions to evaluate real-world performance. Each study used a within-subject design with no compensation. All participants were informed of the study's purpose and procedures and voluntarily agreed to participate. User data was anonymized, and

no photos containing participants' faces were retained. All studies were approved by our university's Institutional Review Board.

4.2.1 Study 1: Platform Accessibility Analysis. Here, we examine device usage patterns to establish the potential user base for earbud-based and glasses-based wearables, providing critical context for earbud accessibility and adoption compared to glasses platforms.

Participants. For the online portion of our study, we recruited 74 participants (48 male, 26 female) between the ages of 19–65 ($\bar{x} = 35.9$, $\sigma = 13.93$) through convenience sampling via personal networks. Participants represented diverse professional backgrounds, including engineers, healthcare professionals, designers, consultants, marketers, students, stay-at-home parents, and retirees. The majority of participants were located in the United States, with additional participants from South Korea, Hong Kong, and Japan.

Methodology. We conducted an online survey examining participants' usage frequency of regular (non-smart) glasses and earbuds. Participants answered two questions about their device usage frequency: 1) *How often do you wear (ordinary) glasses?*, and 2) *How often do you wear earbuds or headphones?* Response options included: Everyday, 5-6 days a week, 3-4 days a week, 1-2 days a week, and Never. We then categorized responses into three usage patterns: high-frequency users (5-7 days per week), occasional users (1-4 days per week), and non-users (never). Chi-square tests compared usage pattern distributions between glasses and earbuds.

Results. Analyzing the usage patterns revealed striking differences in device adoption. For glasses, 45.3% were high-frequency users, 17.3% occasional users, and 37.3% non-users. In contrast, earbuds showed substantially higher engagement: 57.3% high-frequency users, 36% occasional users, and only 6.7% non-users. This translates to 93.3% of participants using earbuds at least occasionally, compared to 62.7% for glasses.

A chi-square test confirmed that usage patterns differed significantly between device types ($\chi^2(2) = 21.98$, $p < 0.001$), with the largest differences in non-user rates. The exclusion rates provide the most compelling insight: over one-third of participants (37.3%) never wear glasses, thereby immediately excluding them from any glasses-based AI system. Conversely, only 6.7% avoid earbuds entirely, suggesting that earbud-based platforms could reach a significantly larger user base.

These findings indicate that camera-integrated earbuds like VueBuds could reach a broader audience than glasses-based systems, particularly among the 37.3% of users who do not wear glasses.

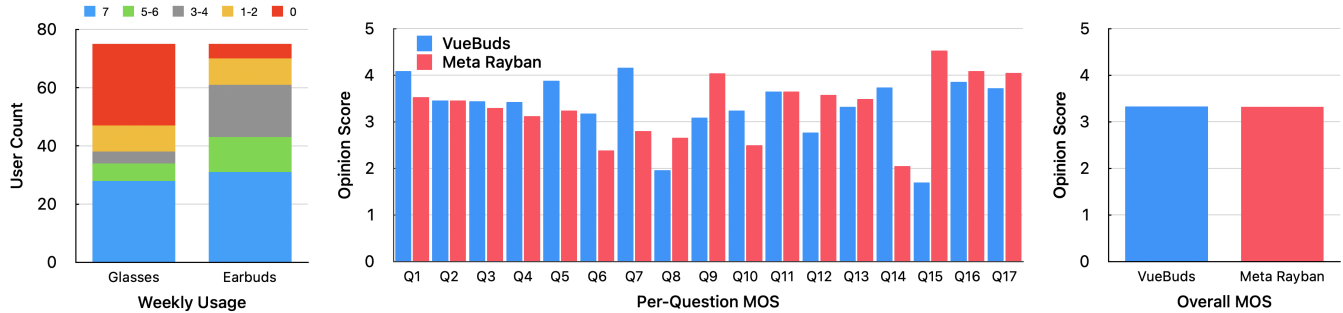


Figure 8: Online user study results. (Left) Weekly usage patterns for regular (non-smart) glasses and earbuds. (Middle) Mean Opinion Score (MOS) of VueBuds against Ray-Ban Meta for each of the 17 visual question answering tasks. (Right) MOS averaged across all 17 tasks.




Visual Scene*	Query	VueBuds Response	Meta Rayban Response
	Q11 "How many calories are in this?"	The image shows a can of LaCroix sparkling water. The label on the can indicates it is a "Calorie Free" drink. Therefore, it contains 0 calories.	The can of La Croix sparkling water has zero calories.
	Q7 "translate this for me"	The text on the package appears to be in Korean. The visible text reads "물냉면," which translates to "Cold Noodles" in English. This suggests the package contains a cold noodle dish.	The text appears to be a product labeled for a Korean cold noodle dish with ingredients and nutritional info listed.
	Q15 "how many brushes do you see?"	In the image, there are six paintbrushes visible.	I see seven brushes.

Figure 9: Samples from the online user study. Visual scenes(*) presented to participants were captured using a mobile phone to eliminate camera quality bias when comparing response quality between VueBuds and Ray-Ban Meta.

While this data reflects comfort with traditional device categories rather than AI acceptance, it establishes that earbud-based systems face fewer baseline accessibility barriers, potentially reducing adoption friction for wearable AI technologies.

4.2.2 Study 2: Response Quality Evaluation Against Ray-Ban Meta. Here, we evaluate the subjective response quality between VueBuds and commercial smart glasses.

Participants. 74 participants from Study 1 completed this study.

Methodology. Participants performed a blind comparative evaluation of VueBuds and Ray-Ban Meta response quality across 17 visual question answering tasks using a 5-point Mean Opinion Score (MOS) scale. We designed 17 evaluation scenarios representing typical smart glasses use cases: scene recognition ("what do you see"), text recognition (book titles), translation, calorie estimation, plant care advice, and object counting. For each scenario, we collected responses from both systems viewing identical real-world scenes. Ray-Ban Meta responses were generated by pointing the device at the scene, issuing verbal queries, and transcribing the audio. VueBuds responses were generated by capturing the scene with our camera system and processing through Qwen2.5-VL 7B.

To ensure fair comparison, we implemented several controls: (1) For questions 1–15, we showed the online participants device-agnostic images captured using a mobile phone pointed at the same

scenes, then paired these images with both systems' responses to eliminate camera quality bias, and (2) randomized response order to prevent position bias. Questions 16–17 presented both device-agnostic images and images captured directly by VueBuds, alongside both systems' responses, allowing participants to compare response quality while also evaluating VueBuds' image quality.

Participants received the instruction: "Imagine you are wearing AI-enabled smart glasses. You will see images with sample questions and two different AI responses." Responses were rated on a 1–5 scale: 1=Poor (inaccurate, confusing, unhelpful), 2=Fair (somewhat relevant but incomplete), 3=Good (reasonably accurate and helpful), 4=Very Good (accurate, clear, useful), and 5=Excellent (highly accurate, very clear, very helpful). Participants were instructed that response length should not influence ratings. After completing the evaluation, participants were given the option to provide open-ended feedback (see Appendix) to elaborate on their rating criteria and explain the factors that drove high and low scores.

Results. The overall MOS scores in Fig. 8 demonstrate near-statistical parity: VueBuds achieved 3.33 compared to Ray-Ban Meta's 3.32. However, per-question analysis revealed task-dependent variations. The largest performance gap favored VueBuds in a translation task (Q7: 4.1 vs 2.8), where users strongly preferred




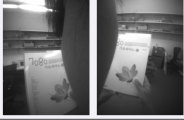


Stereo Images	Query	Response
	"what are some songs off of this album?"	I see a person holding a picture of the Beatles crossing Abbey Road. This is likely a reference to the album "Abbey Road." Some popular songs from this album include "Something," "Here Comes the Sun," and "Something in the Way She Moves."
	"what am I holding?"	I see you're holding a pair of scissors in your right hand.
	"what am I holding?"	I see a box of Cheez-It crackers in the foreground. It looks like you're holding it.
	"can you translate this for me?"	I see a book with Korean text and a flower on the cover. The number "7080" is visible on the book. The text "가요피아노" translates to "Piano Songs" in English.
	"what book is this?"	I see a book titled "Data Structures in C++ Using the Standard Template Library" by Timothy Budd.
	"who is this book by?"	I see the book in the image is titled "What's Our Problem?" and it's written by Tim Urban.

Figure 10: VueBud query and response samples from the in-person user study. VueBuds captures the binocular imagery (left) and user query (middle) which is then provided as input to the vision language model for an auditory response (right).

specific translations over Ray-Ban Meta's general scene description. Conversely, the largest gap favoring Ray-Ban Meta occurred in object counting tasks (Q15: 4.5 vs 1.7), revealing limitations in Qwen2.5VL's numerical accuracy. Most other tasks showed comparable performance, with differences typically under 0.5 points.

The near-identical overall scores suggest that despite variation in individual tasks, both systems provide comparable user experience. This demonstrates that low-resolution monochrome images from VueBuds, when processed by Qwen2.5VL, can achieve response quality comparable to commercial smart glasses.

Analysis of participant feedback revealed five primary criteria driving quality ratings: (1) *Accuracy* - factual correctness, with counting errors particularly penalized; (2) *Directness vs. verbosity* - concise, direct answers preferred over unnecessary elaboration; (3) *Task-appropriate responses* - context-dependent quality expectations, such as actual translations for translation tasks; (4) *Conversational suitability* - responses similar to everyday conversation; and (5) *Confidence calibration* - appropriate uncertainty expression valued over overconfident incorrect responses.

4.2.3 Study 3: In-Person Performance Evaluation with VueBuds. Here, we show the results from an in-person study to evaluate VueBuds under realistic usage conditions, including variation in head shapes, earbud positioning, object holding distances, lighting conditions, and potential visual occlusions (e.g., glare, shadows).

Participants. We recruited 16 participants (10 male, 6 female) aged of 20–44 ($\bar{x} = 27.9$, $\sigma = 6.9$) locally in the United States,

including students (7), healthcare professionals (3), engineers (3), consultants (2), and a professor (1).

Methodology. Participants were tested individually across kitchen, office, and living room environments. Each participant was provided with VueBuds and informed that these were camera-integrated earbuds designed for visual question answering (VQA). After briefing participants on VueBuds' forward-facing camera design, we provided diverse test objects including snack boxes, food packages, kitchen utensils, books, vinyl records, and Korean-language items (signage, piano books, snack packaging).

Participants were instructed to handle objects naturally while asking standardized questions to ensure consistent task categorization. To manage the open-ended nature of VLM capabilities, we focused evaluation on three primary tasks: Object Recognition, Optical Character Recognition (OCR), and End-to-End Translation. These tasks were chosen as they represent foundational and critical real-world utilities (situational awareness [6, 41, 82, 95], text access [14, 56, 95], and multilingual communication [6, 82]) required for current and next-generation wearable vision assistance systems:

- *Object Recognition*: "What am I holding?" tasks testing basic visual identification
- *Optical Character Recognition (OCR)*: Book title and author identification requiring text extraction
- *End-to-End Translation*: Korean-to-English translation combining OCR and linguistic processing

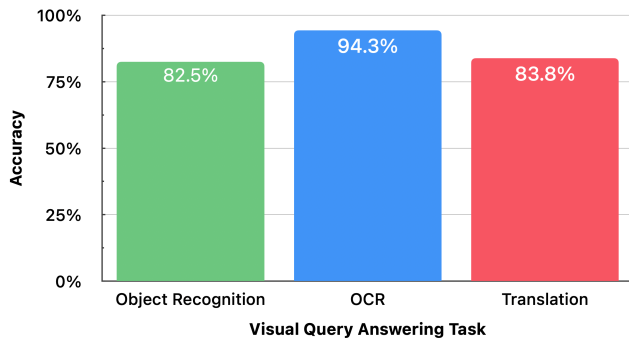


Figure 11: In-person user study results. VueBuds performance across the three visual question answering (VQA) tasks.

All system responses were recorded in real-time and the responses were evaluated by comparing with the ground truth visual scene. Object recognition tasks received binary scoring (correct/incorrect), while OCR accuracy was measured using word error rate (WER). Translation tasks employed a two-stage evaluation: OCR accuracy for text extraction, followed by translation quality given successful text recognition. This approach isolated whether translation failures stemmed from visual processing or linguistic capabilities. After the study, we sent the same 16 participants a short survey (see Appendix) to provide additional qualitative insight regarding their experience with VueBuds.

Results. Across 130 total trials, we achieved accuracies of 82.5% for Object Recognition, 94.3% for OCR, 83.8% for Translation, and an overall accuracy of 86.9% (Fig. 11). Notably, OCR outperformed object recognition, contrary to typical task complexity expectations. We attribute the lower object recognition performance to grayscale imaging limitations and increased glare sensitivity with metallic kitchen utensils. For translation tasks, Korean font stylistic variations primarily impacted OCR accuracy rather than translation quality. Our OCR performance on book titles and authors closely aligns with Qwen’s reported Doc VQA benchmark of 96.4%, demonstrating that wearable camera-integrated earbuds can achieve similar performance in real-world usage scenarios. Finally, we highlight representative trials in Fig. 10. These examples demonstrate robust performance despite challenging conditions, including partial facial occlusion (Rows 4 and 5), objects split between both images (Rows 1 and 3), and minor hair interference (Rows 5 and 6), validating the system’s real-world applicability.

Post-Study Feedback. Participant responses revealed several key themes regarding usability, comfort, and adoption considerations. *Physical comfort and device integration:* Most participants reported that VueBuds felt similar to regular earbuds, with one noting they “felt the same way” and another stating “in terms of how it feels on the ears, it feels like wearing normal earbuds.” Several participants mentioned the devices were “maybe a little heavier” but emphasized this was “not too noticeable” and wouldn’t affect daily use. We note that the weight of the Sony Earbuds at baseline (8.5 grams) is approximately twice that of AirPods (4.3 grams) [11, 93].

Initial thoughts and camera concerns: Participants had predominantly positive initial reactions, with several expressing enthusiasm: “a very fascinating idea because its very similar to the meta glasses but probably more convenient,” “I didn’t really have any concerns! Thought it was a pretty cool idea,” and “I thought they were cool!”

Table 4: VueBuds power consumption results.

Component	IDLE	ACTIVE
SoC (ISP1807)	3.53 mW	19.2 mW
PMIC (MAX77650)	0.148 mW	0.148 mW
Camera (HM01B0)	0.2 mW	1.1 mW
Estimated Total	3.878 mW	20.5 mW
Measured Total	3.8 mW	24.9 mW

One participant appreciated a unique privacy advantage, noting “I think I could put my hair down and I can stop it from recording any visual data. Seems like that can’t be done with glasses.” Concerns centered on privacy regarding cloud data streaming, anticipated bulkiness, and camera positioning accuracy.

Wearer comfort and privacy: Participants expressed generally high comfort levels with wearing VueBuds in social contexts, with two responding “very comfortable” and one rating comfort as 7/10. Visual discreteness emerged as important, with one participant appreciating that camera-integrated earbuds “don’t have much difference than traditional earbuds looks wise.” However, privacy concerns affected comfort levels, with one participant noting “an increased concern in privacy and when the camera is active or what the camera sees.” Comfort was often conditional on utility, depending on whether the device “helps me ask questions about the visual scene rather than just taking pictures.”

Comparison to smartphone alternatives: Participants overwhelmingly favored VueBuds over phone workflows, describing them as providing “a lot more seamless and integrated experience, less steps.” Participants highlighted advantages during physical activities: “when I’m running or biking, it’s a lot easier to just say a wake word and ask a question than to stop, take out my phone, take a photo, etc.” However, one participant made nuanced distinctions, preferring earbuds for “questions that able to be answered within [a few] seconds” but favoring ChatGPT for “problems that require back and forth conversation or additional details.”

Practical use cases: Translation emerged as the most prominent application, particularly for Asian market shopping where participants found it “hard for me to read what it says and the pictures are not sufficient.” Travel applications were frequently cited for “traveling in a foreign country with language barriers,” along with accessibility support, plant identification, hands-free multitasking during “typing, piano, cooking,” and academic support. However, social considerations emerged, with one participant noting voice commands can be “awkward to use in public,” suggesting “gestures or nonverbal cues” as potential alternatives.

4.3 System Evaluations

4.3.1 Power analysis. To evaluate the viability of VueBuds from a power perspective, we conducted comprehensive profiling across varying usage patterns and analyzed battery life impact.

Power Profiling Methodology. We connected a power supply with μA resolution to the battery terminals of the VueBuds camera module, and measured the current draw at 3.8V in both the ACTIVE and IDLE state for at least 20 seconds. For each operating mode, we averaged the current draw and derived the power consumption numbers in Table 4. Based on these measurements, we calculated VueBuds’ power consumption across different query frequencies ranging from standby operation (0 queries/hour) to

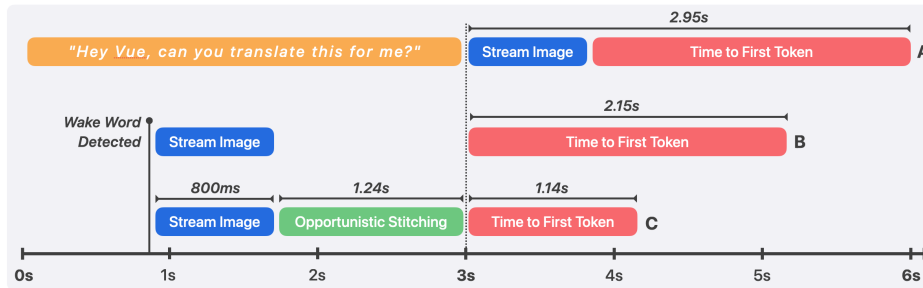


Figure 12: Latency breakdown across three system configurations. (A) Current prototype: 2.95s total latency due to waiting for complete audio processing before camera activation. (B) With on-device wake word detection: 2.15s latency by streaming images during query completion. (C) With opportunistic stitching: 1.14s latency through parallel image stitching while the user speaks, reducing VLM input tokens by 47%. The vertical dotted line indicates typical query completion time.

intensive use (60 queries/hour). Each query involves transitioning from IDLE (3.8mW) to ACTIVE (24.9mW), capturing and streaming visual data for three seconds, and then returning to IDLE. We calculated a weighted average power consumption across each minute, accounting for the proportion of time spent in each state. From this analysis, we report that in IDLE mode, the VueBud camera module adds 3.8mW to the baseline power consumption, while intensive usage of 60 queries per hour adds 4.9mW to the baseline.

Battery Life Impact on Commercial Earbuds. We model the battery life impact when integrated with two commercial wireless earbuds: Sony WF-1000XM3 (65mAh battery capacity) [15] and Apple AirPods Pro 2 (49.7mAh capacity) [5, 24]. Both the devices advertise a 6-hour battery life during streaming with ANC enabled. With the advertised battery life and battery capacity, we calculate the baseline power consumption of both earbuds, which comes out to 40.08mW and 30.65mW, respectively. We then combine this baseline power consumption with our system, and the battery life impact was 11-14%, even under high usage (60 queries per hour). Table 5 shows the battery life numbers for VueBuds using the Sony WF-1000XM3. For AirPods Pro 2, we estimate 5.23 and 5.18 hours of battery life at 5 and 60 queries per hour, respectively.

Table 5: VueBuds battery life across queries per hour.

Queries/hr	VueBuds on Sony Earbuds
0	5.48 hrs
5	5.47 hrs
10	5.46 hrs
20	5.44 hrs
60	5.35 hrs

Always-on Camera Analysis. While wake-word activation suffices for visual question answering, we also evaluated continuous camera operation for potential applications like ambient visual intelligence. With VueBuds constantly in ACTIVE state, battery life decreases to 3.5 hours (Sony WF-1000XM3) and 3.18 hours (AirPods Pro 2), reductions of 42% and 47% respectively. Although significant, we note that emerging earbuds like the AirPods Pro 3 [7] advertise 8-hour battery life, suggesting that advances in silicon may enable always-on visual capture while maintaining acceptable battery life. Regarding thermal safety, continuous camera operation increases baseline power consumption from 40mW to 65mW, which remains

well below thresholds associated with thermal discomfort for on-body wearables [28]. We note that Ray-Ban Meta does not currently support an always-on camera [29]

4.3.2 Latency Analysis. To evaluate VueBuds’ real-time performance, we define end-to-end latency as the time from query completion (t_0) to first token generation by the vision language model (t_1). We start off by characterizing our dual image acquisition latency, and then compare the time to first token using dual images versus stitched images. Finally, we analyze three system configurations: our prototype implementation, a system with on-device wake word detection, and an optimized system that additionally performs opportunistic image stitching.

Image Acquisition Latency. We characterize our dual image acquisition latency by measuring the time it takes for the host to trigger the VueBuds cameras over BLE, and receive images from each earbud. Our pipeline sends a BLE write command to both VueBuds to transition them from IDLE to ACTIVE, captures a single camera frame, and transmits the data back to the host. Across 66 trials, we report dual-image acquisition latency as ($\bar{x} = 800.1ms$, $\sigma = 0.06ms$), which closely aligns with our expectations from §3.1.4.

Time to First Token Latency - Dual and Stitched. To calculate time-to-first-token for dual 324x239 images, we benchmarked multiple trials from our user study data offline. After preloading the model into memory with a dummy query, we measured VLM inference time using paired images with a fixed text query. Direct processing of dual images yielded an average time-to-first-token of 2.15 seconds ($\sigma = 0.002$). We then evaluated stitching runtime by inputting these images through our stitching algorithm, yielding an average of 0.123 seconds ($\sigma = 0.01$). Finally, we measured inference across the set of successfully stitched images, which achieved an average time-to-first-token of 1.14 seconds ($\sigma = 0.23$). This represents a 47% reduction compared to dual image processing.

End-to-End Latency. Our current implementation processes audio using Whisper [81], which operates on 2.3-second chunks before performing fuzzy matching for wake word detection. The full pipeline consists of: (1) audio transmission via Bluetooth HFP, (2) wake word detection from buffered audio, (3) BLE command transmission to activate cameras, (4) image capture and streaming, (5) VLM inference, (6) text-to-speech conversion, and (7) audio playback through earbuds. This implementation using Whisper introduces significant latency, as it processes 2.3 s audio chunks

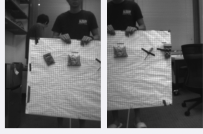
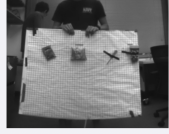



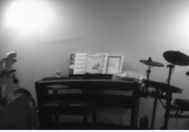
Stereo Images	Response	Stitched Image	Response
	I see a person holding a large grid-like board with two small bags attached to it. The person is wearing a shirt with the word "NAVY" on it. There's a desk and a chair in the background.		I see a person holding a large grid paper with some small items placed on it. There's a pen and what looks like a small box or container on the paper. The person is wearing a shirt with "NAVY" written on it.
	I see a hand holding a small object, possibly a card or a piece of paper, with some text on it. The background shows a room with shelves, a table, and some framed pictures on the wall. There's also a bright light source above, which is casting a glow on the scene.		I see a person holding a book with Korean text on it. The room has a cozy ambiance with warm lighting and some furniture in the background.
	I see a drum set in the foreground with a music book or sheet music on a stand. There's also a guitar to the left side of the frame. The lighting is quite dim, giving the image a dark, moody feel.		I see a room with a piano and a drum set. There's a book on the piano and a poster on the wall. The room looks like it's set up for music practice.

Figure 13: Comparison of VLM responses to dual images (left) versus stitched panoramic images (right). Reducing redundant information reduces the total input tokens, reducing inference latency by 47%.

and wake word detection occurs after the user has completed their query and not at the onset (Fig. 12).

Existing wake word detection models on embedded devices achieve sub-100ms detection latency from the end of the wake word [53]. With proper integration, wake word detection would be integrated with VueBuds hardware, enabling immediate camera activation upon hearing "Hey Vue" or "VueBuds." This would allow images to start streaming while the user completes their query, masking image capture latency. By the time the query ends, the host device has already received and buffered visual data. This optimization would reduce the effective end-to-end latency to the VLM's time to first token (2.15s on M4 Mac Mini Pro) plus audio synthesis overhead.

We also explore optimizing latency through lightweight ORB-based stitching during query completion. While the user speaks, the host device stitches left and right images to reduce redundant visual information. When successful, this reduces VLM input tokens, yielding a 46% improvement in time-to-first-token compared to raw L/R image pairs. However, due to parallax effects from the spatial separation of ear-level cameras, stitching confidence can vary based on scene geometry and how VueBuds are worn by a user. We introduce this technique as Opportunistic Stitching, achieving a time to first token latency of 1.14 seconds.

Fig. 12 illustrates the latency breakdown across all three configurations, while Fig. 13 demonstrates successful panoramic stitching results that enable this optimization. These results show that with proper system integration, one could achieve well under 2-second response times, improving the end-to-end user experience.

5 Discussion and Future Work

As our work presents the first system to integrate cameras and visual intelligence with wireless earbuds, we use this section to outline the limitations of our current prototype and, more importantly, articulate promising research directions. We organize this discussion into three key pillars: (1) viewpoint, deictic grounding, and interaction challenges unique to the ear-level vantage point; (2) social dynamics and privacy implications of the form factor;

and (3) hardware and platform considerations for scaling to future low-power systems.

5.1 Egocentric capture and interaction

Egocentric viewpoint coverage and alignment. In §4.1.1, we empirically validated that ear-level cameras can effectively capture a user's forward field of view through geometric modeling and real-world measurements, with blind spot lengths remaining well below the Harmon distance at camera orientations up to 10°. Notably, VueBuds achieves a 108° horizontal field of view at 10° offset, comparable to Ray-Ban Meta's 100°. In our in-person user study (§4.2.3), participants were instructed to interact with objects naturally by facing their head toward the target. Across 130 trials, no queries failed due to objects being outside the cameras' field of view.

Nevertheless, scenarios exist where ear-level cameras may miss a target. While VueBuds provides 87° of vertical coverage, scenarios involving extreme ocular declination without head movement, such as glancing sharply down at a pill bottle while keeping the head level, may fall outside the frame. While the anatomical limit for downward gaze is approximately 47°±8° [64], ergonomic literature identifies the comfortable resting gaze range as 15-30° [33, 39], which sits safely within VueBuds' vertical field of view. This is demonstrated by a sample from our user study in Figure 14 (Row 4) where VueBuds was able to clearly capture and identify an object the user naturally held close to their chest in a cluttered scene. Additionally, earbuds may not be worn consistently across users or sessions, introducing variability in camera alignment. This could be addressed through improved ergonomic design, and future systems could provide auditory feedback if orientation discrepancies are detected between the two earbuds.

Deictic grounding and interaction modalities. Unlike smart glasses and mixed reality headsets, which can provide both a forward-facing camera and an inward-facing camera for eye-tracking to infer user attention, ear-level cameras represent a different sensing configuration with distinct tradeoffs. Smart glasses can use gaze as an implicit disambiguation signal for referential intent ("look at


Stereo Images	Query	Response
	"what am I holding?"	You're holding a box of Cheez-It crackers.
	"who is this book by?"	I see the book in the image is titled "What's Our Problem?" and it's written by Tim Urban.
	"what am I holding?"	I see you're holding a small object in your hands. It looks like a watch or a similar device.
	"what book is this?"	I see the book titled "Data Structures in C++" by Timothy Budd. It's being held in someone's hand, and there are other books and items on the desk in the background.
	"what am I holding in my hand?"	I see you're holding a pack of Top Ramen noodles in your hand.

Figure 14: Early signs of disambiguation from our in-person study. Each row shows a cluttered scene with multiple candidate objects. Queries like “what am I holding?” or “what book is this?” achieve deictic grounding when users grasp the target object.

this”) [43, 63]. In contrast, cameras placed on the ears do not have direct access to the user’s visual fovea or natural gaze cues.

Our current VueBuds prototype relies primarily on speech commands for intent expression. Because the system lacks access to gaze, users in our study disambiguated references in cluttered environments by physically picking up and holding objects. This strategy worked well for tabletop interactions but becomes limiting in scenarios where objects are too large, distant, or fixed in place. Compared to systems such as GazePointAR [63], which uses eye tracking for referential grounding, the present VueBuds prototype offers fewer implicit channels for resolving ambiguous queries.

Despite these limitations, we see several promising pathways for enhancing disambiguation capabilities in ear-level camera systems. As shown in Figure 14, samples from our in-person study demonstrate that VueBuds can disambiguate objects held in a user’s hand even in cluttered visual environments, suggesting that hand-object relationships serve as a strong cue for referential intent. Future systems could extend this capability by using VLMs to infer pointing gestures directly from egocentric imagery, or by incorporating explicit hand tracking to detect pointing direction and identify likely referents. Such pointing gestures have been explored in the context of smart glasses [43]. VueBuds shares a comparable field of view with Ray-Ban Meta, which already supports pointing for referential intent. Incorporating pointing-based grounding, either implicitly via VLM inference or through dedicated hand-tracking, appears well within reach. More broadly, this opens an opportunity for the research community to explore new multimodal disambiguation strategies tailored specifically to ear-level perspectives, rather than adapting techniques built for head-worn devices.

Finally, voice-based invocation introduces additional interaction challenges. As noted by a participant in our user study (§4.2.3),

voice commands can be “awkward to use in public,” with the participant suggesting “gestures or nonverbal cues” as potential alternatives. The current single-turn query model further limits recovery from ambiguity, as users cannot engage in multi-turn clarification dialogues to refine their request. Future iterations could explore full-duplex conversational interaction [98] to enable more natural back-and-forth exchanges, as well as alternative input modalities such as in-air gestures recognized by the cameras or existing tap-based mechanisms on the earbud surface.

5.2 Social dynamics and privacy

Privacy signaling and bystander transparency. Embedding cameras into wearable devices raises privacy concerns for both wearers and bystanders [18, 46]. Extensive prior work on camera-equipped smart glasses highlights concerns and mitigation strategies that are likewise relevant to Vuebuds. Research on smart-glasses cameras found that wearers feel emotionally burdened with preserving bystander privacy, experiencing discomfort and social pressure even when not actively recording [18]. When considering ethical implications, users typically compare camera wearables to smartphones and CCTV, and view benign use as acceptable while drawing the line at deliberate concealed recording [18]. This presents a challenge for camera-integrated earbuds and glasses which can capture imagery without obvious behavioral cues, compared to smartphones that require a visible holding posture that signals recording intent. Furthermore, prior work found that bystanders do not recognize inconspicuous button taps on smart glasses [18] as recording triggers without verbal audio cues. This complicates the ability of bystanders to provide informed consent [46].

Against this backdrop, VueBuds offers several inherent privacy advantages over existing commercial smart glasses. In contrast to




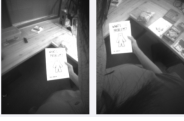
Stereo Images	Query	Response
	"what am I holding?"	You're holding a small object in your hand, but it's hard to tell exactly what it is. It looks like it might be a piece of jewelry or a small electronic device.
	"can you translate this for me?"	I see a person holding a book with Korean text on it. The text reads "창이슬" which translates to "Window Sill" in English. The scene is set in an office with desks, chairs, and shelves in the background.
	"can you translate this for me?"	I see a sign with the text "7080" and some Korean characters underneath. It looks like a label or a sign in a room, possibly an office or a workspace.
	"who is this book by?"	I see the book in the left earbud feed. The text on the book says "What's the Problem?" and there's a simple drawing of a person on it. The author's name isn't visible on the cover.

Figure 15: Explicit failures from our in-person study. The majority of errors stemmed from hardware imaging limitations. Specifically, low sensor resolution and limited dynamic range compromised OCR performance, particularly on fine print and small text. Additionally, specular glare from adverse lighting conditions disrupted object recognition capabilities.

Ray-Ban Meta, all VLM processing for VueBuds occurs on-device, ensuring that captured imagery is never transmitted to the cloud. Additionally, VueBuds captures only low-resolution monochrome images (324×239 pixels) intended for VLM inference rather than photo or video archival, reducing (though not eliminating) the risk of incidental bystander identification. Finally, camera activation explicitly requires a spoken wake-word, providing an audible signal to nearby individuals that visual capture is occurring.

In an attempt to combat these privacy challenges, smart glasses like Ray-Ban Meta incorporate an outward LED indicator for increased bystander transparency. While an LED is integrated into VueBuds hardware, prior work has shown that they are not fully effective under real-world conditions: they are difficult to see in bright sunlight, unnoticeable at a distance or in crowds, and largely unrecognized by the general public [18, 46]. Audio feedback, such as a shutter sound when capturing, offers the wearer an audible cue that VueBuds could readily support through its integrated speakers. The earbud form factor provides an advantage here, as in-ear audio delivery is more robust against ambient noise compared to the open-ear speakers found in smart glasses. However, while such indicators provide some transparency to wearers, we acknowledge they do not offer complete protection for bystanders, given the current low public awareness of these devices [18].

Several technical mitigations could strengthen privacy in future iterations. Recent work demonstrated that gesture-based image cropping, where users define visual boundaries through hand gestures before VLM processing, reduces explicit privacy leakage (e.g., bystander faces) by 70.6% while maintaining response quality [106]. Automatic face detection and blurring could further protect bystander identities. Low-tech solutions also merit consideration: physical lens covers, analogous to webcam covers on laptops, provide a mechanism legible to both wearers and observers [18]. For VueBuds specifically, as noted by a participant in our user study, wearers can naturally occlude the cameras simply by putting their hair down, providing an intuitive physical control unavailable with

glasses-based systems. Prior work has also explored opt-in and opt-out gestures enabling bystanders to signal recording preferences. However, consent-based approaches do not scale to crowded environments and gestures with negative connotations may cause acquiescence effects where bystanders silently accept privacy infringements [18, 58].

Finally, even with careful image cropping, VLMs can infer sensitive attributes from subtle background cues. Prior work found that models correctly inferred location from architectural details, signage, and geographical features in 65.6% of test cases [106]. Addressing such inferential disclosure, particularly when query targets themselves contain identifying features, remains an open challenge for privacy-preserving VLM interaction on wearables.

Form factor and adoption considerations. A central motivation for VueBuds is the observation that wireless earbuds represent a more broadly accessible platform than smart glasses (§4.2.1). For wearers, earbuds benefit from social normalization that smart glasses have struggled to achieve. The negative reception of Google Glass illustrates how prominent camera-equipped eyewear can trigger social friction, referred to as the "Glassholes" effect [32]. In contrast, earbuds are already worn for extended or episodic sessions during commutes, exercise, and work. Participants in our study reported that VueBuds felt similar to regular earbuds, suggesting lower barriers to personal adoption.

However, we note that wearer comfort does not resolve bystander concerns. The visual discreteness that makes earbuds socially acceptable to wear may simultaneously make embedded cameras less detectable, and thus more concerning, to bystanders who cannot easily identify when recording is occurring. This tension between wearer accessibility and bystander awareness represents an important design challenge for camera-integrated earbuds, as discussed in the prior section.

The earbud form factor does offer one distinct advantage: it complements rather than competes with existing eyewear. Users

who require prescription lenses face a particular challenge with camera-equipped glasses: removing the device to protect bystander privacy means sacrificing vision correction. Earbuds, by contrast, are inherently episodic accessories that users routinely remove and stow without functional consequence. For users who depend on prescription eyewear, VueBuds offers visual AI capabilities as a complementary accessory rather than a replacement for something they need. The choice between ear-level and eye-level cameras ultimately reflects a trade-off between adoption accessibility and perceptual fidelity. VueBuds prioritizes the former to explore whether useful visual intelligence can be achieved in this more ubiquitous form factor.

5.3 Hardware and platform considerations

Resolution, depth, and going beyond greyscale. The visual question answering tasks evaluated in our user studies did not require very high resolution and were color-agnostic, i.e., they were proximal tasks and did not involve identifying specific items by color. We opted for monochrome imaging to minimize streaming latency, as color imaging would increase bandwidth requirements by 3x (8-bit) or 2.25x (6-bit depth). At our 324x239 resolution, color streaming alone would require 2.4 seconds (8-bit) or 1.8 seconds (6-bit) per frame over our bandwidth-constrained BLE connection, far exceeding acceptable latency for real-time interaction. However, these constraints introduced concrete failure modes in our in-person study. As shown in Figure 15, the majority of errors stemmed from hardware imaging limitations: low sensor resolution and limited dynamic range compromised OCR performance on fine print and small text, while specular glare from adverse lighting conditions impacted object recognition performance. These failures highlight the tradeoff between streaming latency and resolution in bandwidth-constrained wearable systems. Future implementations could explore on-device JPEG compression or super-resolution [99] to reduce bandwidth but support higher resolution, frame rate, and color imaging. Another noteworthy research direction is integrating wireless time synchronization [25] which could improve stitching success rates. Finally, enabling stereo depth estimation [38, 100] could enhance spatial audio functionality, enabling accurate 3D sound source placement and room acoustic modeling [59].

Adaptive camera positioning. Camera positioning also varies with individual wearing styles, affecting the captured field of view. While our fixed angular orientations (5–10°) generalized well across participants in our user study, individual variations in head/ear anatomy and wearing preferences still introduce some variability. Future systems could incorporate IMU-based pose estimation with micro-servo adjustments [48] to dynamically adapt camera angles, ensuring optimal perspective regardless of user-specific wearing patterns.

Improving panoramic stitching. Finally, while our lightweight stitching achieves 46% improvement in time-to-first-token, traditional panoramic stitching assumes cameras rotate around a fixed point and capture planar scenes, assumptions that are marginally violated by the spatial separation across the head. The resulting parallax effects can exceed homography-based transformation limits, causing our system to fall back to dual independent images. Although

parallax-tolerant stitching methods exist [35, 51, 65], they require 5–40 seconds of GPU processing. Real-time parallax-tolerant stitching for distributed camera systems remains a research challenge.

Computational requirements, deployment trade-offs, and scaling to mobile phones. Our prototype performs VLM inference on desktop-class hardware (Mac Mini M4 Pro). While not available during our development, Apple’s FastVLM has since been released [9]. This framework supports the same Qwen model used in VueBuds in its 7B configuration. While the fastest latencies (800–1200ms time-to-first-token on iPhone 16 Pro) currently require the smaller 0.5B configuration, continued advances in mobile hardware and vision encoders suggest that efficient on-device inference with larger models is increasingly feasible.

Additionally, while we used Qwen2.5-VL and achieved promising results, performance could be enhanced through model fine-tuning on domain-specific training data captured from earbud perspectives. Our choice of on-device processing preserves user privacy but incurs latency penalties compared to cloud-based alternatives. Larger cloud models, such as those powering Ray-Ban Meta, would likely provide faster response times and superior performance on complex reasoning tasks, though at the cost of requiring constant connectivity and raising privacy concerns.

On the hardware front, while our prototype retrofits existing earbuds, emerging SoCs with integrated Bluetooth Classic and BLE [1–3] could enable custom implementations with tighter integration to better optimize for performance. Ultra-low-power AI processors like WiseEye2 [4] demonstrate gesture recognition at 1.2mW, fitting within the power envelope for wireless earbuds. Integrating low-power neural processors could enable on-device preprocessing for in-air gestures for interaction, or human recognition to dynamically control spot-forming radius [26] or vision-aided speech separation [26, 44, 47, 96, 97].

6 Conclusion

This work shows that wireless earbuds can serve as a viable platform for egocentric visual intelligence. By integrating low-power cameras into a familiar ear-worn form factor and pairing them with vision language models, VueBuds enable real-time visual intelligence under strict size, power, and bandwidth constraints. Our results show that binocular ear-level cameras provide comprehensive viewpoint coverage for everyday tasks, and that low-resolution, monochrome imagery is sufficient for scene understanding, text reading, and translation when processed by modern VLMs. Across system benchmarks and user studies, VueBuds achieve response quality comparable to commercial smart glasses while operating within the practical limits of existing earbud hardware. More broadly, VueBuds takes a first step toward bringing visual computing to one of the most widely adopted wearable form factors, positioning earbuds as a viable and broadly accessible platform for visual intelligence.

References

- [1] [n. d.]. FSC-BT631D Nordic nRF5340 BT5.3 Classic and LE Audio Module Transceiver. <https://www.feasycom.com/product/fsc-bt631d/>.
- [2] [n. d.]. Infineon CYW20819/20820. <https://www.infineon.com/part/CYW20819>.
- [3] [n. d.]. Qualcomm QCC5171. <https://www.qualcomm.com/products/internet-of-things/consumer/audio/qcc51xx-series/qcc5171>.
- [4] [n. d.]. WiseEye2 AI Processor (WE2). <https://www.himax.com.tw/products/wiseeye-ai-sensing/wiseeye2-ai-processor/>.

- [5] 2022. *3C certification reveals the Apple AirPods Pro 2's battery capacity*. https://www.gsmarena.com/3c_certification_reveals_the_airpods_pro_2s_battery_capacity-news-56031.php
- [6] 2025. Enhancing Vision, the Smart Way. *Review of Ophthalmology* (may 2025). <https://www.reviewofophthalmology.com/article/enhancing-vision-the-smart-way> Accessed: [Dec 1 2025].
- [7] 2025. Introducing AirPods Pro 3, the ultimate audio experience. <https://www.apple.com/newsroom/2025/09/introducing-airpods-pro-3-the-ultimate-audio-experience/>
- [8] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikołaj Bińkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. 2022. Flamingo: a Visual Language Model for Few-Shot Learning. In *NeurIPS*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Eds.).
- [9] Pavan Kumar Anasosalu Vasu, Fartash Faghri, Chun-Liang Li, Cem Koc, Nate True, Albert Antony, Gokul Santhanam, James Gabriel, Peter Grasch, Oncel Tuzel, and Hadi Pouransari. 2025. FastVLM: Efficient Vision Encoding for Vision Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [10] Apple. [n. d.]. Choose your new Mac mini. <https://www.apple.com/shop/buy-mac/mac-mini/m4-pro>
- [11] Apple. 2025. AirPods 4. <https://www.apple.com/airpods-4/specs/>
- [12] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. 2025. Qwen2.5-VL Technical Report. arXiv:2502.13923 [cs.CV] <https://arxiv.org/abs/2502.13923>
- [13] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. 2025. Qwen2.5-VL Technical Report. *arXiv preprint arXiv:2502.13923* (2025).
- [14] Mirza Samad Ahmed Baig, Syeda Anshrah Gillani, Shahid Munir Shah, Mahmoude Aljawarneh, Abdul Akbar Khan, and Muhammad Hamzah Siddiqui. 2024. AI-based wearable vision assistance system for the visually impaired: Integrating real-time object recognition and contextual understanding using large vision-language models. *arXiv preprint arXiv:2412.20059* (2024).
- [15] BatteryClub. [n. d.]. Batteries for Sony WF-1000XM3, Samsung Galaxy Buds, and Buds Live. https://www.batteryclub.org/Goods/ECN12777_SCA_OTH-Z55-Zenipower-Battery.html
- [16] Abdelkareem Bedri, Richard Li, Malcolm Haynes, Raj Prateek Kosaraju, Ishaan Grover, Temiloluwa Prioleau, Min Yan Beh, Mayank Goel, Thad Starner, and Gregory Abowd. 2017. EarBit: Using Wearable Sensors to Detect Eating Episodes in Unconstrained Environments. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 3, Article 37 (Sept. 2017), 20 pages. doi:10.1145/3130902
- [17] Andrea Bejarano-Carbo, HyoChan An, Kyojin Choo, Shiyu Liu, Dennis Sylvester, David Blaauw, and Hun-Seok Kim. 2022. Millimeter-Scale Ultra-Low-Power Imaging System for Intelligent Edge Monitoring. doi:10.48550/ARXIV.2203.04496
- [18] Divyanshu Bhardwaj, Alexander Ponticello, Shreya Tomar, Adrian Dabrowski, and Katharina Krombholz. 2024. In Focus, Out of Privacy: The Wearer's Perspective on the Privacy Dilemma of Camera Glasses. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 577, 18 pages. doi:10.1145/3613904.3642242
- [19] Laura Boccardo. 2020. Viewing distance of smartphones in presbyopic and non-presbyopic age. <https://pmc.ncbi.nlm.nih.gov/articles/PMC8093538/>
- [20] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of Artificial General Intelligence: Early experiments with GPT-4. arXiv:2303.12712 [cs.CL]
- [21] Runze Cai, Nuwan Janaka, Hyeoncheol Kim, Yang Chen, Shengdong Zhao, Yun Huang, and David Hsu. 2025. AiGet: Transforming Everyday Moments into Hidden Knowledge Discovery with AI Assistance on Smart Glasses. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (CHI '25). Association for Computing Machinery, New York, NY, USA, Article 631, 26 pages. doi:10.1145/3706598.3713953
- [22] Justin Chan, Nada Ali, Ali Najafi, Anna Meehan, Lisa Mancl, Emily Gallagher, Randall Bly, and Shyammath Gollakota. 2022. An off-the-shelf otoacoustic-emission probe for hearing screening via a smartphone. *Nature Biomedical Engineering* 6 (10 2022), 1–11. doi:10.1038/s41551-022-00947-6
- [23] Justin Chan, Sharat Raju, Rajalakshmi Nandakumar, Randall Bly, and Shyammath Gollakota. 2019. Detecting middle ear fluid using smartphones. *Science Translational Medicine* 11 (05 2019), eaav1102. doi:10.1126/scitranslmed.aav1102
- [24] Hartley Charlton. 2022. *AirPods Pro 2 Earbud Batteries 15% Bigger, But Charging Case Sees Little Improvement*. https://www.gsmarena.com/3c_certification_reveals_the_airpods_pro_2s_battery_capacity-news-56031.php
- [25] Ishan Chatterjee, Maruchi Kim, Vivek Jayaram, Shyammath Gollakota, Ira Kelmacher, Shwetak Patel, and Steven M Seitz. 2022. ClearBuds: wireless binaural earbuds for learning-based speech enhancement. In *MobiSys*.
- [26] Tuochao Chen, Malek Itani, Sefik Eskimez, Takuya Yoshioka, and Shyammath Gollakota. 2024. Hearable devices with sound bubbles. *Nature Electronics* (2024).
- [27] Tuochao Chen, Benjamin Steeper, Kinan Alsheikh, Songyun Tao, François Guimbretière, and Cheng Zhang. 2020. C-Face: Continuously Reconstructing Facial Expressions by Deep Learning Contours of the Face with Ear-mounted Miniature Cameras. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology* (Virtual Event, USA) (UIST '20). Association for Computing Machinery, New York, NY, USA, 112–125. doi:10.1145/3379337.3415879
- [28] COMSOL Blog. 2016. Design Safe Wearable Technology with Heat Transfer Modeling. COMSOL Blog. <https://www.comsol.com/blogs/design-safe-wearable-technology-with-heat-transfer-modeling> Accessed: [Insert Date]. This article finds a maximum safe power dissipation of 1.75 W before skin temperature exceeds the damage threshold of 43°C..
- [29] ConnectSafely. 2024. Guide to Ray-Ban Meta Smart Glasses. <https://connectsafely.org/rayban/>
- [30] 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* (Pittsburgh, PA, USA) (UIST '24). Association for Computing Machinery, New York, NY, USA, Article 19, 15 pages. doi:10.1145/3654777.3676379
- [31] DuckyBlender. 2024. moondream returns no response. <https://github.com/ollama/ollama/issues/4063>
- [32] Brian L Due. 2015. The social construction of a Glasshole: Google Glass and multiactivity in social interaction. *Psychology Journal* 13, 2 (2015).
- [33] Renzo D'Amato, Fabrizio Cutolo, Giovanni Badiali, Marina Carbone, Hao Lu, Harm Hogenbirk, and Vincenzo Ferrari. 2022. Key ergonomics requirements and possible mechanical solutions for augmented reality head-mounted displays in surgery. *Multimodal Technologies and Interaction* 6, 2 (2022), 15.
- [34] Xiaoran Fan and Trausti Thormundsson. 2023. Design Earable Sensing Systems: Perspectives and Lessons Learned from Industry. In *Adjunct Proceedings of the 2023 ACM International Joint Conference on Pervasive and Ubiquitous Computing and the 2023 ACM International Symposium on Wearable Computers*.
- [35] Feng Liu Fan Zhang. 2014. Parallax-tolerant Image Stitching. https://www.cv-foundation.org/openaccess/content_cvpr_2014/papers/Zhang_Parallax-tolerant_Image_Stitching_2014_CVPR_paper.pdf
- [36] Andrea Ferlini, Dong Ma, Robert Harle, and Cecilia Mascolo. 2021. EarGate: gait-based user identification with in-ear microphones. In *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking* (New Orleans, Louisiana) (MobiCom '21). Association for Computing Machinery, New York, NY, USA, 337–349. doi:10.1145/3447993.3483240
- [37] Google. [n. d.]. Google Glasses: Never miss a moment. <https://www.google.com/glass/photography/>
- [38] Xianda Guo, Chenming Zhang, Youmin Zhang, Dujun Nie, Ruilin Wang, Wenzhao Zheng, Matteo Poggi, and Long Chen. 2024. Stereo Anything: Unifying Stereo Matching with Large-Scale Mixed Data. arXiv:2411.14053 [cs.CV] <https://arxiv.org/abs/2411.14053>
- [39] Herbert Heuer and D Alfred Owens. 1989. Vertical gaze direction and the resting posture of the eyes. *Perception* 18, 3 (1989), 363–377.
- [40] Jason Hong. 2013. Considering privacy issues in the context of Google glass. *Commun. ACM* 56, 11 (Nov. 2013), 10–11. doi:10.1145/2524713.2524717
- [41] Hein Min Htike, Tom H Margrain, Yu-Kun Lai, and Parisa Eslambolchilar. 2020. Ability of head-mounted display technology to improve mobility in people with low vision: A systematic review. *Translational Vision Science & Technology* 9, 10 (2020), 26–26.
- [42] Changshuo Hu, Qiang Yang, Yang Liu, Tobias Röddiger, Kayla-Jade Butkow, Mathias Ciliberto, Adam Luke Pullin, Jake Stuchbury-Wass, Mahbub Hassan, Cecilia Mascolo, and Dong Ma. 2025. A Survey of Earable Technology: Trends, Tools, and the Road Ahead. arXiv:2506.05720 [cs.HC] <https://arxiv.org/abs/2506.05720>
- [43] Xiyun Hu, Dizhi Ma, Fengming He, Zhengzhe Zhu, Shao-Kang Hsia, Chenfei Zhu, Ziyi Liu, and Karthik Ramani. 2025. GesPrompt: Leveraging Co-Speech Gestures to Augment LLM-Based Interaction in Virtual Reality. In *Proceedings of the 2025 ACM Designing Interactive Systems Conference (DIS '25)*. Association for Computing Machinery, New York, NY, USA, 59–80. doi:10.1145/3715336.3735769
- [44] Jeremy Zhengqi Huang, Jaylin Herskovitz, Liang-Yuan Wu, Cecily Morrison, and Dhruv Jain. 2025. Weaving Sound Information to Support Real-Time Sensemaking of Auditory Environments: Co-Designing with a DHH User. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (CHI '25). Association for Computing Machinery, New York, NY, USA, Article 75, 18 pages. doi:10.1145/3706598.3714268

- [45] International Data Corporation. 2024. Worldwide Quarterly Wearable Device Tracker. <https://www.idc.com/promo/wearable/vendor/>. Accessed: 2025-01-23.
- [46] Muhammad Zahid Iqbal and Abraham G. Campbell. 2023. Adopting smart glasses responsibly: potential benefits, ethical, and privacy concerns with Ray-Ban stories. *AI and Ethics* 3 (2023), 325–327. doi:10.1007/s43681-022-00155-7
- [47] Malek Itani, Tuochao Chen, Arun Raghavan, Gavriel Kohlberg, and Shyamnath Gollakota. 2025. Wireless Hearables With Programmable Speech AI Accelerators. *ACM MOBICOM* (2025).
- [48] Vikram Iyer, Ali Najafi, Johannes James, Sawyer Fuller, and Shyamnath Gollakota. 2020. Wireless steerable vision for live insects and insect-scale robots. *Science Robotics* 5 (07 2020), eabb0839. doi:10.1126/scirobotics.abb0839
- [49] Vikram Iyer, Ali Najafi, Johannes James, Sawyer Fuller, and Shyamnath Gollakota. 2020. Wireless steerable vision for live insects and insect-scale robots. *Science Robotics* 5, 44 (2020), eabb0839. doi:10.1126/scirobotics.abb0839
- [50] Suyao Ji, Jing Pu, Byong Chan Lim, and Mark Horowitz. 2016. A 220pJ/pixel/frame CMOS image sensor with partial settling readout architecture. (2016).
- [51] Feipeng Da Jian Yu. 2024. Parallax-Tolerant Image Stitching with Epipolar Displacement Field. <https://arxiv.org/html/2311.16637v2>
- [52] Lavender Jiang, Xujin Liu, Nima Nejatian, Mustafa Nasir-Moin, Duo Wang, Anas Abidin, Kevin Eaton, Howard Riina, Ilya Laufer, Paawan Punjabi, Madeline Miceli, Nora Kim, Cordelia Orillac, Zane Schnurman, Christopher Livia, Hannah Weiss, David Kurland, Sean Neifert, Yosef Dastagirzada, and Eric Oermann. 2023. Health system-scale language models are all-purpose prediction engines. *Nature* (2023).
- [53] Christin Jose, Yuriy Mishchenko, Thibaud Sénéchal, Anish Shah, Alex Escott, and Shiv Vitaladevuni. 2020. Accurate detection of wakedword start and end using a CNN. (2020). <https://www.amazon.science/publications/accurate-detection-of-wake-word-start-and-end-using-a-cnn>
- [54] Colleen Josephson, Lei Yang, Pengyu Zhang, and Sachin Katti. 2019. Wireless Computer Vision Using Commodity Radios. In *IPSN*.
- [55] Fahim Kawsar, Chulhong Min, Akhil Mathur, and Alessandro Montanari. 2018. Earables for Personal-Scale Behavior Analytics. *IEEE Pervasive Computing* 17, 3 (2018), 83–89. doi:10.1109/MPRV.2018.03367740
- [56] Muiz Ahmed Khan, Pias Paul, Mahmudur Rashid, Mainul Hossain, and Md Atiqur Rahman Ahad. 2020. An AI-based visual aid with integrated reading assistant for the completely blind. *IEEE Transactions on Human-Machine Systems* 50, 6 (2020), 507–517.
- [57] Maruchi Kim, Antonio Glenn, Bandhav Veluri, Yunseo Lee, Eyoel Gebre, Aditya Bagaria, Shwetak Patel, and Shyamnath Gollakota. 2024. IRIS: Wireless ring for vision-based smart home interaction. In *UIST*.
- [58] Marion Koelle, Swamy Ananthanarayan, Simon Czupalla, Wilko Heuten, and Susanne Boll. 2018. Your smart glasses' camera bothers me! exploring opt-in and opt-out gestures for privacy mediation. In *Proceedings of the 10th Nordic Conference on Human-Computer Interaction (Oslo, Norway) (NordCHI '18)*. Association for Computing Machinery, New York, NY, USA, 473–481. doi:10.1145/3240167.3240174
- [59] Shoichi Koyama, Enzo De Sena, Prasanga Samarasinghe, Mark R. P. Thomas, and Fabio Antonacci. 2025. Past, Present, and Future of Spatial Audio and Room Acoustics. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 1–5. doi:10.1109/ICASSP49660.2025.10890366
- [60] Kvgng. 2025. Pipeline (computing). [https://en.m.wikipedia.org/wiki/Pipeline_\(computing\)](https://en.m.wikipedia.org/wiki/Pipeline_(computing))
- [61] Pupil Labs. [n. d.]. Pupil Invisible Glasses: Eye tracking glasses for the real-world. <https://pupil-labs.com/products/invisible>
- [62] Chae Young Lee, Pu, Yi, Maxwell Fite, Tejus Rao, Sara Achour, and Zerina Kapetanovic. 2025. HyperCam: Low-Power Onboard Computer Vision for IoT Cameras. *ACM MobiCom* (2025).
- [63] 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 (Honolulu, HI, USA) (CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 408, 20 pages. doi:10.1145/3613904.3642230
- [64] Won June Lee, Ji Hong Kim, Yong Un Shin, Sunjin Hwang, and Han Woong Lim. 2019. Differences in eye movement range based on age and gaze direction. *Eye* 33, 7 (2019), 1145–1151.
- [65] Tianli Liao, Ce Wang, Lei Li, Guangen Liu, and Nan Li. 2024. Parallax-tolerant Image Stitching via Segmentation-guided Multi-homography Warping. <https://arxiv.org/abs/2406.19922>
- [66] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual Instruction Tuning. In *NeurIPS*.
- [67] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual Instruction Tuning. arXiv:2304.08485 [cs.CV] <https://arxiv.org/abs/2304.08485>
- [68] Dong Ma, Andrea Ferlini, and Cecilia Mascolo. 2021. OESense: Employing Occlusion Effect for in-Ear Human Sensing. In *MobiSys*.
- [69] Cognitive market Research. [n. d.]. Wireless Earbuds Market Report 2025 (Global Edition). <https://www.cognitivemarketresearch.com/wireless-earbuds-market-report>
- [70] maxi1134. 2025. Ollama Moondream: Latest returns no answer. <https://github.com/valentinfrlch/ha-llmvision/issues/201>
- [71] Meta. [n. d.]. Ray-Ban Meta Glasses. <https://www.meta.com/ai-glasses/>
- [72] Yurina Mizuho, Yohei Kawasaki, Takashi Amesaka, and Yuta Sugiura. 2024. EarAuthCam: Personal Identification and Authentication Method Using Ear Images Acquired with a Camera-Equipped Hearable Device. In *AHS*. 119–130.
- [73] 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 (Hamburg, Germany) (CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 529, 12 pages. doi:10.1145/3544548.3581392
- [74] moondream.ai. [n. d.]. Moondream.
- [75] Fukashi Morishita, Norihito Kato, Satoshi Okubo, Takao Toi, Mitsuru Hiraki, Sugako Otani, Hideaki Abe, Yuji Shinohara, and Hiroyuki Kondo. 2021. A CMOS Image Sensor and an AI Accelerator for Realizing Edge-Computing-Based Surveillance Camera Systems. In *2021 Symposium on VLSI Circuits*. 1–2.
- [76] Saman Naderiparizi, Mehrdad Hesar, Vamsi Talla, Shyamnath Gollakota, and Joshua R. Smith. 2018. Towards Battery-Free HD Video Streaming. In *NSDI*.
- [77] NovelBits. 2023. Bluetooth 5 speed: How to achieve maximum throughput for your BLE application. <https://novelbits.io/bluetooth-5-speed-maximum-throughput/>
- [78] Ollama. 2024. Gemma2 models with Ollama run extremely slow. <https://github.com/ollama/ollama/issues/9683> GitHub Issue #9683.
- [79] Jay Prakash, Zhijian Yang, Yu-Lin Wei, Haitham Hassanieh, and Romit Roy Choudhury. 2020. EarSense: Earphones as a Teeth Activity Sensor. In *MobiCom*.
- [80] Tech Radar. 2025. samsung and meta are looking into earbuds with cameras.
- [81] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning (Honolulu, Hawaii, USA) (ICML '23)*. JMLR.org, Article 1182, 27 pages.
- [82] T.P. Rani, S Susila Sakthy, P Kalaichelvi, Vignesh T, and Priyadharshan M. 2023. Visual Information Translator Using Smart Glasses for Blind. In *2023 Intelligent Computing and Control for Engineering and Business Systems (ICCEBS)*. 1–6. doi:10.1109/ICCEBS58601.2023.10449230
- [83] Grand View Research. [n. d.]. Smart Glasses Market Size, Share & Trends Analysis Report By Type (Binocular, Audio), By Operating System (Android, Linux), By Glass Tinting Technology, By Application, By Connectivity, By Region, And Segment Forecasts, 2025 - 2030. <https://www.grandviewresearch.com/industry-analysis/smart-glasses-market-report>
- [84] Tobias Röddiger, Tobias King, Dylan Ray Roodt, Christopher Clarke, and Michael Beigl. 2023. OpenEarable: Open Hardware Earable Sensing Platform (*UbiComp/ISWC '22 Adjunct*).
- [85] Tobias Röddiger, Michael Küttner, Philipp Lepold, Tobias King, Dennis Moschina, Oliver Bagge, Joseph A. Paradiso, Christopher Clarke, and Michael Beigl. 2025. OpenEarable 2.0: Open-Source Earphone Platform for Physiological Ear Sensing. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* (2025).
- [86] Davide Rossi, Francesco Conti, Manuel Eggiman, Stefan Mach, Alfio Di Mauro, Marco Guermandi, Giuseppe Tagliavini, Antonio Pullini, Igor Loi, Jie Chen, Eric Flamand, and Luca Benini. 2021. 4.4 A 1.3TOPS/W @ 32GOPS Fully Integrated 10-Core SoC for IoT End-Nodes with 1.7uW Cognitive Wake-Up From MRAM-Based State-Retentive Sleep Mode. In *ISSCC*.
- [87] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. 2011. ORB: An efficient alternative to SIFT or SURF. In *2011 International Conference on Computer Vision*. 2564–2571. doi:10.1109/ICCV.2011.6126544
- [88] Ali Saffari, Mehrdad Hesar, Saman Naderiparizi, and Joshua R. Smith. 2019. Battery-Free Wireless Video Streaming Camera System. In *RFID*. 1–8.
- [89] Samsung. [n. d.]. Discover a new perspective with Bixby Vision. <https://www.samsung.com/us/explore/bixby/sees/>
- [90] Nordic Semiconductor. 2022. *Things You Should Know About Bluetooth Range*. Nordic Semiconductor. <https://blog.nordicsemi.com/getconnected/things-you-should-know-about-bluetooth-range>
- [91] Sander Smets, Toon Goedemé, Anurag Mittal, and Marian Verhelst. 2019. 2.2 A 978GOPS/W Flexible Streaming Processor for Real-Time Image Processing Applications in 22nm FDSOI. In *ISSCC*.
- [92] Jake Stuchbury-Wass, Andrea Ferlini, and Cecilia Mascolo. 2023. Multimodal Attention Networks for Human Activity Recognition From Earable Devices. (2023), 258–260. doi:10.1145/3544793.3563422
- [93] Sony Support. 2025. Specifications. <https://www.sony.com/electronics/support/wireless-headphones-bluetooth-headphones/wf-1000xm3/specifications>
- [94] Gemma Team. 2025. Gemma 3 Technical Report. arXiv:2503.19786 [cs.CL] <https://arxiv.org/abs/2503.19786>
- [95] Ankitkumar Varshney, Maryam Chougale, Chetna Patel, and Mahendrasinh Chauhan. 2025. Evaluating usability of “the smart vision glasses” for individuals who are visually impaired and totally blind. *Saudi Journal of Ophthalmology* 39 (02 2025), 1–6. doi:10.4103/sjopt.sjopt_241_24

- [96] Bandhav Veluri, Malek Itani, Justin Chan, Takuya Yoshioka, and Shyamnath Gollakota. 2023. Semantic Hearing: Programming Acoustic Scenes with Binaural Hearables. In *ACM UIST*.
- [97] Bandhav Veluri, Malek Itani, Tuochao Chen, Takuya Yoshioka, and Shyamnath Gollakota. 2024. Look Once to Hear: Target Speech Hearing with Noisy Examples. In *ACM CHI*.
- [98] Bandhav Veluri, Benjamin N Peloquin, Bokai Yu, Hongyu Gong, and Shyamnath Gollakota. 2024. Beyond Turn-Based Interfaces: Synchronous LLMs as Full-Duplex Dialogue Agents. arXiv:2409.15594 [cs.CL] <https://arxiv.org/abs/2409.15594>
- [99] Bandhav Veluri, Collin Pernu, Ali Saffari, Joshua Smith, Michael Taylor, and Shyamnath Gollakota. 2023. *NeuriCam: Key-Frame Video Super-Resolution and Colorization for IoT Cameras*. Association for Computing Machinery, New York, NY, USA, Chapter 25, 1–17. <https://doi.org/10.1145/3570361.3592523>
- [100] Jialiang Wang, Daniel Scharstein, Akash Bapat, Kevin Blackburn-Matzen, Matthew Yu, Jonathan Lehman, Suhub Alsisan, Yanghan Wang, Sam Tsai, Jan-Michael Frahm, Zijian He, Peter Vajda, Michael F. Cohen, and Matt Uyttendaele. 2023. A Practical Stereo Depth System for Smart Glasses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 21498–21507.
- [101] Tianyi Wang, Xun Qian, Fengming He, Xiyun Hu, Ke Huo, Yuanzhi Cao, and Karthik Ramani. 2020. CAPturAR: An Augmented Reality Tool for Authoring Human-Involved Context-Aware Applications. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology (Virtual Event, USA) (UIST '20)*. Association for Computing Machinery, New York, NY, USA, 328–341. doi:10.1145/3379337.3415815
- [102] Yuntao Wang, Jiexin Ding, Ishan Chatterjee, Farshid Salemi Parizi, Yuzhou Zhuang, Yukang Yan, Shwetak Patel, and Yuanchun Shi. 2022. FaceOri: Tracking Head Position and Orientation Using Ultrasonic Ranging on Earphones. In *ACM CHI*.
- [103] Xreal. [n. d.]. XREAL: Building Augmented Reality for everyone. <https://www.xreal.com/us/aura/>
- [104] Han Xu, Zirui Li, Ningchao Lin, Qi Wei, Fei Qiao, Xunzhao Yin, and Huazhong Yang. 2021. MACSen: A Processing-In-Sensor Architecture Integrating MAC Operations Into Image Sensor for Ultra-Low-Power BNN-Based Intelligent Visual Perception. *IEEE Transactions on Circuits and Systems II: Express Briefs* (2021).
- [105] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, Qianyu Chen, Huarong Zhou, Zhensheng Zou, Haoye Zhang, Shengding Hu, Zhi Zheng, Jie Zhou, Jie Cai, Xu Han, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2024. MiniCPM-V: A GPT-4V Level MLLM on Your Phone. arXiv:2408.01800 [cs.CV] <https://arxiv.org/abs/2408.01800>
- [106] Ziyang Zhang, Chong Bao, Xiaokun Pan, Chia-Ming Chang, Takeo Igarashi, and Guofeng Zhang. 2025. Through the Lens of Privacy: Exploring Privacy Protection in Vision-Language Model Interactions on Smart Glasses. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25)*. Association for Computing Machinery, New York, NY, USA, Article 530, 8 pages. doi:10.1145/3706599.3720234

A User Study Questionnaire

This appendix contains the full questionnaire used in the user study. The study was conducted online using Google Forms.

Part 1: Introduction and Platform Accessibility Questions

Introductory Text: You will be participating in a research study which will ask you about your usage frequency of earbuds and glasses, and then have you rate the overall response quality from a visual AI system on a scale from 1–5.

1. What is your age?

- 10–19
- 20–29

- 30–39
- 40–49
- 50–59
- 60+

2. What is your gender?

- Male
- Female
- Non-binary
- Prefer not to say

3. How often do you wear glasses?

- Everyday
- 5–6 days a week
- 3–4 days a week
- 1–2 days a week
- I don't wear glasses

4. How often do you wear earbuds or headphones?

- Everyday
- 5–6 days a week
- 3–4 days a week
- 1–2 days a week
- I don't use headphones

Part 2: Evaluating Visual AI Response Quality

Instructions: Imagine you are wearing AI-enabled smart glasses. You will be shown a series of images along with a sample question you might ask your glasses. For each image, you will be shown two different AI-generated responses. Please rate each response independently on a scale from 1 (Poor) to 5 (Excellent). Note that the length of the response should NOT be a factor in your rating.

Scenario 1

User Query: "What do you see?"

[Image displayed: Hand holding a book titled "Normal People" by Sally Rooney]

Response A:

The image shows a person holding a book titled "Normal People" by Sally Rooney. The background appears to be a kitchen, with a stove and a pot visible. A checkered cloth hangs on the oven handle.

Response B:

I see a book titled Normal People by Sally Rooney.

Scenario 2

User Query: "What do you see?"

[Image displayed: Tray with two coasters, scotch tape, two Nintendo Switch controllers, bag of cookies, and a hair clip]

Response A:

A tray with various items on it. Including nintendo switch controllers, a hair clip, and a bag of snacks.

Response B:

The image shows a wooden tray with several items on it. There are two Nintendo Switch Joy-Con controllers, a small box labeled "Mikrotik," a small cube-shaped object, and a rectangular object with a pattern that appears to be a comb or a similar item. The objects are arranged in a somewhat scattered manner within the tray.

Scenario 3

User Query: "Read any text you can see"

[Image displayed: Book "Normal People" on a sofa]

Response A:

The image shows a book titled "Normal People" by Sally Rooney. The book is partially covered by a textured, fuzzy object, possibly a blanket or scarf. The text is readable and the book appears to be resting on a dark surface.

Response B:

Normal people, a novel by Salley Rooney

Scenario 4

User Query: "Translate this for me"

[Image displayed: Korean Jinro Chamisul bottle with Korean characters on it.]

Response A:

JINRO CHAMISUL FRESH
SOJU Jinro Chamisul

Response B:

The text on the bottle appears to be in Korean. The readable part says "INRO ..." which translates to "INRO Chonggi Sul" in English. "Chonggi Sul" is a type of Korean rice wine.

Scenario 5

User Query: "Read any text you can see"

[Image displayed: Book cover "Yellowface" by R.F. Kuang on a sofa]

Response A:

The image shows the cover of a book titled "Yellowface" by R.F. Kuang.

Response B:

Yellowface R.F. Kuang

Scenario 6

User Query: "Translate this for me"

[Image displayed: Korean cream soup food packaging]

Response A:

This appears to be a package of Korean instant noodles.

Response B:

The text in the image appears to be in Korean. The readable text is "creams" which translates to "Creams" in English.

Scenario 7

User Query: "Translate this for me"

[Image displayed: Korean cold noodle dish package on a stovetop]

Response A:

The text on the package appears to be in Korean. The visible text reads "mulnaengmyeon," which translates to "Cold Noodles" in English. This suggests the package contains a cold noodle dish.

Response B:

The text appears to be a product labeled for a Korean cold noodle dish with ingredients and nutritional info listed.

Scenario 8

User Query: "How many screws do you see"

[Image displayed: Five screws on a table]

Response A:

I see several screws on the table.

Response B:

In the image, there are six screws visible.

Scenario 9

User Query: "Read any text you can see"

[Image displayed: Hand holding a Nature Valley chocolate chunk bar]

Response A:

The text on the object in the image reads "Nature Valley."

Response B:

Nature Valley chewy chocolate chunk.

Scenario 10

User Query: "How many calories are in this?"

[Image displayed: Hand holding Nature Valley chocolate chunk bar]

Response A:

The image shows a hand holding a Nature Valley granola bar. The text on the packaging indicates it is a "Nature Valley Granola Bar." To determine the calorie count, you would need to refer to the nutrition label on the packaging, which is not fully visible in this image.

Response B:

I guess there are around 200 to 250 calories in this nature valley chew chocolate chunk bar.

Scenario 11

User Query: "How many calories are in this?"

[Image displayed: Hand holding a can of LaCroix sparkling water]

Response A:

The image shows a can of LaCroix sparkling water. The label on the can indicates it is a "Calorie Free" drink. Therefore, it contains 0 calories.

Response B:

The can of La Croix sparkling water has zero calories.

Scenario 12

User Query: "What do you see"

[Image displayed: Wooden tray with various items, including two

nintendo switch controllers, a paintbrush, a pair of scissors, tape, and two coasters]

Response A:

I see a wooden tray with various items on it. I see nintendo switch controls, a paint brush, and scissors.

Response B:

The image shows a close-up view of a wooden surface with a rectangular tray containing several items. Inside the tray, there is a small, patterned object that appears to be a coaster or a small mat, a white object with a circular pattern that could be a remote control or a similar device, and a pair of scissors. The scissors are partially open, and the blade is visible.

Scenario 13

User Query: "How many calories are in this"

[Image displayed: Can of Spam on a couch]

Response A:

The number of calories in this image is not explicitly stated. However, it appears to be a can of spam which typically contains 180 to 200 calories per serving.

Response B:

The image shows a can of Spam. To determine the calorie content, you would need to refer to the nutrition label on the can, which is not visible in this image. Spam's calorie count can vary depending on the specific variety and serving size.

Scenario 14

User Query: "Read any text you can see"

[Image displayed: Hand holding a can of LaCroix]

Response A:

The text visible in the image reads "La Croix."

Response B:

Naturally based calorie free.

Scenario 15

User Query: "How many brushes do you see"

[Image displayed: Seven Paintbrushes on a coffee table]

Response A:

In the image, there are six paintbrushes visible.

Response B:

I see seven brushes.

Scenario 16

Note: Throughout this study, you have been rating the response quality between two different AI wearable devices. In these last two examples, we will show you the images from each device as the scene is slightly different between the two.

User Query: "What kind of plant is this?"

[Image displayed: Fiddle leaf fig]

Response A:

The plant in the image is a fiddle leaf fig.

Response B:

The image shows a small white dog sniffing a dark green leafy plant in a white pot. The background includes a metallic object, possibly a trash bin or a piece of furniture, and a light-colored floor. The scene is in black and white.

Scenario 17

User Query: "What do you see?"

[Image displayed: Cluttered countertop with a can of spam, dog treats, can of LaCroix, Lysol wipes, Lunchables, and a roll of paper towels]

Response A:

The image shows a cluttered countertop with various items. In the foreground, there are a can of beer, a bottle of Lysol disinfectant, a roll of paper towels, and a box of tea or coffee...

Response B:

I see a kitchen counter with various food items and cleaning supplies including a bag of bone broth bites, a can of spam, a package of lunchables, and a container of lysol wipes.

Part 3: Conclusion

Please briefly explain your rating methodology.

[Open-ended text field]

Please explain why you gave responses a high rating versus a low rating. (Optional)

[Open-ended text field]

B In-Person Follow-Up Survey

Thank you for completing our in-person user study! We'd like to follow up on your experience by asking a few questions.

1. How did wearing camera-enabled earbuds feel compared to regular earbuds?

[Open-ended text field]

2. What were your initial thoughts and/or concerns about cameras in earbuds?

[Open-ended text field]

3. How comfortable would you be wearing camera-integrated earbuds?

[Open-ended text field]

4. How does this compare to taking a photo with your phone and asking ChatGPT?

[Open-ended text field]

5. What situations would you actually use this in?

[Open-ended text field]