

Extending Human Capabilities with Deep Learning Powered Wearables

Maruchi Kim

A dissertation submitted in partial fulfillment
of the requirements for the degree of

Doctor of Philosophy

University of Washington
2025

Reading Committee:
Shyamnath Gollakota, Chair
Steve Seitz
Ira Kemelmacher-Shlizerman

Program Authorized to Offer Degree:
Computer Science and Engineering

© Copyright 2025
2025 Maruchi Kim

University of Washington

Abstract

Extending Human Capabilities with
Deep Learning Powered Wearables

Maruchi Kim

Chair of the Supervisory Committee:
Shyamnath Gollakota
Computer Science and Engineering

Deep learning-powered wearables have the potential to seamlessly extend human capabilities by enhancing perception and interaction in everyday environments. In this dissertation, I present three wearable systems that integrate deep learning neural networks with custom hardware to enable real-time audio enhancement, vision-based smart interactions, and visual intelligence through wireless earbuds.

First, I present ClearBuds, a wireless earbud system that performs real-time speech enhancement using a synchronized binaural microphone array and a lightweight dual-channel neural network. The system achieves high-precision synchronization and low-latency processing on mobile devices, enabling robust noise suppression and background speech removal in diverse real-world conditions.

Second, I introduce IRIS, a vision-enabled smart ring that fits within the size and power constraints of the ring form factor to enable context-aware smart home interactions. By combining scene semantics with detected objects, IRIS achieves instance-level device recognition and outperforms voice commands in speed, precision, and social acceptability.

Third, and as the final contribution, I present VueBuds, the first vision-enabled wireless earbuds integrating low-power cameras with vision language model interaction. VueBuds addresses fundamental challenges in embedding cameras into earbuds—strict power and form-factor constraints, facial occlusion from ear-level positioning, and real-time multimodal processing over Bluetooth. Through a stereo camera system operating at under 5 mW and end-to-end system optimizations, VueBuds achieves visual question-answering performance comparable to commercial smart glasses while leveraging a significantly more ubiquitous form factor.

Together, these systems demonstrate how deep learning powered wearables can extend human capabilities with on-the-go intelligence, establishing new platforms for intuitive, responsive, and enhanced human-computer interaction.

Chapter 1

Introduction

In an era of ubiquitous smart devices, the way we interact with technology is evolving rapidly. From seamless telephony in noisy environments to intuitive control of smart home devices to AI-powered visual assistance, there is a growing need for wearable systems that enhance human-device interaction while remaining unobtrusive and efficient. However, designing such systems presents significant challenges, including real-time operation, low-latency processing, strict power constraints, and the ability to generalize to diverse real-world scenarios. In this work, I present three wearable systems that address these challenges: ClearBuds, wireless binaural earbuds for real-time, learning-based speech enhancement; IRIS, a vision-enabled smart ring for intuitive, context-aware smart home control; and VueBuds, the first camera-integrated wireless earbuds for visual intelligence.

ClearBuds tackles the long-standing problem of speech enhancement in noisy environments by leveraging wireless earbuds as a binaural microphone array. Unlike conventional speech separation techniques that rely on single-channel processing or cloud-based models with privacy concerns, ClearBuds introduces a lightweight neural network designed to run efficiently on mobile devices. The system integrates custom-designed synchronized wireless earbuds with a novel cascaded neural network that performs real-time source separation, significantly improving speech intelligibility for telephony and voice interfaces. I demonstrate that ClearBuds outperforms existing commercial solutions, while operating within strict end-to-end latency constraints.

IRIS reimagines the way users interact with smart home devices by introducing a camera-integrated smart ring that supports real-time instance-based object detection. Traditional interaction methods, such as voice commands or smartphone applications, often suffer from usability issues in social settings or require cumbersome navigation. IRIS offers a natural and intuitive alternative by enabling users to simply point and click at a device to control it. By combining a lightweight, low-power wearable camera with an optimized instance-level classification pipeline, IRIS ensures seamless and efficient interaction with smart home devices. IRIS significantly reduces inference latency while maintaining high accuracy in distinguishing between multiple instances of the same device class, making it a practical solution for real-world deployment.

VueBuds, the core contribution of this dissertation, extends visual intelligence to wireless earbuds, a significantly more ubiquitous form factor than commercial smart glasses. Despite their ubiquity, wireless earbuds have remained limited to audio-centric functionality due to strict size, weight, and power constraints. VueBuds demonstrates the feasibility of integrating low-power cameras into earbuds while enabling real-time interaction with vision language models. This contribution addresses three fundamental research challenges: supporting camera hardware within the stringent power budget of wireless earbuds, mitigating facial occlusion inherent to ear-level camera positioning through binocular imaging, and achieving real-time multimodal

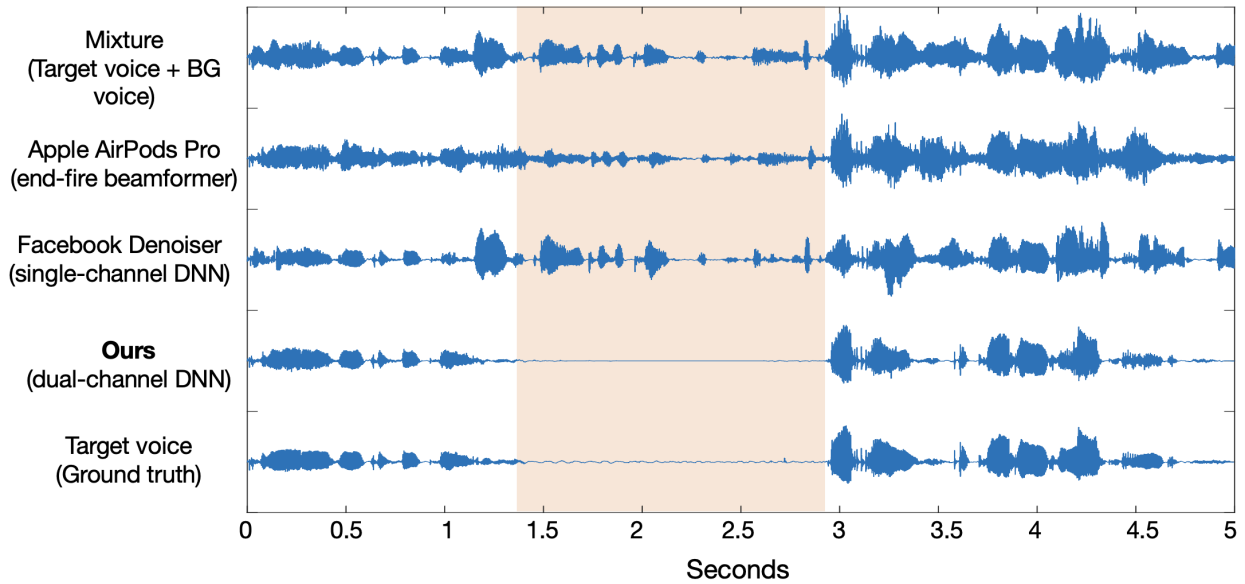


Figure 1.1: ClearBuds Bystander Speech Removal. We use spatial cues to separate background voices from the target speaker, even when the background voice is louder than the target voice. This is evident when the target speaker is silent but background voice continues to talk (highlighted in orange). Apple AirPods Pro uses an endfire beamformer to partially suppress background voice. The mono-channel Facebook Denoiser (Demucs) is unable to suppress the background voice. Clearbud’s network removes the background voice, approaching ground truth.

inference over bandwidth-constrained Bluetooth protocols. Through custom hardware operating at under 5 mW and end-to-end system optimizations, VueBuds achieves visual question-answering performance comparable to commercial smart glasses like Ray-Ban Meta while leveraging the social acceptability and widespread adoption of everyday earbuds.

Together, ClearBuds, IRIS, and VueBuds share a common goal: enabling real-time, context-aware interaction through wearable systems that are computationally efficient and user-friendly. By addressing the challenges of real-time inference, low-power hardware constraints, and generalization to diverse environments, this work pushes the boundaries of wearable technology across audio, vision, and multimodal domains. Through extensive evaluations, including user studies and real-world deployments, I demonstrate that these systems not only improve user experience but also establish new platforms for future research in wearable computing and human computer interaction.



Figure 1.2: Smart home interaction with IRIS. (A) A user unlocks the front door by pointing and clicking IRIS at the smart lock. (B) Another user points IRIS at a television and rotates their hand to adjust its volume. (C) The user points IRIS at their living room lights to turn them off before leaving home. (D) A user points and clicks IRIS at the blinds to lower them for privacy.

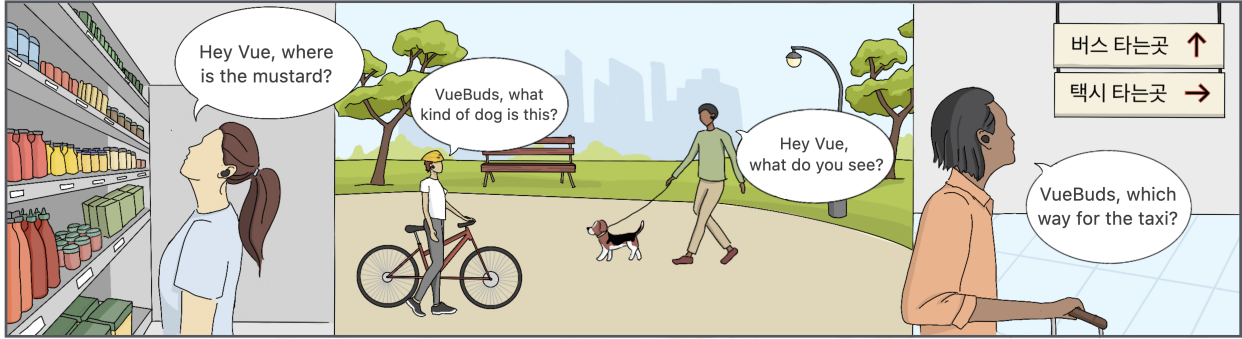


Figure 1.3: Applications of VueBuds, wireless vision-enabled earbuds with visual intelligence. Users can query their surroundings in natural language to accomplish everyday tasks, such as locating items in a store, identifying objects, obtaining scene-level descriptions, and interpreting foreign text for navigation.

Chapter 2

ClearBuds: Wireless Binaural Earbuds for Learning-Based Speech Enhancement

With the rapid proliferation of wireless earbuds (100 million AirPods sold in 2020 air [b]), more people than ever are taking calls on-the-go. While these systems offer unprecedented convenience, their mobility raises an important technical challenge: environmental noise (e.g., street sounds, people talking) can interfere and make it harder to understand the speaker. We therefore seek to enhance the speaker’s voice and suppress background sounds using speech captured across the two earbuds.

Source separation of acoustic signals is a long-standing problem where the conventional approach for decades has been to perform beamforming using multiple microphones. Signal processing-based beamformers that are computationally lightweight can encode the spatial information but do not effectively capture acoustic cues Van Veen and Buckley [1988]; Krim and Viberg [1996]; Chhetri et al. [2018]. Recent work has shown that deep neural networks can encode both spatial and acoustic information and hence can achieve superior source separation with gains of up to 9 dB over signal processing baselines Subakan et al. [2021]; Luo and Mesgarani [2019]. However, these neural networks are computationally expensive. None of the existing binaural (i.e., using two microphones) neural networks can meet the end-to-end latency required for telephony applications or have been evaluated with real earbud data. Commercial end-to-end systems, like Krisp kri, use neural networks on a cloud server for single-channel speech enhancement, with implications to cost and privacy.

We present the first mobile system that uses neural networks to achieve real-time speech enhancement from binaural wireless earbuds. Our key insight is to treat wireless earbuds as a binaural microphone array, and exploit the specific geometry – two well-separated microphones behind a proximal source – to devise a specialized neural network for high quality speaker separation. In contrast to using multiple microphones on the same earbud to perform beamforming, as is common in Apple AirPods air [a] and other hearing aids, we use microphones across the left and right earbuds, increasing the distance between the two microphones and thus the spatial resolution.

To achieve this system, we make three technical contributions spanning earable hardware and neural networks.

- **Synchronized binaural earables.** We designed a binaural wireless earbud system (Fig. 2.1) capable of streaming two time-synchronized microphone audio streams to a mobile device. This is one of the first systems of its kind, and we expect our open-source earbud hardware and firmware to be of wider interest as a research and development platform. Existing earable platforms such as eSense Kawsar et al. [2018] do not support time-synchronized audio transmission from two earbuds to a mobile device.



Figure 2.1: ClearBuds hardware inside 3D-printed enclosure and when placed beside a quarter.

- **Lightweight cascaded neural network.** We introduce a lightweight neural network that utilizes binaural input from wearable earbuds to isolate the target speaker. To achieve real-time operation, we start with the Conv-TasNet source separation network Luo and Mesgarani [2019] and redesign the network to achieve a 90% re-use of the computed network activations from the previous time step for each new audio segment. While these optimizations make this network real-time, they also introduce artifacts in the audio output (i.e., crackling, static). These artifacts are often visible in a frequency representation of the audio. To address this, we combine our mobile temporal model with a real-time spectrogram-based frequency masking neural network. We show that by combining the two networks and creating a lightweight cascaded network, we can reduce artifacts and improve the audio quality further.
- **Network training for in-the-wild generalization.** Training the network in a supervised way requires clean ground truth speech samples as training targets. This is difficult to obtain in fully natural settings since the ground truth speech is corrupted with background noise and voices. Training a network that generalizes to in-the-wild scenarios also requires the training data to mimic the dynamics of real speech as closely as possible. This includes reverb, voice resonance, and microphone response. Our training methodology uses large amounts of synthetic data simulated in software, small amounts of hardware data with speakers embedded into a foam mannequin head and small amounts of data from human speakers wearing the earbuds in an anechoic chamber to create a neural network that generalizes to users and multi-path environments not in the training data.

We combine our wireless earbuds and neural network to create ClearBuds, an end-to-end system capable of (1) source separation for the intended speaker in noisy environments, (2) attenuation and/or elimination of both background noises and external human voices, and (3) real-time, on-device processing on a commodity mobile phone paired to the two earbuds. We believe that this paper bridges state-of-the-art deep learning for blind audio source separation and in-ear mobile systems.

2.1 ClearBuds Design

We first introduce our lightweight neural network architecture. We then describe system design of our hardware platform and our synchronization algorithm.

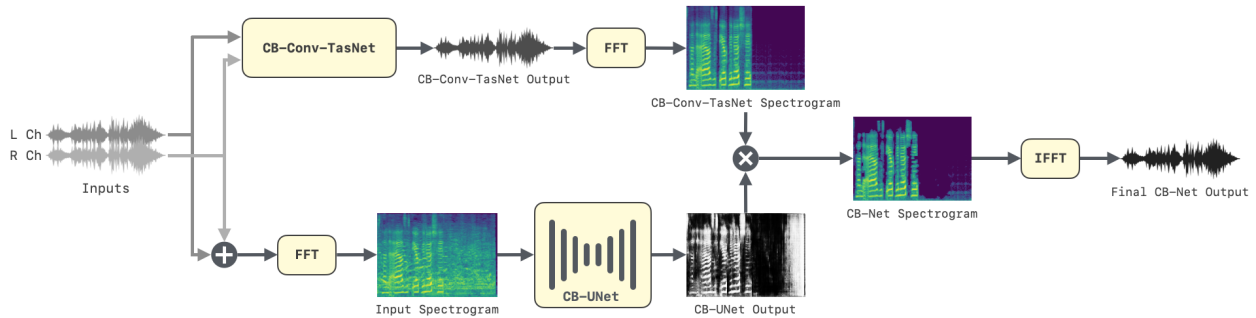


Figure 2.2: ClearBuds Network Architecture.

2.1.1 Neural Network Architecture

Our network needs to perform in real-time on a mobile device with minimal latency. This is challenging for several reasons. First, the processing device has a much lower compute capacity, especially compared to cloud GPUs. Additionally, the network should separate non-speech noises as well as unwanted speech. To do this, it must learn spatial cues and human voice characteristics. Finally, the resulting output should maximize the quality from a human experience perspective while minimizing any artifacts the network might introduce.

Our network, which we call *ClearBuds-Net* or *CB-Net*, is a cascaded model that operates in both time and frequency domains. The full network architecture is illustrated in Fig. 2.2 and contains two main sub-components: A dual-channel time domain network called *CB-Conv-TasNet*, and a frequency based network called *CB-UNet*.

CB-Conv-TasNet

The first component of separation method is a time domain network that is based on a multi-channel extension of Conv-TasNet Luo and Mesgarani [2019]. This is a network in the waveform domain that has a Temporal Convolution Network (TCN) structure, lending itself to a causal implementation with intermediate layer caching Paine et al. [2016]. We use depthwise separable convolutions Howard et al. [2017] to further reduce the number of parameters and make the design real-time. We call this network *CB-Conv-TasNet* since it is an optimized version of the original Conv-TasNet.

A key feature of the time domain approach is that it can easily capture spatial cues in the network. In our application, the desired source is always physically between two microphones, thus the voice signal will reach the microphones roughly at the same time. In contrast, background or other speakers are typically not temporally aligned and will reach one microphone earlier or later. By feeding two time synchronized channels into the neural network, this spatial alignment of the sources can be learned from time differences in the signal.

CB-UNet

The output of our lightweight *CB-Conv-TasNet* often contains audible artifacts (i.e., crackling, static) that reduce the listening experience. Interestingly, these artifacts have little effect on traditional metrics, like Signal-to-Distortion Ratio (SDR), but have a noticeable effect on subjective listening scores. These artifacts are often visible in a frequency representation of the audio. Fig. 2.3 shows how *CB-Conv-TasNet* alone contains noticeable artifacts when compared to the ground truth. To address this, we cascade a lightweight causal UNet Ronneberger et al. [2015] which operates on the mel-scale spectrogram of the input audio. This network, which we call *CB-UNet*, produces a binary mask which is applied to the output of *CB-Conv-TasNet*.

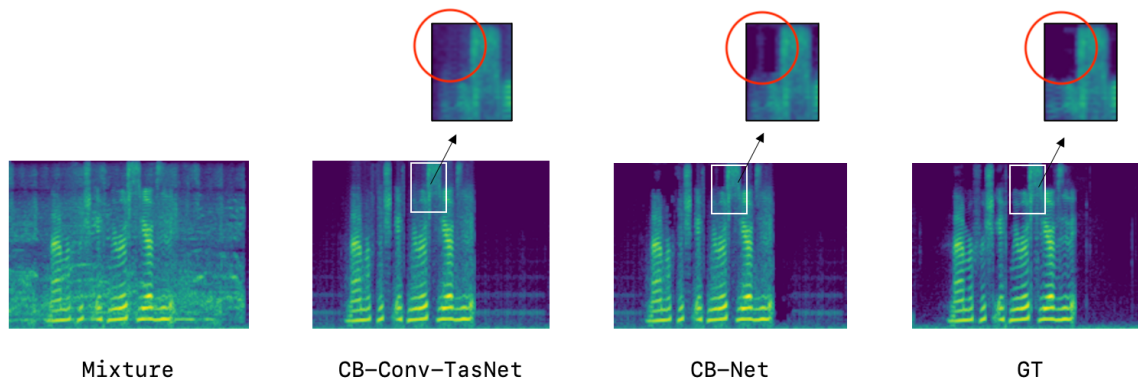


Figure 2.3: The spectrograms above show the motivation behind a combined time and frequency domain method. The output of the time-domain component, CB-Conv-TasNet, contains artifacts, particularly at high frequencies.

The combined output, shown in Fig. 2.3 as CB-Net, reduces these artifacts. The mean opinion scores in our evaluation shows the strength of the cascaded CB-Net when compared to the time-domain component only.

2.1.2 Synchronized wireless earbuds

We seek to capture speech from the target speaker’s mouth which sits on the sagittal plane roughly equidistant to the ears. Given an ear-to-ear spacing of 17.5cm, to effectively isolate this central plane we require a distance precision on the order of a few centimeters. An interaural time difference of $100\mu s$ would correspond to source maximally 3.43 cm off this central plane, therefore we target a synchronization accuracy under $100\mu s$.

Hardware

Our custom hardware design contains a pulse-density modulated (PDM) microphone (Invensense ICS-41350) and a Bluetooth Low Energy (BLE) microcontroller (Nordic nRF52840). The system is powered off of a CR2032 coin cell battery and programmed via SWD over a Micro-USB connector. Each ClearBud has an integrated PDM microphone set to a clock frequency of 2MHz. With an internal PDM decimation ratio of 64, this provides us a sampling frequency of 31.25kHz. As most HD voice applications and wideband codecs are limited to 16kHz C. and M.H. [2009], we decimate further in firmware by a factor of 2, giving us a final sampling frequency of 15.625kHz.

Two 16-bit 180 sample size Pulse-Code Modulation (PCM) buffers are round-robin: one is filled with incoming PCM data while the other is processed. The DMA is responsible for both clocking in the PDM data and converting it into PCM. One buffer is always connected to the DMA, while the other is freed for processing for the rest of the data pipeline. When the buffer connected to the DMA fills, the buffers switch roles and we begin processing data on the newly freed buffer, and connect the other buffer back to the DMA. With this design we always have a continuous PCM stream to operate on. Both ClearBuds transmit the PCM microphone data to a mobile phone for input into our neural network. To maximize throughput, we use the highest Bluetooth rate and packet sizes supported by iOS, which is 2Mbps and 182 bytes, respectively. We design a lightweight wireless protocol where the first 2 bytes represent a monotonically-increasing sequence number, while the other 180 bytes are reserved for the 16-bit PCM audio samples.

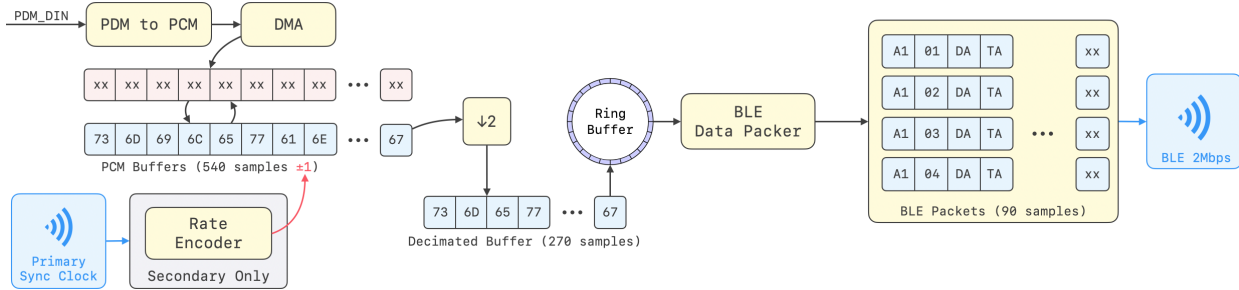


Figure 2.4: Time Sync Design

Microphone synchronization

Three components are necessary for maintaining microphone synchronization: (1) As each of our earbuds has its own local clock source, we need to establish a common clock between them so that they have the same reference of time, (2) a synchronized startup so each earbud starts recording from their respective microphone at the exact same time, and (3) a rate encoding scheme to control the earbud’s sampling rate to match each other.

In our system, each earbud has its own respective 32MHz clock source with a total +/- 20ppm frequency tolerance budget. So, in the worst case scenario, the earbuds will have 2.4 milliseconds of drift each minute. We use the Nordic’s TimeSlot API [tim \[2015\]](#), which grants us access to the underlying radio hardware in between Bluetooth transmissions. This provides us a transport to transmit and receive accurate time sync beacons [wir \[2016\]](#). One ClearBud is assigned as the timing master while the other ClearBud will synchronize its free-running timer to the master’s. The primary ClearBud (timing master) transmits time sync packets at a rate of 200 Hz. These packets contain the value of the free-running timer at the time of the radio packet transmission. When the secondary ClearBud receives this packet, it can then add or subtract an offset to its own free-running timer for a common clock.

Once each ClearBud is connected to the mobile phone, the phone sends a `START` command to both ClearBuds over BLE. Each ClearBud contains firmware which arms a programmable peripheral interconnect (PPI) to launch the PDM bus once the 16MHz free-running timer wraps around at 800,000. By using this method, we bypass the CPU and trigger a synchronized startup entirely at the hardware layer.

The final component to keeping the audio streams aligned is to create a rate encoding scheme between the ClearBuds. With the time sync beacons from the primary ClearBud, the other ClearBud now has both its local clock and the common clock (primary ClearBud’s local clock). With these two clocks, the secondary ClearBud can identify how much faster or slower its PDM clock is running in relation to the primary ClearBud. We note that with a 2MHz PDM clock and a PDM decimation ratio of 64, each audio sample occupies 32 us. The non-primary ClearBud can then add or remove a sample to its PDM buffer every time the difference between the clocks exceeds a multiple of 32 us. By doing this, the secondary ClearBud ensures that its PDM buffer starts filling up at the exact same time as the primary ClearBud’s PDM buffer, with a tolerance of 32 us.

2.1.3 Training methodology

Training the network in a supervised way requires clean ground truth speech samples as training targets. This is difficult to obtain in fully natural settings since the ground truth speech is corrupted with background noise and voices. Training a network that generalizes to in-the-wild scenarios also requires the training data to

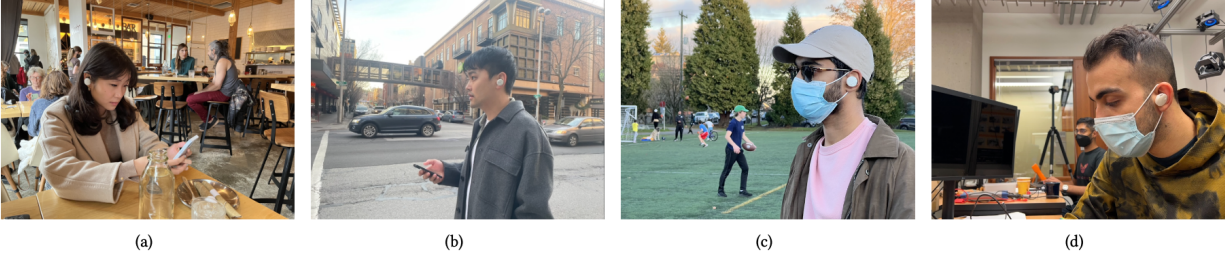


Figure 2.5: In-the-wild experiments in various scenarios (crowded cafe, busy intersection, outdoor plaza, classroom) were conducted across 8 users and indoor and outdoor environments, all unseen in our training dataset.

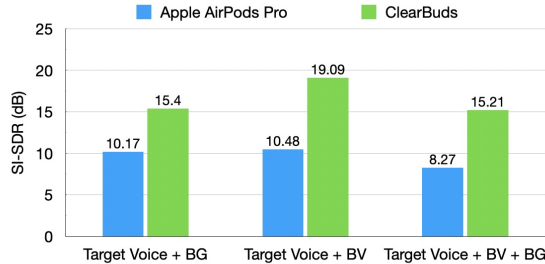


Figure 2.6: Comparison with AirPods Pro. ClearBuds exceeds in three conditions: target voice plus background noise (BG), target voice plus background voice (BV), target voice plus background voice and noise.

mimic the dynamics of real speech as closely as possible. We adopt a hybrid training methodology where we first train on a large amount of synthetic data and fine-tune on real data recorded with our hardware. Our training method is based on the commonly used mix-and-separate framework Zhao et al. [2018], where clean speech and noise samples are recorded separately and combined randomly to form noisy mixtures. Our results show that our network trained this way generalizes to naturally recorded noisy data in real-world environments.

2.2 ClearBuds Results

We first compare our end-to-end system performance against a commercial wireless earbud system. We then present in-the-wild evaluation of our system. Next, we compare numerical results against various speech enhancement baselines. Finally, we present system-level evaluations.

2.2.1 Comparison with Beamforming Earbuds

We evaluate our end-to-end system against the Apple AirPods Pro headset connected to a iPhone 12 Pro in a repeatable physical set up. In our evaluation, as is typical, there is no overlap between training and test datasets.

Procedure. We use the popular metric *scale-invariant signal-to-distortion ratio* (SI-SDR) Roux et al. [2018]. While SI-SDR provides a repeatable metric used in the acoustic community, it requires a clean, sample-aligned ground truth (target voice) as the basis for evaluation. Therefore, we create a repeatable soundscape for our test setup where a sample-aligned ground truth can be obtained. A foam mannequin head with a speaker (Sony SBS-XB12) inserted into its artificial mouth uttered one hundred VCTK samples with identities and samples unseen in the training set. The mannequin wore ClearBuds and AirPods Pro

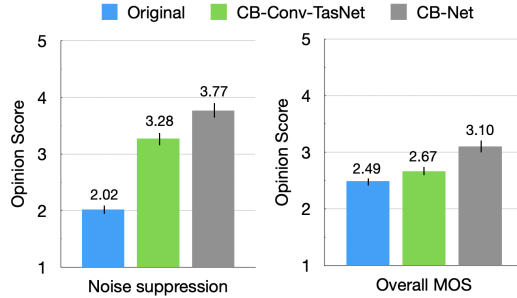


Figure 2.7: In-the-wild study results. Noise suppression indicates perceived quality of background noise reduction (higher is less intrusive). Overall MOS indicates overall perceived quality. Error bars are 95% CI.

in subsequent experiments, and the outputs of the two systems could be directly compared. Ambient environmental sound (from WHAM! dataset) was played via four monitors (PreSonus Eris E3.5) positioned to fill 3 meter by 4 meter room, and background voice (also VCTK) was played from a monitor positioned 0.4 meters from head on the right. All speakers were driven through a common USB interface (PreSonus 1810c) ensuring the same time-alignment and loudness between the two test conditions.

Results. We report output SI-SDR from the two systems in Fig. 2.6. To calculate output SI-SDR, we align individual one second chunks and take the logarithmic mean across 250 chunks. We find that ClearBuds achieves higher output SI-SDR across all test conditions when compared to the beamforming utilized by the Apple AirPods Pro.

2.2.2 In-the-Wild Evaluation

We perform in-the-wild evaluation in indoor and outdoor scenarios as well as users not in the training data. The procedure and results are described in the following sections.

In-the-wild experiments. Eight individuals (four male, four female, mean age 25) with a variety of accents wore a pair of ClearBuds and read excerpts from Project Gutenberg gut while in four noisy environments: a coffee shop, a noisy intersection, an outdoor plaza, and a classroom (see Fig. 2.5). The environments featured ringing phones, cross-talk from other people, ambient music, a crying baby, opening/closing doors, driving vehicles, and street noise, amongst other common sounds. These experiments were uncontrolled in that the background voices and noise were naturally occurring sounds that are typical to these real-world scenarios and were mobile.

Evaluation procedure. In-the-wild evaluation precludes access to clean, sample-aligned truth to compute SI-SDR. Instead, the common (and expensive) procedure is to perform a user study and compute the mean opinion score. Since this is a time-consuming process, prior works on binaural networks, e.g., Luo et al. [2020]; Tan et al. [2019]; Jenrungrot et al. [2020], avoid in-the-wild evaluation. Since our goal is to design and evaluate an in-ear system in real scenarios, we recruit thirty-seven participants (11 female, 26 male, mean age 29) for a user study. Each participant listened to between 6 and 11 in-the-wild audio samples (avg. 9.38 samples, each between 10–60 seconds). Each speech sample was processed and presented three ways: (1) the original input, (2) CB-Conv-TasNet, and (3) CB-Net, yielding a total of $37 \times 9.38 \times 3 = 1,041$ rating samples.

Results. Fig. 2.7 shows the noise intrusiveness and MOS values for the original microphone, CB-Conv-TasNet, and CB-Net. As expected, applying CB-Conv-TasNet to the original audio helped suppress noise dramatically, increasing opinion score from 2.02 (slight better than 2 - *Somewhat intrusive*) to 3.28 (between 3 - *Noticeable, but not intrusive* and 4 - *Slightly noticeable*) ($p < 0.01$). The light-touch, spectrogram-masking clean up method featured in CB-Net increased noise suppression opinion score significantly ($p < 0.001$) to 3.77, indicating the method did indeed further suppress perceptually annoying noise artifacts. Importantly,

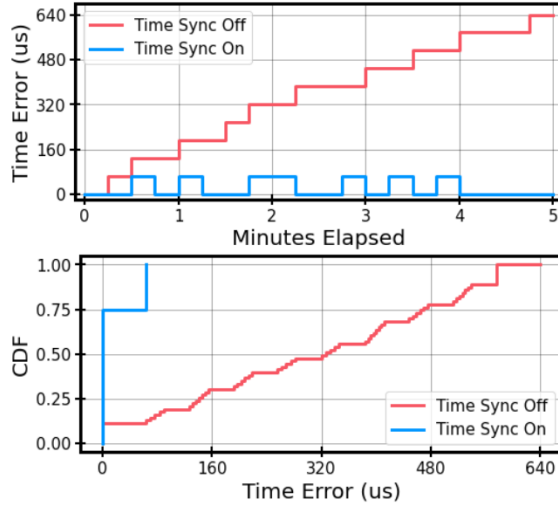


Figure 2.8: Time Synchronization Validation. Without time synchronization (red), microphone samples drift apart and lose alignment at about $128\mu\text{s}/\text{min}$.

this step also increased overall MOS. While users only slightly preferred ($p < 0.05$) CB-Conv-TasNet (2.67) to the original input (2.49) due to artifacts introduced, they more significantly ($p < 0.001$) preferred our CB-Net (3.10), an increase of 0.61 opinion score points from the input. For context, in the flagship ICASSP 2021 Deep Suppression Noise Challenge Reddy et al. [2021], with state-of-the-art, real-time algorithms run on a quad-core desktop CPU, the winning submission increased MOS by 0.57 ica from input.

2.2.3 Synchronization Evaluation

In order to evaluate this, we place both ClearBuds roughly equidistant from a speaker. A click tone is played every 15 seconds for 5 minutes, and recorded on both ClearBuds with time sync disabled and enabled. We calculate the sample error on each recorded click offline and convert it into time error with a sampling rate of 15.625kHz. Fig. 2.8(a) shows the synchronization results across a five minute interval. With time sync enabled, the sample error never exceeds 1 sample at 15,625 kHz, or $64\mu\text{s}$. Fig. 2.8(b) also shows the CDF of the timing error across experiments of 5 minutes each conducted with other Bluetooth devices in the environment, with and without time synchronization.

Chapter 3

IRIS: Wireless Ring for Vision-Based Smart Home Interaction

As households transition into smart homes, they are outfitted with an array of interconnected devices, like smart speakers, door locks, and other smart home appliances Dive [2024]; Oberlo [2024]. While these devices promise improved convenience, the means by which users control them remain ripe for improvement. Issues such as social discomfort in using voice commands and the unreliability of voice input in noisy environments hinder seamless interaction Jain et al. [2023]; Rodrigues et al. [2019]. Despite the availability of smartphone apps for direct control, it is often more convenient to resort to traditional methods, such as light switches or dedicated remotes, rather than unlocking one’s phones, locating the appropriate application, and navigating through its interface Newman [2022]; Sapienza [2022].

We introduce IRIS, short for Interactive Ring for Interfacing with Smart home devices. IRIS is an end-to-end wireless ring system that supports real-time object instance detection using contextual scene semantics. The underlying principle is rooted in the age-old adage that a “picture is worth a thousand words,” asserting that capturing images is far more efficient than verbalizing lengthy auditory commands. IRIS enables users to control smart home devices by simply pointing at the target device and performing a corresponding gesture, offering an intuitive alternative to traditional interaction methods.

Achieving this is challenging for three key reasons. First, while sensors integrated in today’s ring devices, such as IMUs, are low-power, camera hardware can generate significantly more data, leading to orders of magnitude higher power consumption Veluri et al. [2023]. Secondly, while object detection systems excel in real-time detection, their limitations become evident when confronted with multiple instances of the same object class. For instance, merely determining that a user pointed at a set of blinds is inadequate; the system must also discern which specific set of blinds in the home the user intends to control. Therefore, a system capable of precisely identifying and distinguishing between individual devices of the same class is necessary. Finally, the entire end-to-end system should operate in real-time under around one second to be considered a seamless interaction modality Nielsen [1993]; Endo et al. [1996].

In IRIS, we address these challenges by making technical contributions spanning hardware, software, and system design. At a high level, the wearer points at the device of interest and presses the button on the ring, as shown in Fig. 1.2. This wakes up the on-board camera which captures a few frames of the target device, which are streamed to a nearby smartphone for processing via Bluetooth. The wearer also performs a gesture (e.g. rotation) to control the device, the intent of which is captured using the on-board IMU. This is also transmitted to the phone which then, in real-time, controls the target device. Specifically, we make three key contributions.

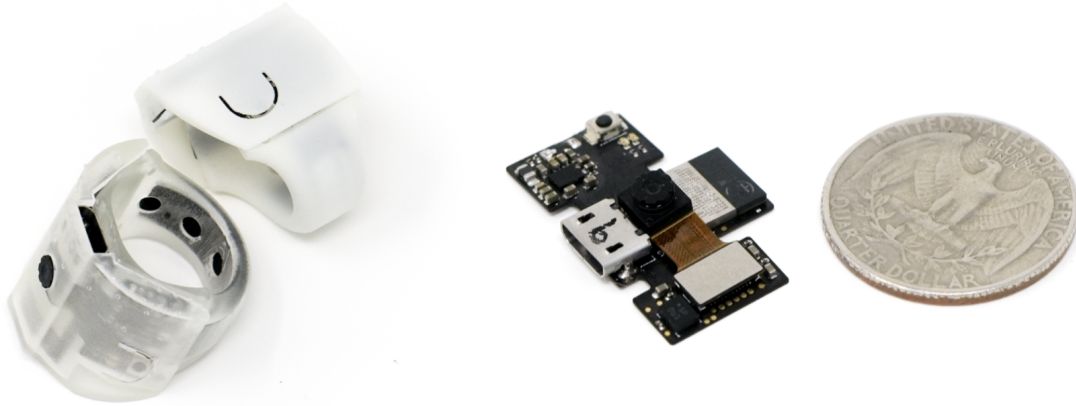


Figure 3.1: IRIS hardware inside 3D-printed enclosure and when placed beside a quarter. The battery sits inside the band of the ring. The ring diameter and band thickness are 17.5 and 2.9 mm.

- **SWAP-constrained wireless camera ring.** We designed the first wireless ring form-factor device for vision-based smart home interaction (Fig. 3.1). Our hardware is equipped with a camera, Bluetooth radio, IMU, and an on-board battery, while meeting the small size, weight and power (SWaP) requirements expected in ring form-factor devices. We present cross-layer optimization methods across wireless camera hardware and firmware, and a user interface that enables extremely low-power states. The PCB, battery, and enclosure weigh 4 grams. IRIS is the first ring to stream camera data wirelessly, and operate for over 16 hours on a single charge.
- **Instance-level classification based on scene semantics.** To achieve instance-level classification, we start by utilizing a self-supervised vision transformer model, DINOv2 Oquab et al. [2024], to generate scene-level embeddings. These embeddings capture not only the object itself but also the surrounding environment, providing valuable context and scene semantics. However, DINOv2’s search runtime increases linearly with the size of the embedding database, introducing potential latency issues for an interactive system like IRIS. To address this challenge, we reduce the search space of DINOv2 by utilizing YOLO Redmon et al. [2016] to first detect the smart home device closest to the image center. Our results demonstrate that this optimization effectively reduces DINOv2’s query runtime by hundreds of milliseconds.
- **End-to-end system optimization for real-time operation and low-power.** Real-time operation is critical for interactive mobile systems Nielsen [1993]; Endo et al. [1996]. We optimize our wireless ring’s streaming performance to maximize camera throughput, and show that the end-to-end system can control devices within one second, delivering near real-time feedback to users by confirming successful gesture recognition. Furthermore, we extensively optimized IRIS’s low-power design for all day use, despite the limited on-board battery capacity.

Put together, we design an end-to-end wearable ring system capable of (1) smart home device interaction that complements voice commands, (2) instance-based object detection to correctly distinguish between individual instances of the same class, and (3) real-time operation, on-device processing on a standard mobile phone. We evaluate our system in five different homes from various angles and lighting conditions. Our real-world dataset also includes 98 examples of blinds, 57 doors, 34 door handles, 228 lights, 24 smart locks, 73 speakers, 64 televisions, 37 windows, and 161 background instances.

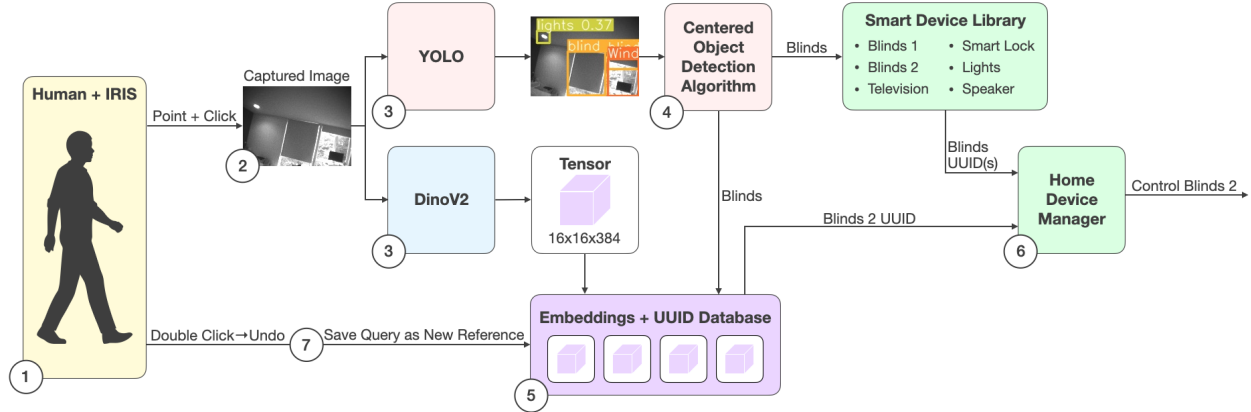


Figure 3.2: IRIS Pipeline

3.1 IRIS Design

We present our wireless ring hardware system and next describe our real-time neural network pipeline.

3.1.1 Wireless vision-enabled ring hardware

Our custom hardware design is comprised of an ultra-low-power 1/11" 320x320 QVGA CMOS image sensor (Himax HM01B0), a 6-axis inertial measurement unit (Bosch BMI270), and a Bluetooth Low Energy (BLE) microcontroller (Nordic nRF52840). The system is powered by a 27mAh battery and programmed via SWD over a Micro-USB connector. Our design is fully rechargeable through an on-board power management integrated circuit and provides the system with all necessary voltage rails (Maxim MAX77650). A single-pull single-throw (SPST) switch is used for gesture initiation, and images are streamed to a mobile phone for input into our neural network.

Wireless Latency

The first challenge with our design is meeting a <500ms image acquisition latency target. IRIS utilizes the maximum supported data rate of 2 Mbps over BLE Semiconductor [2022]. To maximize throughput, we utilize the shortest connection interval available to iOS (15 ms) while transmitting 4 packets per interval (maximum supported by iOS) Through [2022]. With a packet size of 247 bytes the effective data rate is 526,933 bits per second. The effective BLE throughput can be written as:

$$\text{Throughput} = \frac{1000 \text{ ms} \times \text{packets_per_interval} \times \text{packet_size} \times 8}{\text{connection_interval (ms)}}$$

The challenge is that a full 320x320 image is 819,200 bits, translating to 1562.5ms of latency (0.64fps) for the full image to be transmitted to the phone. So, we need to reduce the image size while preserving as much information as possible. To do this, first, we enable the QVGA window readout on the image sensor, which reduces the resolution to 320x240. This is still insufficient, so we utilize pixel binning to improve image acquisition time. Pixel binning is the concept of combining electrical charges from multiple adjacent pixels into a single "superpixel" Yoo et al. [2015]; Li et al. [2012]. Pixel binning essentially increases the effective pixel size, providing an improvement in SNR (signal-to-noise ratio) and low light. Furthermore, pixel binning

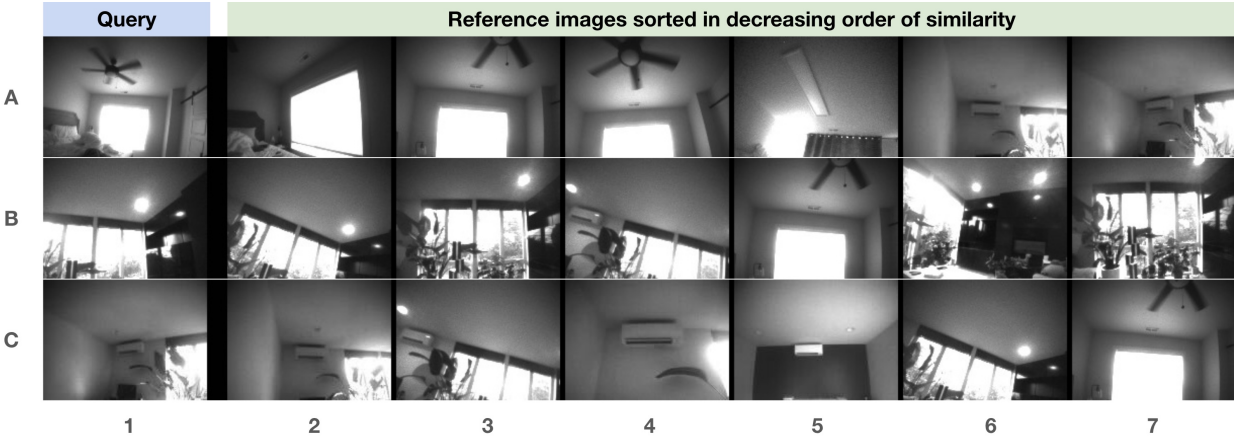


Figure 3.3: DINOv2 queries based on semantic similarity.

also enables faster frame rates. By combining pixel data, the sensor can read out information from a smaller number of "superpixels" compared to the original number of individual pixels. This reduced readout time effectively leads to faster frame rates. Thus, while binning reduces the resolution of our system by a factor of four to QQVGA (160x120), it offers two key advantages: (1) a 4x improvement in signal-to-noise ratio, and (2) a 4x increase in frame rate. In this configuration, our image resolution is now 160x120 bringing the frame rate to 3.43fps, or about 290ms of end-to-end latency. This meets our design goal of less than 500ms of latency with some margin to spare.

Low-power design

IRIS manages power consumption through its three distinct power states: SLEEP, IDLE, and ACTIVE. In the SLEEP state, IRIS conserves energy by deactivating all hardware components except for an internal timer within the nRF52840 chip. This timer periodically awakens IRIS to check for the presence of a home WiFi network by receiving data from the connected mobile phone.

In the IDLE state, IRIS enables all power rails and readies all peripherals, but clock-gates the camera and suspends the IMU. This ensures responsiveness while minimizing energy consumption. IRIS exits the IDLE state with a single button press via a hardware interrupt, transitioning it into ACTIVE mode. During this mode of operation, IRIS continuously streams data such as button state, IMU data, and camera pixels to the connected mobile device. After 3 seconds of streaming and inactivity from the button, IRIS exits ACTIVE and returns to IDLE to optimize low-power performance.

3.1.2 Neural Network Pipeline

Our pipeline consists of a fused model comprised of YOLOv8 Ultralytics [2024] and DinoV2 Oquab et al. [2024]. A YOLOv8 model was fine-tuned on a custom dataset of 10 classes to control a set of five common smart home devices. We augment YOLO with Dinov2 to distinguish between two instances of the same class. IRIS provides a scanning feature, which enables users to take pictures when they have multiple instances of the smart device in their home. We create an embedding database with these collected pictures, and users can map reference images to a specific instance of a smart home device.

Human-in-the-loop learning with semantic similarity

YOLO’s bounding boxes are combined with the centered object detection algorithm to disambiguate input frames with multiple objects, and provide us with a single object classification. This classification is sufficient in scenarios where there is one object (device) per class that the user wishes to control. However, if a user seeks to control multiple devices belonging to the same class (e.g., two TVs, one in the living room and the other in a bedroom), we would need to classify specific instances of the objects; beyond just object-level classification. Taking this into account, in addition to our out-of-the-box solution based on a pre-defined set of classes, we implemented a human-in-the-loop model that allows users to define a separate class for each device.

The intuition behind such a capability is that a device in a home is characterized by not just its own visual features but that of the surroundings it is placed in as well. For example, two smart speakers of the same make could be distinguished by the visual characteristics of the room it’s placed in. This necessitates us to extract not just the features of the object of interest in the frame, but the features of the entire frame. We leverage advances in large-scale self-supervised learning of visual features such as CLIP Radford et al. [2021] and DINOv2 Oquab et al. [2024] to compute semantic features at the image level. Specifically, we use the open-source DINOv2 Oquab et al. [2024] model in our implementation, to obtain image-level semantic features.

This approach is illustrated in Fig. 3.3. On the left most column, we show query images that the ring captured during its regular use. The rest of the columns show different reference images captured during the setup. In each row, the reference images are sorted in decreasing order of similarity with the query image on the left.

3.1.3 Training Methodology

We fine-tune YOLOv8 by collecting pictures using a scraper on Google Images. While these web-scraped images were easy to collect, they fail to model our camera’s characteristics like field-of-view, dynamic range, and low-light sensor noise. Thus, we also collected and created a dataset directly from IRIS’s camera. We adopt a hybrid training methodology where we first train on our web-scraped dataset, and then fine-tune on real data from our hardware. Our results show that our network trained in this way generalizes to real-world images captured from IRIS’s camera.

3.2 IRIS Results

3.2.1 User Study Evaluation

We recruited 23 participants (16 male, 7 female) aged 18-35 ($\sigma=4.72$) for a user study. Our study involved participants interacting in real-time with smart devices with IRIS and their voice. The participants were also provided a short questionnaire asking them about their overall experience and to compare IRIS against voice interaction.

Procedure

The goal of one part of our study was to compare IRIS against a commercial voice assistant. As our neural network implementation and home device management were implemented on iOS, we opted to compare against Siri. Benchmarking against Siri guarantees that the communication under the hood (i.e., HomeKit) remains the same between the two interfaces.

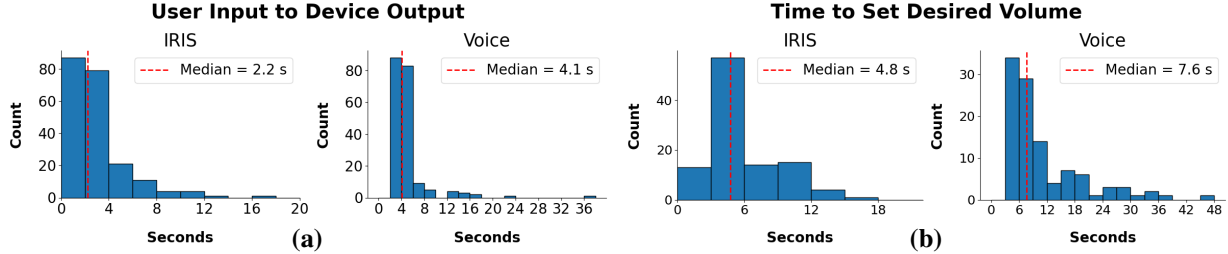


Figure 3.4: Histogram of User Input to Device Output.

This next phase of our study begins with introducing IRIS to the participants. To familiarize the participants with IRIS, we first play a short 30-second demonstration video. Next, we permitted them to try out IRIS by controlling a pair of smart devices, lights and speaker. After confirming their comfort and confidence using IRIS, participants engaged in a series of trials.

The first trial involved a proctor who directed the participants to control either the lights or the speaker, toggling the device’s state in response to cues. 10 of these trials were conducted for both IRIS and voice. The second trial evaluated the granular control performance between IRIS and voice. Participants were told to set the speaker volume to that of a loud gathering or a pleasant ambient level based on if the proctor said "loud" or "quiet". In total, we collected 690 IRIS and voice trials.

Quantitative Results

We conducted a quantitative analysis of the user input to device output (UIDO) time between IRIS and Siri for toggling the state (on/off) of a smart-light and smart-speaker as well as setting the volume for the smart-speaker.

The histograms in Fig. 3.4 show the UIDO time for state toggling and volume control across all participant trials for IRIS and Siri control. Fig. 3.4(a) shows that most participants were able to control the device state within 0 to 2 seconds using IRIS compared to 2 to 4 seconds when using Siri. Fig. 3.4(b) shows that for granular volume control a majority of participants were able to set their desired volume within 3 to 6 seconds when using both IRIS and Siri. However, when using Siri there was a wider distribution of times, with a wider positive skew.

For both control scenarios, the median UIDO and time to set desired volume was lower for IRIS (2.2 and 4.7 seconds respectively) than Siri (4.1 and 7.6 seconds respectively). This can be attributed to the fact that it took longer for participants to utter a full voice command than perform the button press and rotation gestures.

3.2.2 Vision Pipeline Evaluation

We assess the performance of YOLO+CODA and DINOv2. These models were evaluated independently as their joint performance primarily reduces runtime by reducing the search space in the DINOv2 (Embedding) database (see Fig. 3.2). Classification accuracies of our models are independent of one another. We later evaluate the latency of jointly running the models.

For evaluating our instance-based detection method, we collect a user-defined test set of 96 images. These are all collected with the ring and contain 18 unique instances (devices): 2 blinds, 1 door, 4 lighting systems, 2 smart locks, 5 speakers, 2 TVs, and 2 HVACs. Each instance has 3-7 images associated with it taken from different perspectives. We sample one of these images *without replacement* and use it as a query image. The rest of the images are considered references. We then predict the instance associated with the query image using the semantic similarity based search algorithm described earlier. This prediction is compared against the ground-truth instance this query was sampled from, and marked as a correct or incorrect detection. This

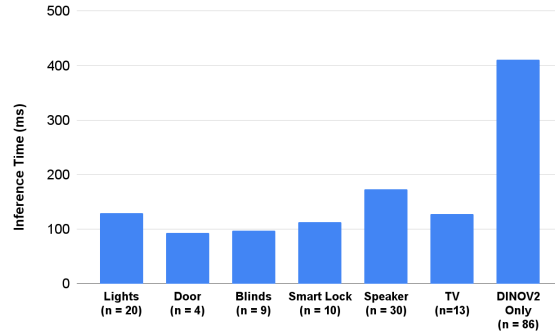


Figure 3.5: Query times for specific classes against the entire embedding database.

process is repeated across all 18 unique instances in the test set. Accuracy is computed over the entire set as the ratio of correct predictions over the total number of queries.

To evaluate how many reference images a user needs to collect, we repeat this process while limiting the number of reference images per instance. This experiment allows us to gain insight by measuring the accuracy as a function of the number of reference images available per instance. Our results indicate that 2 reference images are required per instance for 95% accuracy and 3 images reaches an upper bound of 98% accuracy. This approach results in very few inaccurate predictions (2%); failures with semantic search primarily occur due to a lack of sufficiently diverse references.

3.2.3 Latency Evaluation

End-to-end latency in our system is defined as the time elapsed between a participant completing a gesture (e.g. clicking the button) and the corresponding HomeKit command (e.g. toggling lights) being transmitted via an iPhone. We start our measurement by first putting IRIS into the IDLE state (camera clock-gated and IMU suspended) as this is the default state of IRIS when not in use. A hardware interrupt triggers upon image transfer completion through the nRF52840’s BLE stack. We measured the hardware latency using an oscilloscope, calculating the delta between the button’s falling edge and the image completion interrupt pin. This experiment yielded a hardware latency of 293ms, closely aligning with the expected value discussed in the design section.

The other contributors to latency in our system are YOLO and DinoV2. We measured YOLO latency on the iPhone and observed values between 24-28ms. For Dinov2, we first measured the embedding computation time, which was 7.69ms. Next, we evaluated the time required to query the embedding database, noting that this latency scales linearly with the database size. We measured query times across a range of database sizes (4 to 100 embeddings), resulting in values ranging from 9ms to 423ms, respectively. Even in the scenario where IRIS queries 100 embeddings, the total system latency remains at 752ms.

To further optimize inference latency, we utilized the YOLO+ CODA output to reduce the search space for relevant embeddings that solely align with the classified device. We measure the runtime using this technique across an embedding database size of 86 images, corresponding to 16 unique device instances (2 blinds, 1 door, 4 lights, 2 smart locks, 5 speakers, and 2 TVs). We show in Fig. 3.5, that reducing the search space in this way reduces the query runtime on the order of hundreds of milliseconds.

Chapter 4

VueBuds: Visual Intelligence with Wireless Earbuds

The emergence of large language models has transformed human-computer interaction, enabling natural language conversations with intelligent systems across diverse applications Bubeck et al. [2023]; Jiang et al. [2023]. This has accelerated further with the integration of visual intelligence, giving rise to visual language models that can comprehend and reason about images alongside text Liu et al. [2023]; Alayrac et al. [2022]. These advances have found rapid adoption in consumer devices, with smartphones leveraging on-device visual intelligence for enhanced photography and accessibility features Samsung, while smart glasses like Ray-Ban Meta have demonstrated the potential for wearable visual computing in everyday contexts Meta.

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, health-related sensors, and speakers Mollyn et al. [2023]; Bedri et al. [2017]. 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 150–200× larger than that of smart glasses Research; market Research.

We introduce VueBuds, the first wireless earbud system integrating low-power cameras with visual language model interaction. VueBuds allow users to capture visual context from their surroundings and engage with visual language models through a familiar, everyday wearable platform, without the requiring specialized eyewear. Our binaural system integrates dual, forward-facing wireless cameras, delivering stereo vision that can both capture the wearer’s perspective and the surrounding environment.

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 typically consume far more power than conventional earbud components, making it unclear whether wireless earbuds can support camera hardware while preserving acceptable battery life and form factor. Furthermore, camera data has significantly higher bandwidth requirements than audio, raising questions about whether the low-power wireless protocols used in earbuds can reliably transmit a binaural 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 align closely with the user’s eyes, ear-level cameras feature a posterior-lateral offset. This placement introduces potential facial occlusion, raising an unresolved question of whether such a vantage point can support effective egocentric perception.



Figure 4.1: VueBuds hardware 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).

• *RQ3: Can a fully wireless, Bluetooth-based pipeline deliver real-time multimodal interaction with vision–language models?* Answering user queries (e.g., “Where are my keys?”) requires capturing contextually relevant imagery, streaming it to a companion device via low-bandwidth Bluetooth,¹ performing multimodal inference using an on-device vision–language model, and returning an audio response. Whether this end-to-end pipeline can meet real-time latency constraints remains an open systems challenge.

In VueBuds, we address these questions through four core contributions:

• **Camera-integrated wireless earbud hardware.** We develop the first dual-camera earbud prototype, maintaining practical size, weight, and power (Fig. 4.1). The custom camera module attaches to commodity Sony WF-1000XM3 earbuds and operates at under 5 mW, adding only 11–14% battery overhead even under heavy use of 60 visual queries per hour.

• **Stereo vision for facial occlusion.** We introduce an ear-level stereo 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 (see 4.1.2).

• **End-to-end system optimizations for real-time operation and VLM integration.** In 4.1.1, we detail optimizations that maximize wireless throughput and minimize latency for concurrent binaural 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 (4.2.3) 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. 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) stereo ear-level capture effectively mitigates facial occlusion to provide egocentric views; (3) modern VLMs can operate effectively on low-resolution earbud imagery in real time; and (4) camera-integrated earbuds are perceived as highly accessible, achieving utility competitive with Ray-Ban Meta glasses. Together, these findings establish earbuds as a promising and socially accepted platform for

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.

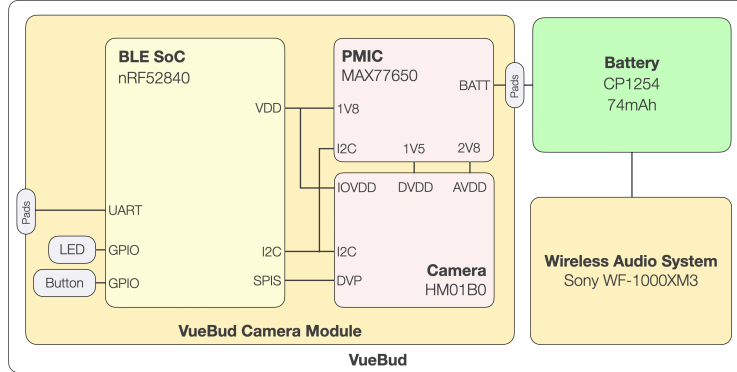


Figure 4.2: Hardware block diagram. Each VueBud integrates a Himax CMOS Imaging Camera and a Power Management Integrated Circuit (PMIC). Power is delivered through the onboard battery inside a Sony WF-1000XM3 wireless earbud.

everyday egocentric visual intelligence.

4.1 System Design

This section presents the design and implementation of VueBuds, addressing the hardware, stereo vision, and end-to-end integration with vision language models.

4.1.1 VueBuds Hardware

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 the original audio functionality while adding visual capabilities (Fig. 4.2). 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 scratch. Modern wireless earbuds incorporate sophisticated components, including digital signal processors, multiple microphones and 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 real-world earbud designs rather than simplified prototypes.

Following 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 the original earbuds while accommodating the additional camera hardware. Camera positioning optimization and angular field of view considerations are detailed in 4.1.2.

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,

Table 4.1: Latency and frame rate tradeoff.

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

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.

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 Semiconductor [2022]. This configuration achieves approximately 992 kbps of effective throughput NovelBits [2023].

We evaluate two camera configurations from the HM01B0 sensor: QQQVGA (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 QQQVGA and 1.6 fps for QVGA operation. However, image acquisition introduces additional latency through the 8MHz SPI interface, requiring 19.3ms for QQQVGA 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 QQQVGA and QVGA configurations, respectively (see Table. 4.1). These rates could be improved through pipeline optimization techniques that overlap image acquisition with wireless transmission Kvng [2025].

4.1.2 Stereo Vision for Earbud-Mounted Cameras

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 presents significant challenges when positioned at ear-level on wireless earbuds. Unlike temple-mounted cameras on smart glasses that benefit from forward positioning relative to the user’s face, earbud-positioned cameras suffer from substantial facial obstruction, where the

Table 4.2: Impact of camera angle on VueBud field of view.

Angular Orientation	Forward Blind Spot Length	Added Field of View	Stereo 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

user’s face blocks a considerable portion of each camera’s field of view, particularly in the inward-facing regions.

Facial obstruction severely impacts the effective field of view and can cause issues with visual 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.

However, this 25% reduction in captured visual information 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.

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 two purposes: (1) further reducing potential facial obstruction and (2) expanding the combined field of view coverage through stereo coordination.

In the completely forward-facing orientation (0°), windowed readout enables a pair of 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 Boccardo [2020]. 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 stereo 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. 4.2 demonstrates that camera orientations of 5° and 10° maintain forward blind spot 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 4.2.1.

Stereo Image Stitching for Redundancy Reduction

One approach is to feed dual independent cameras images from both the left and right earbuds into the visual-language model, along with context indicating that they represent the wearer’s perspective from each ear. We also explore stereo image stitching to reduce redundant visual information and improve processing efficiency for visual language model input.

Overlaps exist between left and right stereo 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 lightweight stitching using ORB (Oriented FAST and Rotated BRIEF) feature detection, selected for computational efficiency Rublee et al. [2011]. At a high level, these techniques work by detecting repeatable keypoints 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 visual language model and minimize processing latency.

4.2 Evaluation

First, 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.

4.2.1 Vision Capability Benchmarks

Here, we evaluate VueBuds’ tolerance to blind spots against camera angle, and then describe VLM accuracy and latency results with images from our camera to explain why we chose Qwen2.5VL 7B out of 5 selected visual language models.

Blind Spot Evaluation

As discussed in 4.1.2, the usability of camera-integrated earbuds depends critically on capturing the user’s forward field of view while maintaining natural interaction patterns. Users should not need to hold objects at awkward angles or step backward to accommodate system limitations.

In 4.1.2, we developed a geometric model to project the stereo camera fields of view from ear-mounted positions. Using measurements from a mannequin wearing VueBuds, we determined the camera-to-head distance as 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 user’s

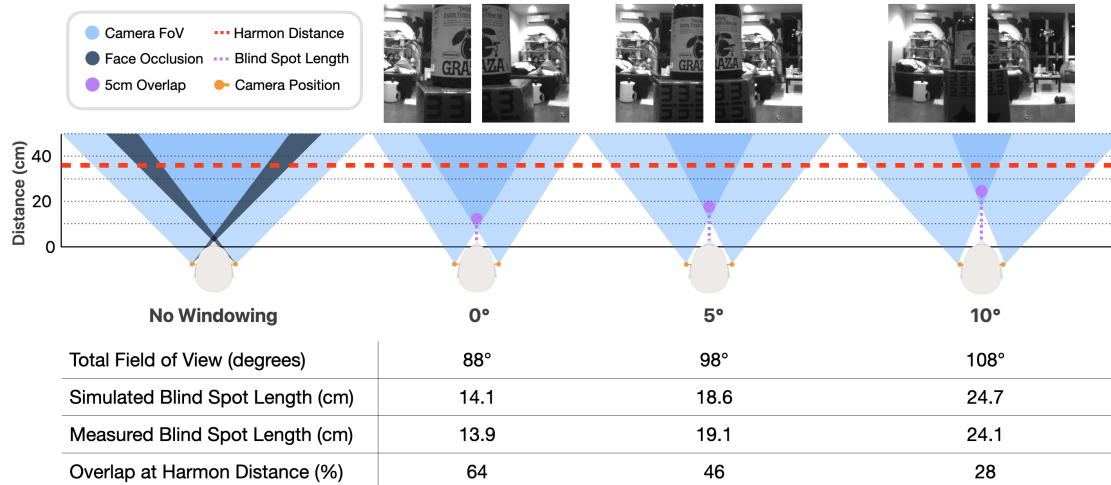


Figure 4.3: Blind spot evaluation. Geometric modeling of stereo camera field of view (blue) showing facial occlusion effects and blind spot regions at different angular orientations. The red line represents the Harmon distance. Empirically measured forward blind spot lengths closely align with the analytical model. (Top) Samples from the empirical measurements.

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. 4.3, our empirical measurements closely aligned with the geometric model calculations across all tested orientations. This strong correspondence (within 3% error) validated our geometric modeling approach for system design.

Understanding 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°. However, this expanded coverage comes at the cost of increasing blind spot length. The increasing blind spot 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, VueBuds prototypes were designed with angular orientations between 5° and 10°, balancing expanded field of view coverage with maintaining intuitive visual queries where users can simply look toward objects of interest without needing to adjust their natural interaction distances.

VLM Accuracy and Latency Performance

Selecting an appropriate visual language model for VueBuds requires balancing accuracy, latency, and on-device deployment constraints. We evaluated five models under 8B parameters.

Methodology. We evaluated Qwen2.5-VL (7B), Moondream (1.8B), MiniCPM-V (8B), LLaVA (7B), and Gemma3 (4B), all deployable through Ollama for on-device inference. 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 Apple, with human evaluators verifying response accuracy.

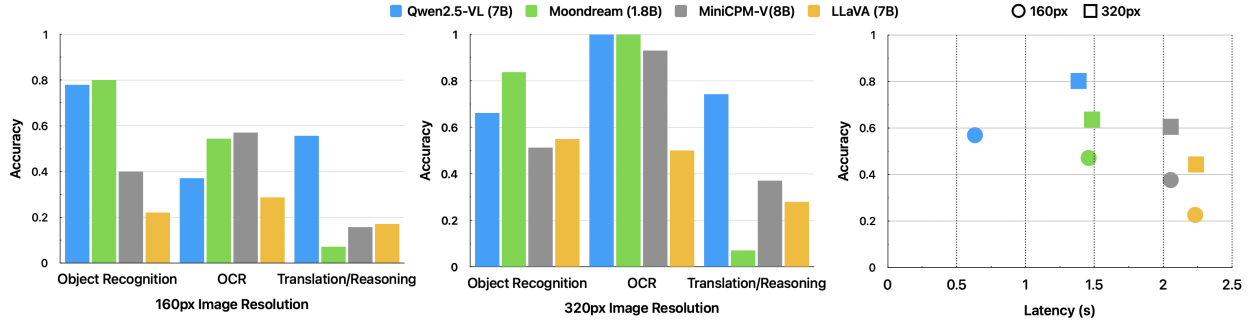


Figure 4.4: VLM comparison. Task-specific accuracy at 160px (left) and 320px (center), and accuracy-latency tradeoffs (right).

Resolution Impact. Fig. 4.4 reveals that increasing resolution from 160px to 320px, a 4× 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 that OCR capabilities degrade rapidly below certain resolution thresholds, as character boundaries become indistinguishable in lower-resolution captures. For queries like "What does this sign say?" or "Can you translate this for me?", the higher resolution proves essential. These findings validated our decision to use HM01B0’s 324×239 resolution mode despite the increased streaming latency.

Accuracy-Latency Trade-offs. Response latency directly impacts user experience in conversational interfaces. Fig. 4.4 illustrates this trade-off 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 Ollama [2024] 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 Bai et al. [2025], enable efficient processing despite its dynamic resolution adaptation. While Qwen2.5-VL shows the largest relative latency increase when scaling from 160px to 320px, this reflects its sophisticated dynamic resolution processing that adjusts visual token counts based on input complexity. Crucially, even at 320px, Qwen2.5-VL remains the fastest model in our evaluation.

Model Selection Rationale. Moondream, while matching Qwen2.5-VL on object recognition and OCR tasks, failed on reasoning queries (7.1% accuracy), often returning empty responses for complex questions DuckyBlender [2024]; maxi1134 [2025]. This limitation disqualifies it for queries requiring inference, such as "How many calories are in this?" or comparative questions. MiniCPM-V exhibited overfitting to our system prompts, generating templated responses that failed to adapt to varied user inputs. LLaVA, despite its widespread adoption, 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.

4.2.2 User Studies

To comprehensively evaluate VueBuds, we conducted user studies with 90 total 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,

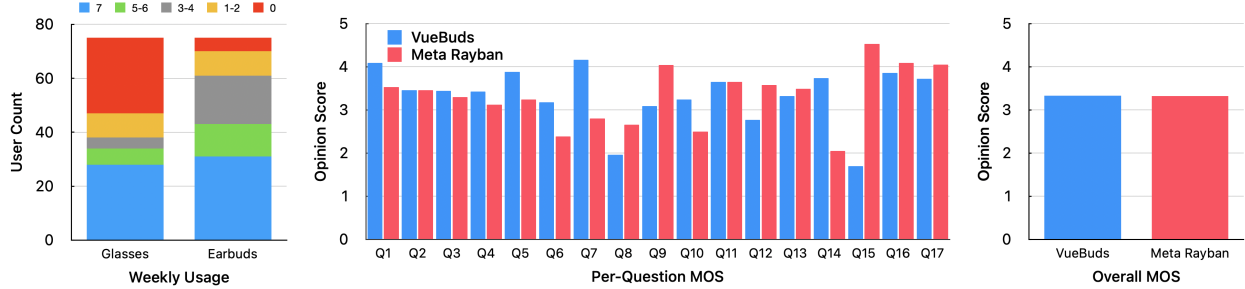


Figure 4.5: Online user study results. (Left) Weekly usage patterns for regular (non-smart) glasses and earbuds. (Middle) Mean Opinion Score of VueBuds in comparison with Meta Ray-Ban for each of the 17 visual-language question answering tasks. (Right) Mean opinion scores averaged across all the 17 tasks.

while the third was conducted through in-person sessions to evaluate real-world performance with physical hardware. Each study was a within-subject design, and no compensation was provided for participation. 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 captured or retained. All studies were approved by our university’s Institutional Review Board.

User Study 1: Platform Accessibility Analysis

Here, we examine device usage patterns to establish the potential user base for earbuds-based and glasses-based wearable systems. This analysis provides critical context for earbuds 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 regular 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 occurring in non-user rates. The exclusion rates present the most compelling insight for wearable AI platform accessibility: over one-third of participants (37.3%) never wear ordinary glasses, immediately excluding them from any glasses-based AI system. Conversely, only 6.7% avoid earbuds entirely, suggesting that earbud-based platforms could address a significantly larger potential user base.

These findings indicate that camera-integrated earbuds like VueBuds could reach a broader audience than




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 4.6: Samples from the online user study. *The visual scenes provided to participants were device-agnostic images captured with a mobile phone to ensure that the participants do not know the camera differences between the Meta Rayban and VueBuds.

glasses-based alternatives, particularly among the 37.3% of users who do not wear glasses. 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.

Study 2: Response Quality Evaluation Against Ray-Ban Meta

Here, we evaluate the subjective response quality between VueBuds and commercial smart glasses.

Participants. The same 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 vision-language tasks using a 5-point Mean Opinion Score (MOS) scale. We designed 17 evaluation scenarios representing typical smart glasses use cases: visual 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 responses. 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 comparative evaluation, participants were given the option to provide open-ended feedback (see Appendix)

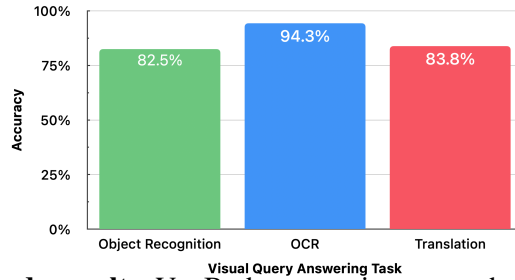


Figure 4.7: In-person user study results. VueBuds accuracies across the three visual question answering (VQA) tasks.

to elaborate on their rating criteria, specifically explaining the factors that drove high versus low scores.

Results. The overall MOS scores in Fig. 4.5 demonstrate near-statistical parity: VueBuds achieved 3.33 compared to Ray-Ban Meta’s 3.32. However, per-question analysis revealed notable task-dependent variations. The largest performance gap favored VueBuds in a translation task (Q7: 4.1 vs 2.8), where users strongly preferred specific translations over Ray-Ban Meta’s general scene descriptions. 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 using images from VueBuds. Most other tasks showed comparable performance, with differences typically under 0.5 points.

The near-identical overall scores (difference of 0.01) suggest that while individual task performance varies, both systems provide comparable user experience across diverse vision-language challenges. This parity validates that low-resolution grayscale images from VueBuds, when processed by Qwen2.5VL, can achieve response quality comparable to commercial smart glasses systems.

Analysis of participant feedback revealed five primary criteria driving quality ratings: (1) *Accuracy as foundation* - factual correctness was paramount, 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 appropriate for wearable, everyday conversation contexts; and (5) *Confidence calibration* - appropriate uncertainty expression valued over overconfident incorrect responses.

User Study 3: In-Person Performance Evaluation with VueBuds

Here, we perform an in-person study to evaluate VueBuds’ performance under realistic usage conditions, including natural variation in head shapes, earbud positioning, object holding distances, lighting conditions, and potential visual occlusions (e.g., glare, shadows).

Participants. We conducted an in-person user study with 16 participants (10 male, 6 female) between the ages of 20–44 ($\bar{x} = 27.9$, $\sigma = 6.9$). Participants were recruited locally in the United States and included 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 Htike et al. [2020]; Sys [2025]; Varshney et al. [2025]; Rani et al. [2023], text access Khan et al.




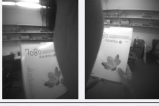


Stereo Images	Query	Response
	"what are some songs off of this album?"	I see a person holding a picture of the Beatles crossing Abbey Road. This is likely a reference to the album "Abbey Road." Some popular songs from this album include "Something," "Here Comes the Sun," and "Something in the Way She 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 4.8: VueBud question-response samples from the in-person user study. VueBuds captures the stereo images, which are then provided as input along with the query to the VLM model.

[2020]; Baig et al. [2024]; Varshney et al. [2025], and multilingual communication Sys [2025]; Rani et al. [2023]) 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

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. 4.7). 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 aligns closely 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. 4.8. These examples demonstrate robust performance despite challenging conditions, including partial facial occlusion (Row 4 and 5), objects split between stereo images (Row 1 and 3), and minor hair interference (Row 5 and 6), validating the system’s real-world applicability.

Table 4.3: VueBuds power consumption.

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

Table 4.4: Battery life (hours) across query rates for VueBuds integrated into Sony WF-1000XM3 earbuds.

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

4.2.3 System Evaluations

Power analysis

To evaluate the viability of VueBuds from a power perspective, we conducted comprehensive profiling across varying usage patterns and analyzed the 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 4.3. Based on these measurements, we calculated VueBuds’ power consumption across different query frequencies ranging from standby operation (0 queries/hour) to 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 introduces 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) BatteryClub and Apple AirPods Pro 2 (49.7mAh capacity) Air [2022]; Charlton [2022]. 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 in intensive scenarios (60 queries per hour). Table 4.4 shows the battery life numbers for VueBuds using the Sony WF-1000XM3. For the AirPods Pro 2, we estimate a battery life of 5.23 hours with 5 queries per hour and 5.18 hours with 60 queries per hour.

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 visual 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. We create an isolated pipeline which first sends a BLE write command to both VueBuds and mark this time in software. Upon receiving this command, each VueBud transitions from IDLE to ACTIVE, captures a single camera frame, and transmits the data back to the host device. Once both images are received, we calculate the

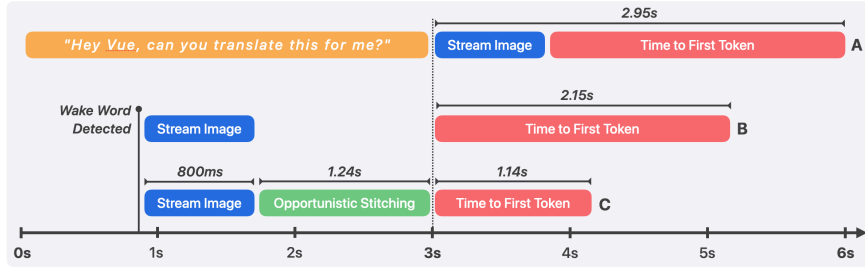


Figure 4.9: Latency breakdown across three system configurations. (A) Current prototype: 2.95s total latency due to waiting for complete audio processing before camera activation. (B) With on-device wake word detection: 2.15s latency by streaming images during query completion. (C) With opportunistic 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.

difference from the start time. Across 66 trials, we report dual image acquisition latency as ($\bar{x} = 800.1ms$, $\sigma = 0.06ms$), closely aligning with our expectations from 4.1.1.

Time to First Token Latency - Dual and Stitched. To calculate time-to-first-token for dual 324×239 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 Radford et al. [2023], 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 and wake word detection occurs after the user has completed their query and not at the onset (Fig. 4.9).

Existing wake word detection models on small embedded devices have lower detection latencies of sub-100ms from the end of the wake word Jose et al. [2020]. Thus, with proper integration, wake word detection would occur directly on the VueBuds hardware, enabling immediate camera activation upon hearing "Hey Vue" or "VueBuds." This allows images to stream in parallel with the user’s query, effectively hiding image capture latency. By the time the user completes their query, 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 a M4 Mac Mini Pro) plus audio synthesis overhead.

We also explore optimizing latency by performing 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 pre-processing decreases the number of input tokens processed by the VLM, resulting in an 46% improvement in time-to-first-token compared to processing raw L/R image pairs. However, due to parallax effects from the spatial separation of earbud-mounted cameras, stitching confidence can vary based on scene geometry and how VueBuds are worn by a user. We introduce this technique as Opportunistic Stitching, and the complete optimized pipeline achieves a time to first token latency of 1.14 seconds.

Fig. 4.9 illustrates the latency breakdown across all three configurations, while Fig. 4.10 demonstrates

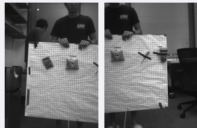




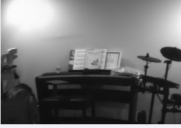
Stereo Images	Response	Stitched Image	Response
	I see a person holding a large grid-like board with two small bags attached to it. The person is wearing a shirt with the word "NAVY" on it. There's a desk and a chair in the background.		I see a person holding a large grid paper with some small items placed 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 4.10: Comparison of VLM responses to dual stereo images (left) versus stitched panoramic images (right) captured by VueBuds. Reducing redundant information reduces the total input tokens, reducing inference latency by 47%.

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.

Chapter 5

Conclusion

Together, ClearBuds, IRIS, and VueBuds demonstrate the potential of real-time, wearable AI systems that seamlessly integrate sensing, communication, and neural processing to enhance human-computer interaction. ClearBuds pushes the boundaries of real-time speech enhancement using binaural wireless earbuds, IRIS pioneers vision-based interaction in a wireless ring form factor, and VueBuds establishes wireless earbuds as a viable platform for visual intelligence—extending the capabilities of one of the most ubiquitous wearable devices. Each system required innovations in hardware design, wireless communication, and efficient neural inference to operate within the strict constraints of their respective form factors. Collectively, this work demonstrates the utility of deep learning-powered wearables, showing that visual, auditory, and multimodal intelligence can be seamlessly embedded into familiar, socially acceptable form factors. By bridging state-of-the-art neural networks into ubiquitous wearable devices, this dissertation paves the way for future intelligent wearables that are more responsive, context-aware, and seamlessly integrated into everyday life.

Bibliography

- a. Apple airpods. <https://www.apple.com/airpods/>.
- b. <https://appleinsider.com/articles/21/03/30/apple-airpods-beats-dominated-audio-wearable-market-in-2020>.
<https://www.microsoft.com/en-us/research/academic-program/deep-noise-suppression-challenge-interspeech-2021/>.
- Project gutenber. <https://www.gutenberg.org/>. Accessed: 2021-12-20.
- www.krisp.ai.
2015. Setting up the timeslot api. <https://devzone.nordicsemi.com/nordic/short-range-guides/b/software-development-kit/posts/setting-up-the-timeslot-api>.
2016. Wireless timer synchronization among nrf5 devices. <https://devzone.nordicsemi.com/nordic/short-range-guides/b/bluetooth-low-energy/posts/wireless-timer-synchronization-among-nrf5-devices>.
2022. 3c certification reveals the apple airpods pro 2's battery capacity.
2025. Enhancing Vision, the Smart Way. *Review of Ophthalmology*. Accessed: [Dec 1 2025].
- 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*.
- Apple. Choose your new mac mini.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibó Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yanzhi 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*.
- Mirza Samad Ahmed Baig, Syeda Anshrah Gillani, Shahid Munir Shah, Mahmoud 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*.
- BatteryClub. Batteries for sony wf-1000xm3, samsung galaxy buds, and buds live.

- 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).
- Laura Boccardo. 2020. Viewing distance of smartphones in presbyopic and non-presbyopic age.
- 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.
- Cox R. Neto S.F.de C. Lamblin C. and Sherif M.H. 2009. Itu-t coders for wideband, superwideband, and fullband speech communication. IEEE.
- Hartley Charlton. 2022. AirPods pro 2 earbud batteries 15% bigger, but charging case sees little improvement.
- Amit Chhetri, Philip Hilmes, Trausti Kristjansson, Wai Chu, Mohamed Mansour, Xiaoxue Li, and Xianxian Zhang. 2018. Multichannel audio front-end for far-field automatic speech recognition. In *2018 EUSIPCO*, pages 1527–1531. IEEE.
- Retail Dive. 2024. 27% increase in smart home adoption since 2020: Yougov report.
- DuckyBlender. 2024. moondream returns no response.
- Yasuhiro Endo, Zheng Wang, J Bradley Chen, and Margo I Seltzer. 1996. Using latency to evaluate interactive system performance. *ACM SIGOPS Operating Systems Review*, 30(si):185–199.
- Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. 2017. Mobilenets: Efficient convolutional neural networks for mobile vision applications.
- 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):26–26.
- Shilpi Jain, Sriparna Basu, Arghya Ray, and Ronnie Das. 2023. Impact of irritation and negative emotions on the performance of voice assistants: Netting dissatisfied customers’ perspectives. *International Journal of Information Management*, 72:102662.
- Teerapat Jenrungrot, Vivek Jayaram, Steve Seitz, and Ira Kemelmacher-Shlizerman. 2020. The cone of silence: Speech separation by localization.
- 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*.
- Christin Jose, Yuriy Mishchenko, Thibaud Sénéchal, Anish Shah, Alex Escott, and Shiv Vitaladevuni. 2020. Accurate detection of wakeword start and end using a cnn.
- Fahim Kawsar, Chulhong Min, Akhil Mathur, and Alessandro Montanari. 2018. Earables for personal-scale behavior analytics. *IEEE Pervasive Computing*, 17(3):83–89.

- 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):507–517.
- Hamid Krim and Mats Viberg. 1996. Two decades of array signal processing research: the parametric approach. *IEEE signal processing magazine*, 13(4):67–94.
- Kvng. 2025. Pipeline (computing).
- Xiaoyu Li, Shuqin Zeng, Yanwei Zhang, Ping Wan, and Jun Wang. 2012. Analysis and processing of pixel binning for color image sensor. *EURASIP Journal on Advances in Signal Processing*, 2012(1):81.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *NeurIPS*.
- Yi Luo, Zhuo Chen, Nima Mesgarani, and Takuya Yoshioka. 2020. End-to-end microphone permutation and number invariant multi-channel speech separation.
- Yi Luo and Nima Mesgarani. 2019. Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.
- maxi1134. 2025. Ollama moondream:latest returns no answer.
- Meta. Ray-ban meta glasses.
- 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*, CHI '23, New York, NY, USA. Association for Computing Machinery.
- Jared Newman. 2022. The smart home is flailing as a concept because it sucks.
- Jakob Nielsen. 1993. Smart Home Statistics. <https://www.nngroup.com/articles/response-times-3-important-limits/>.
- NovelBits. 2023. Bluetooth 5 speed: How to achieve maximum throughput for your ble application.
- Oberlo. 2024. Smart Home Statistics. <https://www.oberlo.com/statistics/smart-home-market>.
- Ollama. 2024. Gemma2 models with ollama run extremely slow. GitHub Issue #9683.
- Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. 2024. Dinov2: Learning robust visual features without supervision.
- Tom Le Paine, Pooya Khorrami, Shiyu Chang, Yang Zhang, Prajit Ramachandran, Mark A. Hasegawa-Johnson, and Thomas S. Huang. 2016. Fast wavenet generation algorithm.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision.

- 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, ICML'23*. JMLR.org.
- 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)*, pages 1–6.
- Chandan K. A. Reddy, Harishchandra Dubey, Vishak Gopal, Ross Cutler, Sebastian Braun, Hannes Gamper, Robert Aichner, and Sriram Srinivasan. 2021. Iccasp 2021 deep noise suppression challenge. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6623–6627.
- Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788, Seattle, WA, USA. IEEE.
- Cognitive market Research. Wireless earbuds market report 2025 (global edition).
- Grand View Research. 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.
- Ana Rodrigues, Rita Santos, Jorge Abreu, Pedro Beça, Pedro Almeida, and Sílvia Fernandes. 2019. Analyzing the performance of asr systems: The effects of noise, distance to the device, age and gender. In *Proceedings of the XX International Conference on Human Computer Interaction, Interacción '19*, New York, NY, USA. Association for Computing Machinery.
- Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation.
- Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John R. Hershey. 2018. SDR - half-baked or well done? *CoRR*, abs/1811.02508.
- 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*, pages 2564–2571.
- Samsung. Discover a new perspective with bixby vision.
- Mia Sapienza. 2022. Are you still relying on your phone to control your home?
- Nordic Semiconductor. 2022. Things you should know about bluetooth range.
- Cem Subakan, Mirco Ravanelli, Samuele Cornell, Mirko Bronzi, and Jianyuan Zhong. 2021. Attention is all you need in speech separation.
- Ke Tan, Xueliang Zhang, and DeLiang Wang. 2019. Real-time speech enhancement using an efficient convolutional recurrent network for dual-microphone mobile phones in close-talk scenarios. In *ICASSP 2019*, pages 5751–5755.
- Punch Through. 2022. Maximizing ble throughput on ios and android.

- Ultralytics. 2024. Yolov8. <https://github.com/ultralytics/yolov8>. Accessed: 2024-03-31.
- Barry D Van Veen and Kevin M Buckley. 1988. Beamforming: A versatile approach to spatial filtering. *IEEE assp magazine*, 5(2):4–24.
- 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*, 39:1–6.
- 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*, chapter 25. Association for Computing Machinery, New York, NY, USA.
- Yoonjong Yoo, Jaehyun Im, and Joonki Paik. 2015. Low-light image enhancement using adaptive digital pixel binning. *Sensors*, 15(7):14917–14931.
- Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. 2018. The sound of pixels.