

Extending Human Capabilities with Deep Learning Powered Wearables

Doctoral Defense

Maruchi Kim

Shyam Gollakota

Computer Science and Engineering

University of Washington



Thank you to my collaborators

Shyam Gollakota

Steve Seitz

Ira Kemelmacher-Shlizerman

Vikram Iyer

Shwetak Patel

Srdjan Jelacic

Kelly Michaelson

Antonio Glenn

Bandhav Veluri

Anran Wang

Ishan Chatterjee

Vivek Jayaram

Eyoel Gebre

Jeanju Sung

Rasya Fawwaz

Zhi Yang Lim

Brinda Moudgalya

Hexi Wang

Yuanhao Zeng

Jose Jaime

Lucas Lee

Agenda

Introduction - *5min*



ClearBuds

Wireless Binaural Earbuds for
Speech Enhancement

5min



IRIS

Wireless Ring for Vision-based
Smart Home Interaction

5min



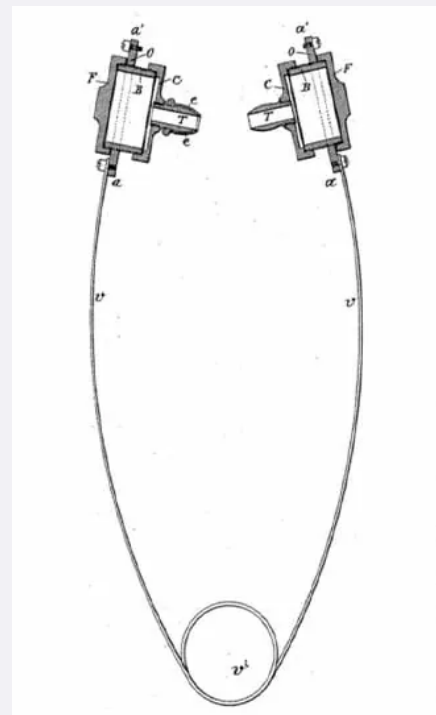
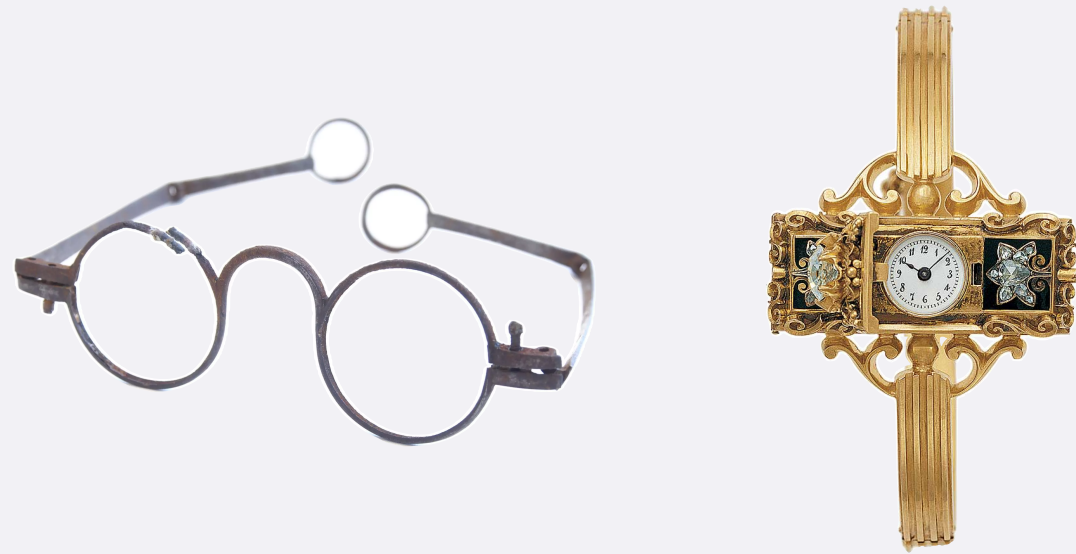
VueBuds

Visual Intelligence with
Wireless Earbuds

20min

Conclusion - *5min*

A brief history of wearables



Analog Era

Glasses (1727)
Wristwatch (1868)
Earbuds (1891)



Digital Era

Fitbit (2009)
Google Glass (2013)
AirPods (2016)
Oura Ring (2024)



AI Era

Rayban Meta (2023)
Humane AI Pin (2024)

Unlocking the potential of wearable systems with deep learning

Wearables

Passive sensing
Positioned for all-day context
Tiny batteries, limited compute

Deep Learning

Intelligence and understanding
Resource-intensive
Trained on massive datasets

Deep Learning Powered Wearables

Active intelligence
Utilizes all-day context
Enabling new interactions

Understanding and enhancement of human capabilities

Key Challenges

Resource constraints (power, compute, memory)
Real world deployment (latency, privacy, generalization)
User experience (form factor, all-day comfort, seamless interaction)

ClearBuds

Wireless Binaural Earbuds for
Learning-based Speech Enhancement

Ishan Chatterjee*, **Maruchi Kim***, **Vivek Jayaram***

Shyam Gollakota, Ira Kemelmacher, Shwetak Patel, Steve Seitz

University of Washington

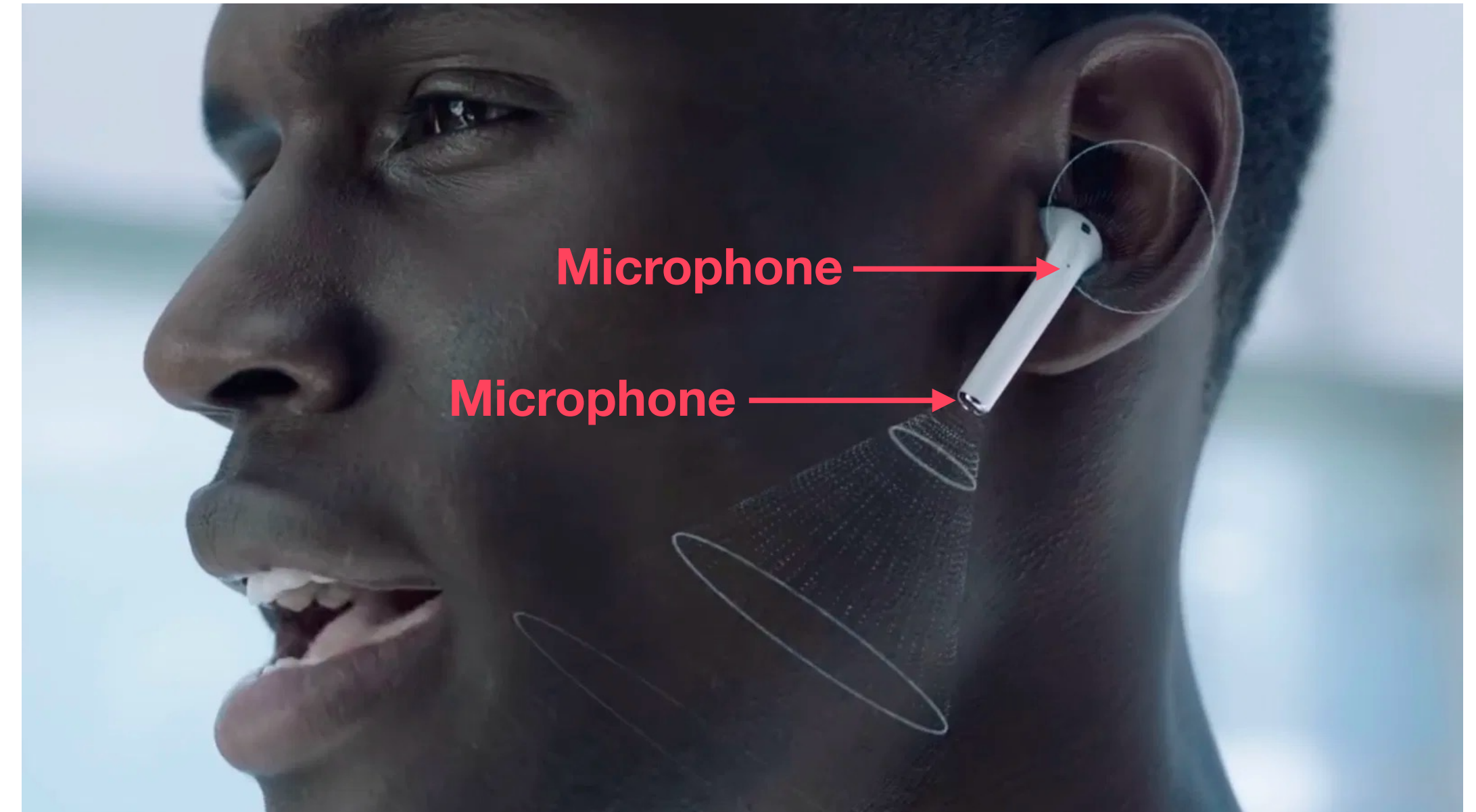
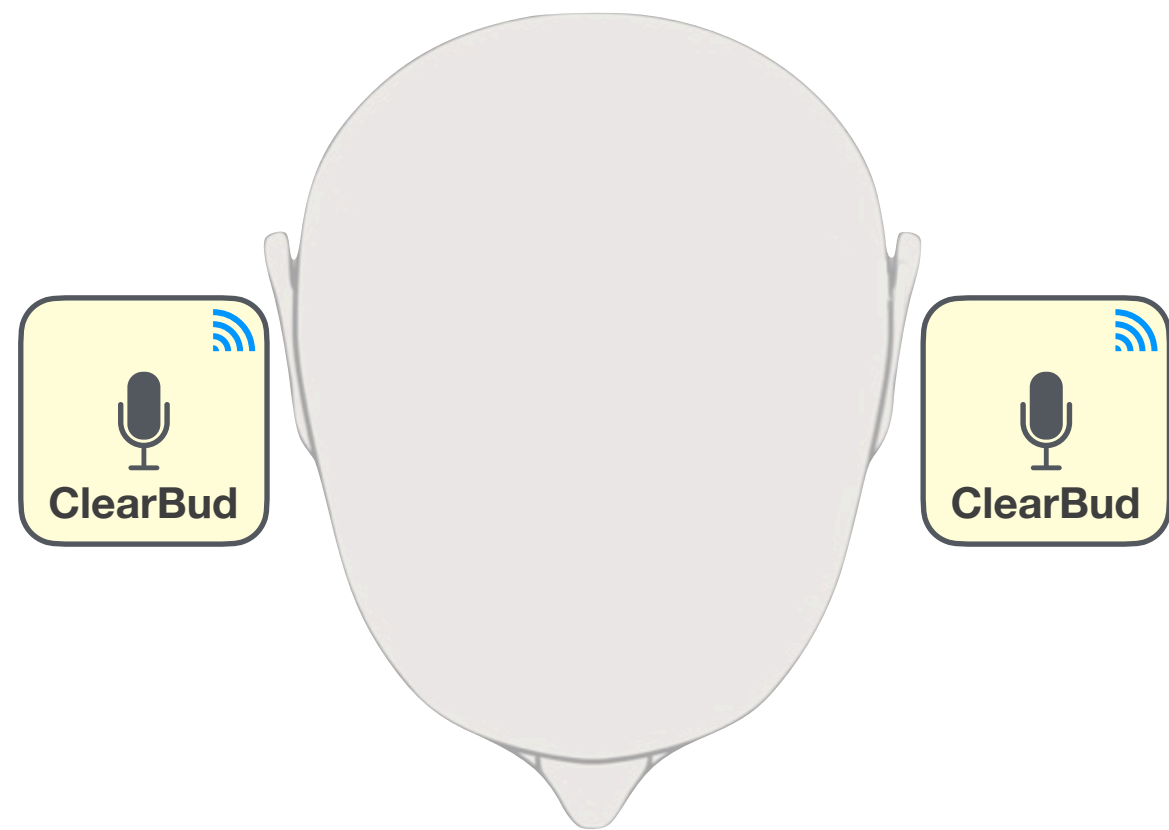
*equal contribution



Millions of people taking calls in noisy environments



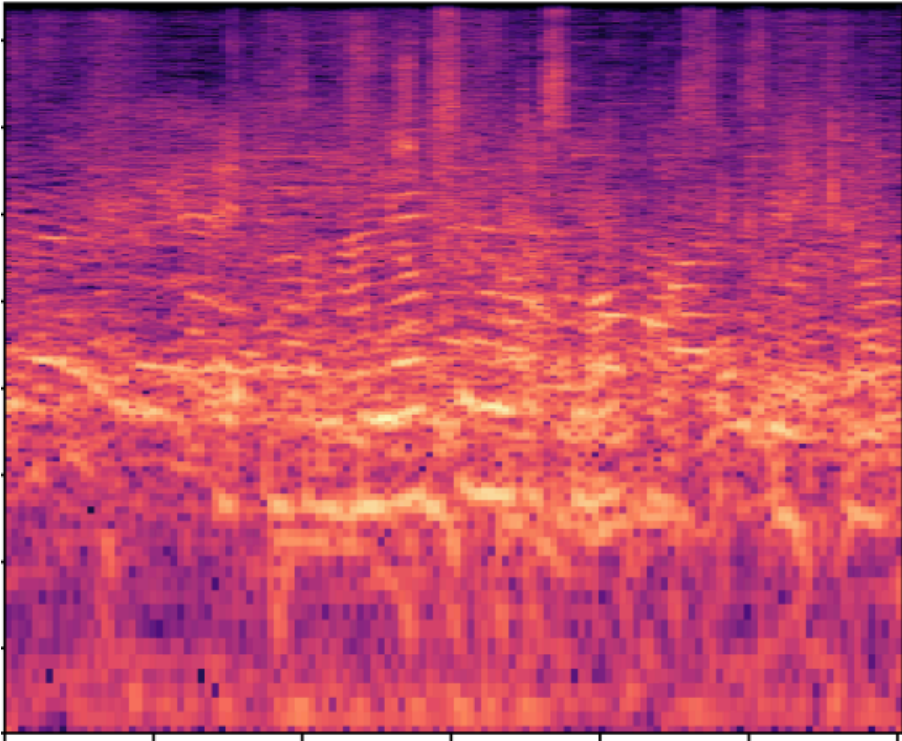
Observation 1: Microphones across the head provide better spatial resolution



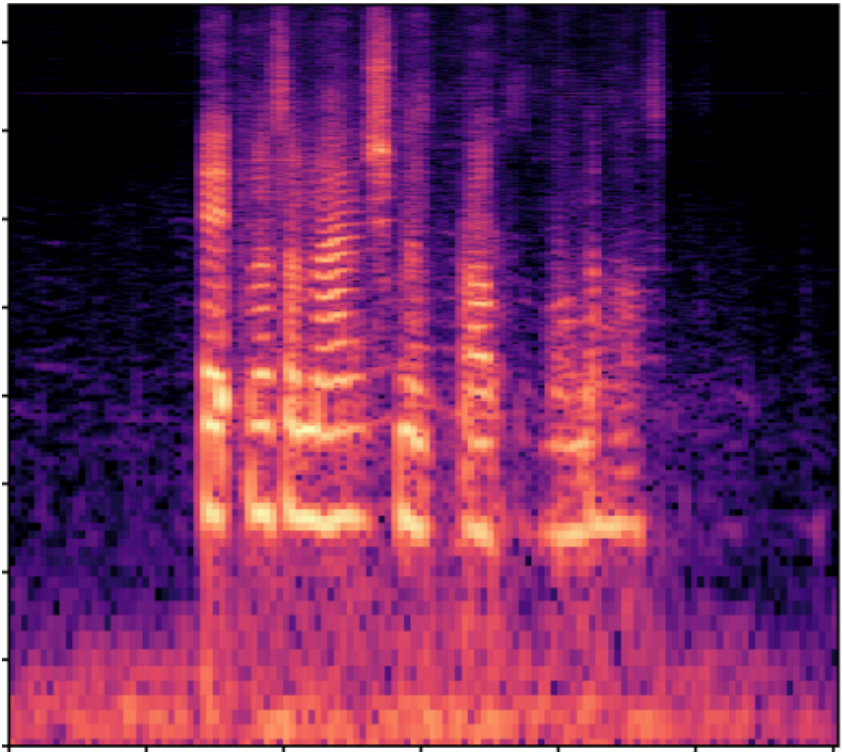
Challenge: Creating a wireless synchronized microphone array

Observation 2: Neural networks outperform signal processing by 8-10 dB

“Hey Joe, can you get me a coffee?”



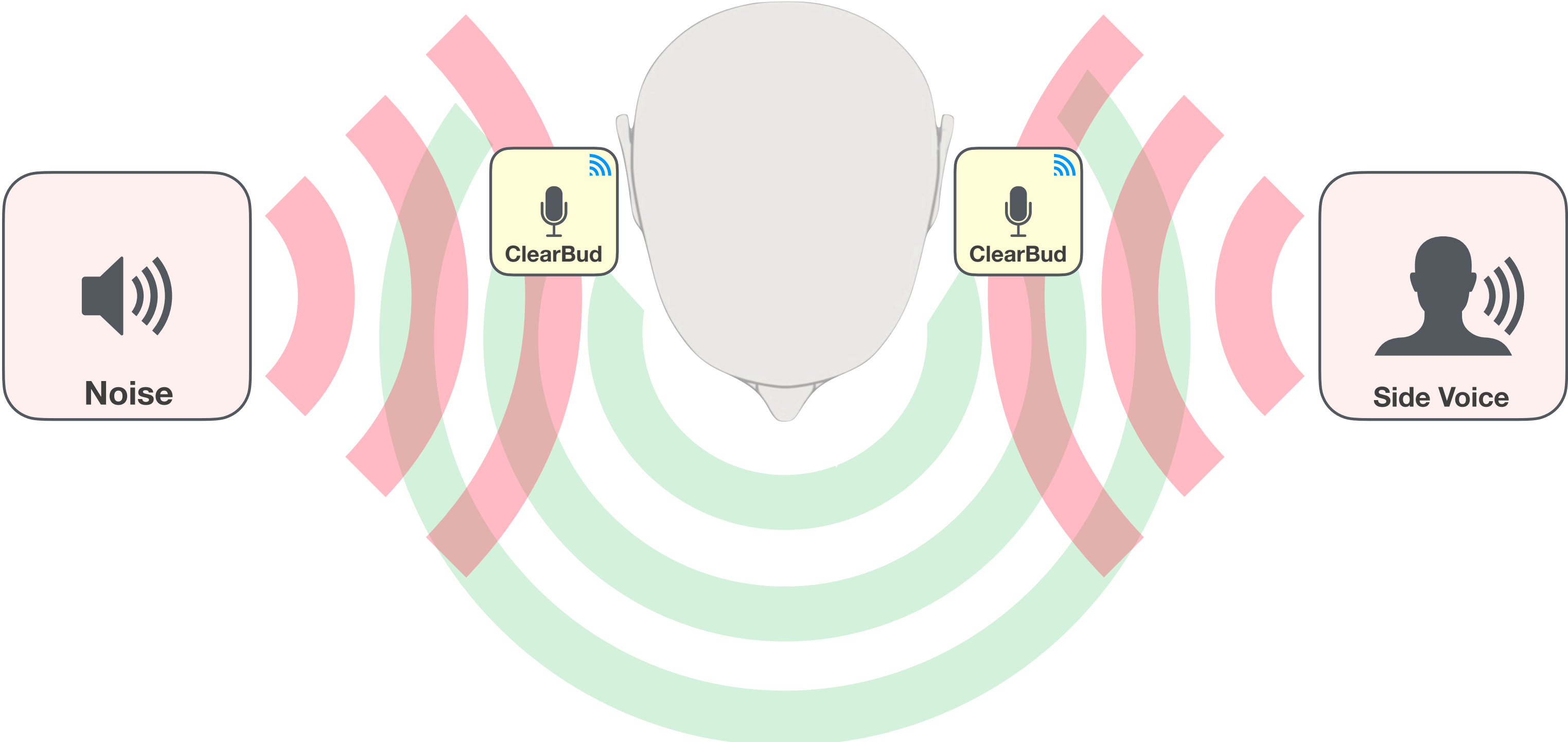
Speech + Noise



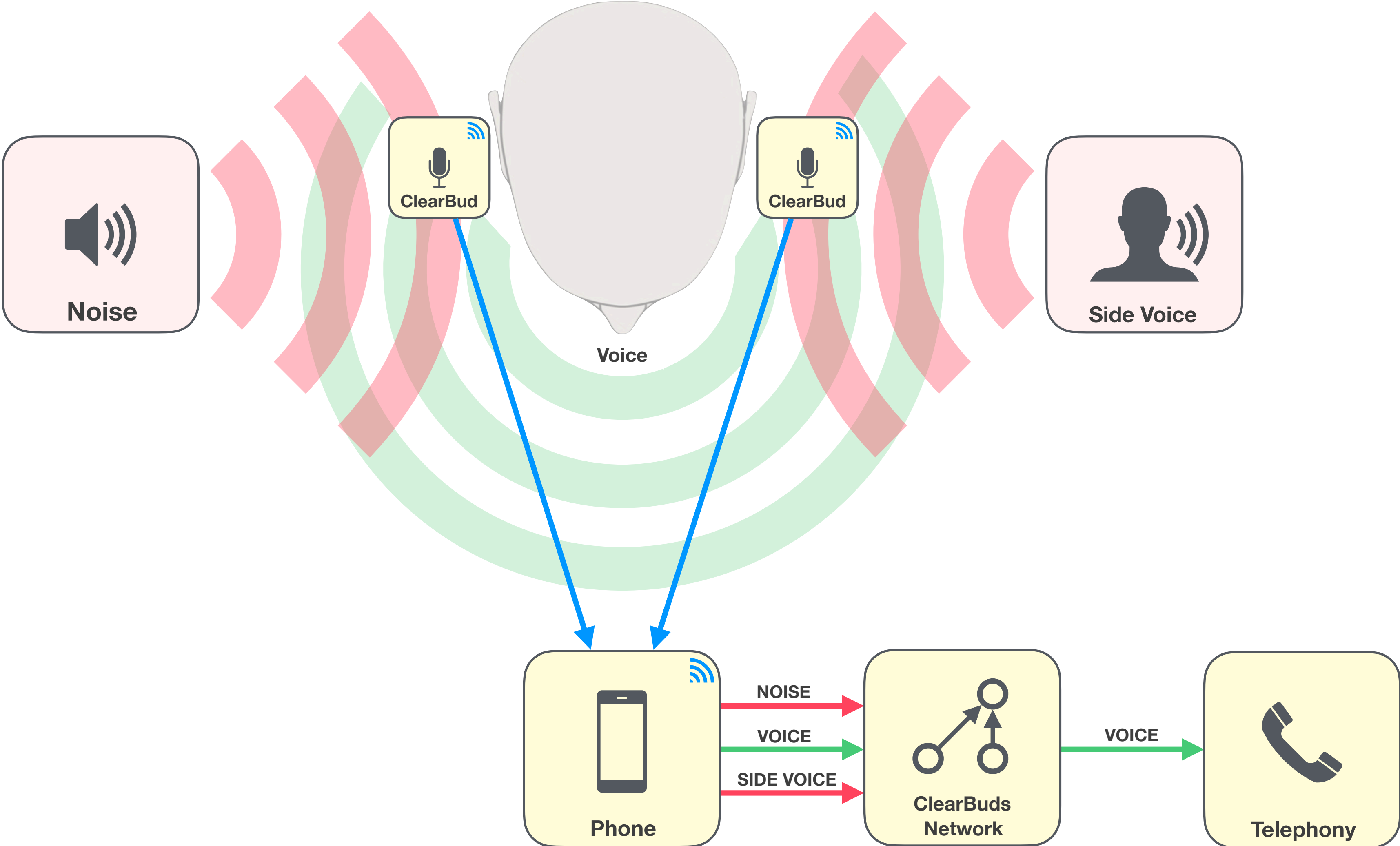
Isolated Speech

Challenge: Neural networks are computationally expensive

ClearBud System Overview



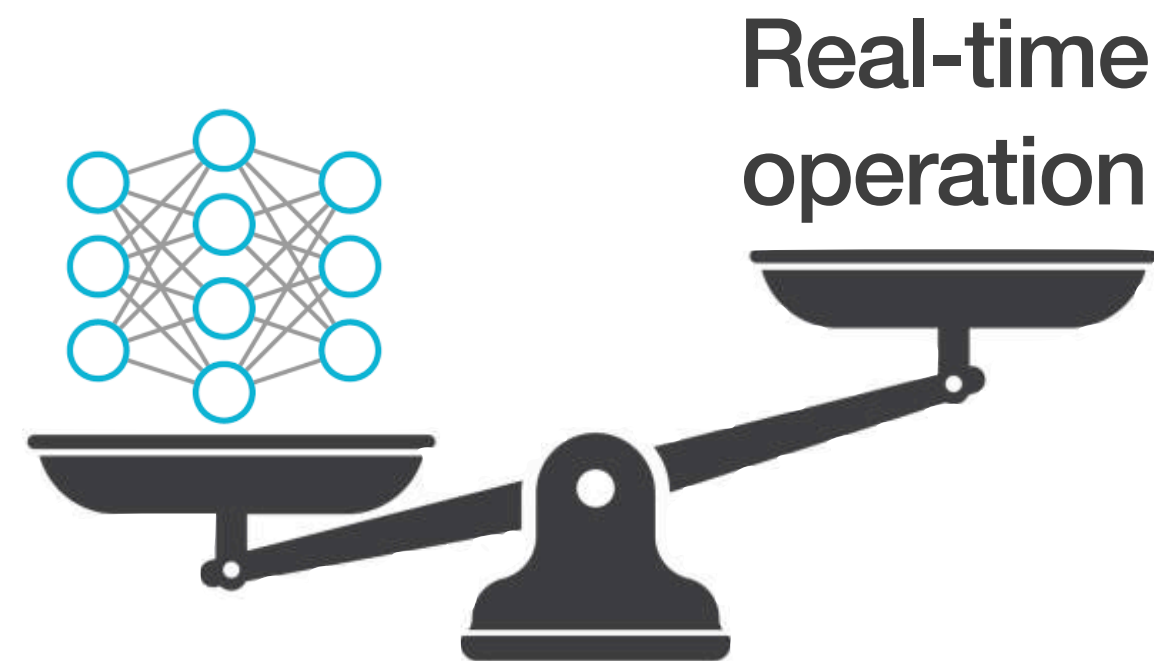
ClearBud System Overview



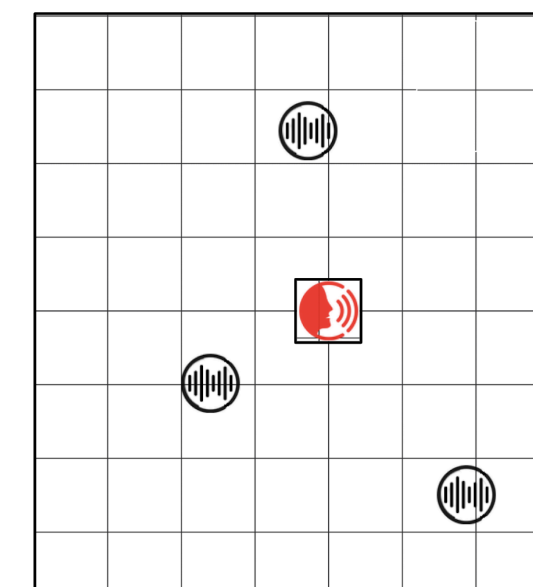
What are the key technical challenges?



Creating a synchronized wireless microphone array



Designing a lightweight neural network for speech enhancement



≠

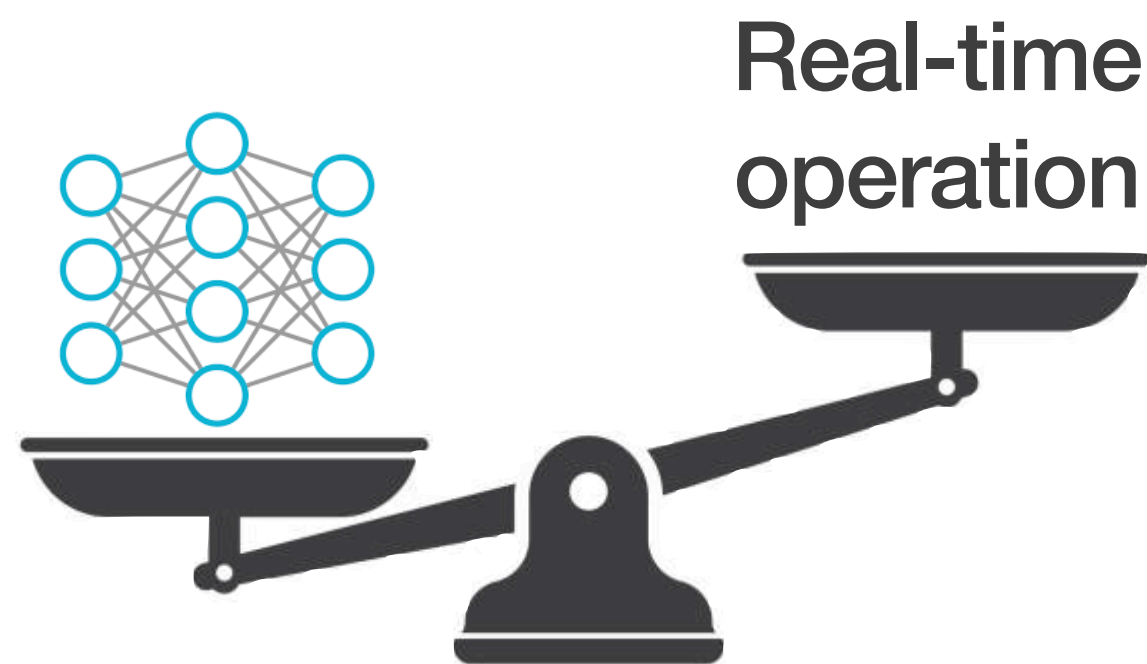


Generalization to real-world multi-path environments

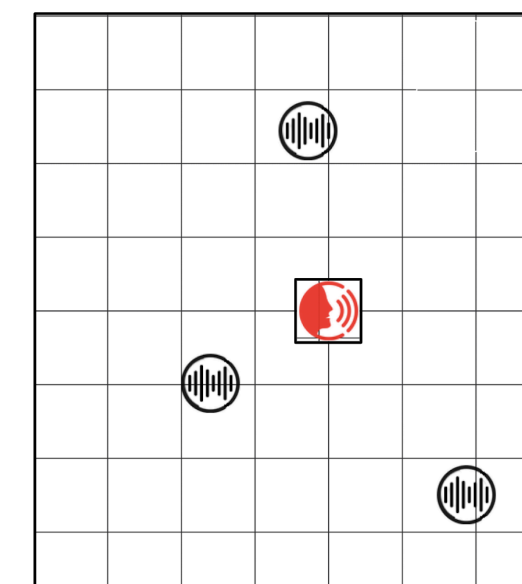
What are the key technical challenges?



Creating a synchronized wireless microphone array



Designing a lightweight neural network for speech enhancement



≠

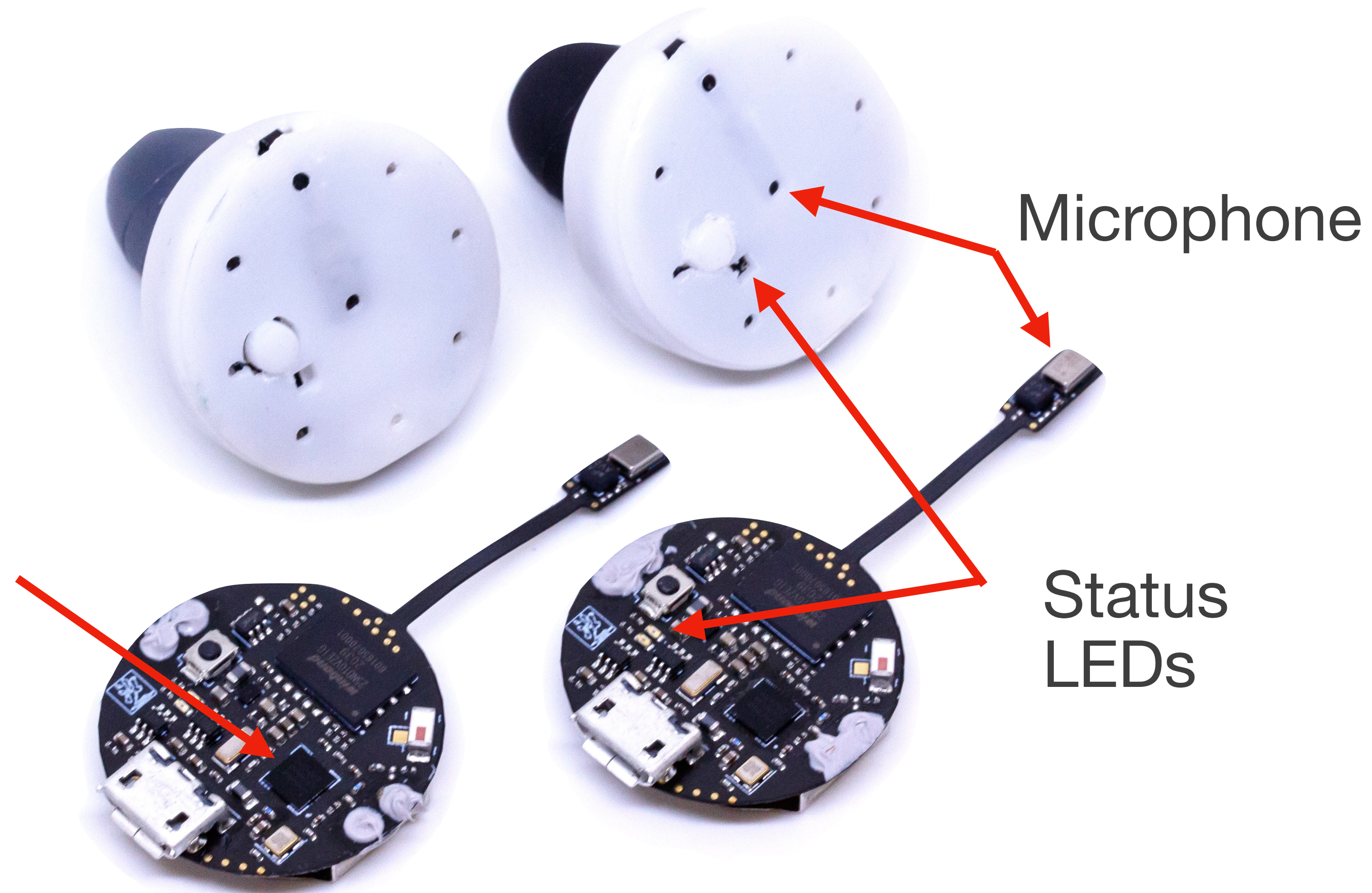


Generalization to real-world multi-path environments

No existing wireless earbuds hardware support binaural microphone streaming



World's first wireless binaural microphone streaming earbuds



Bluetooth LE
radio (nrf52840)

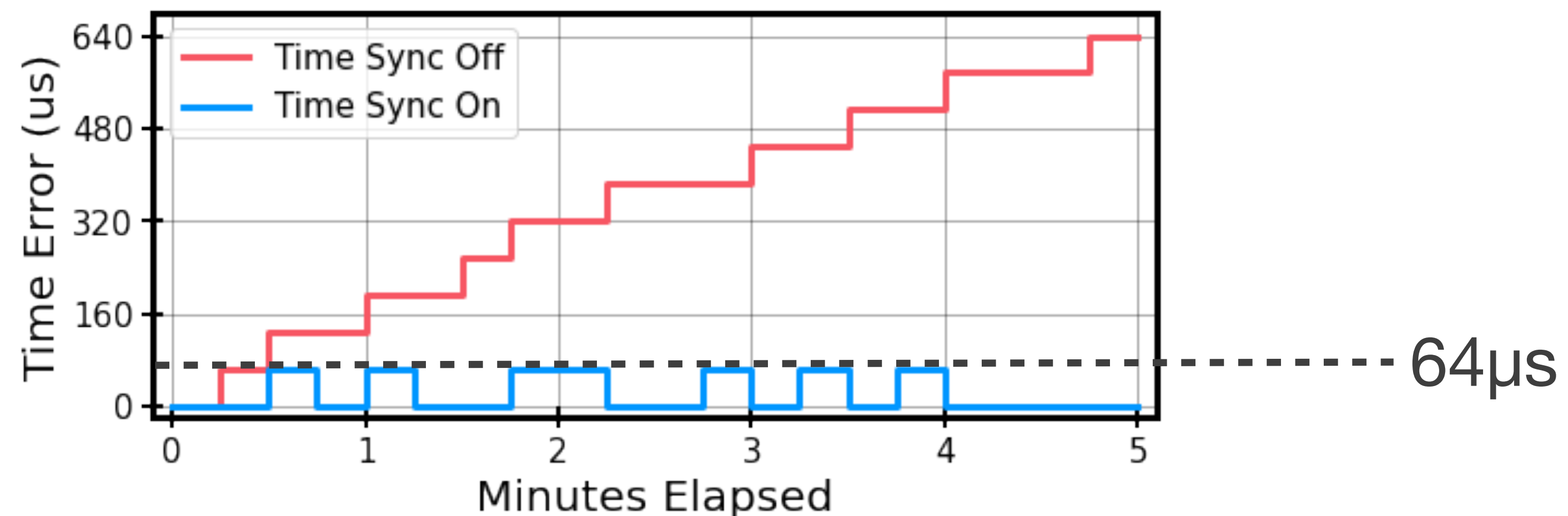
Microphone

Status
LEDs

40 hours of operation on a coin
cell battery (underneath)

Wireless microphone synchronization requires three components

1. A common clock for a shared reference of time
2. A synchronized startup so both mics start recording at the exact same time
3. Autocorrecting sampling rate algorithm to address hardware clock drift



SIGLENT

SDS 1104X-E
Digital Storage Oscilloscope

SPO
Super Phosphor Oscilloscope

100 MHz

Dual 1 GSa/s
Quad 500 MSa/s

SIGLENT **Time** M 50.0ns/ Delay:0.00s

f = 199.999Hz

Sa 1.00GSa/s

Curr 700pts

Edge

f DC

L 1.76V

1

1X 2.00V/

3.92V

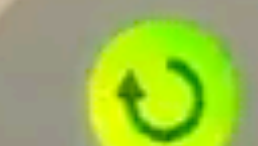
3

1X 2.00V/

0.0V

$\Delta X = 300.0\text{ns}$
 $1/\Delta X = 3.333\text{MHz}$
 $X2 = 149.0\text{ns}$
 $X1 = -151.0\text{ns}$

50.0ns/

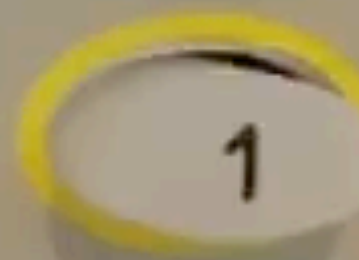


Intensity
Adjust

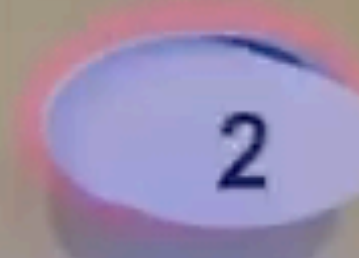


Select

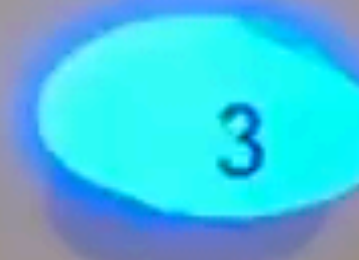
Ve



1



2

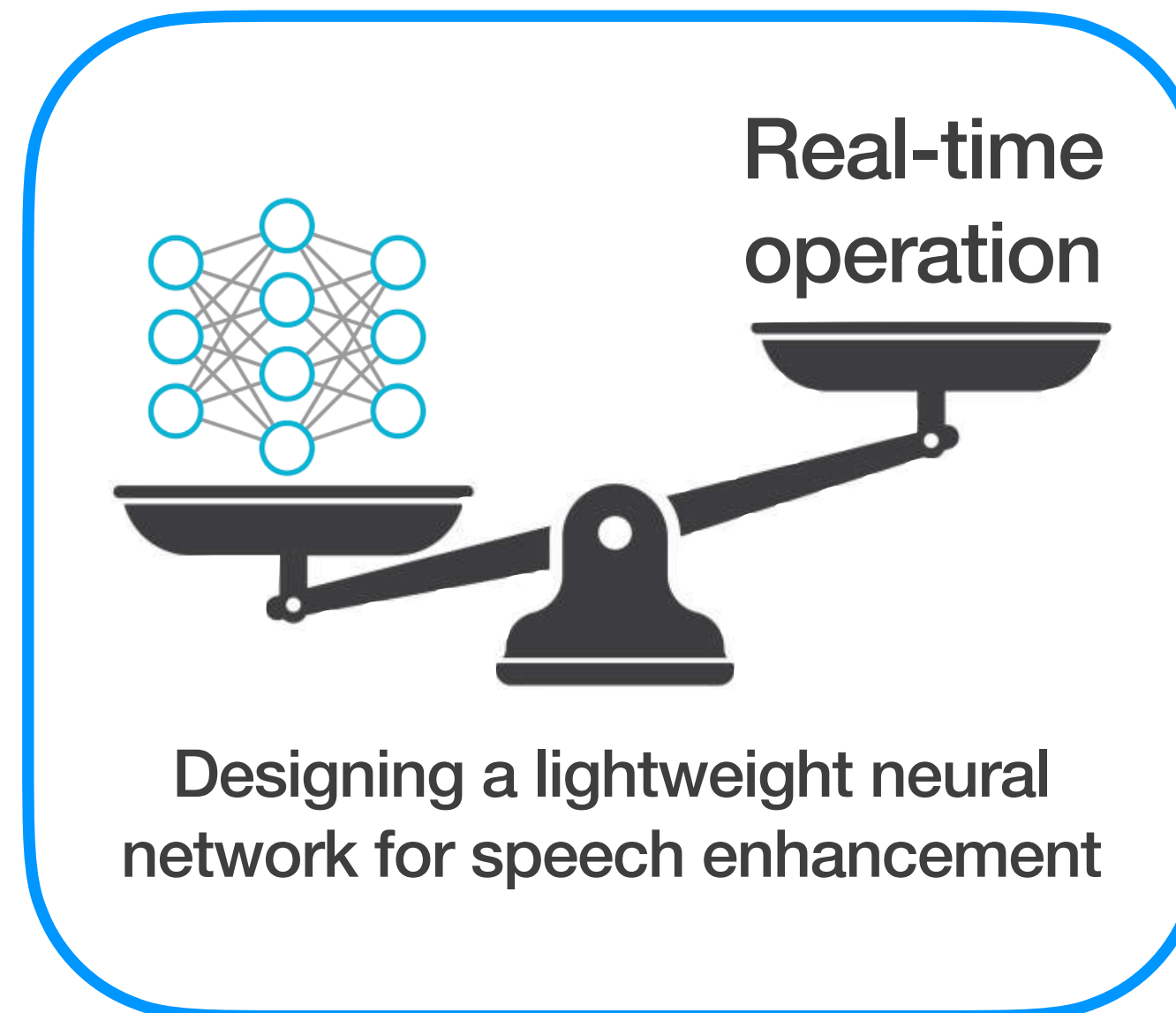


3

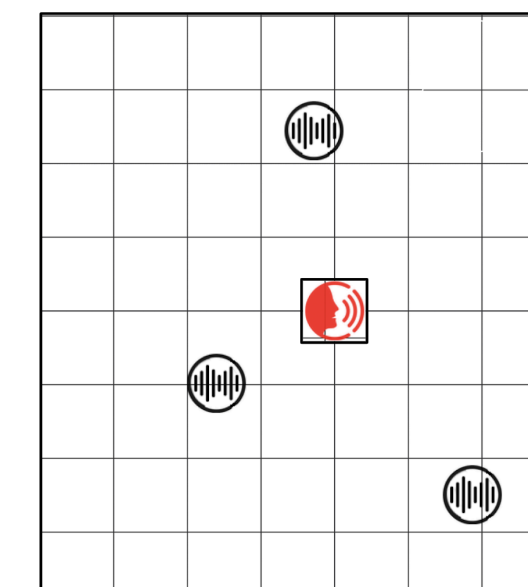
What are the key technical challenges?



Creating a synchronized wireless microphone array



Designing a lightweight neural network for speech enhancement

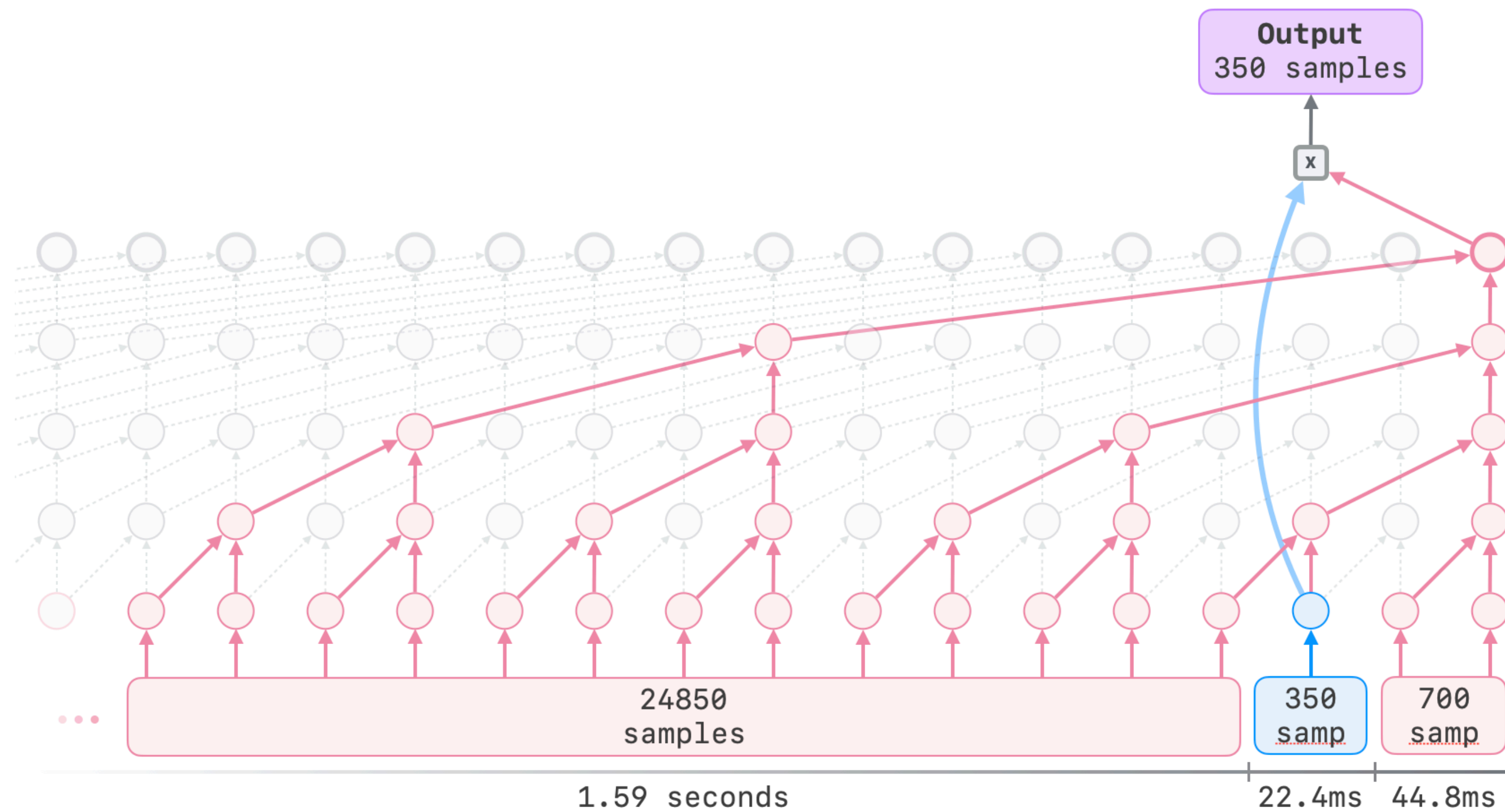


≠



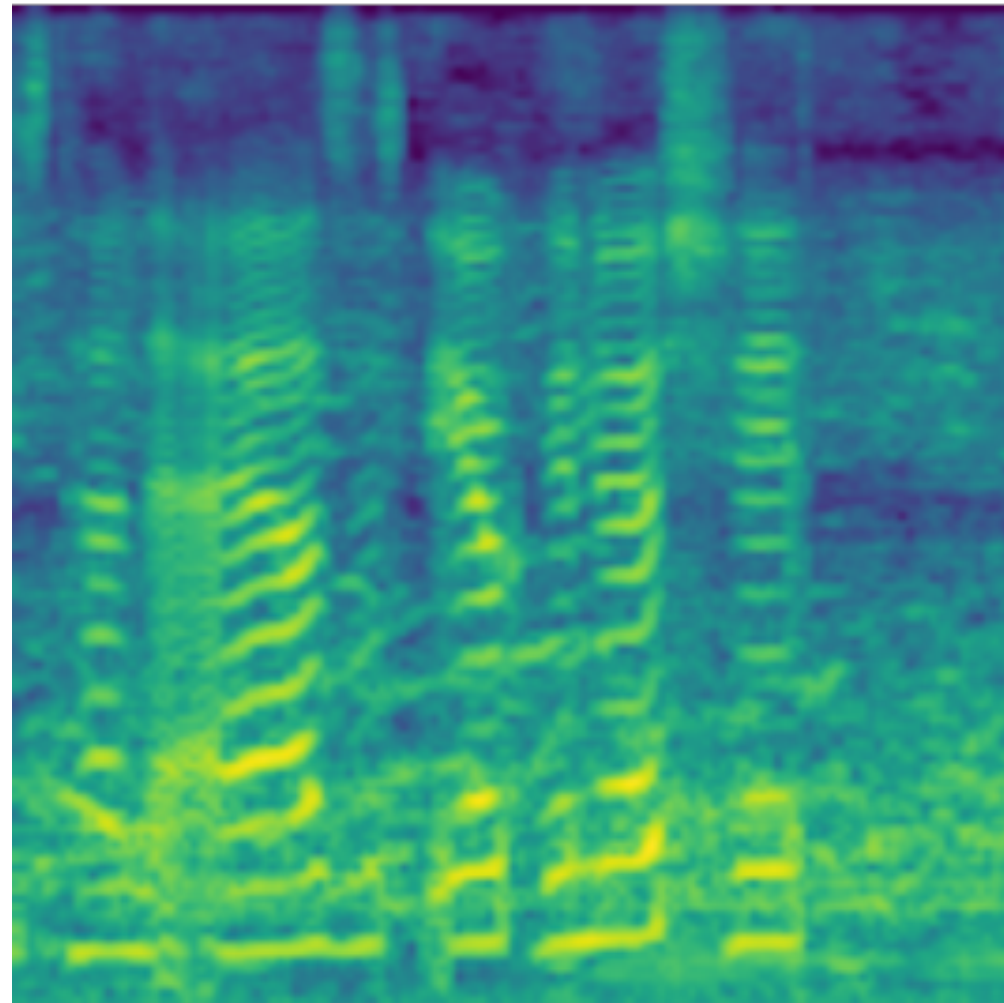
Generalization to real-world multi-path environments

Naive solution: Reduce the number of layers and features

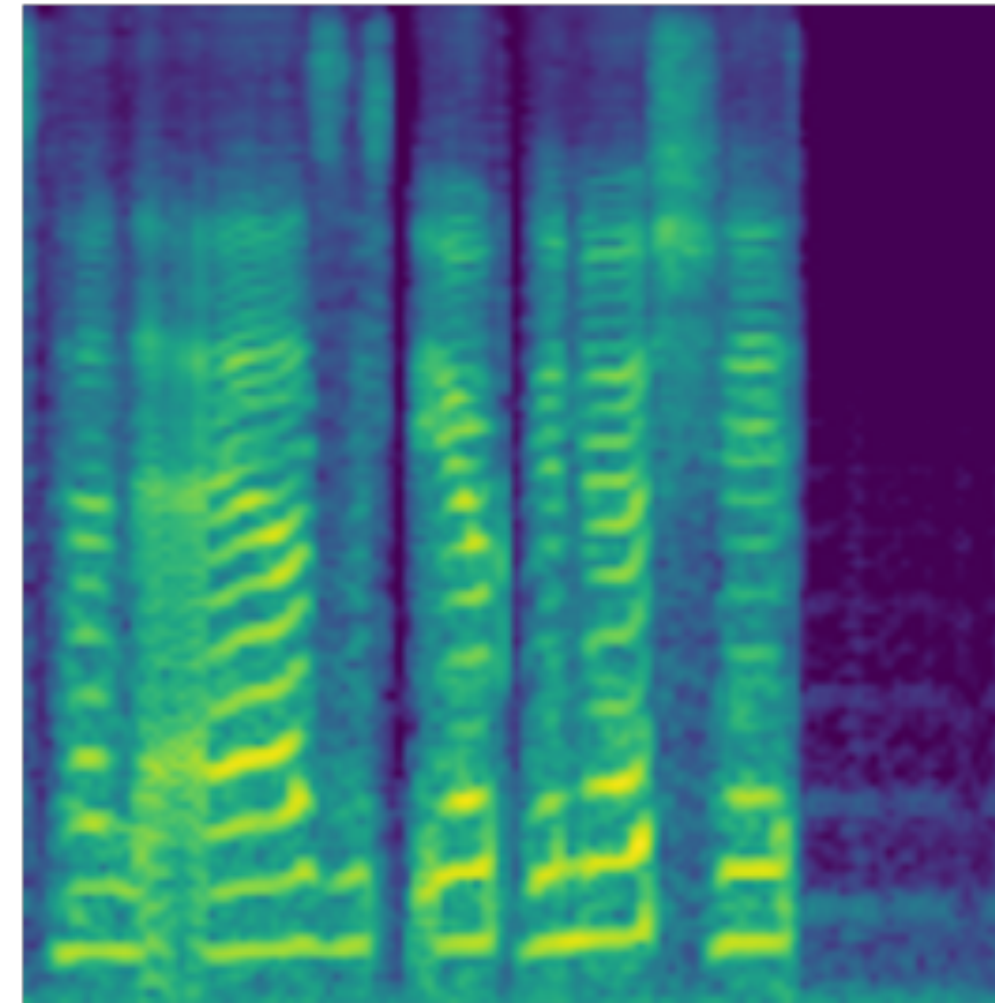


- Time domain network
- Reduce number of layers and number of features
- Cache intermediate activations

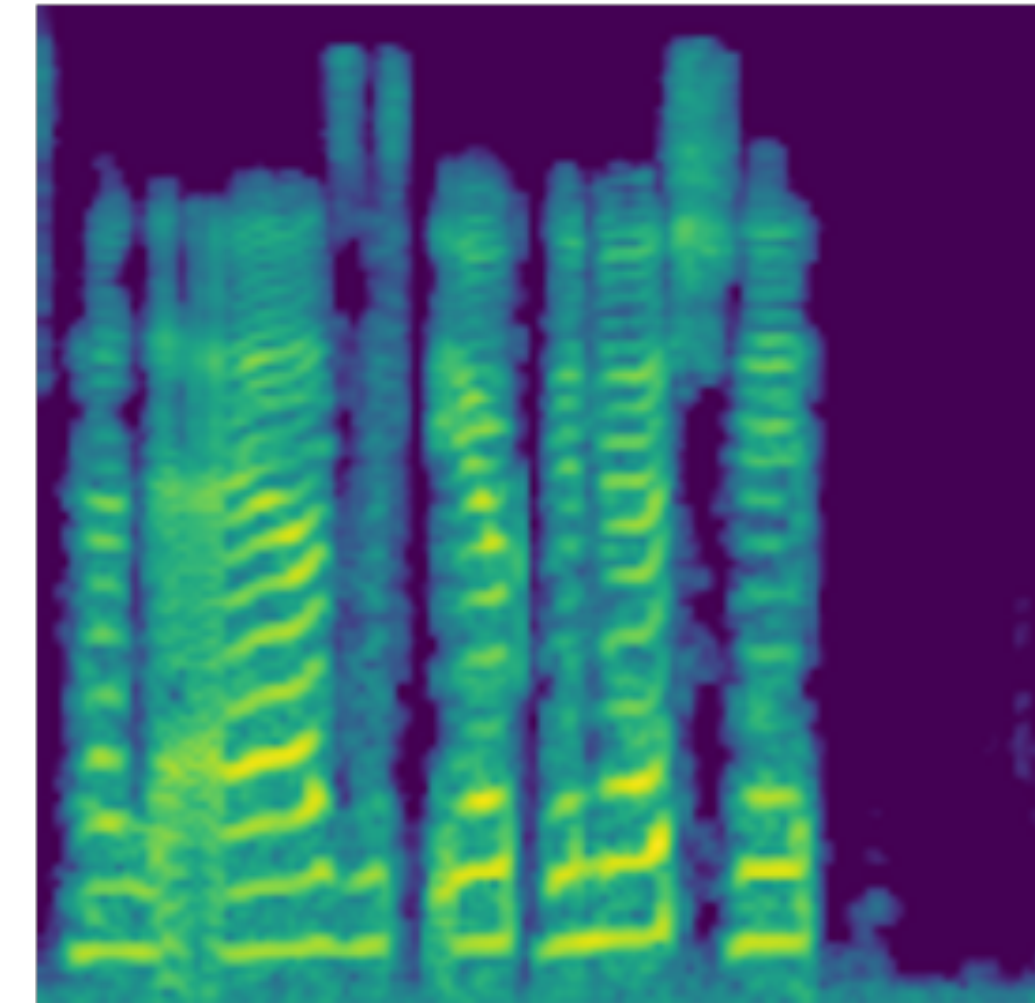
Problem: Reducing network size introduces artifacts



Mixture

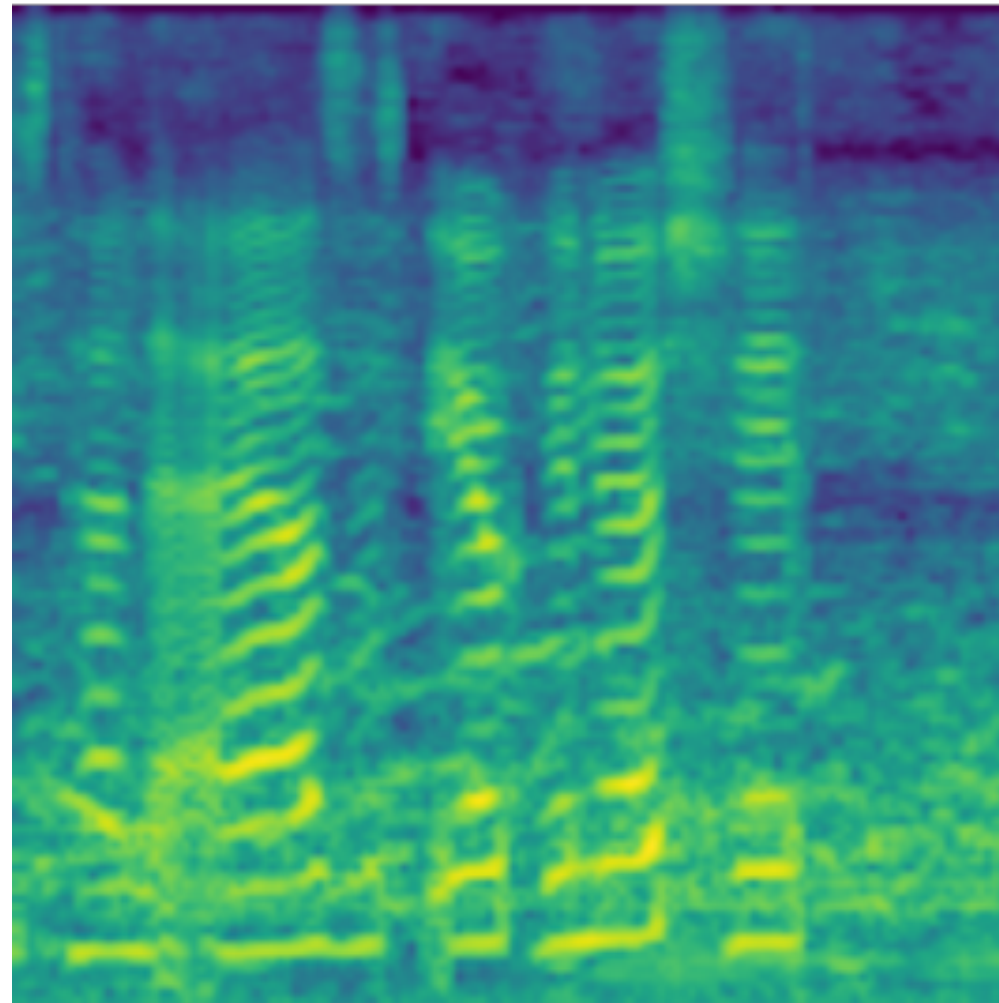


Time Domain Network

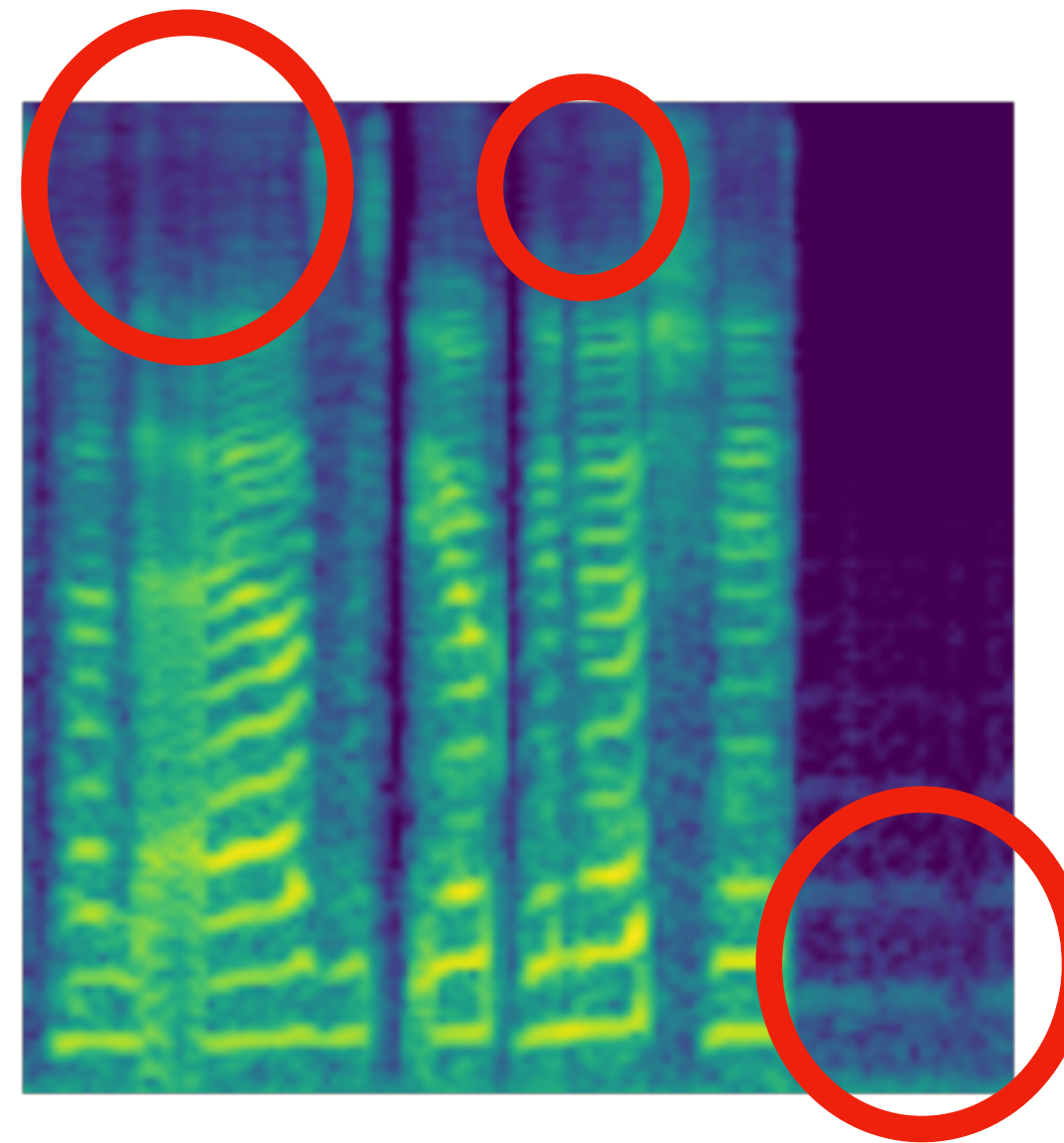


Ground Truth

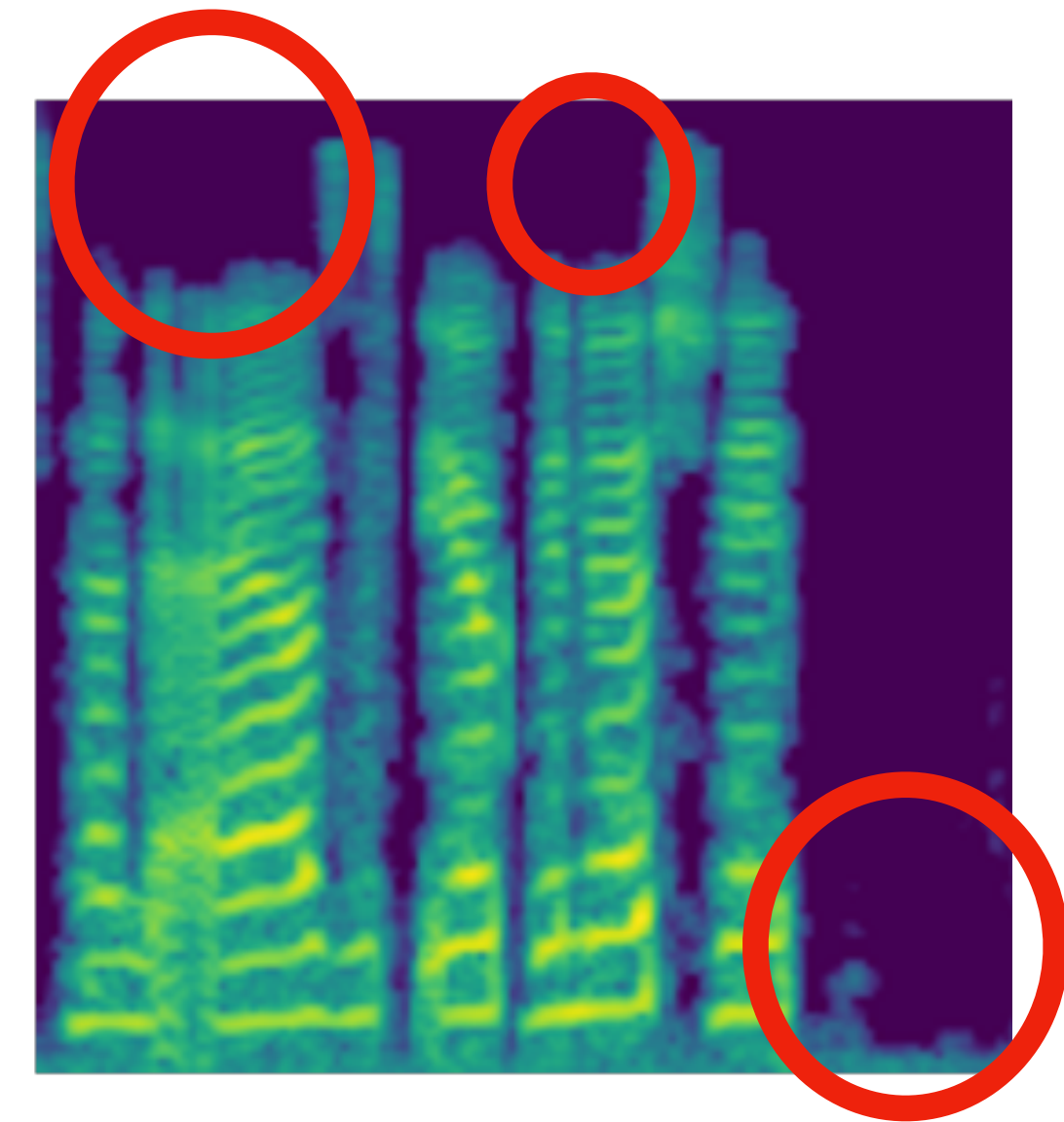
Problem: Reducing network size introduces artifacts



Mixture



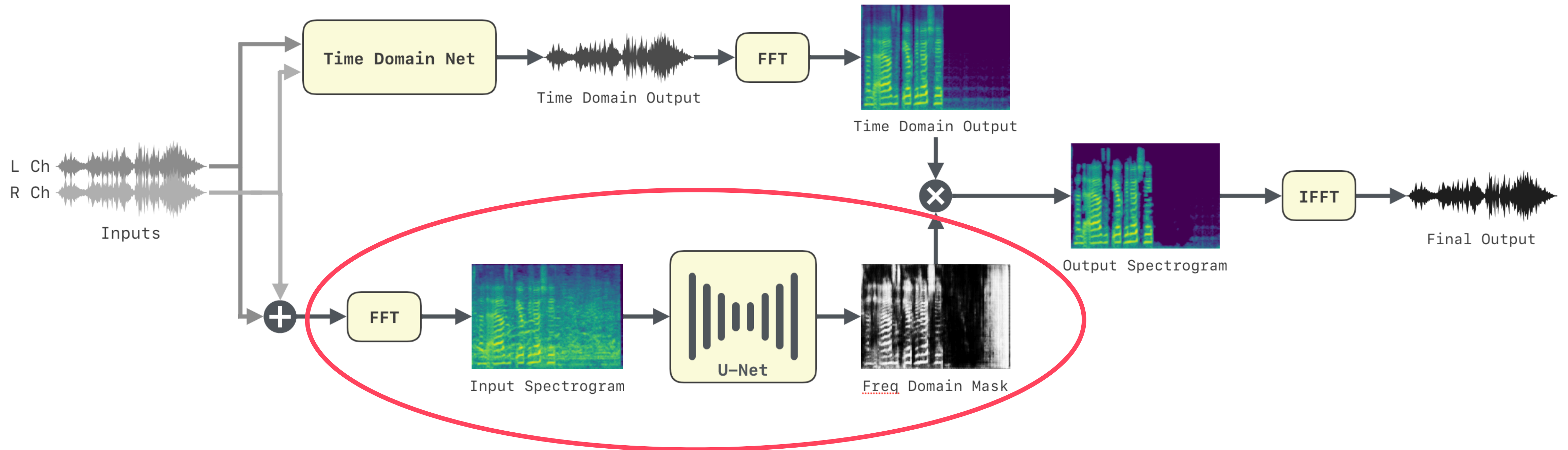
Time Domain Network



Ground Truth

**How do we reduce the artifacts without adding
too much extra computation?**

Our solution: Cascade network that masks artifacts



- Lightweight Spectrogram Network to reduce artifacts
- Full network - 21 ms on a iPhone 12 Pro

Results

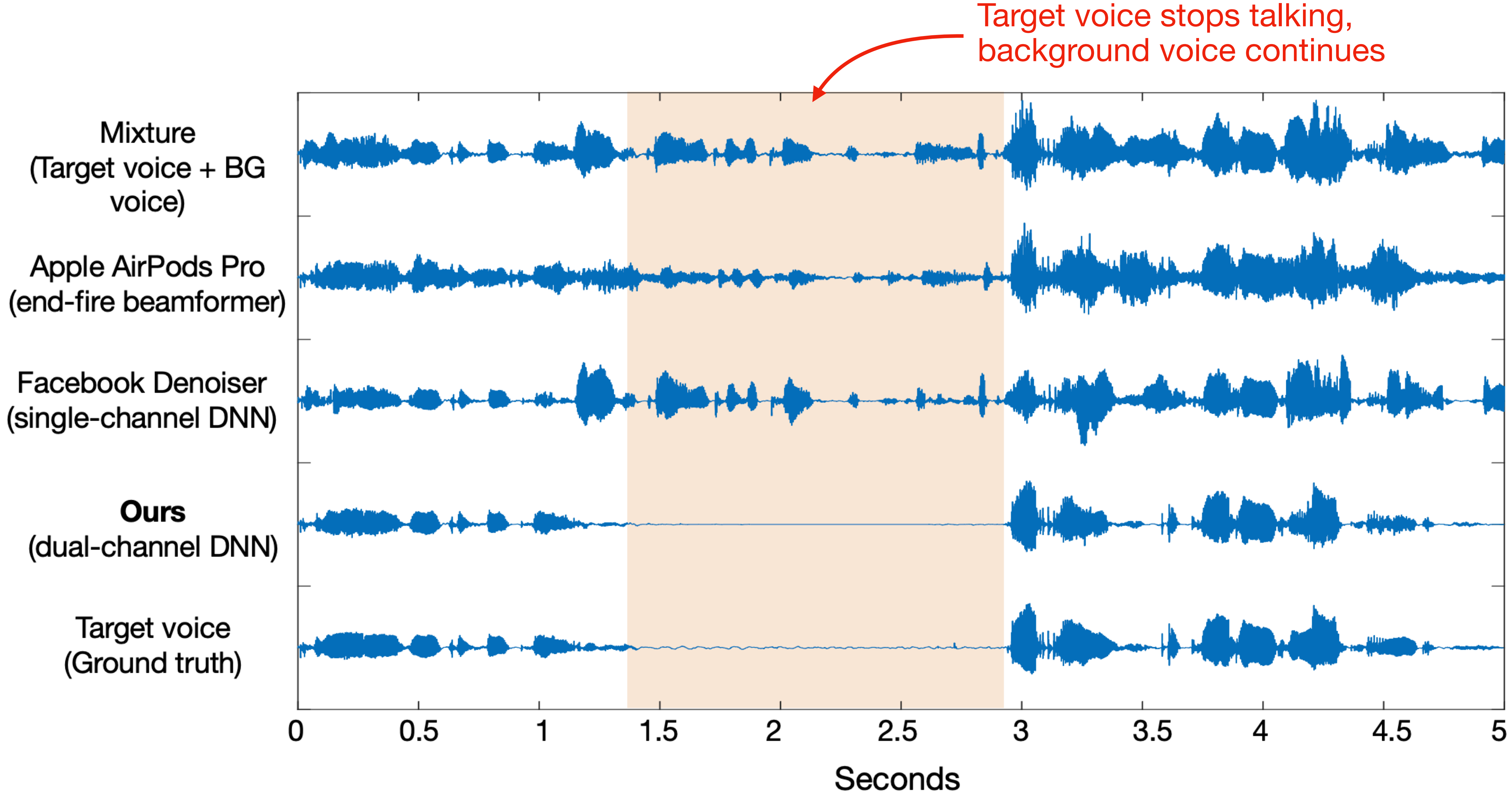
ClearBuds Demo: Comparison between

Noisy Environment

ClearBuds Noise Suppression

AirPods Pro Noise Suppresion

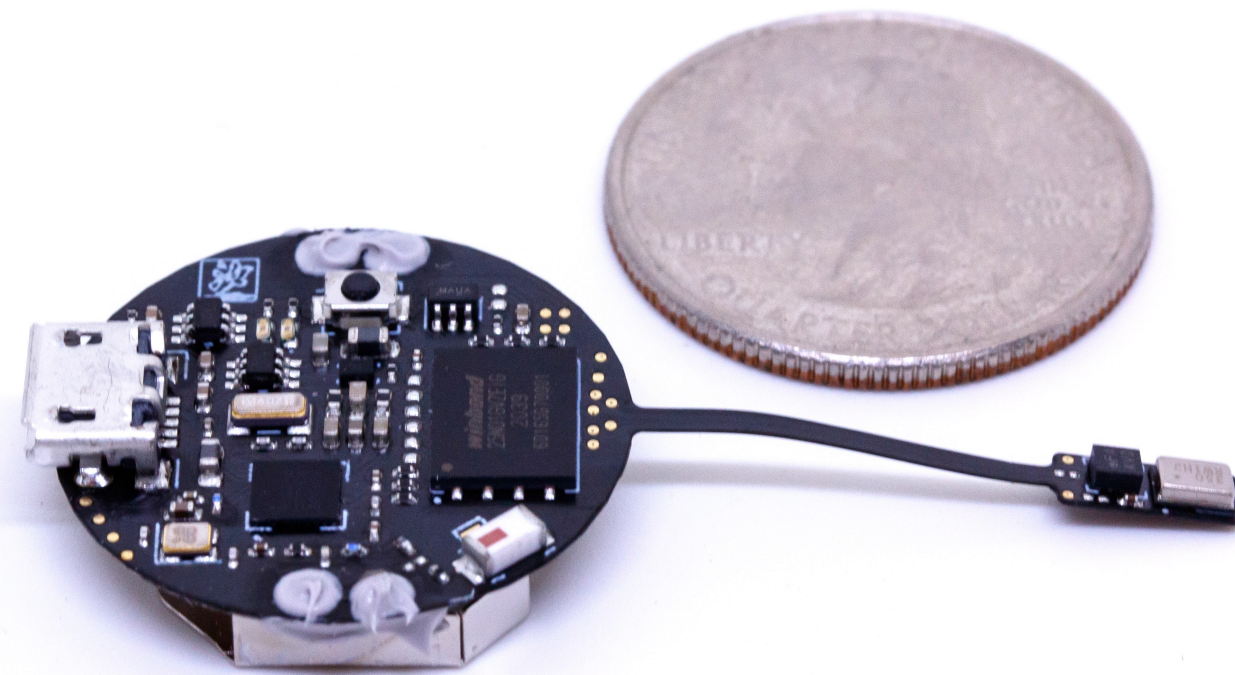
Performance Against Other Systems



Conclusion

- **First wireless earbuds that stream binaural synchronized microphone data**
- **Novel light-weight hybrid neural network that performs speech enhancement in real-time on a mobile device**
- **Extensive benchmarking and user studies that demonstrate generalization to real-world environments**

Brief Q&A



IRIS

Wireless Ring for Vision-based
Smart Home Interaction

Maruchi Kim*, **Antonio Glenn***, **Bandhav Veluri***

Yunseo Lee, Eyoel Gebre, Aditya Bagaria

Shwetak Patel, Shyamnath Gollakota

University of Washington



Smart home adoption is on the rise

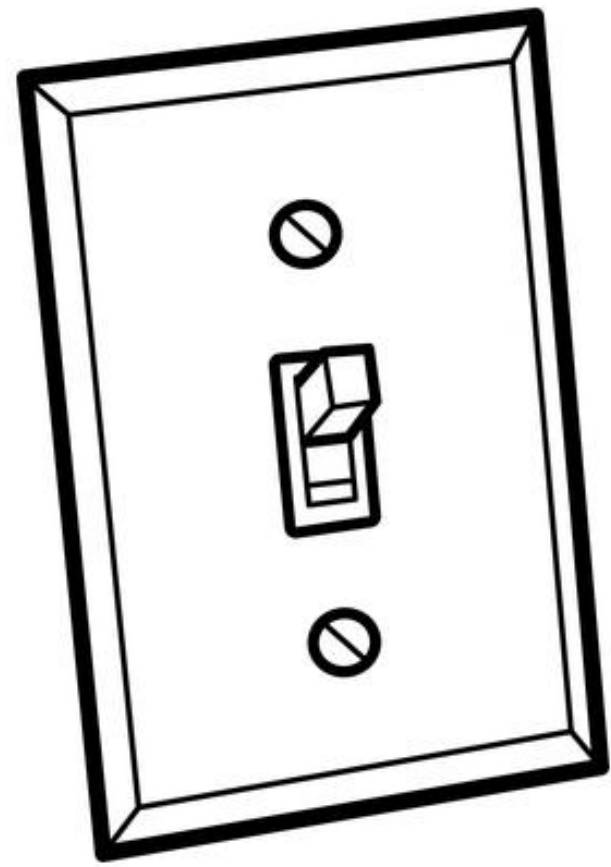






IRIS: Wireless ring for vision-based smart home interaction

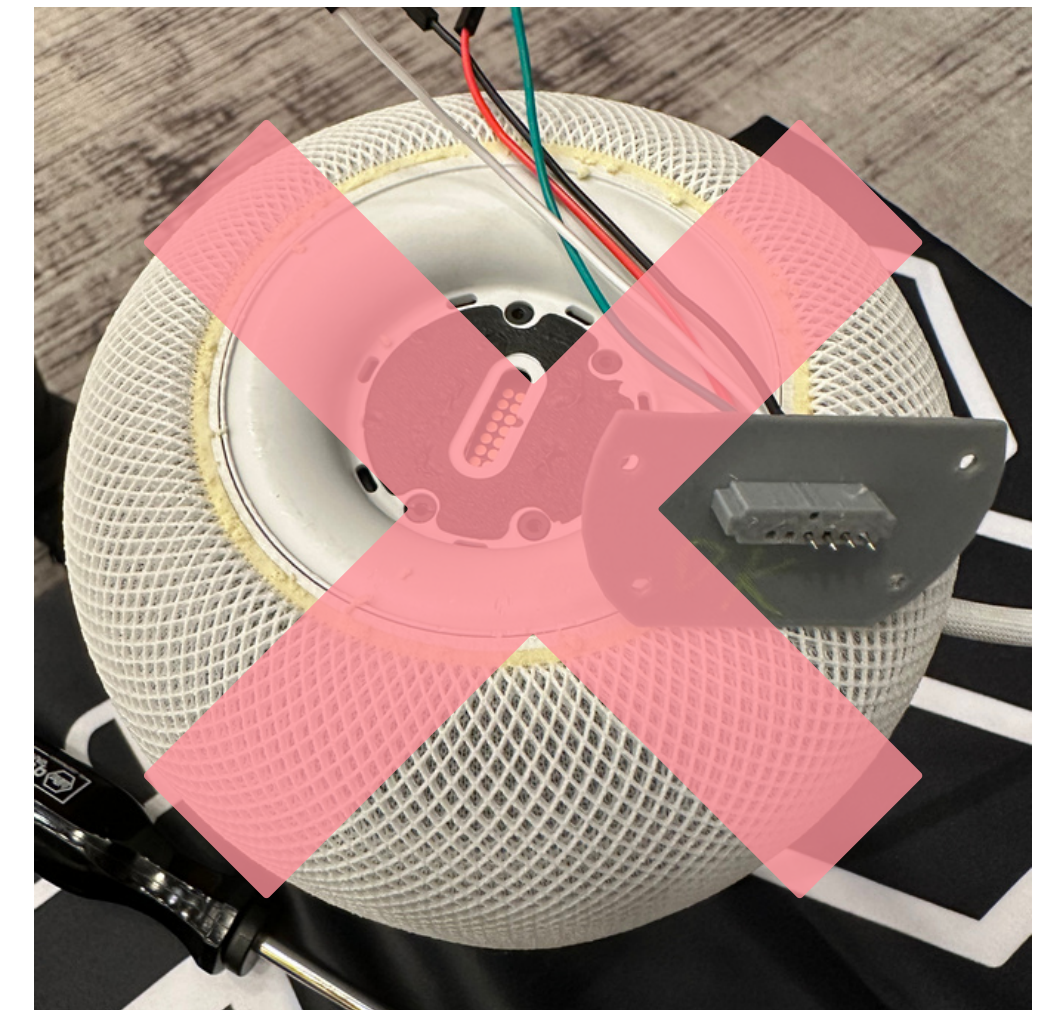
Design Goals



Simple and intuitive



Ubiquitous and context-aware

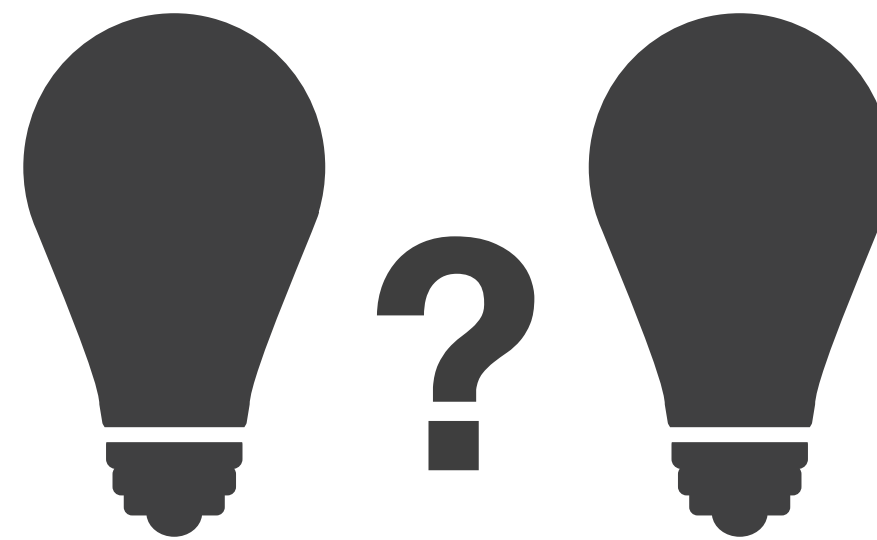


Zero modifications

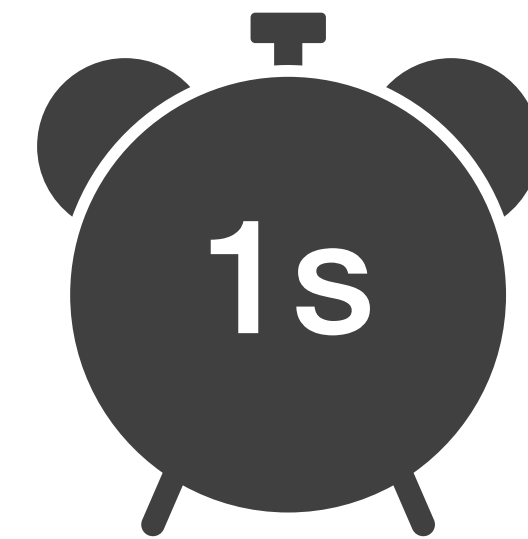
What are the key technical challenges?



A wireless, vision-enabled ring with all-day battery life



Instance-based object detection

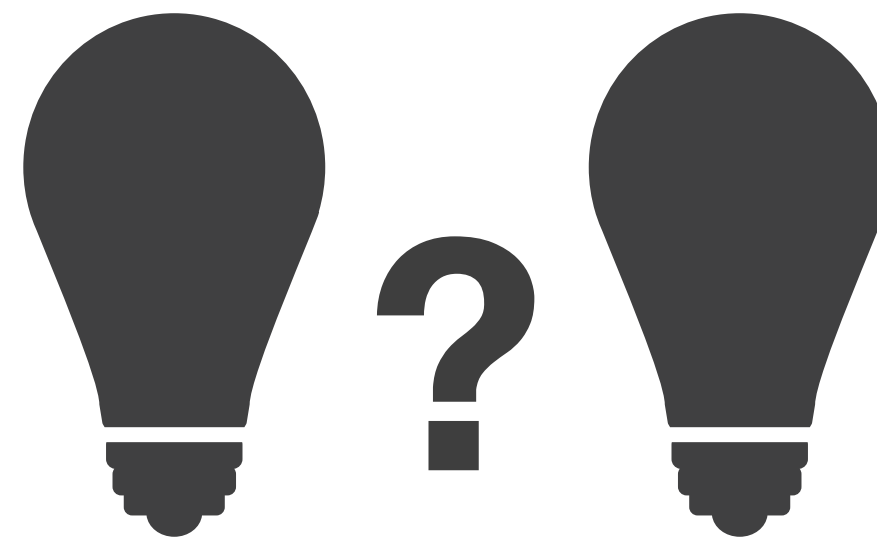


Real-time operation, single second latency

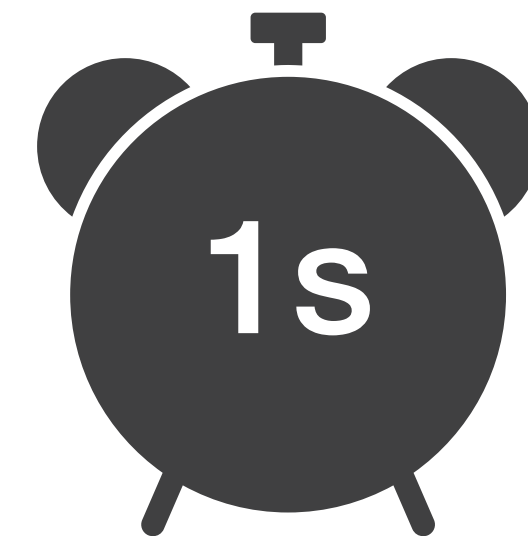
What are the key technical challenges?



A wireless, vision-enabled ring with all-day battery life



Instance-based object detection

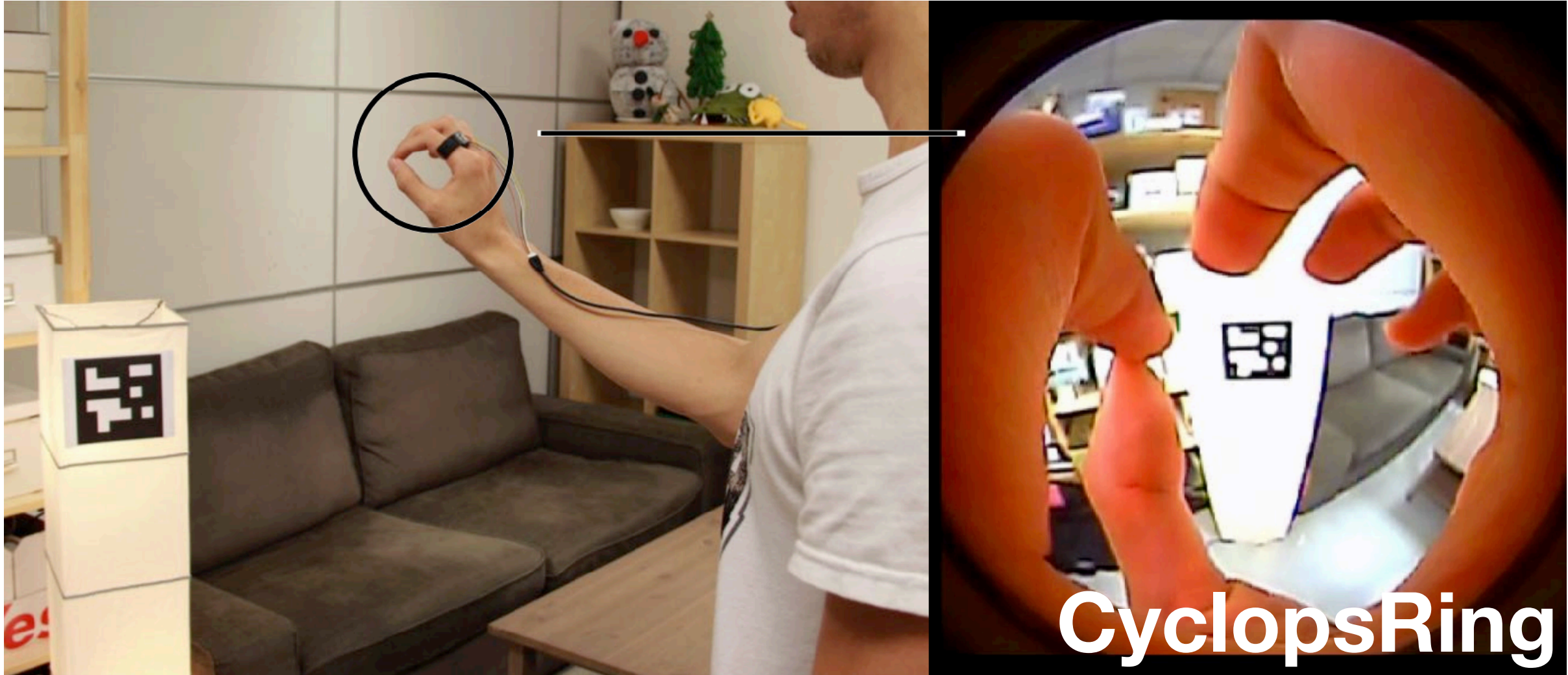


Real-time operation, single second latency

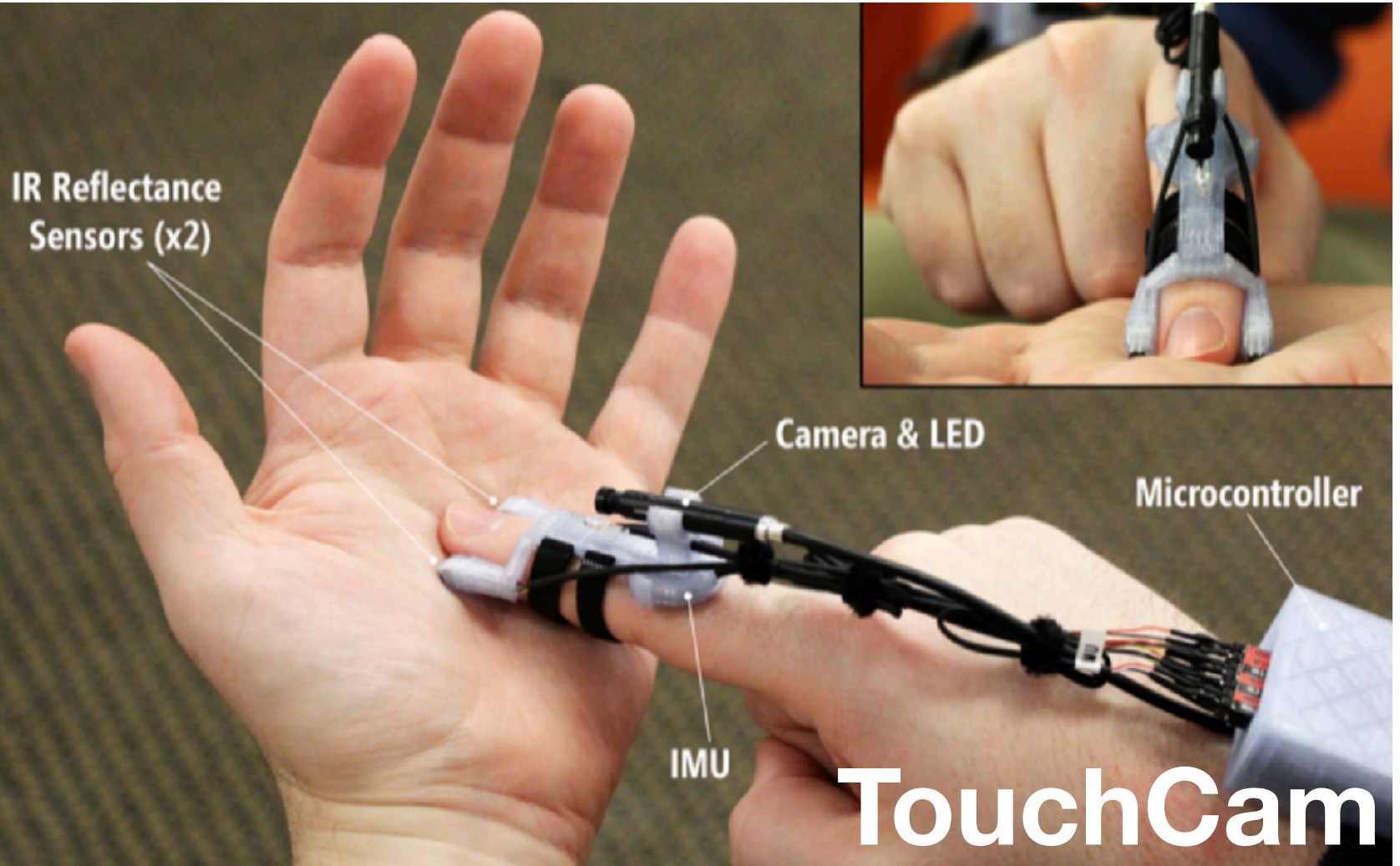
No existing camera-enabled ring that is wireless and/or meets SWaP requirements expected in this form-factor



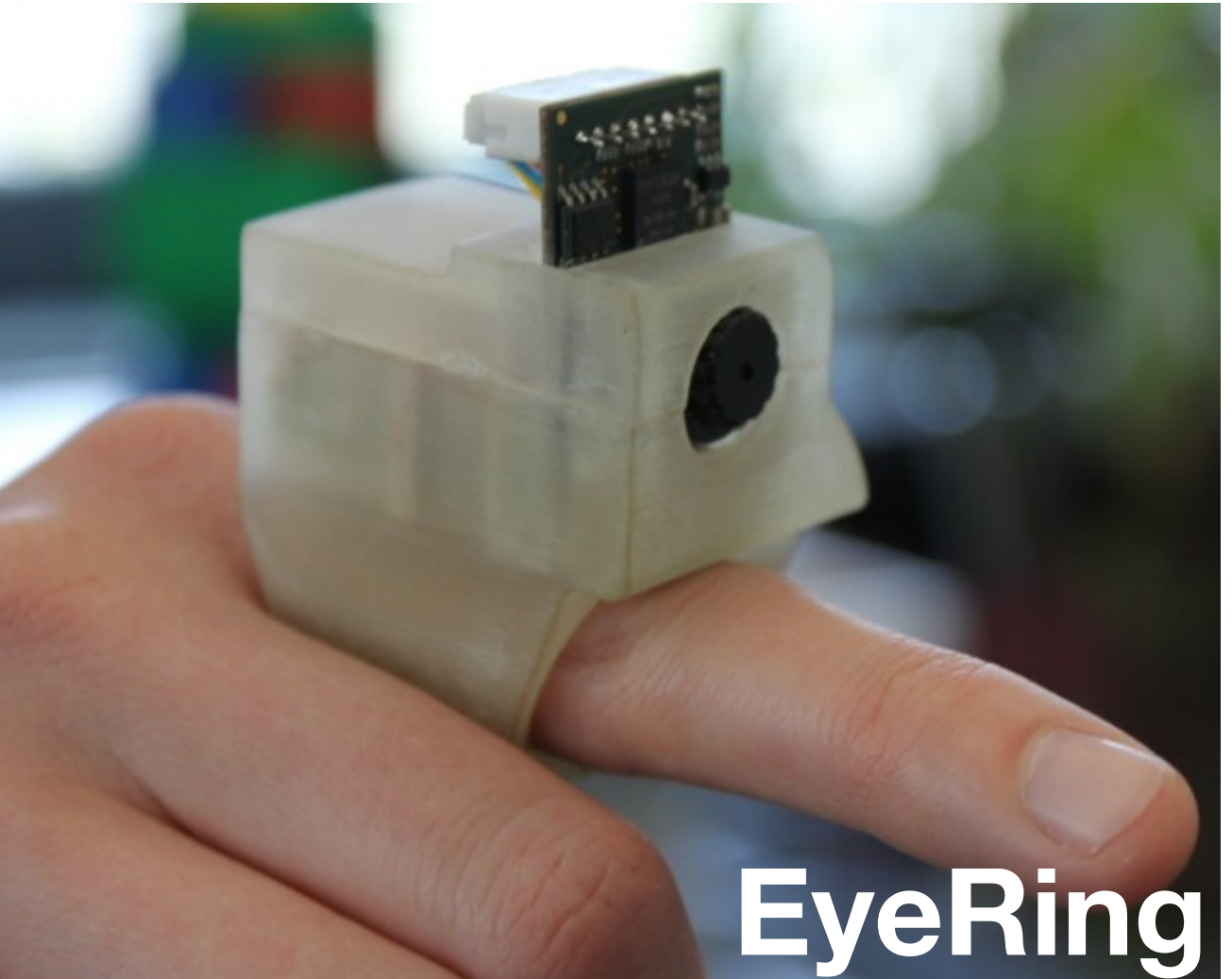
FingerReader



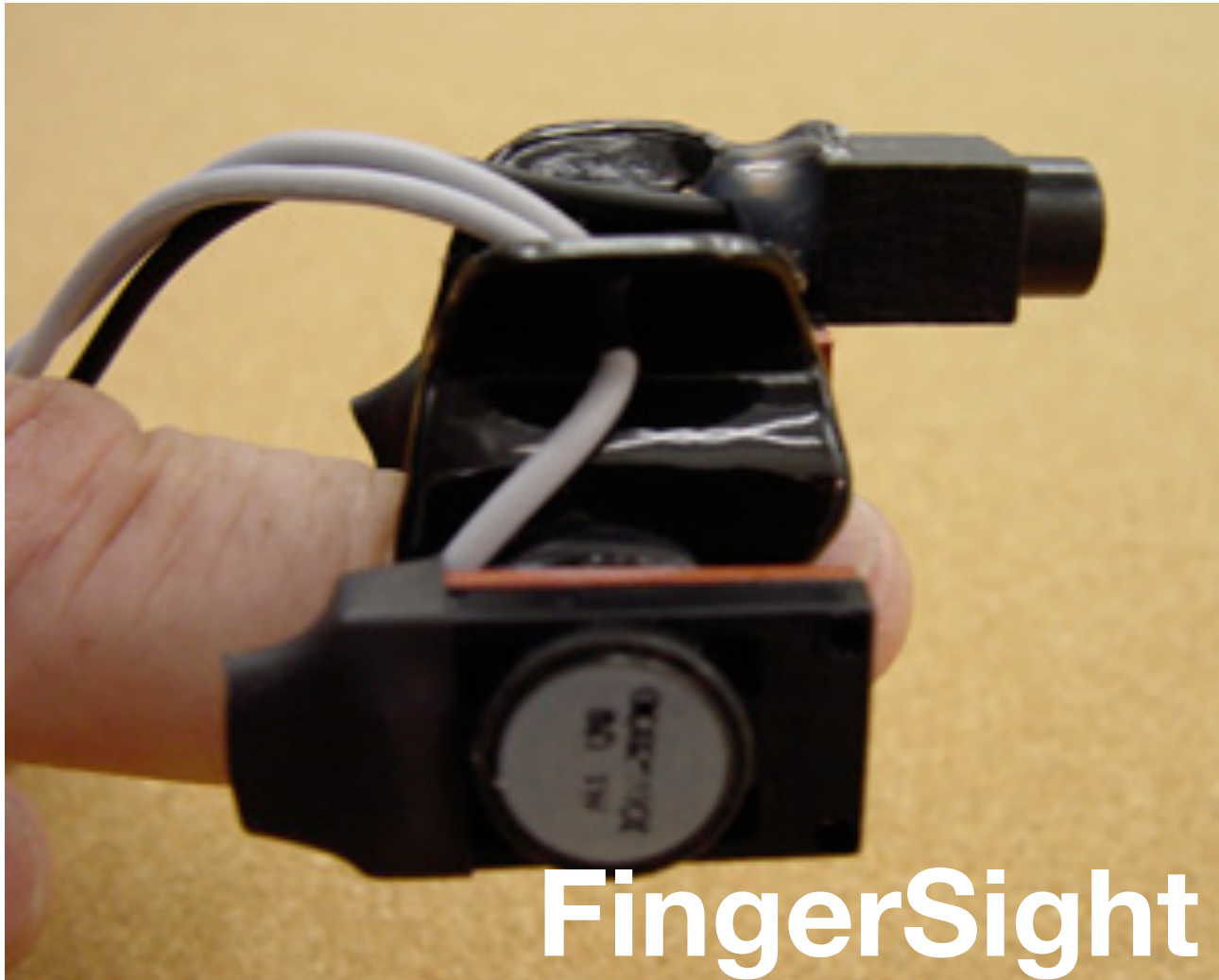
CyclopsRing



TouchCam

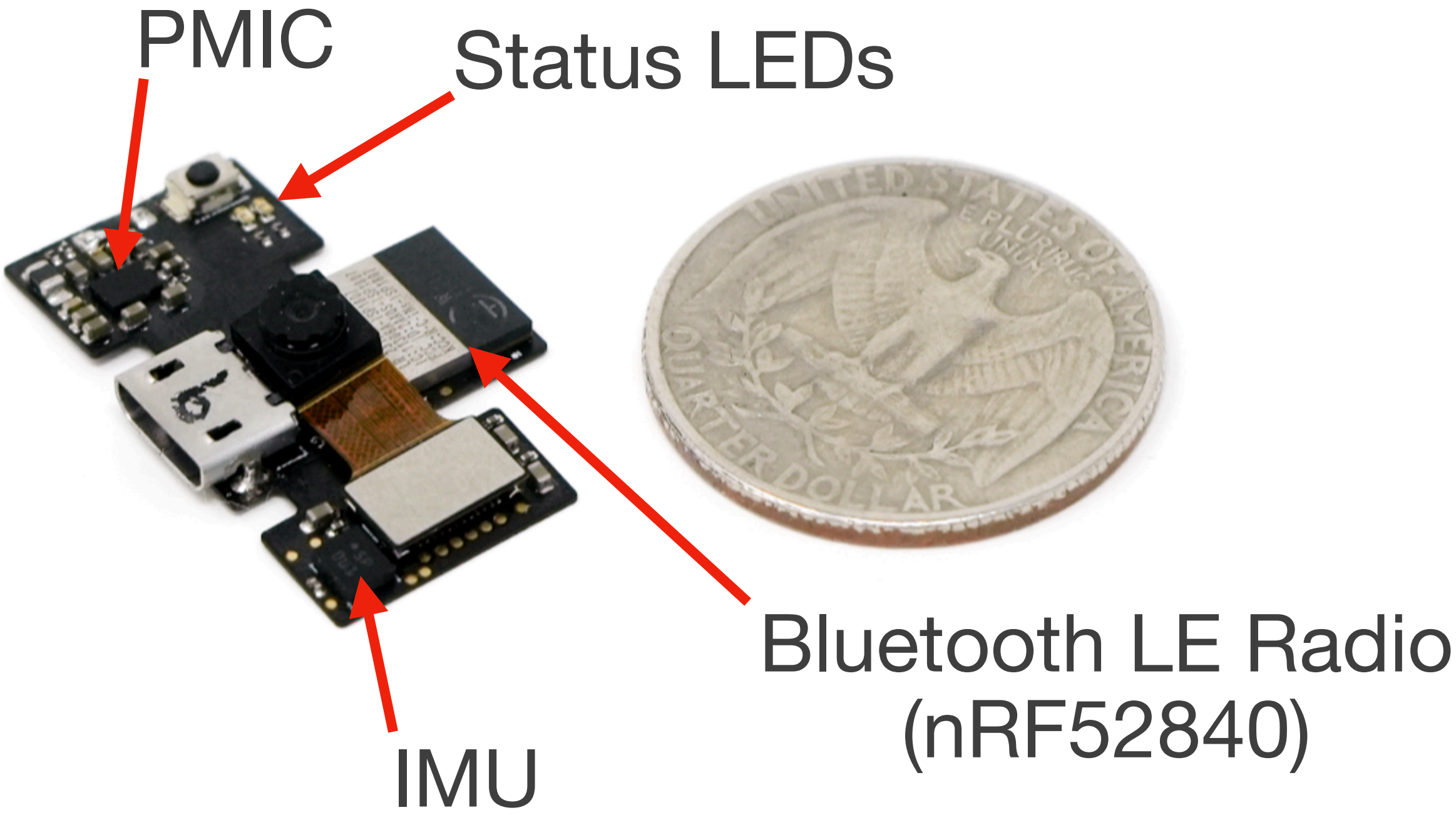
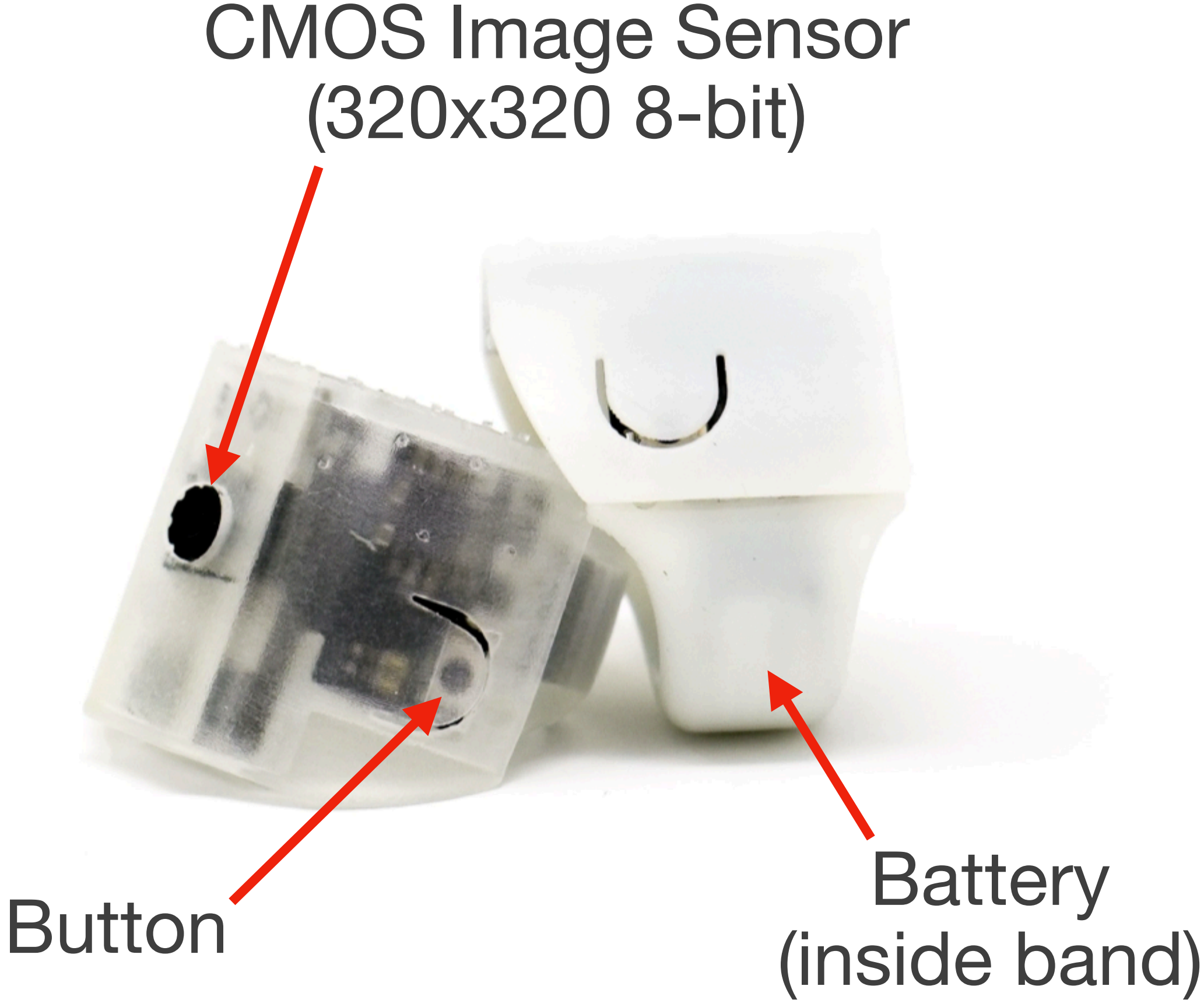


EyeRing



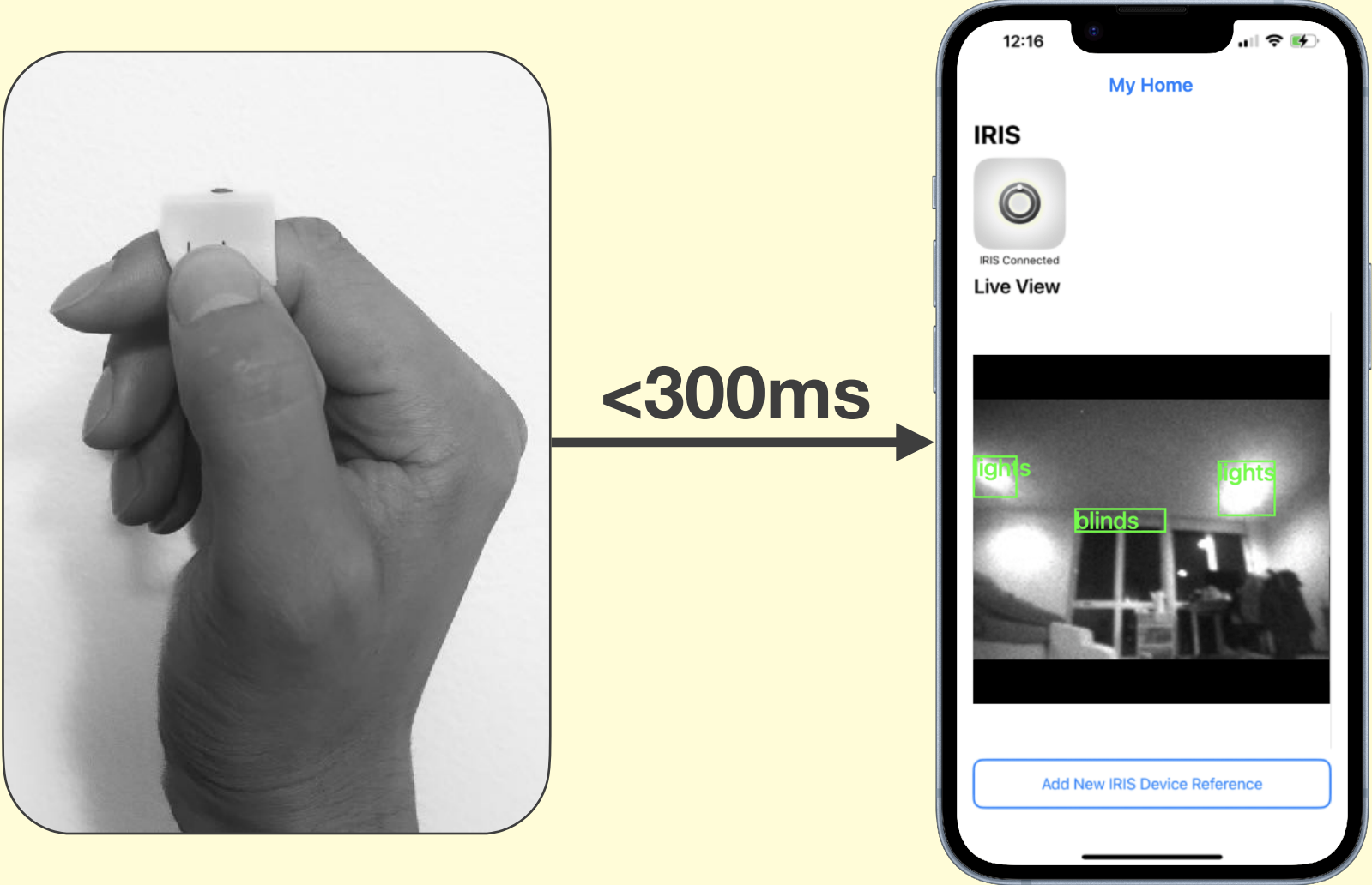
FingerSight

IRIS Hardware

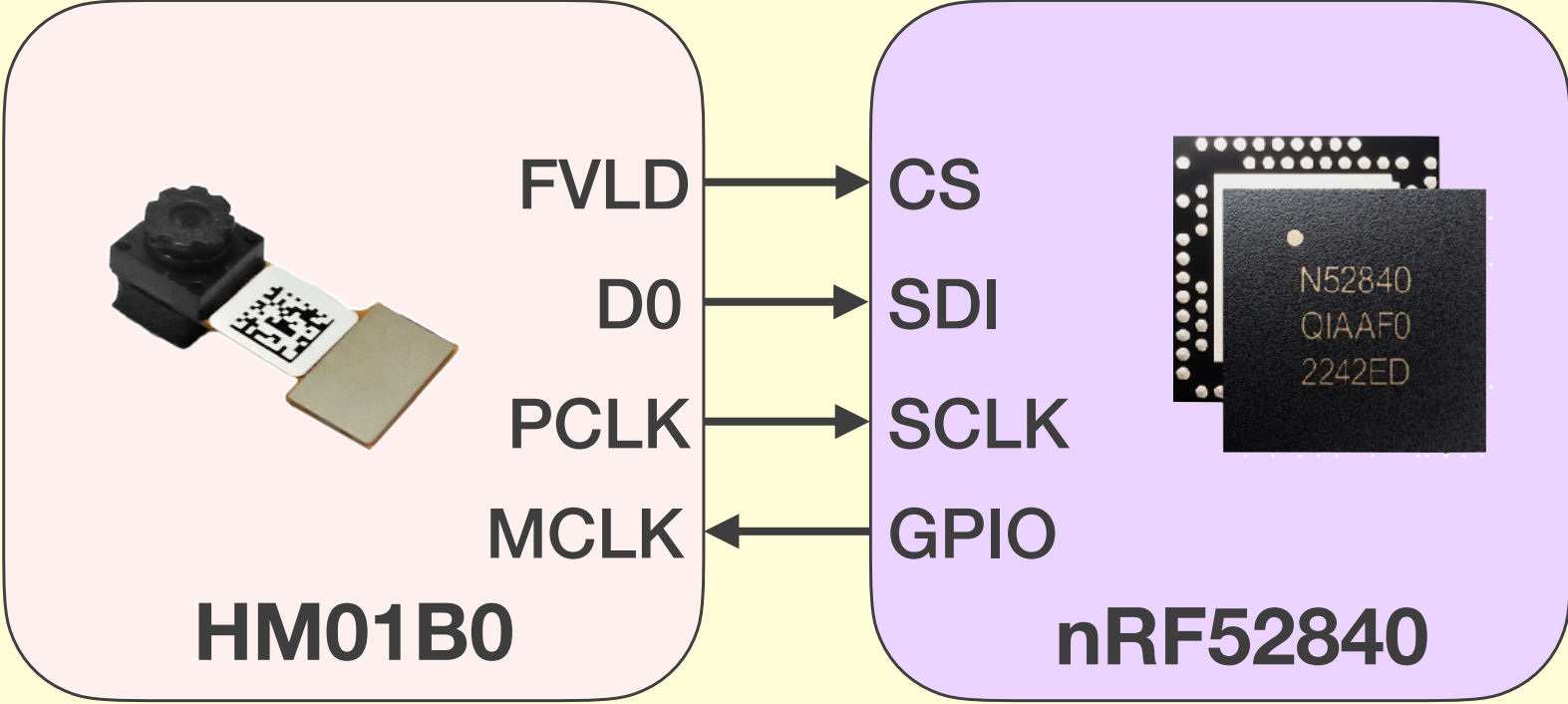


IRIS Hardware

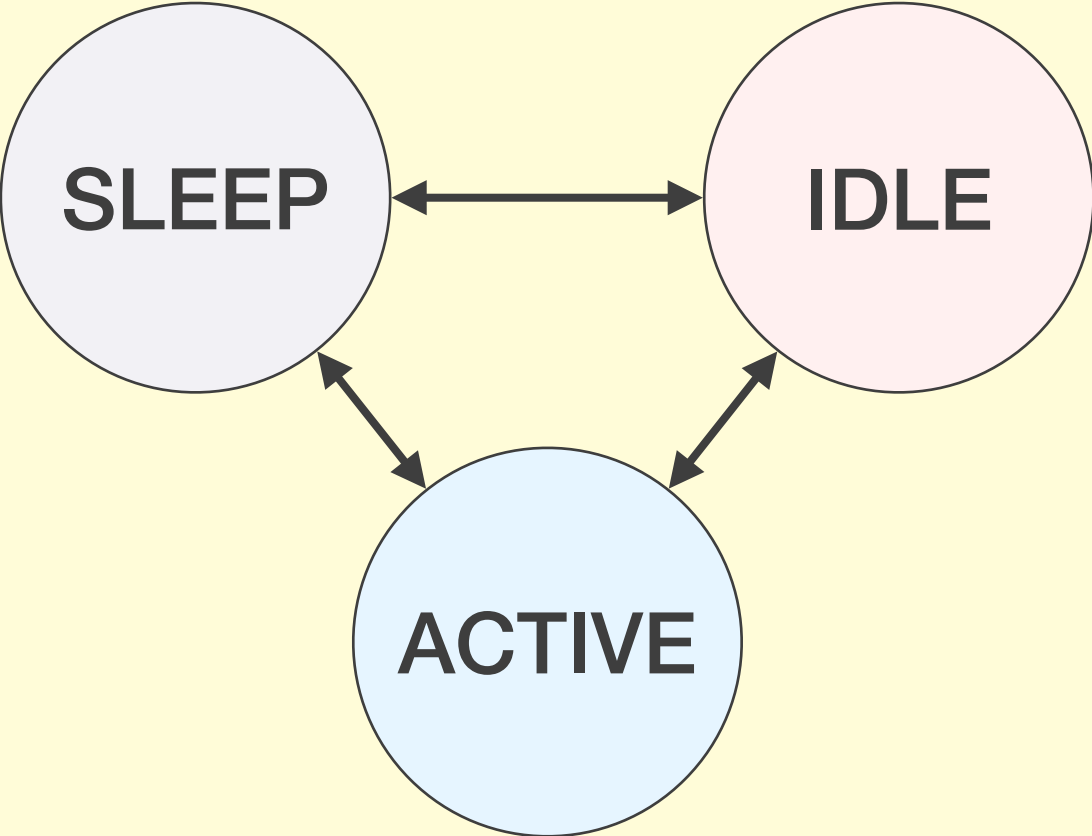
Low Latency



CMOS Integration over SPI



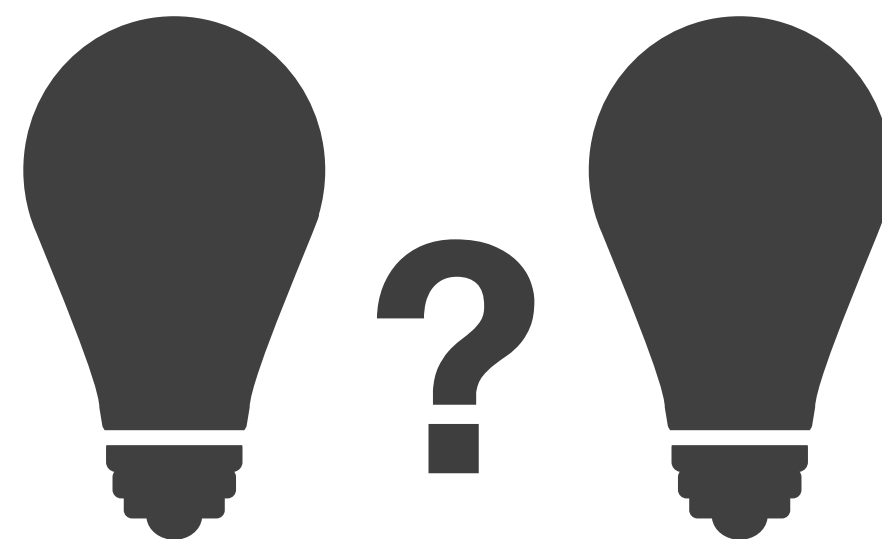
Low-power State Machine



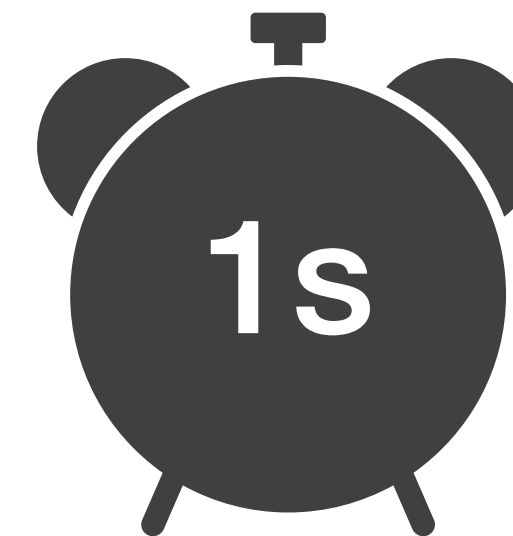
What are the key technical challenges?



A wireless, vision-enabled ring with all-day battery life



Instance-based object detection

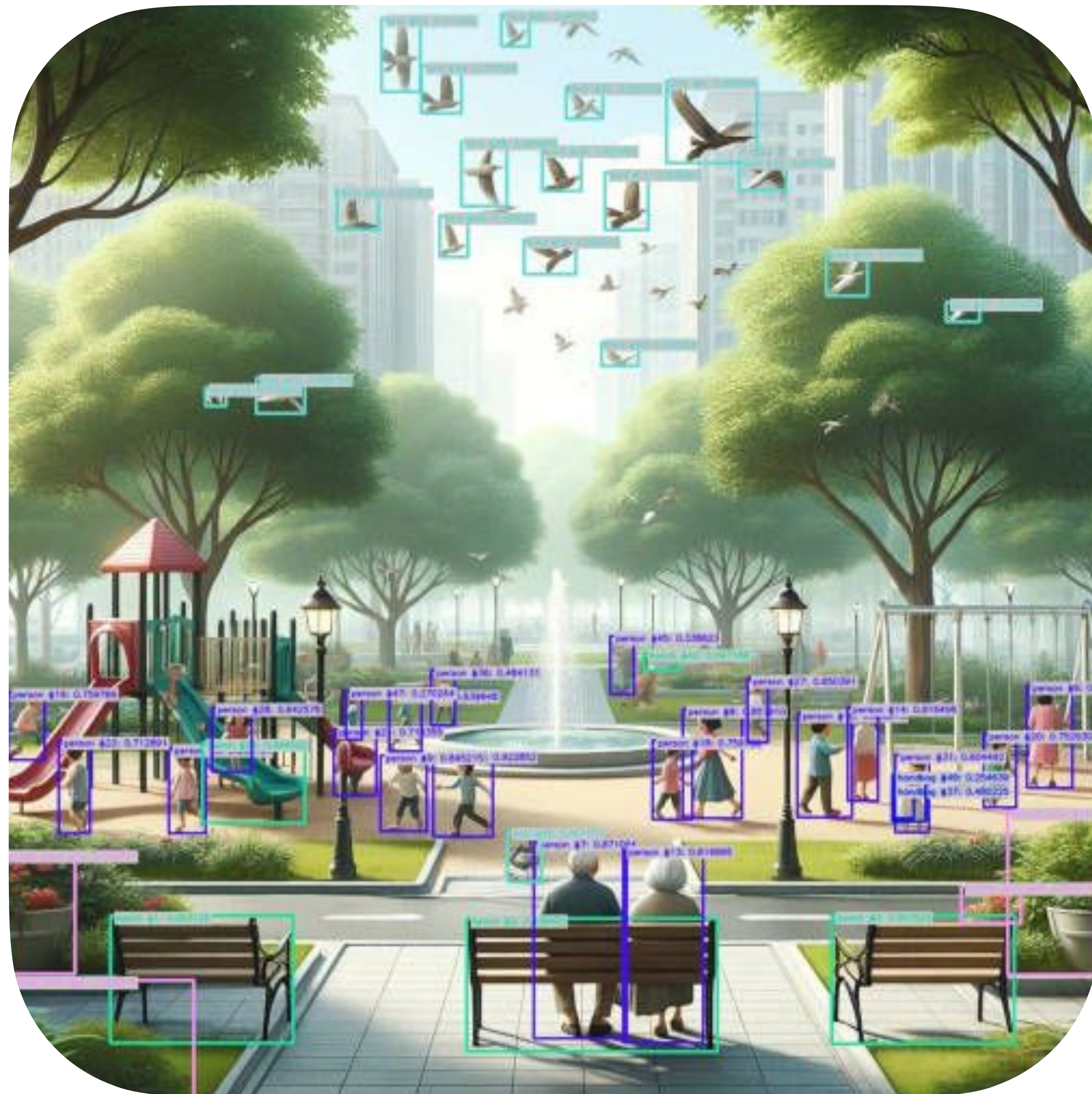


Real-time operation, single second latency

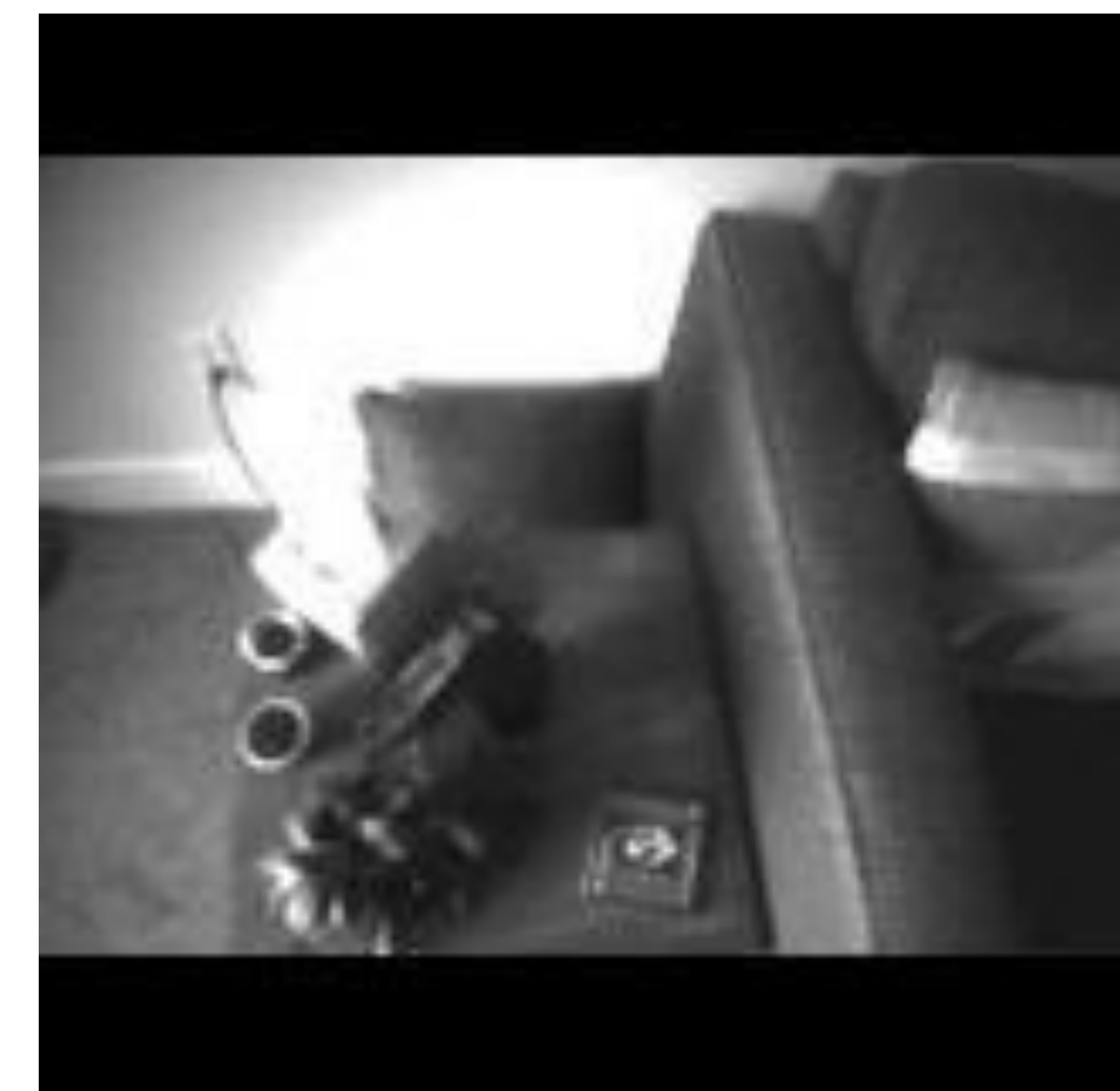
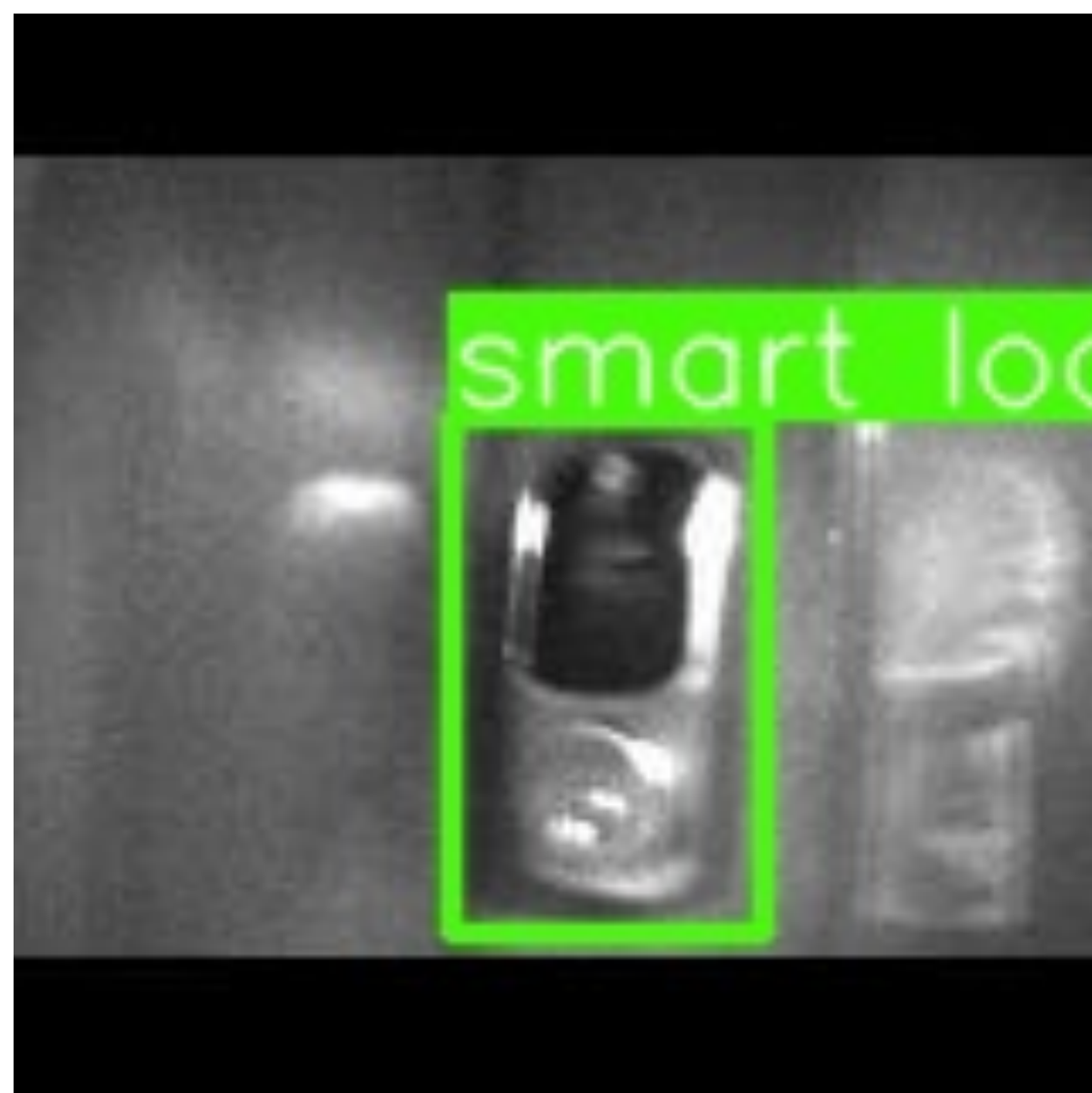
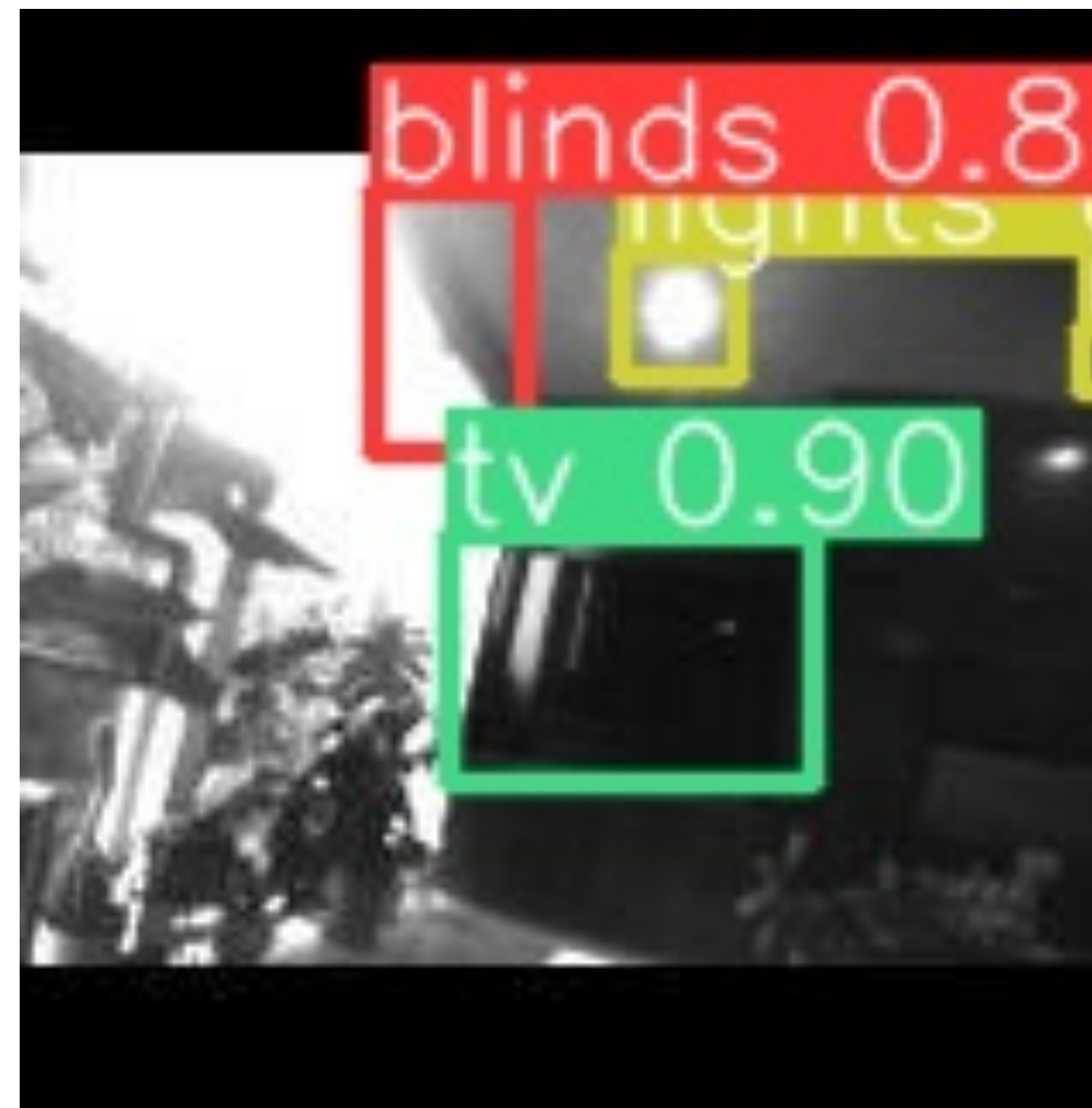
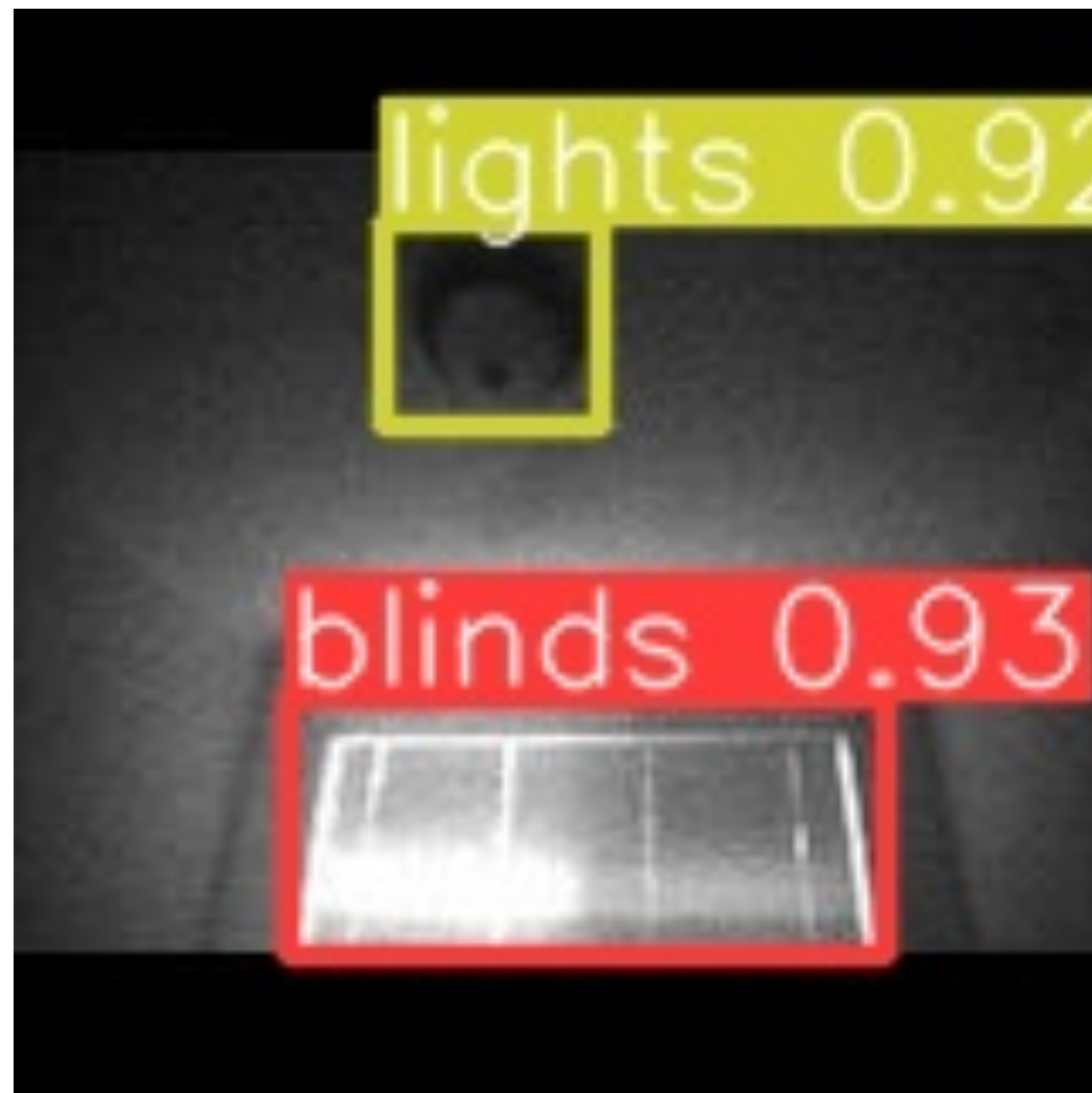


+


















Fine-tuned YOLO for smart home interaction



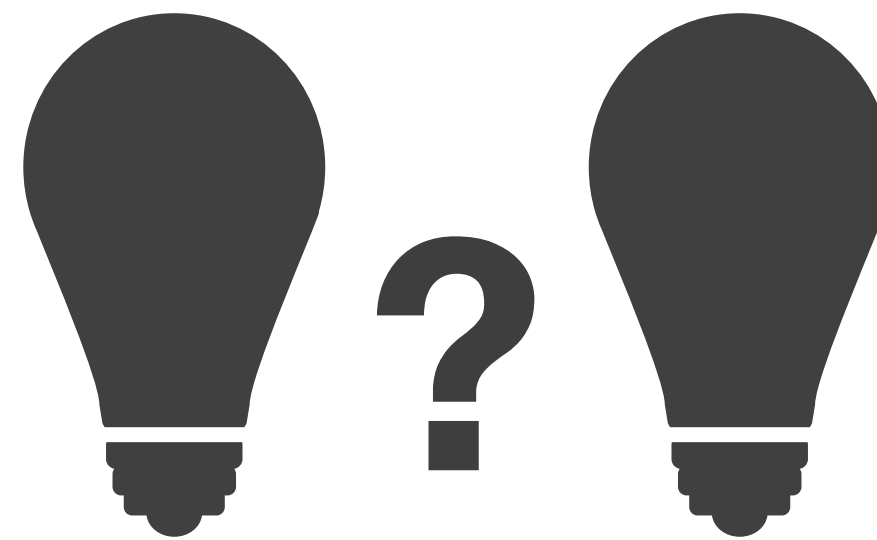
DINOv2 queries based on semantic similarity

Query	Reference images sorted in decreasing order of similarity						
							
							
							
1	2	3	4	5	6	7	

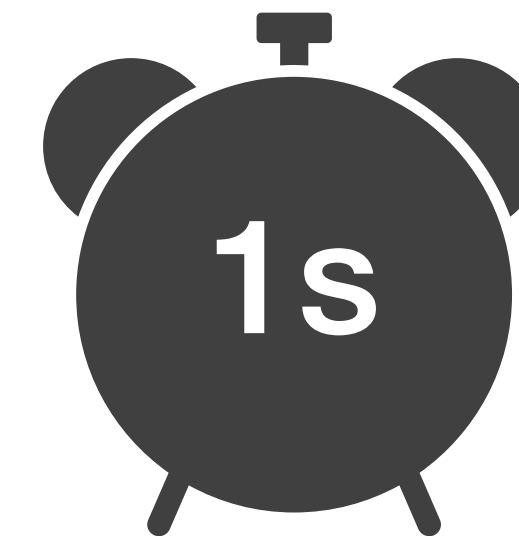
What are the key technical challenges?



A wireless, vision-enabled ring with all-day battery life

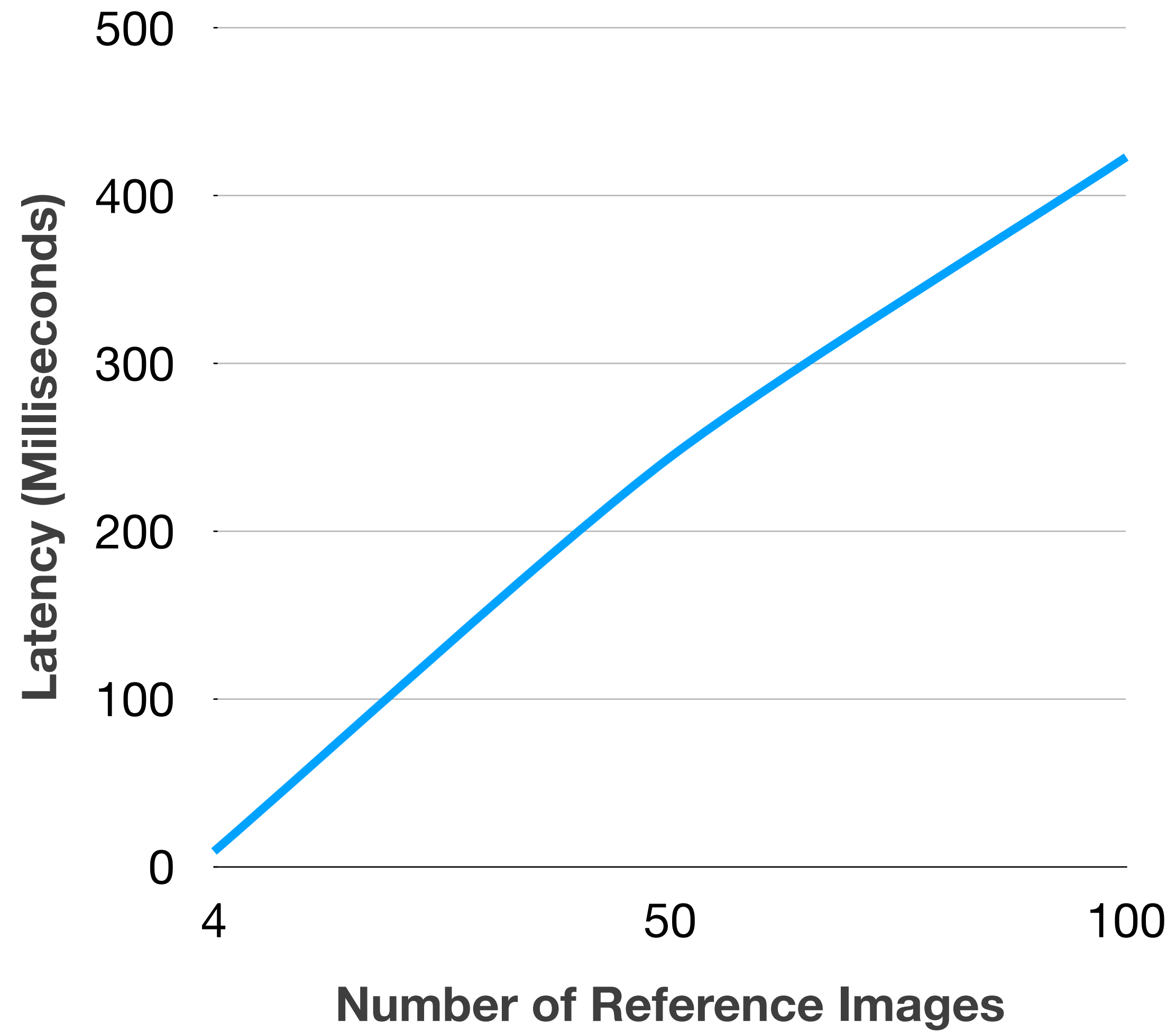


Instance-based object detection

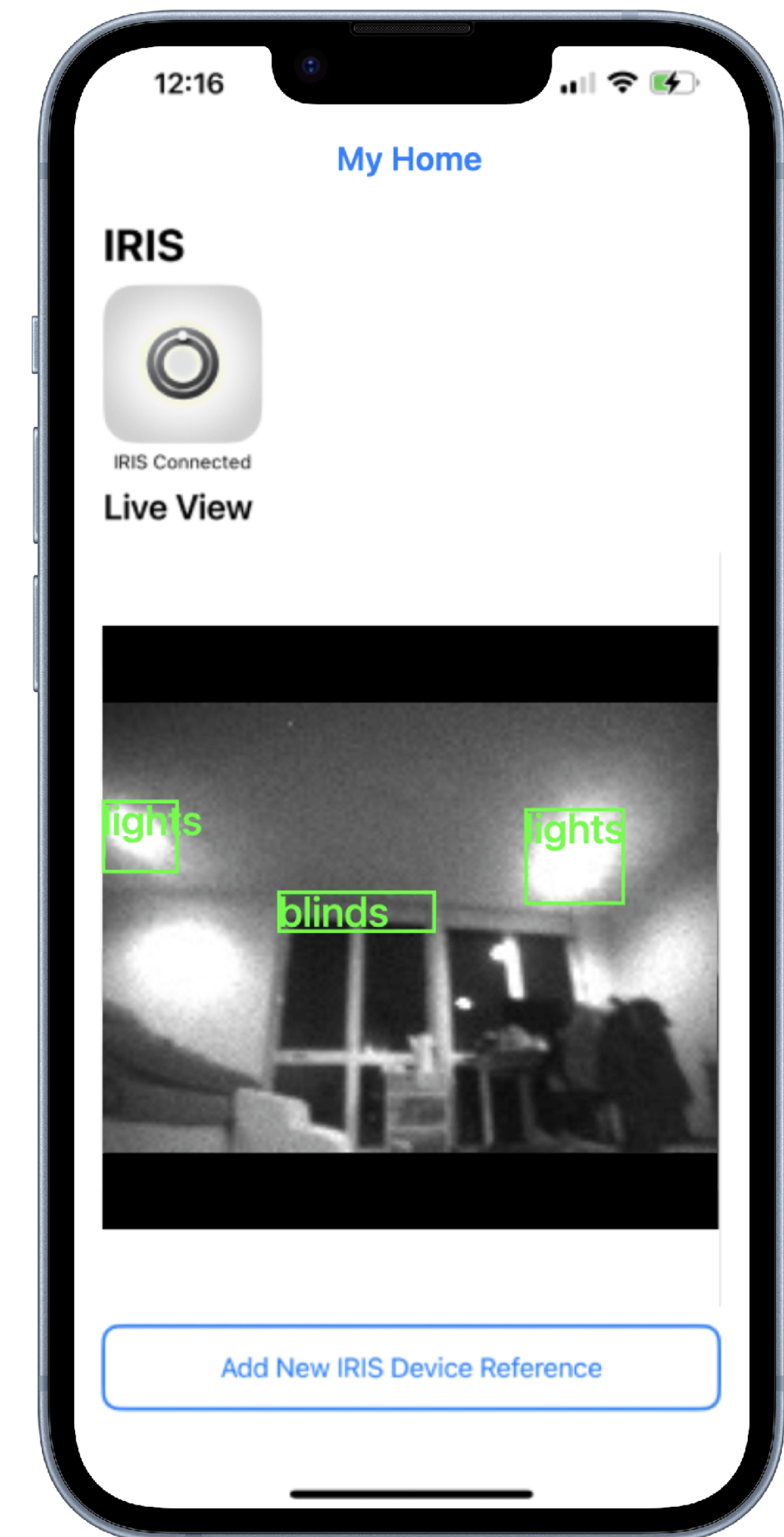
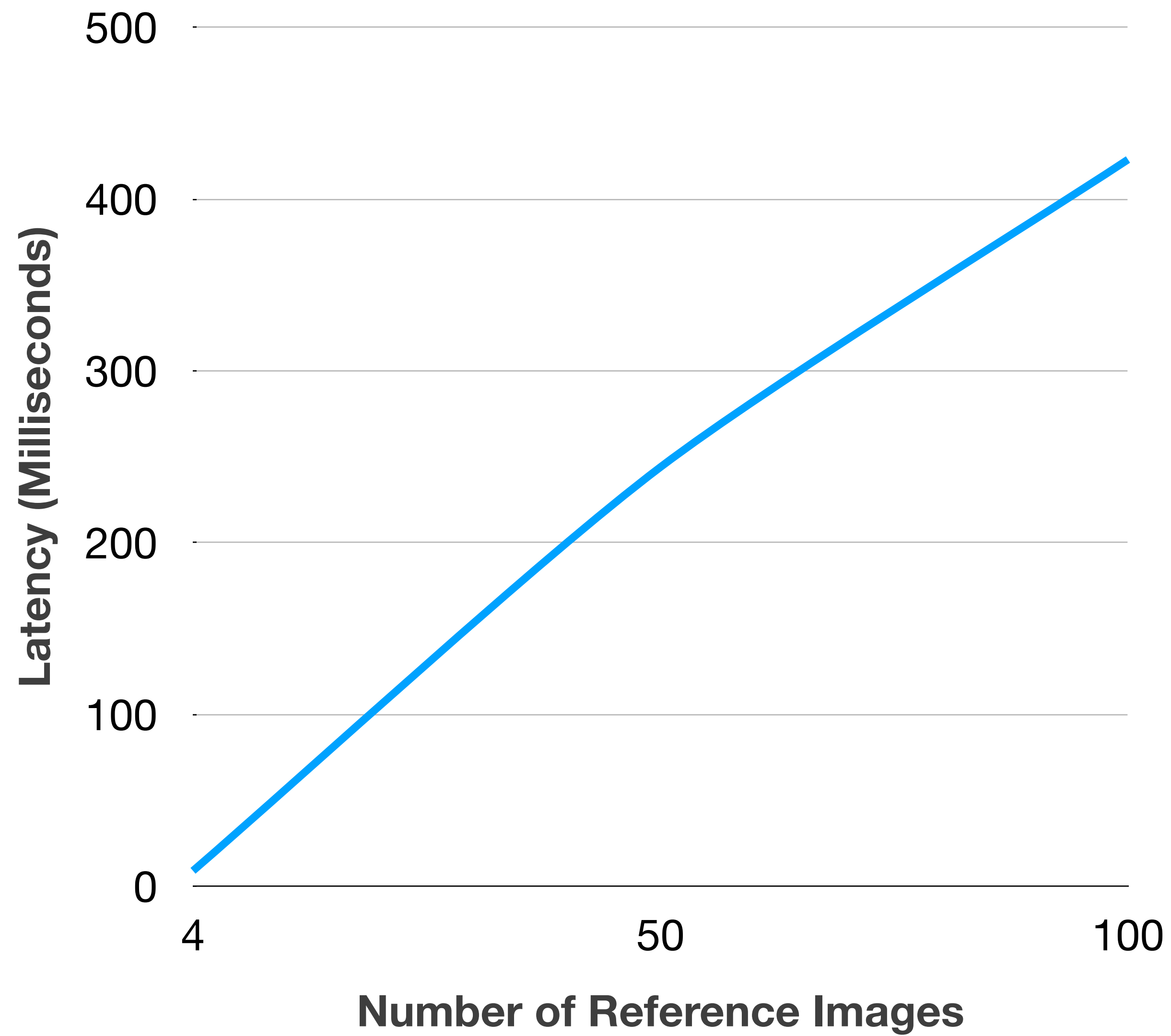


Real-time operation, single second latency

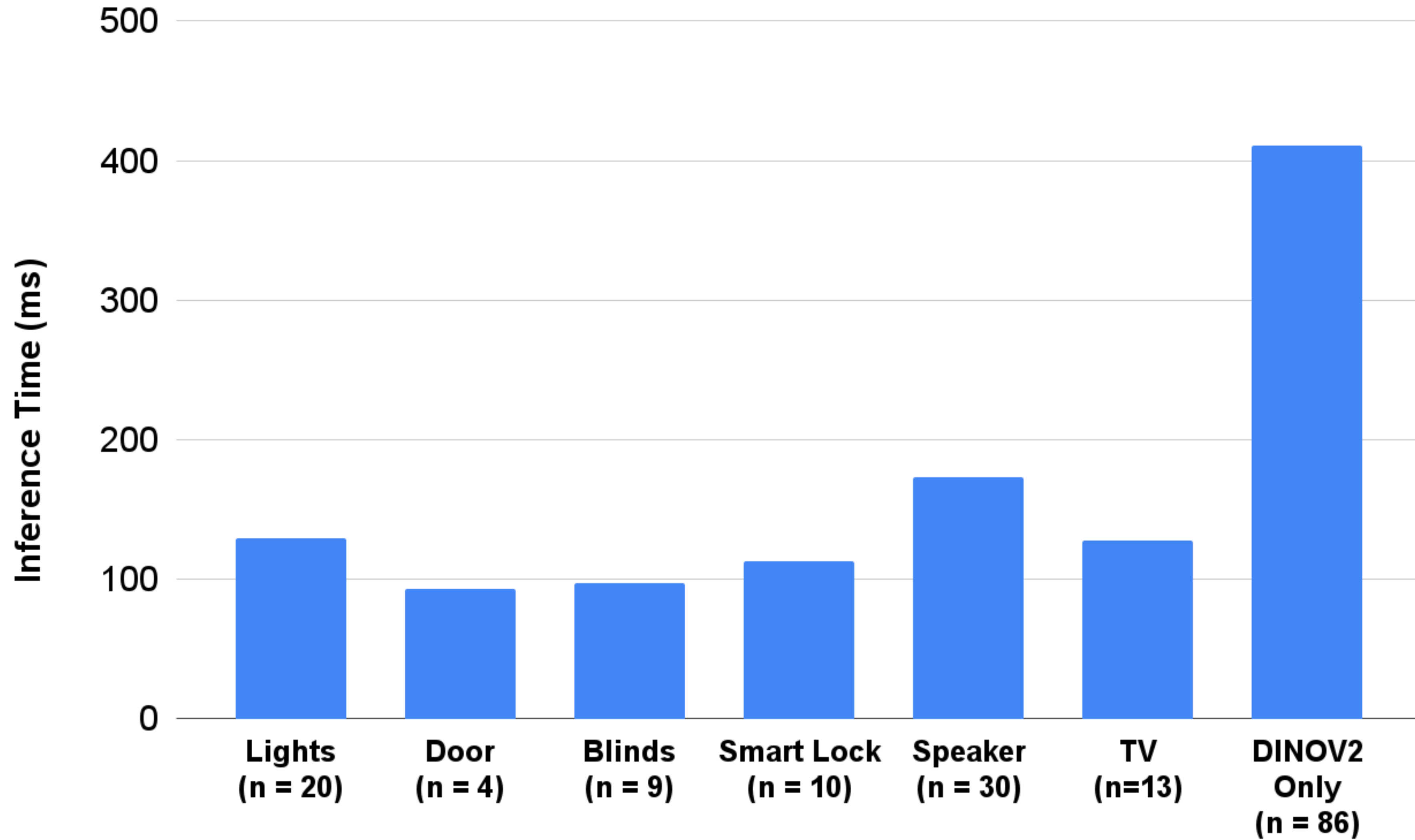
DINO Inference Latency



DINO Inference Latency



DINO+YOLO Inference Latency

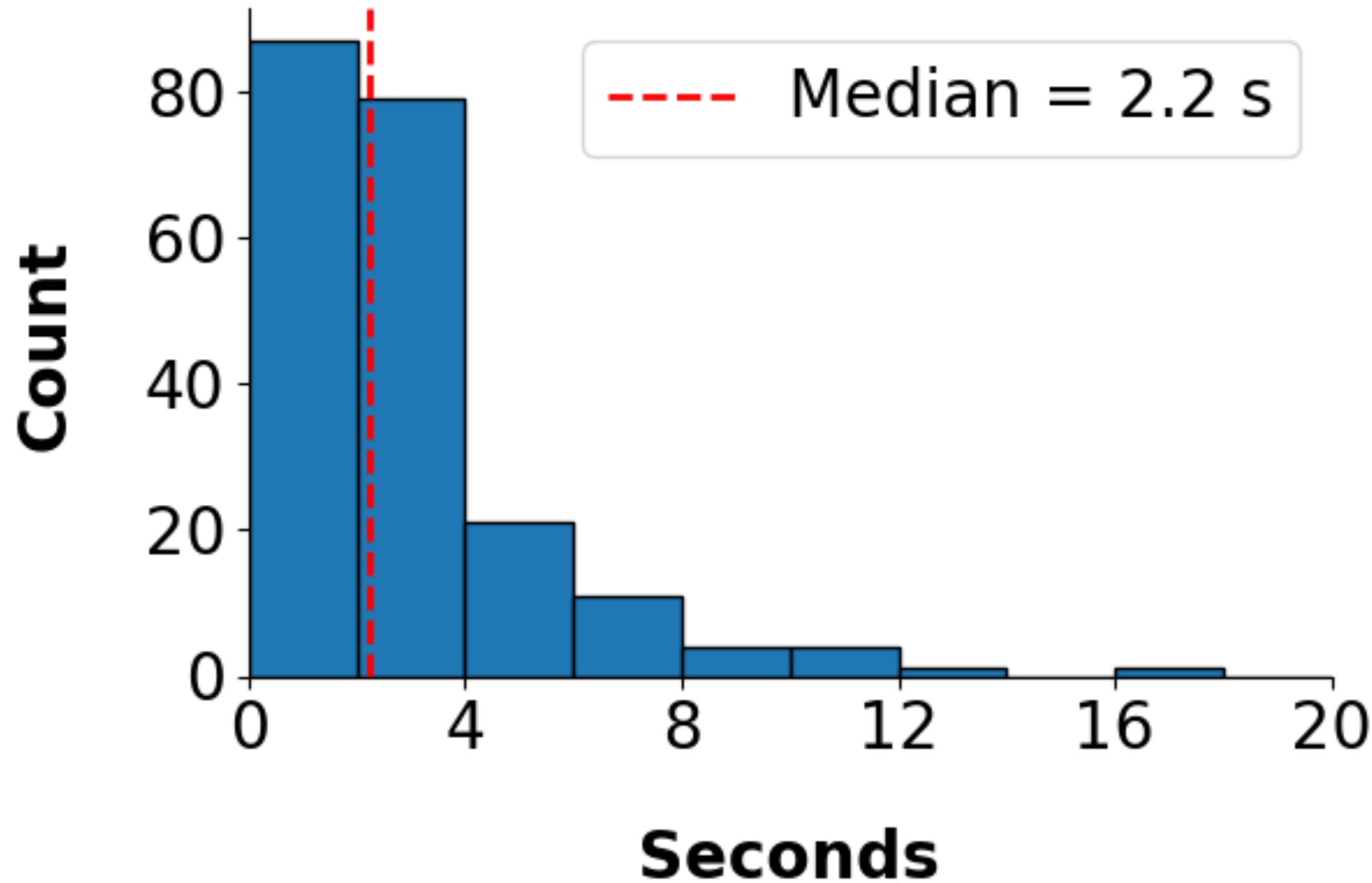


User Study

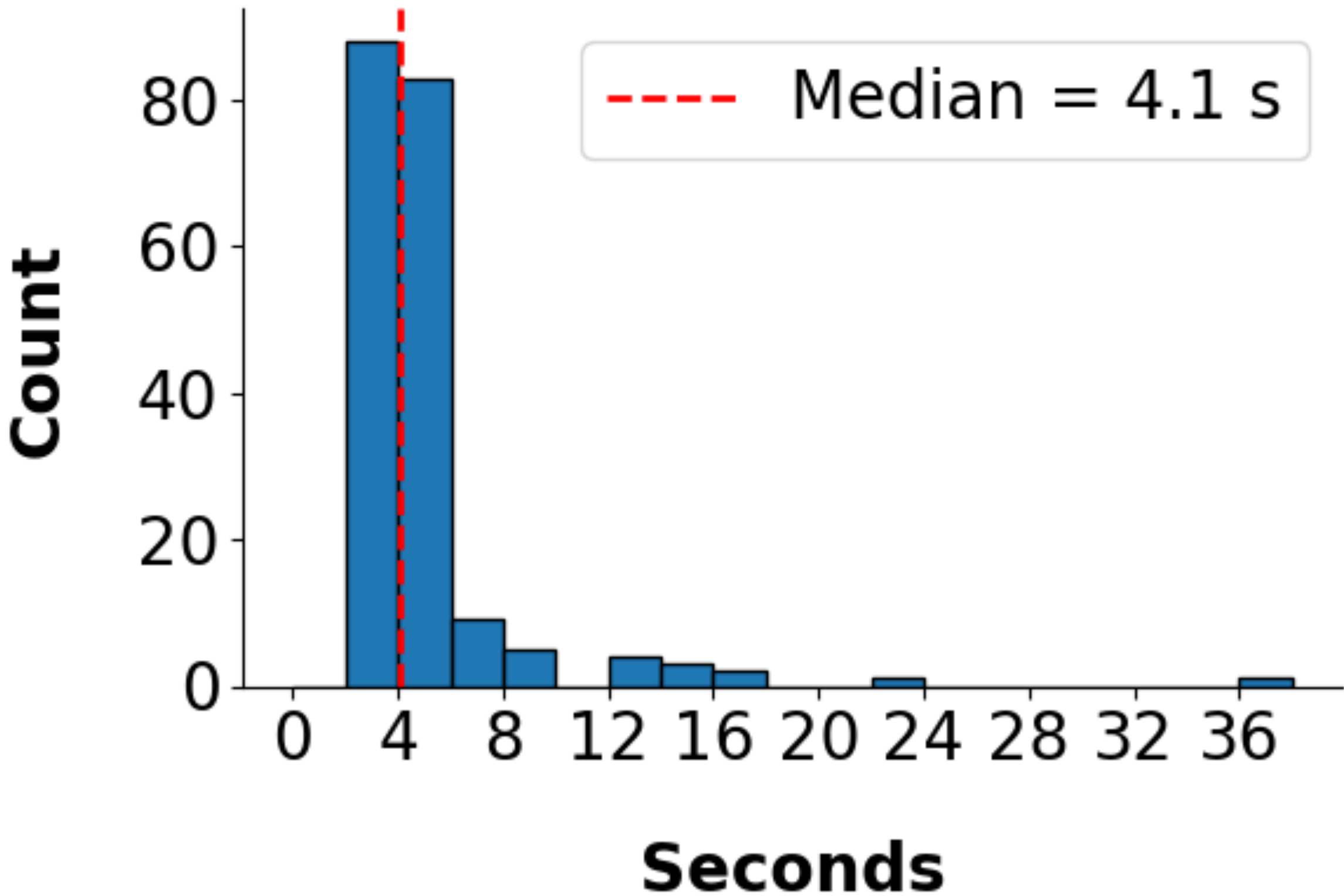


User Input to Device Output

IRIS

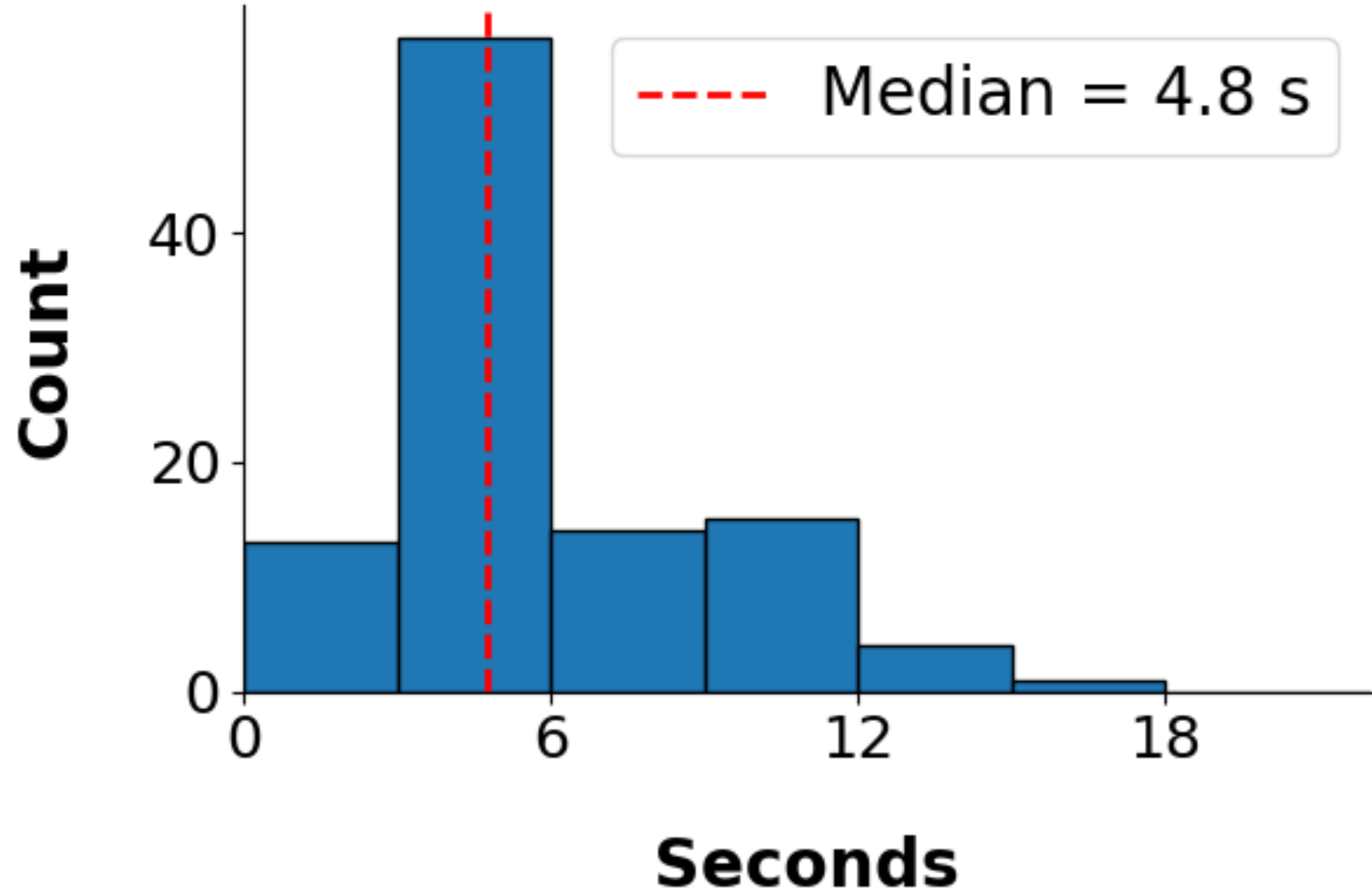


Voice

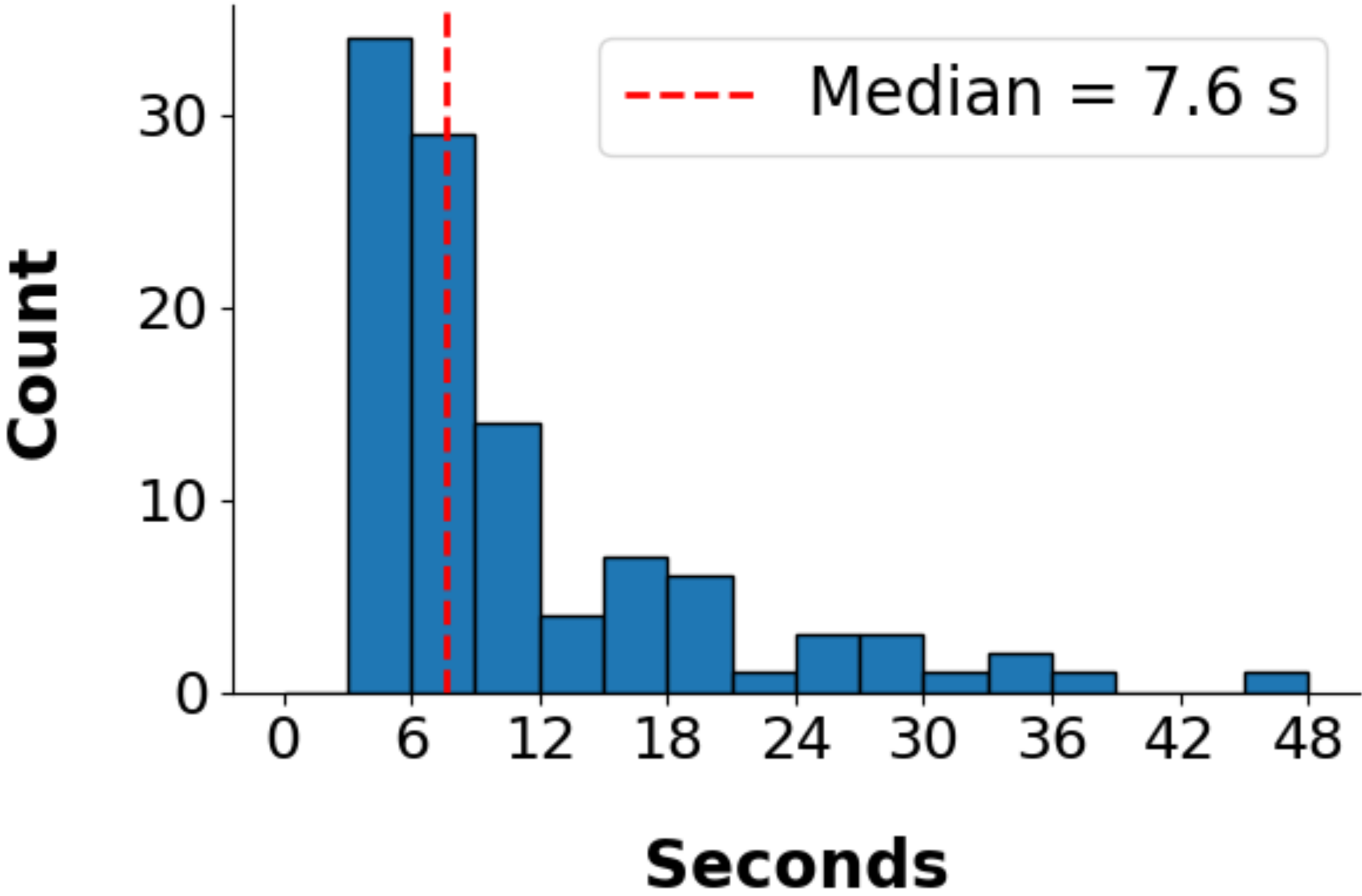


Time to Set Desired Volume

IRIS

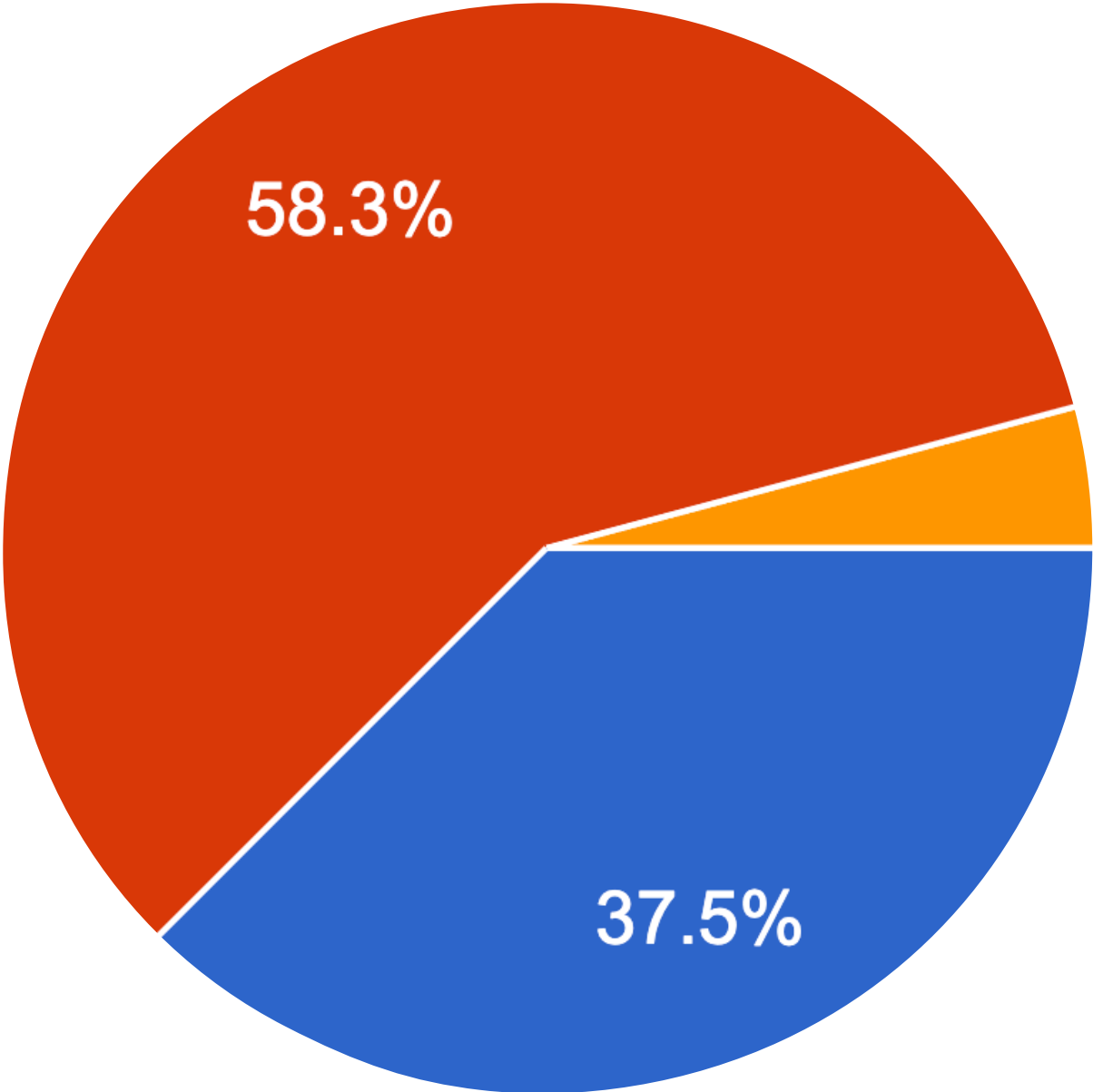


Voice



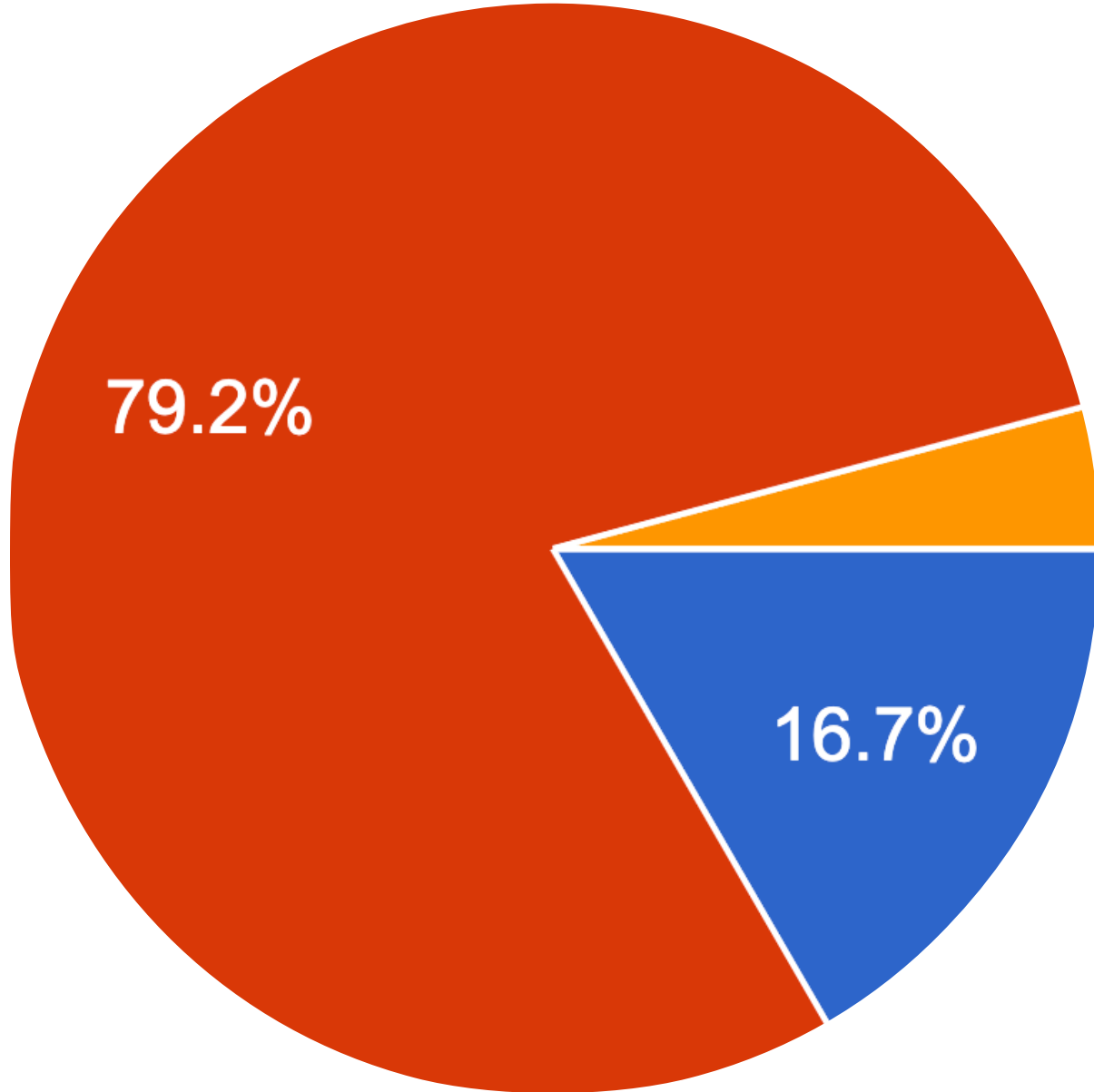
User Study | Questionnaire Results

Toggle device state



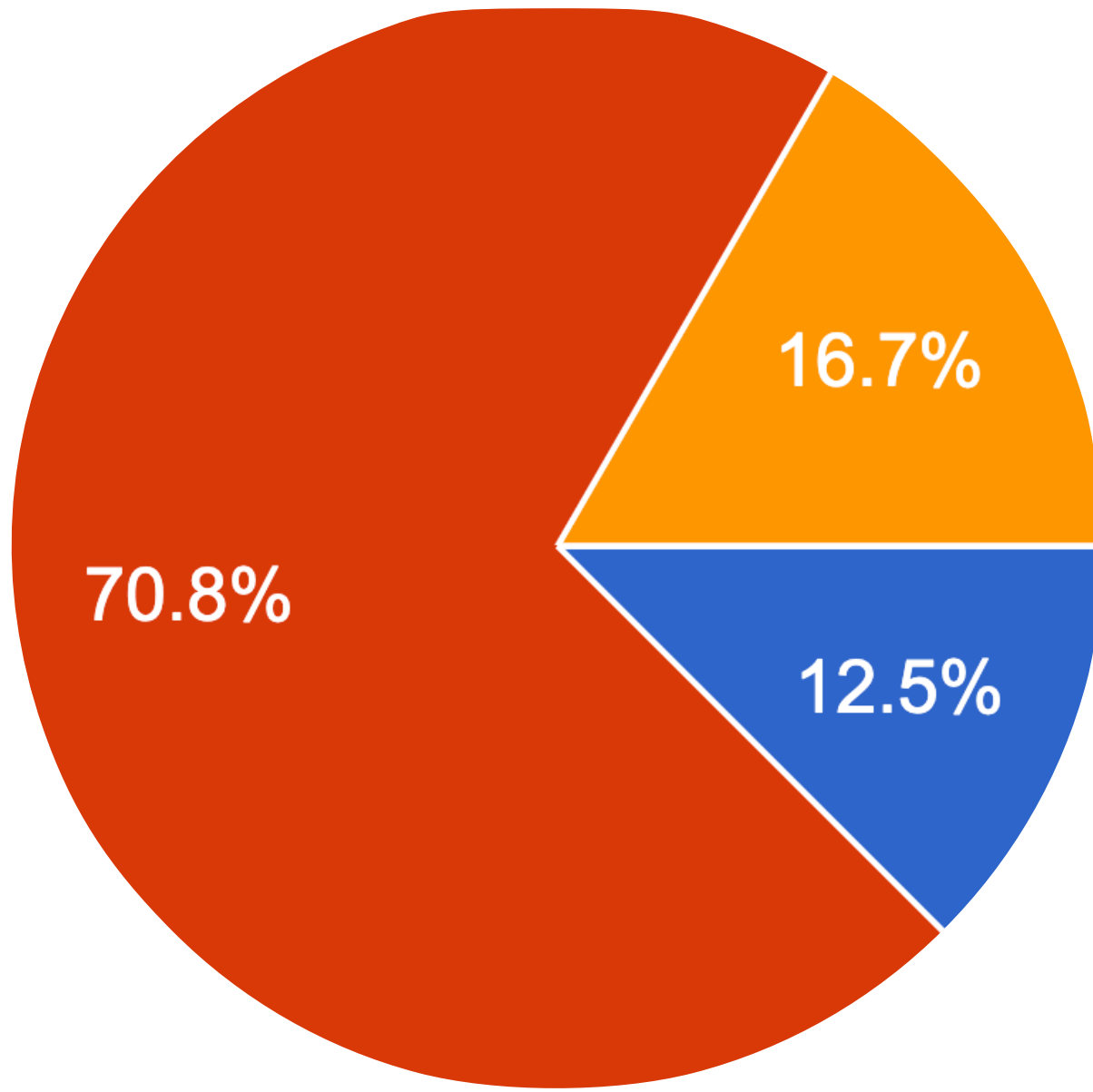
p-value: 0.202

Granular control



p-value: 0.001

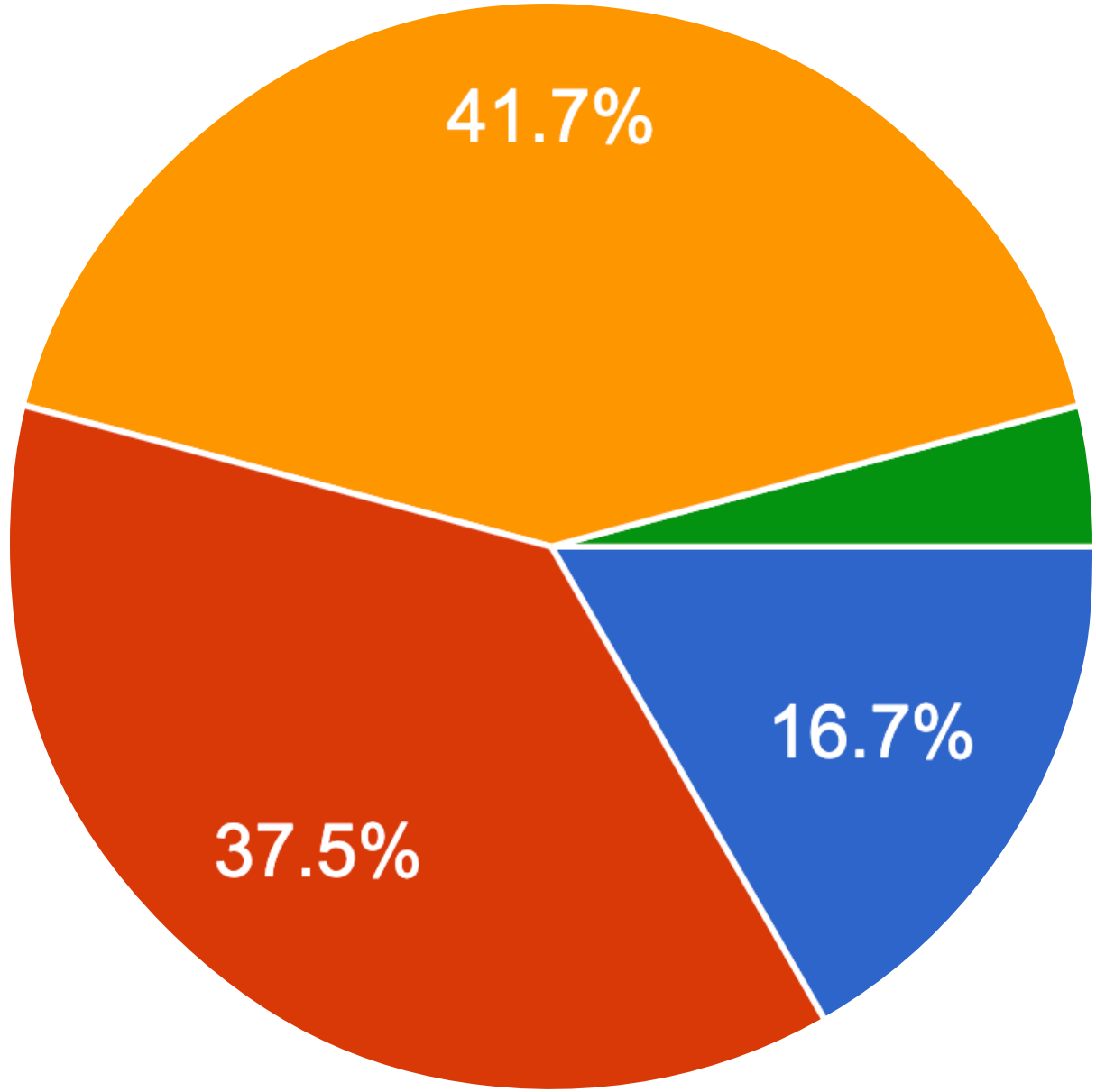
More socially acceptable



p-value: 0.017

● Voice ● IRIS ● No difference

Day-to-day use



p-value: 0.196

- Voice
- IRIS
- Both
- Neither

System Evaluation

Power - 16 hours

Accuracy - 98% Accurate after 3 images

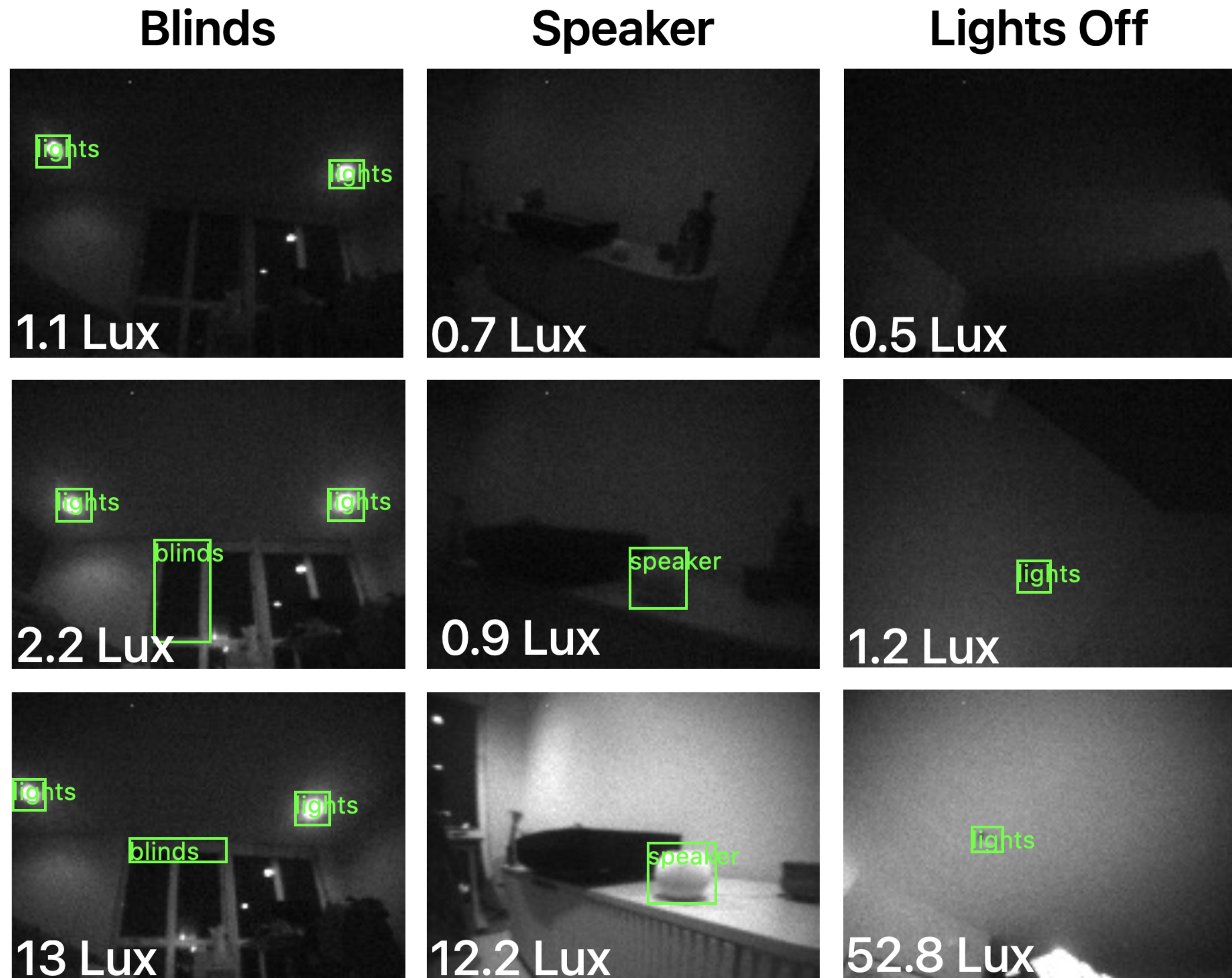
Distance - 2-5 meters

Low-Light - >1 Lux



lights: 100.0

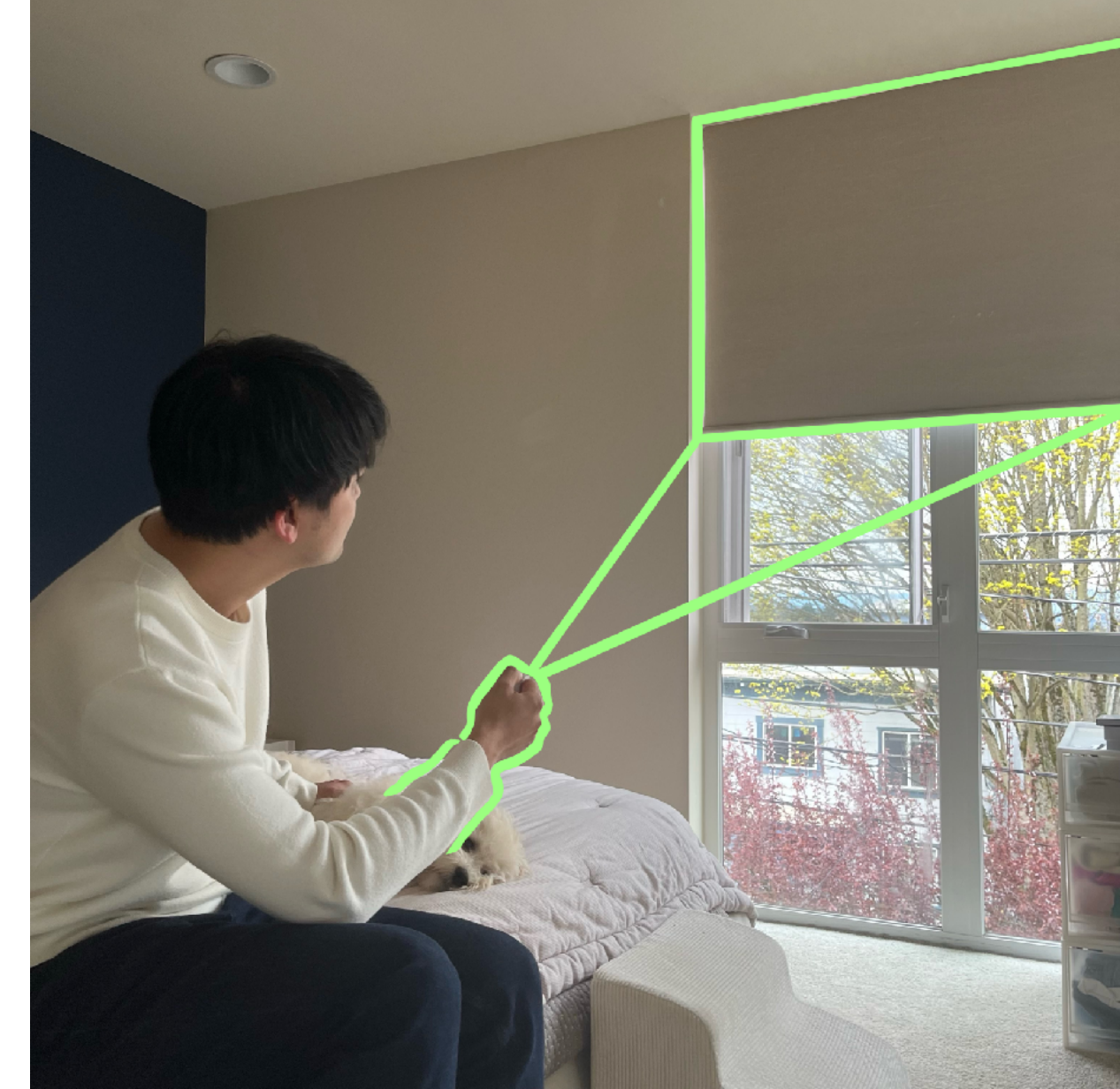
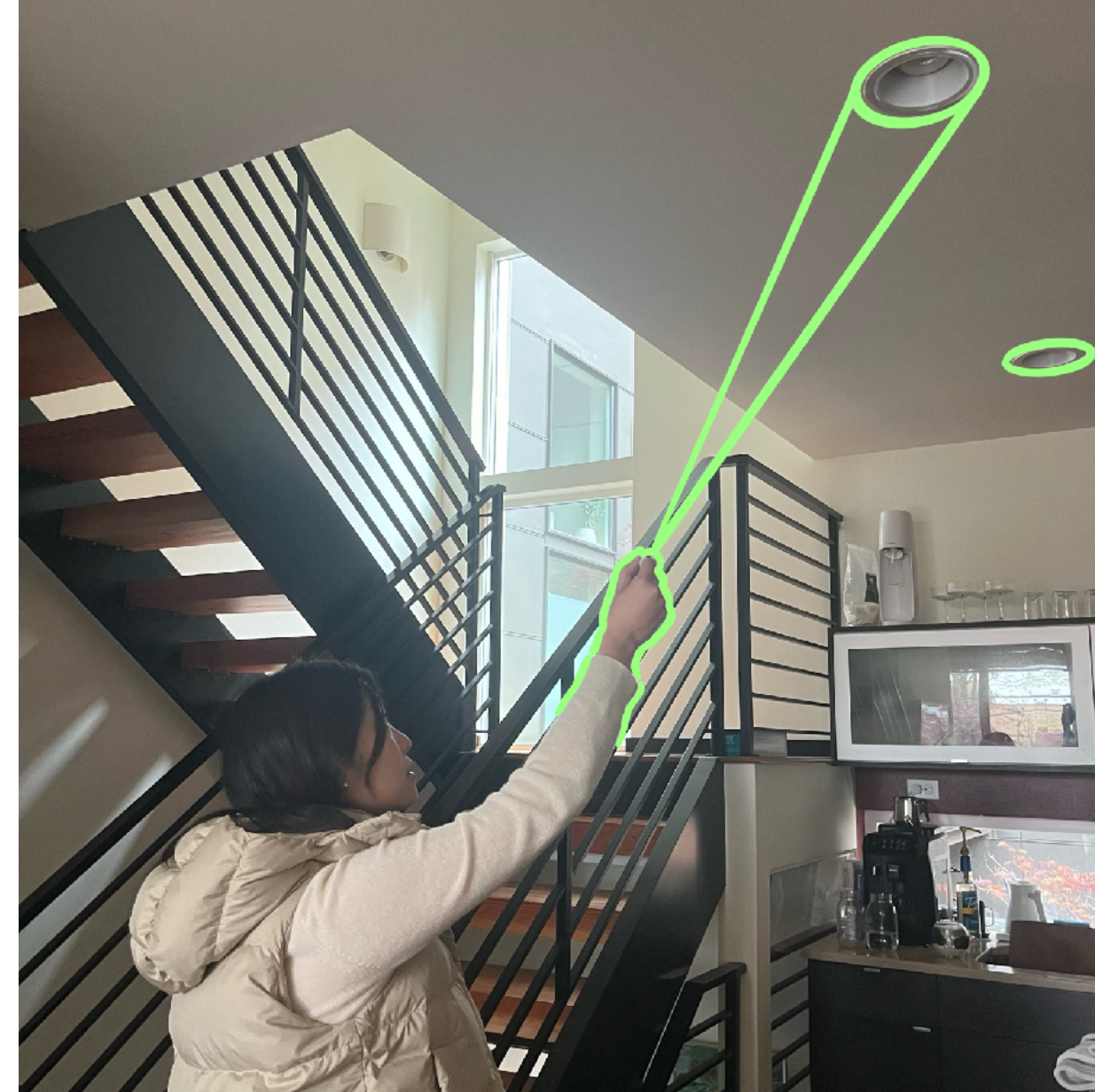
Evaluation | Low-light performance



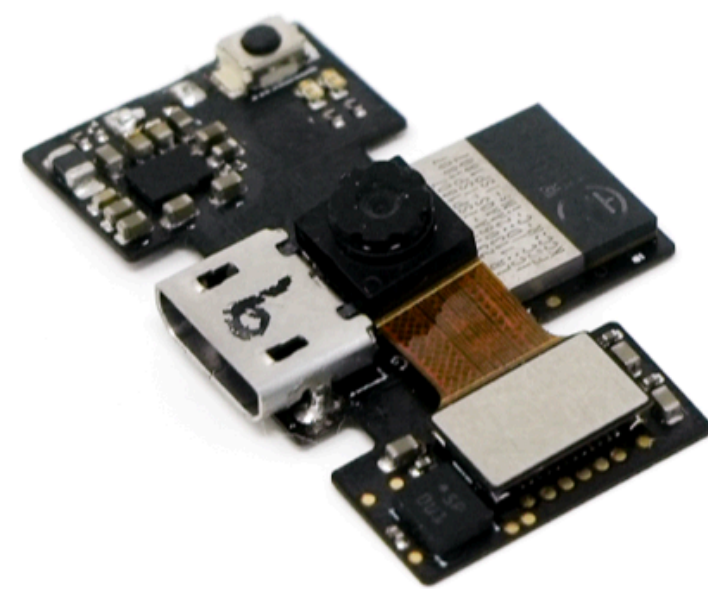
Lighting controlled through ceiling lights

- 50% brightness - 12 lux
- 20% - 2 lux
- 0% - 1 lux
- All examples were collected at night

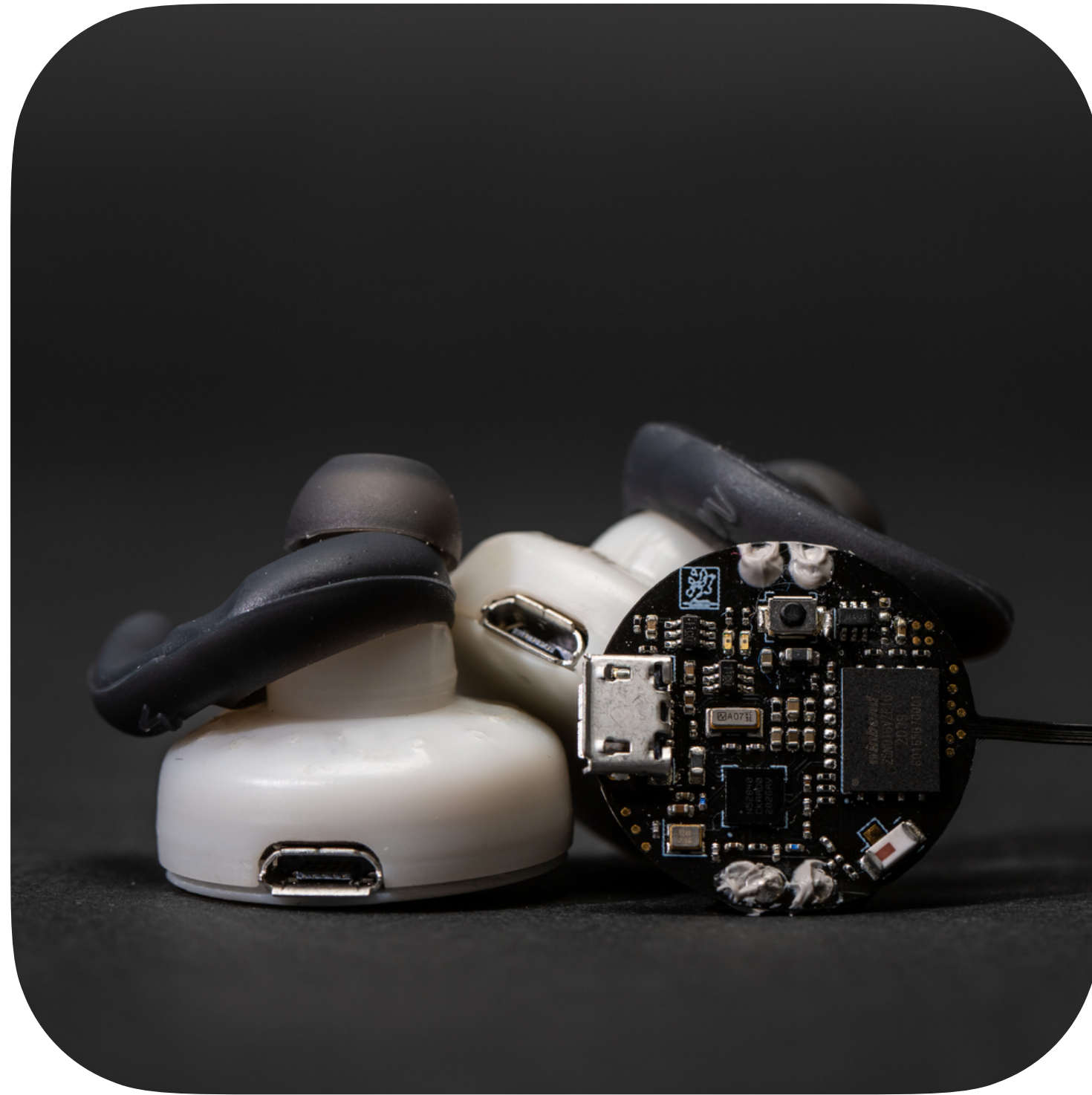




Brief Q&A



What's Next?



ClearBuds
Synchronized Wireless Earbuds
for Speech Enhancement



IRIS
Camera-Integrated Ring for
Vision-Based Interaction



Multimodal Earbuds
Visual Intelligence on
Wireless Earbuds

VueBuds

Visual Intelligence with Wireless Earbuds

Maruchi Kim

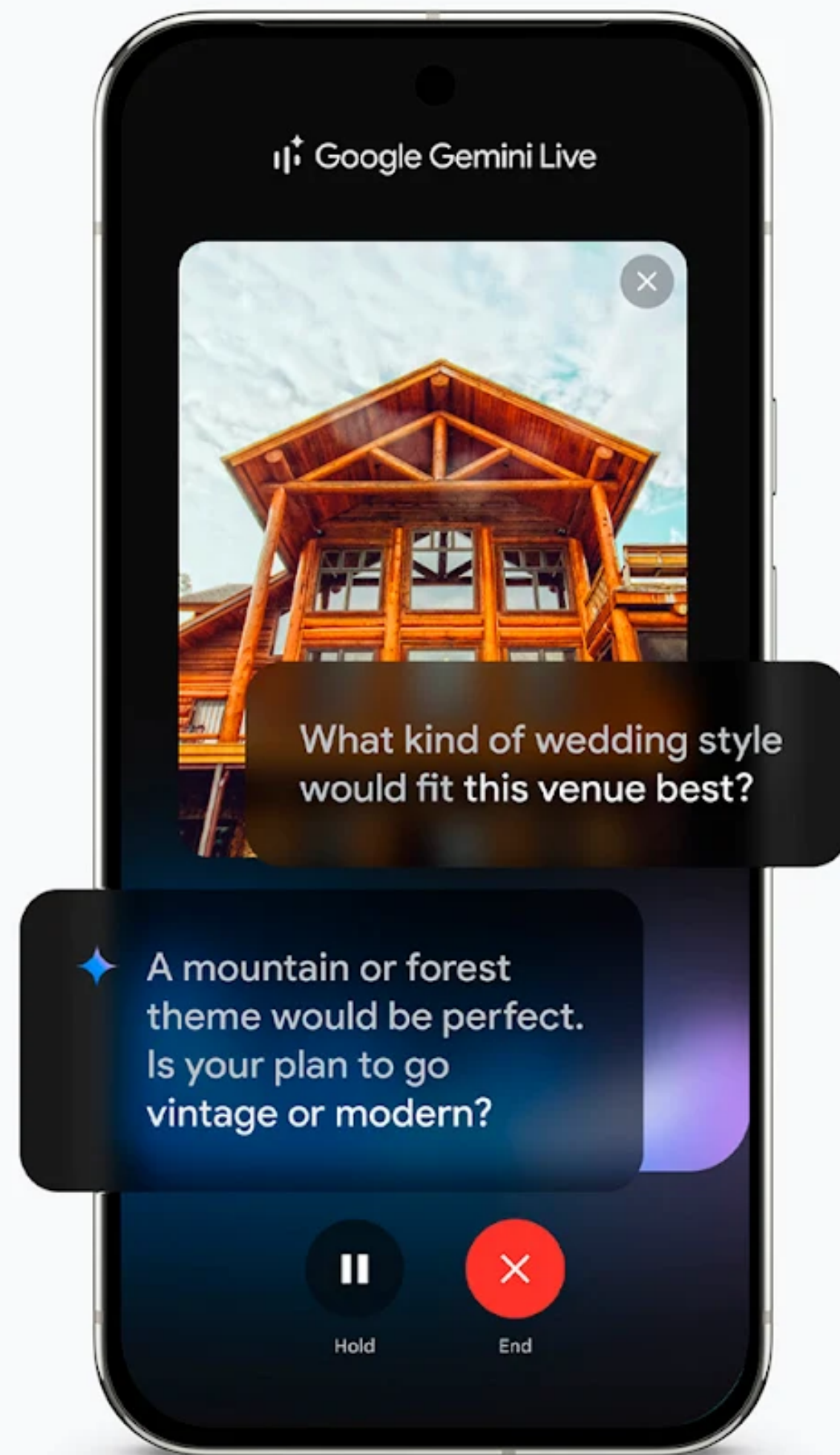
R. Fawwaz, Z. Lim, B. Moudgalya, H. Wang, Y. Zeng

Shyam Gollakota

University of Washington



Visual AI has exploded, but it's stuck in phones and smart glasses



Displayless smart glasses and earbuds share many of the same core features



Displayless Smart Glasses

Hardware

- Stereo Speakers
- Microphone Array
- **Cameras**

Features

- Audio Playback
- Telephony
- Voice Input
- **AI Integration**
- **Photo and Video**



Wireless Earbuds

Hardware

- Stereo Speakers
- Microphone Array

Features

- Audio Playback
- Telephony
- Voice Input
- **ANC + Transparency**



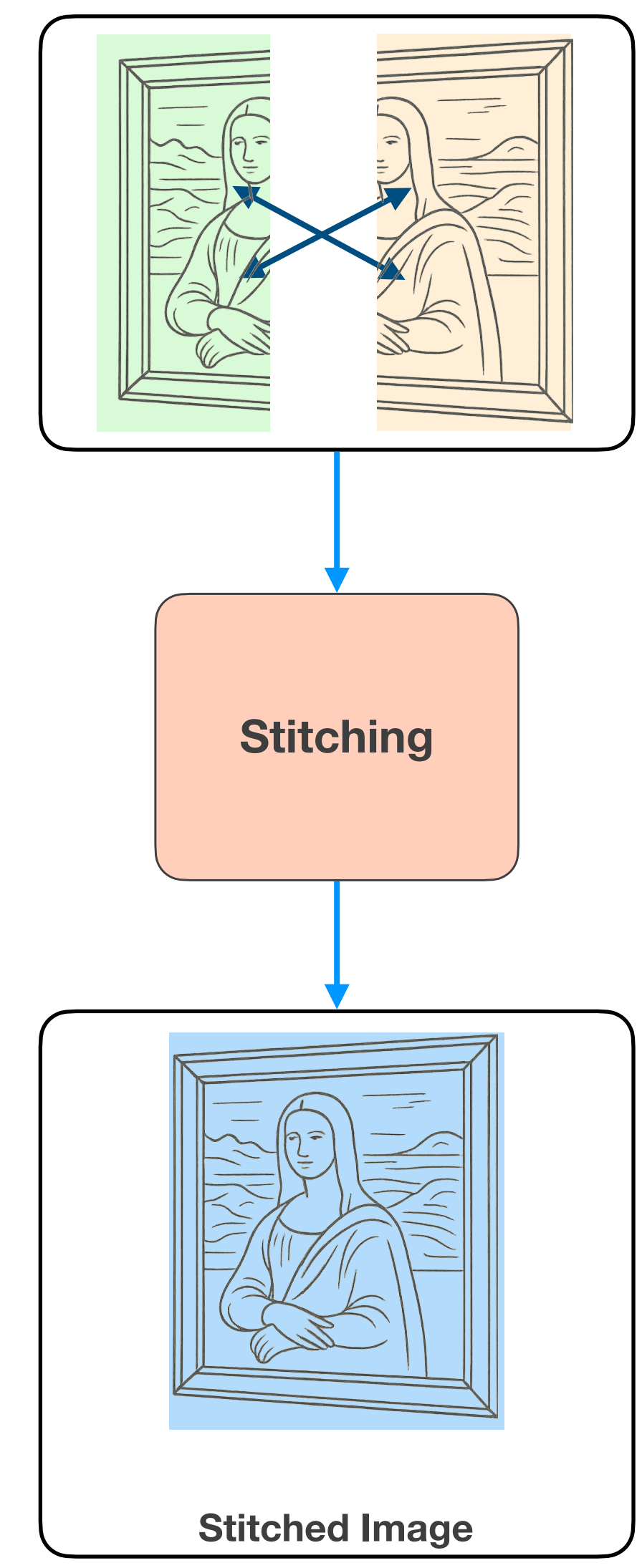
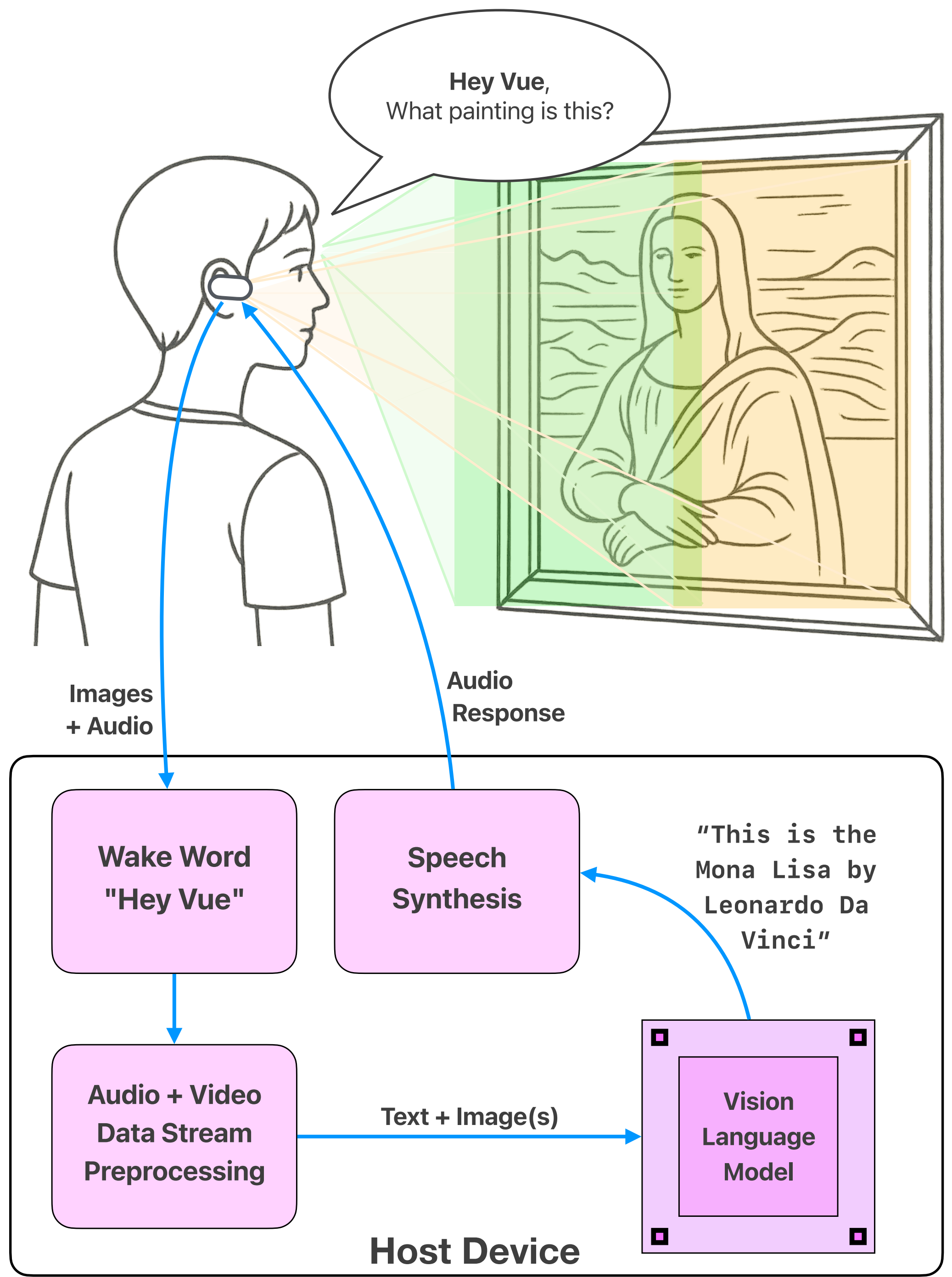
Over a billion people wear earbuds everyday.

What if they could see?

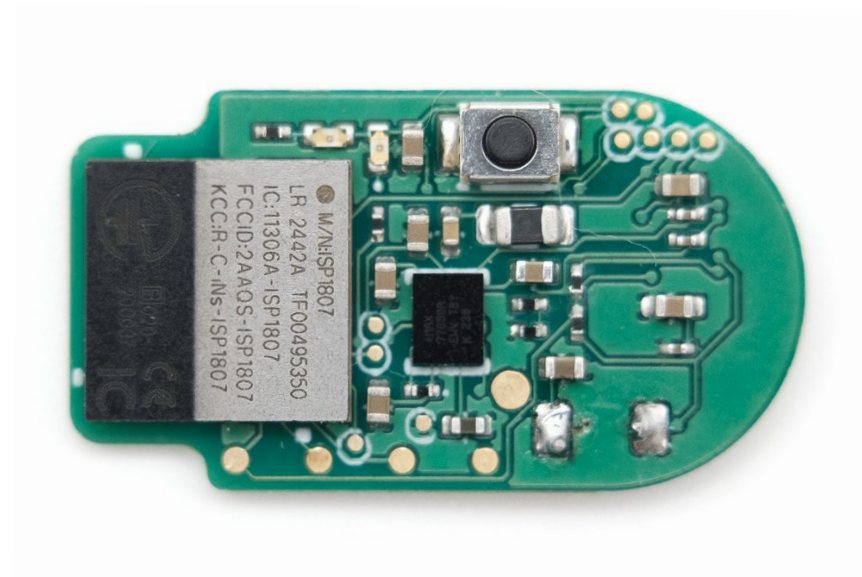
First camera-integrated earbuds to integrate with visual AI



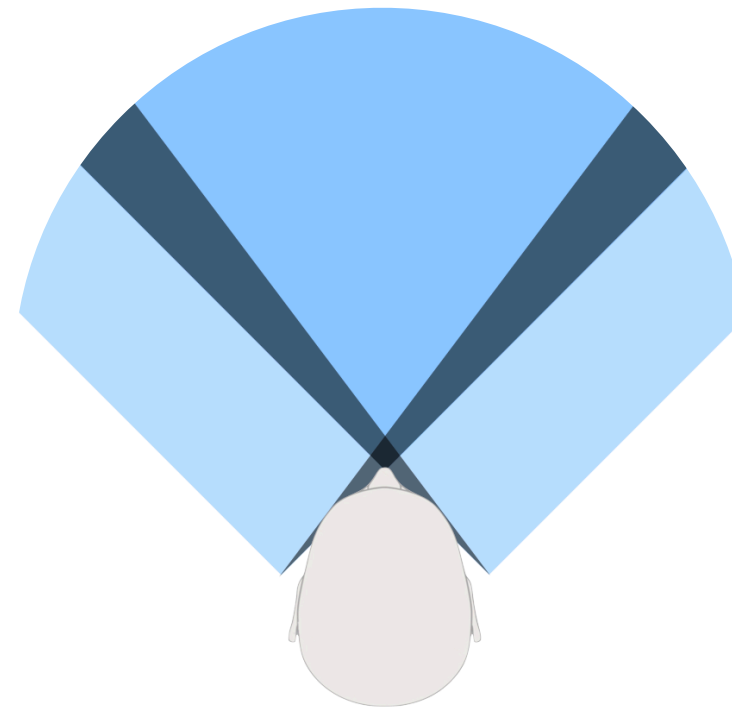




What are the key technical challenges?



Creating a low-power camera module compatible with wireless earbuds



Overcoming facial occlusion through binaural stereo vision

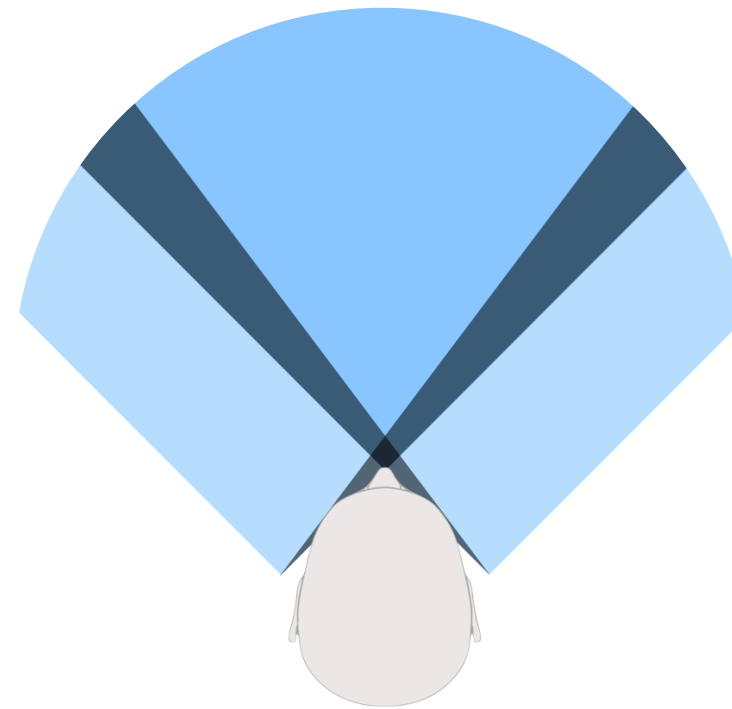


Integrating vision capabilities with real-time performance

What are the key technical challenges?



Creating a low-power camera module compatible with wireless earbuds

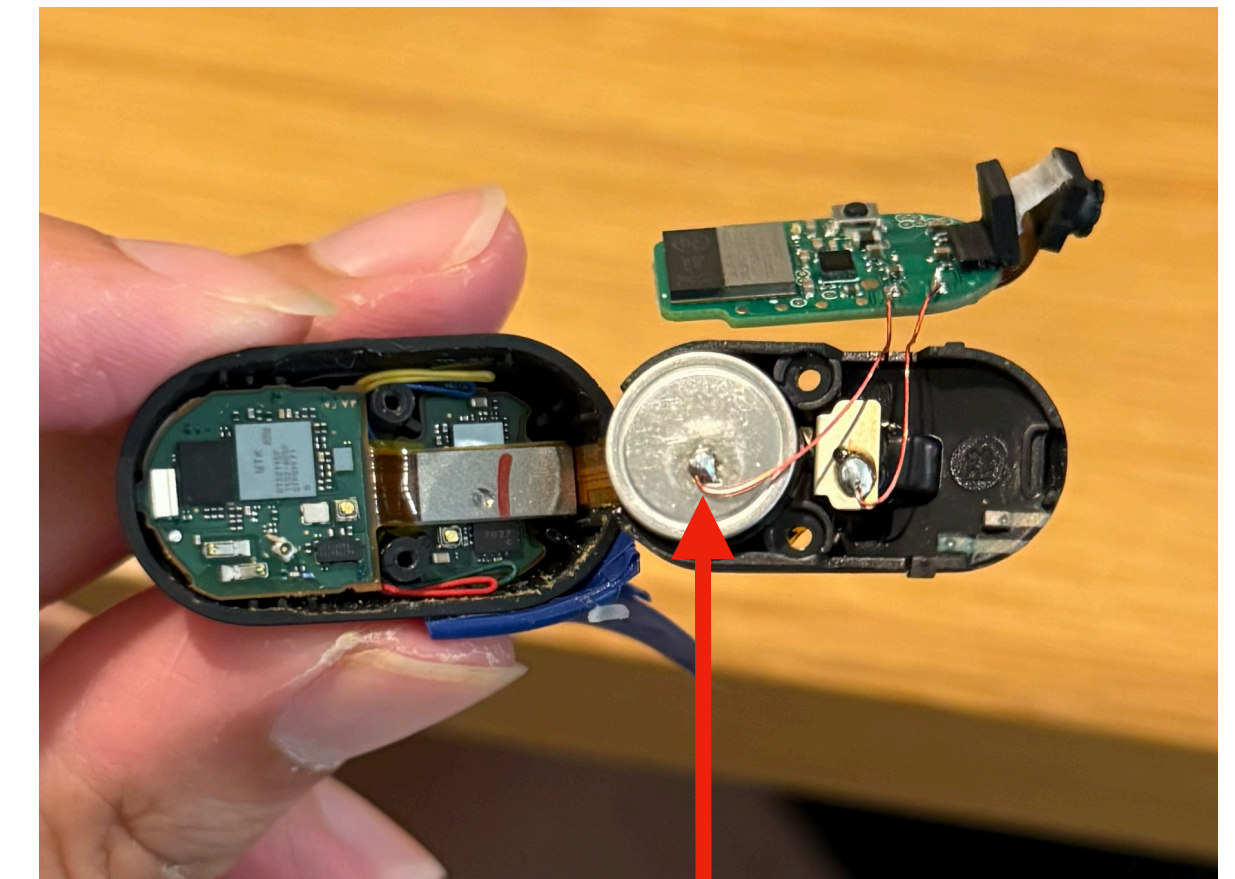
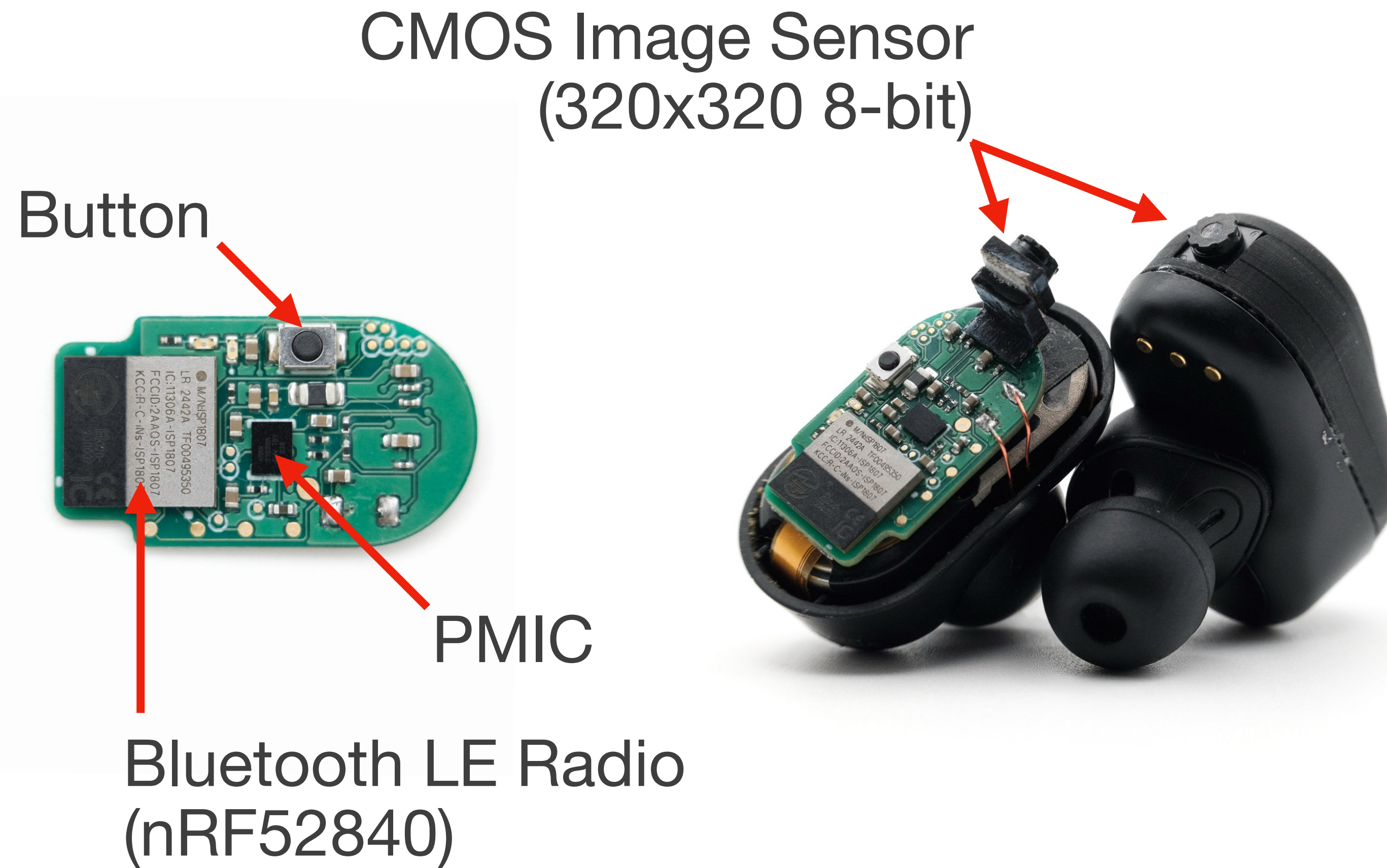


Overcoming facial occlusion through binaural stereo vision



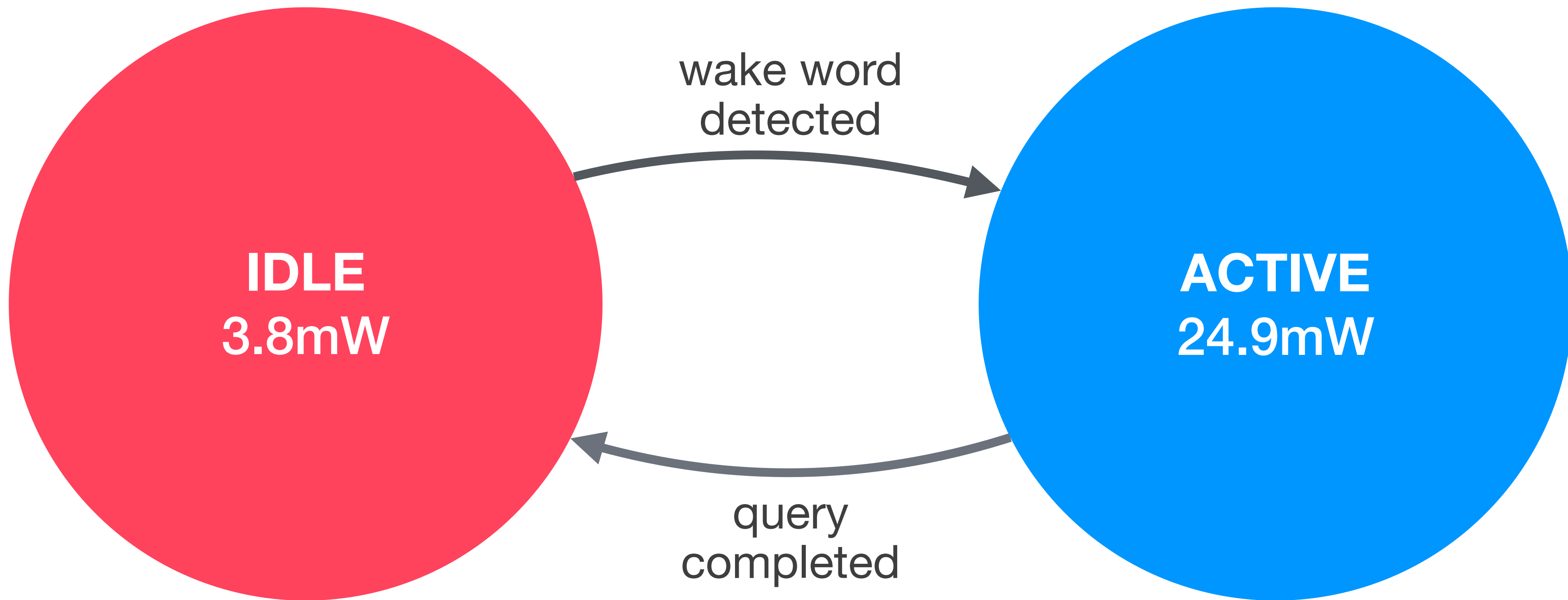
Integrating vision capabilities with real-time performance

VueBuds Hardware

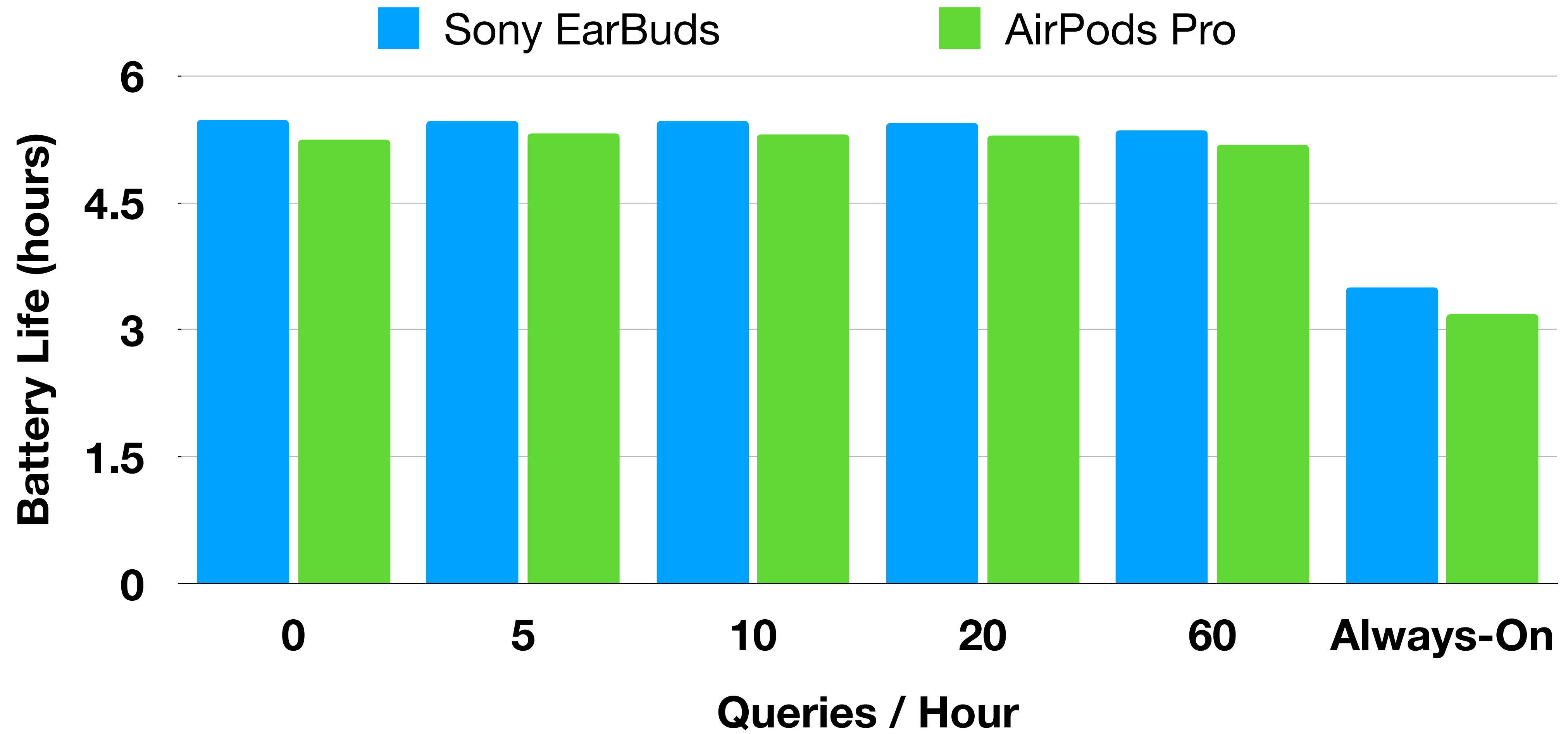


Battery

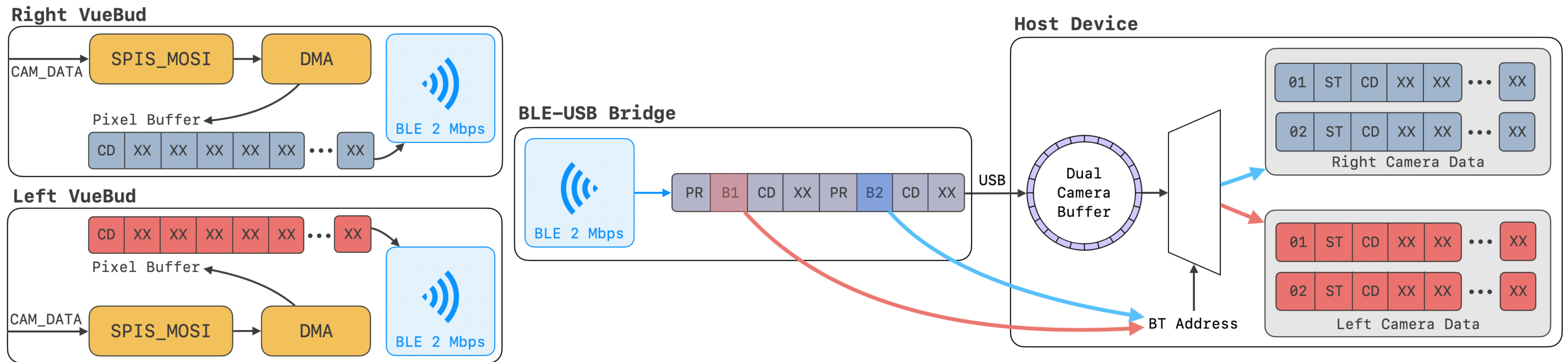
VueBuds Hardware



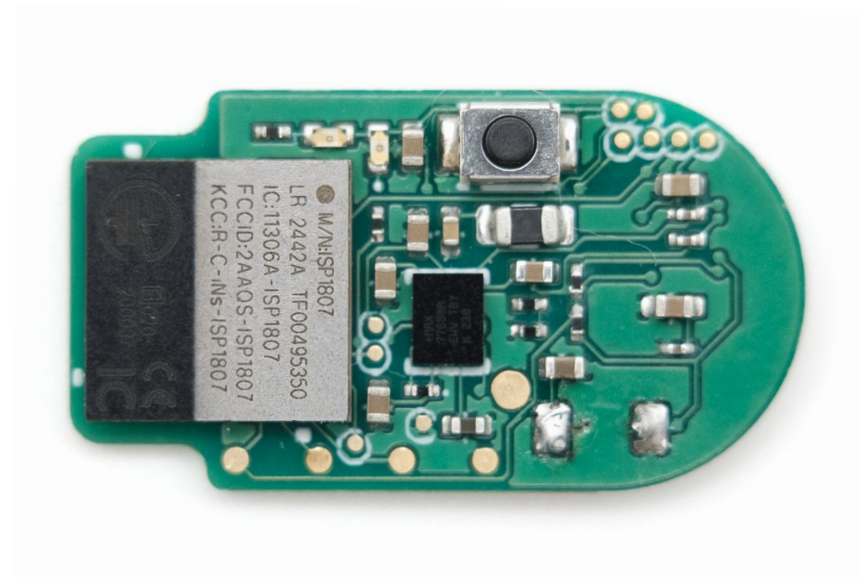
VueBuds Hardware



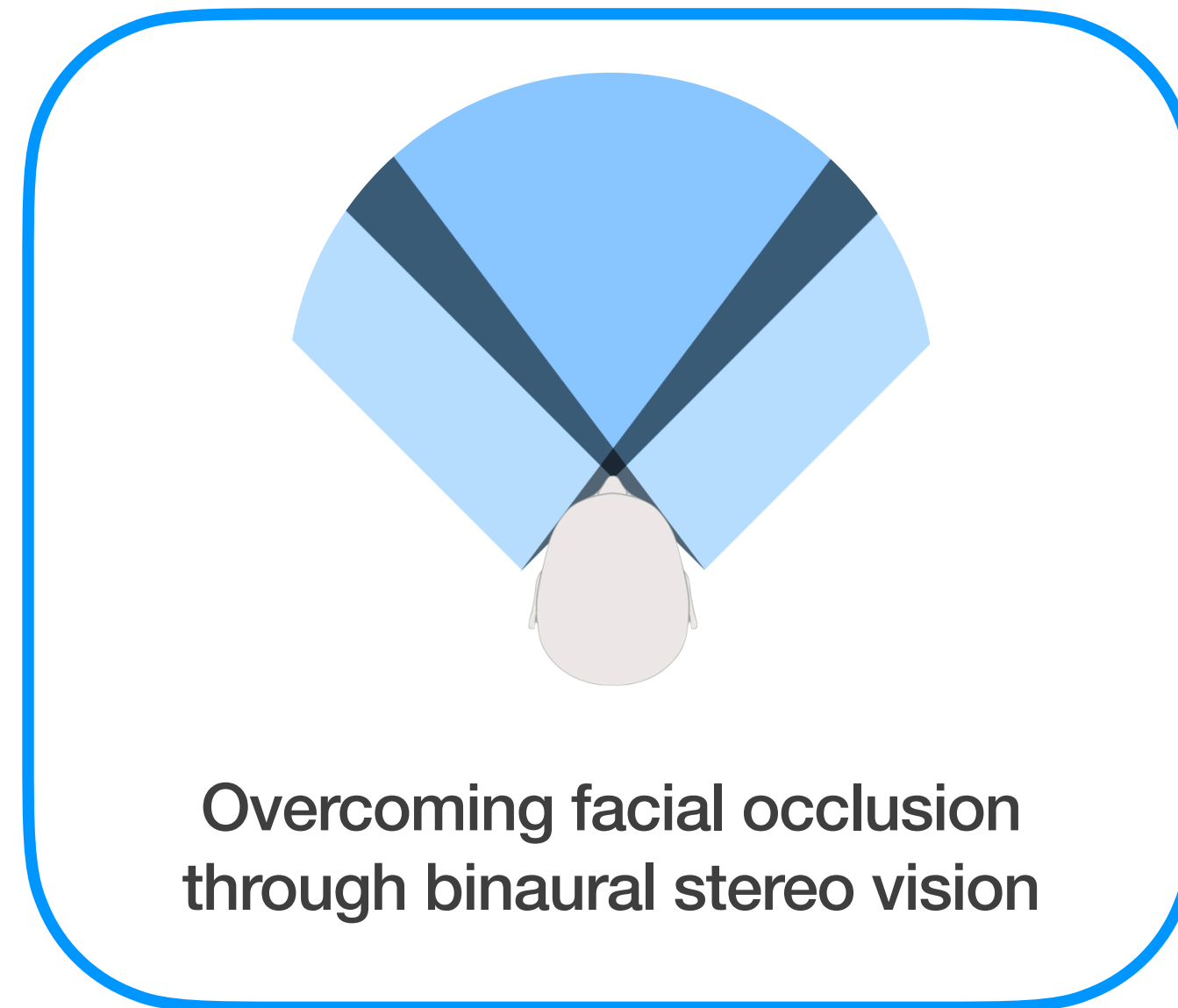
VueBuds Hardware



What are the key technical challenges?



Creating a low-power camera module compatible with wireless earbuds

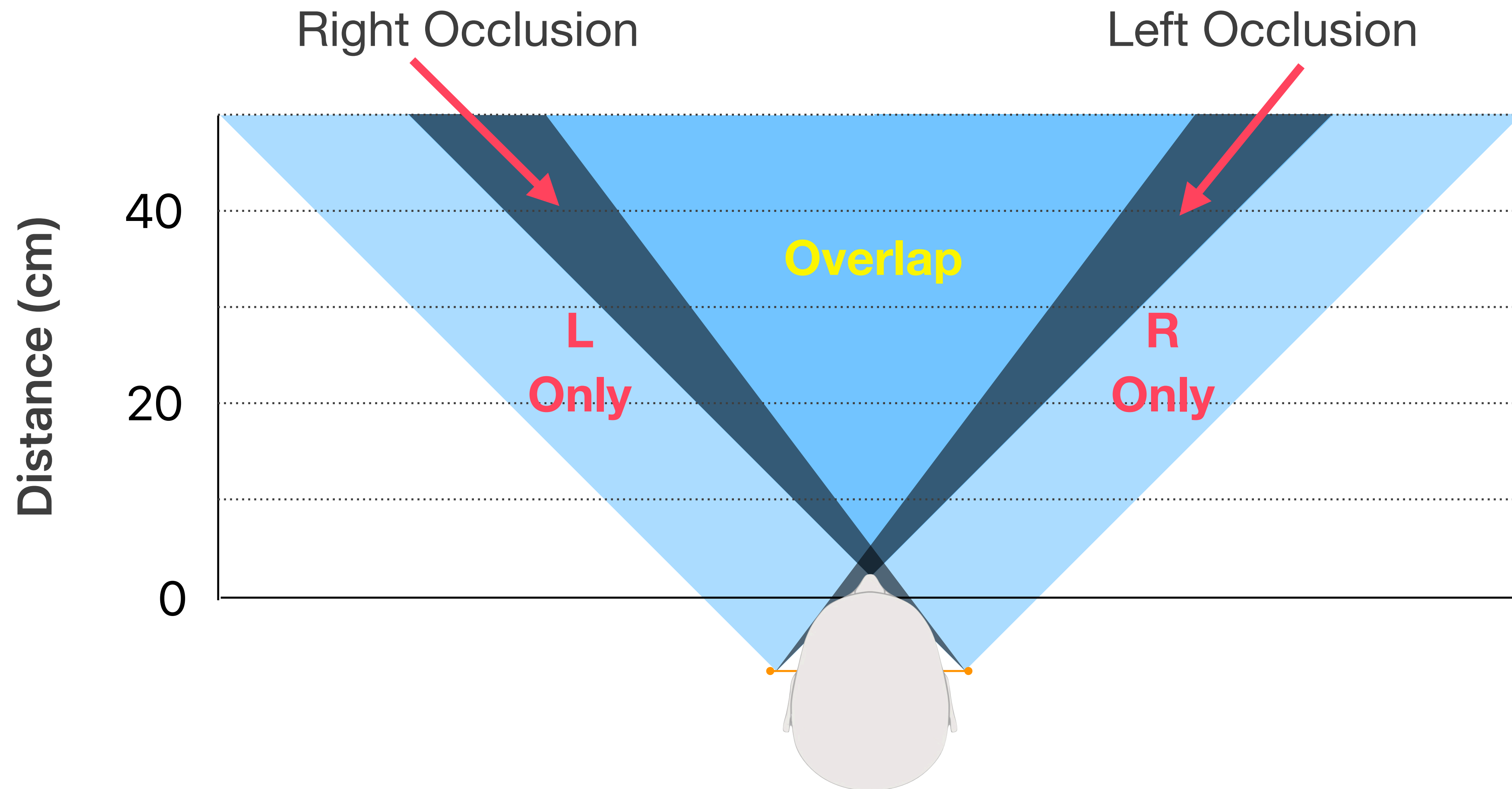


Overcoming facial occlusion through binaural stereo vision

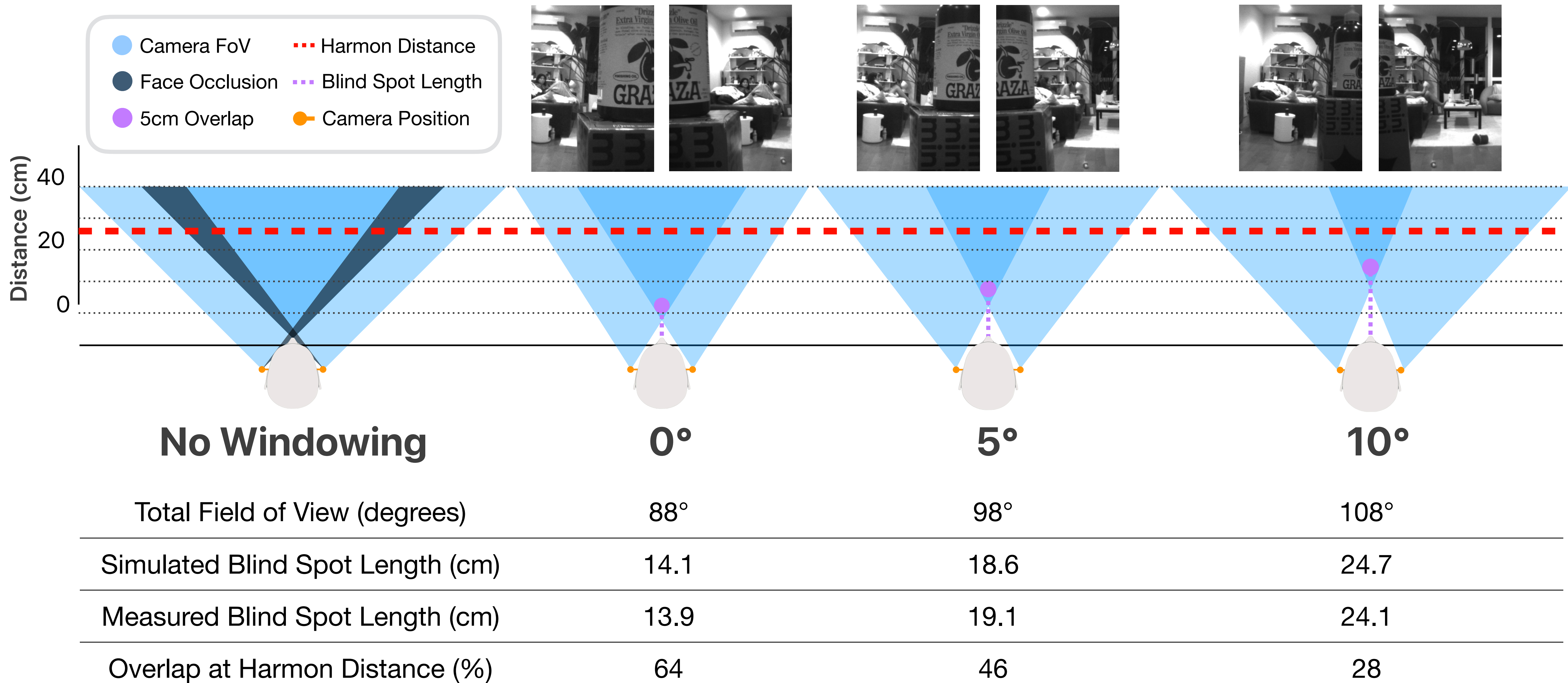


Integrating vision capabilities with real-time performance

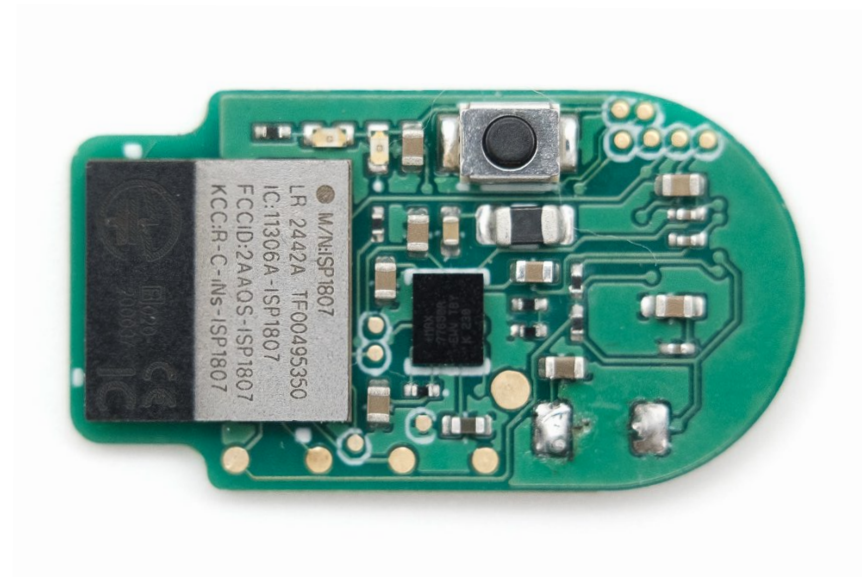
VueBuds Stereo Vision



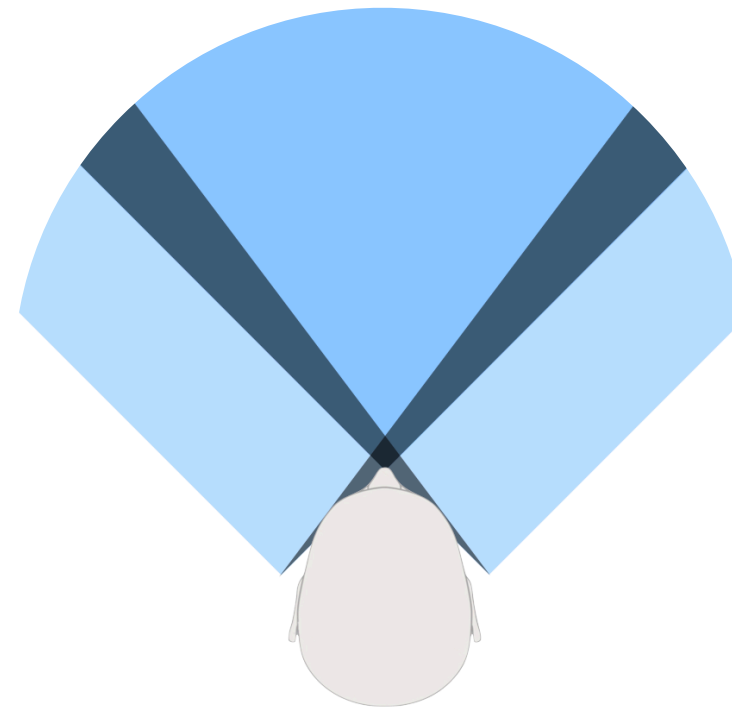
VueBuds Stereo Vision



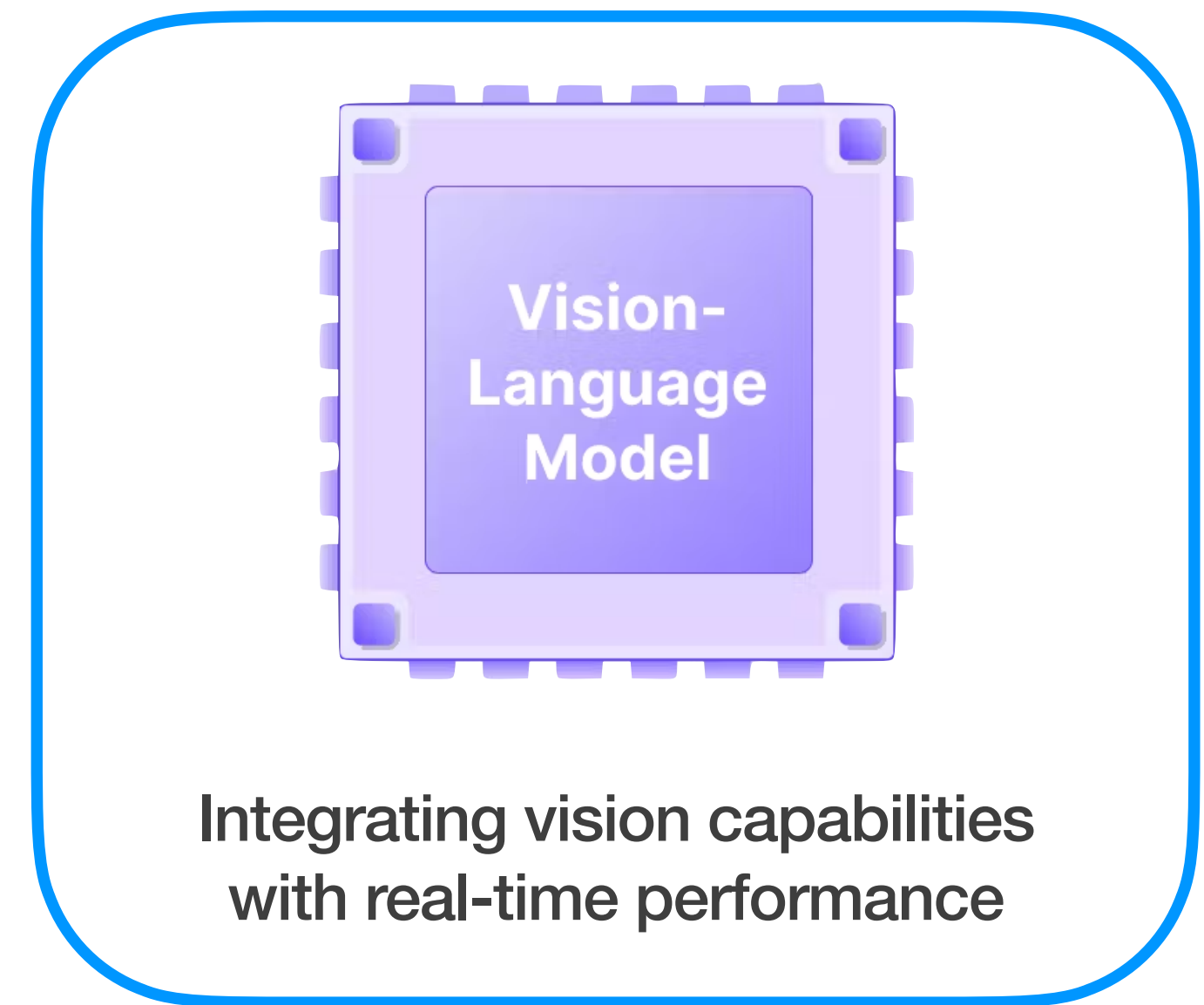
What are the key technical challenges?



Creating a low-power camera module compatible with wireless earbuds

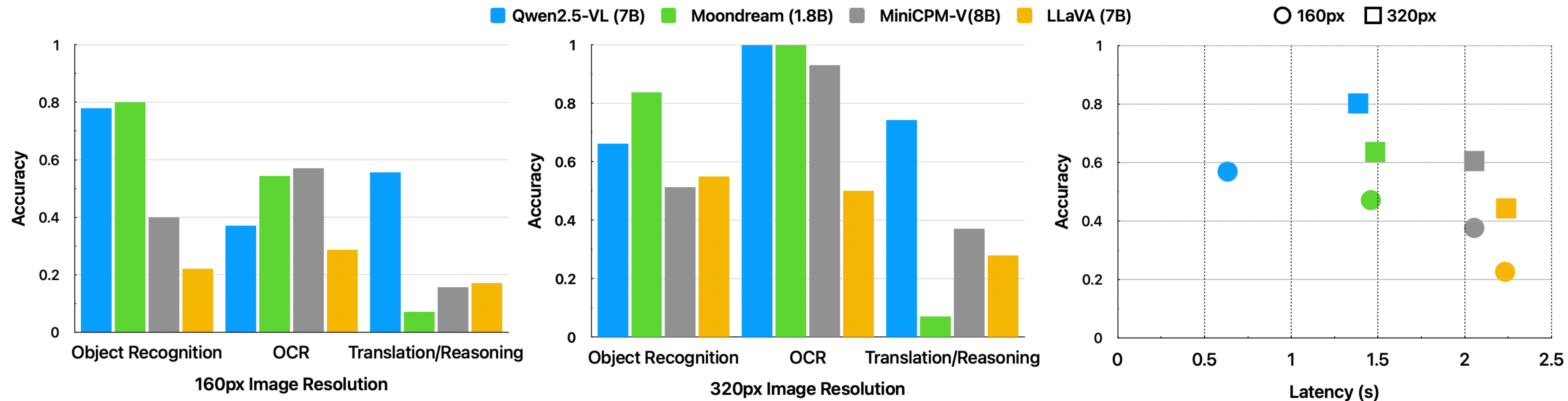


Overcoming facial occlusion through binaural stereo vision


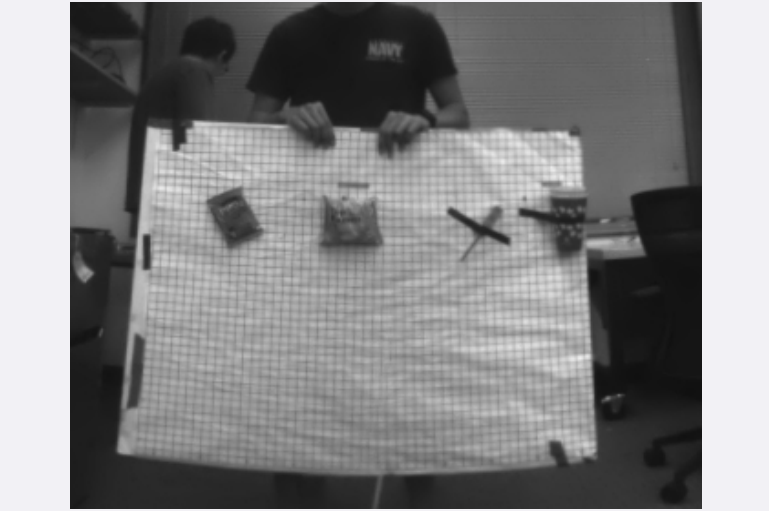






Integrating vision capabilities with real-time performance

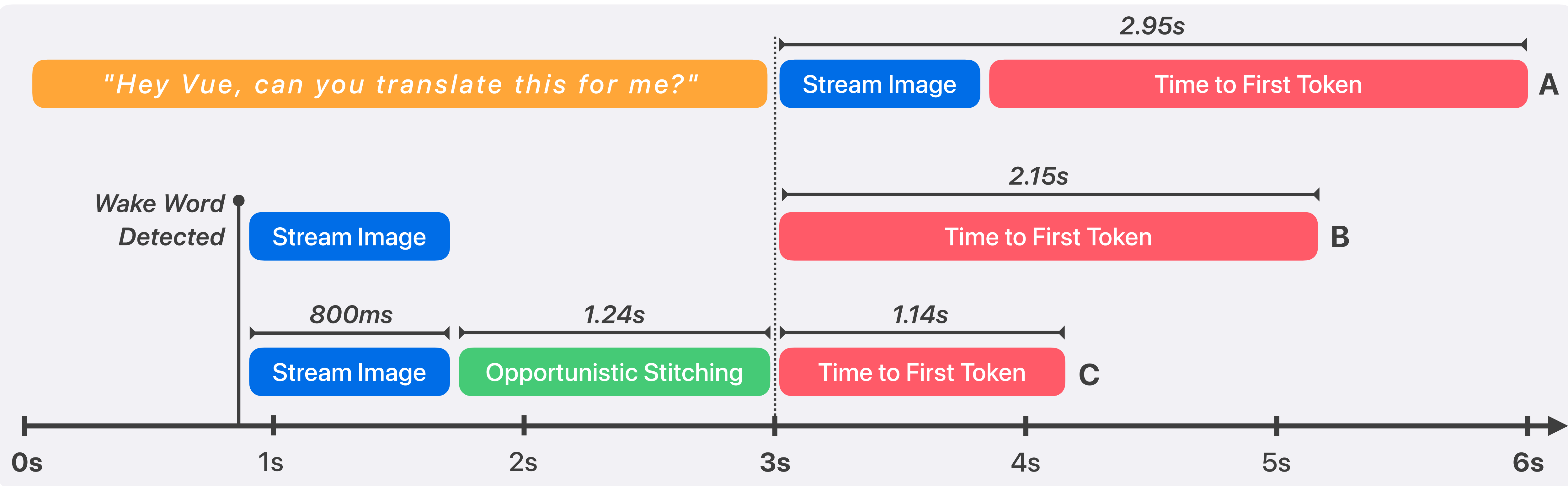
VueBuds Vision Language Model



VueBuds Opportunistic Stitching

Stereo Images	Response	Stitched Image	Response
	<p>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.</p>		<p>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.</p>
	<p>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.</p>		<p>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.</p>
	<p>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.</p>		<p>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.</p>

VueBuds End-to-End Latency





```

read display)
at do you see?'
'few buds' in 'Few buds, what do you see?'
'what do you see?'
on: 'what do you see?'
dual_ts: 1757059862.139478 combined_ts: 1757059862.139478
stitched image

```



```

Processing 1 image(s): [(676, 330)]
User question: what do you see?
Images encoded in 0.241s (Total Base64 size: 81.2 KB)

```

Question: what do you see?

Response:

Camera: 1.3 FPS (sending to main thread display)

I see a dimly lit room with a computer monitor displaying some text. There's a desk with various items on it, and a door is visible in the background.

Performance: 2.7 tok/s, First token: 11.061s

=====

LLM RESPONSE:

=====

I see a dimly lit room with a computer monitor displaying some text. There's a desk with various items on it, and a door is visible in the background.

=====

Query saved to: queries/query_20250905_011116_126

Camera: 1.2 FPS (sending to main thread display)

Camera: 1.3 FPS (sending to main thread display)

□

User Study

User Study | Methodology

90 Participants (n=90)

Online Survey (n=74)

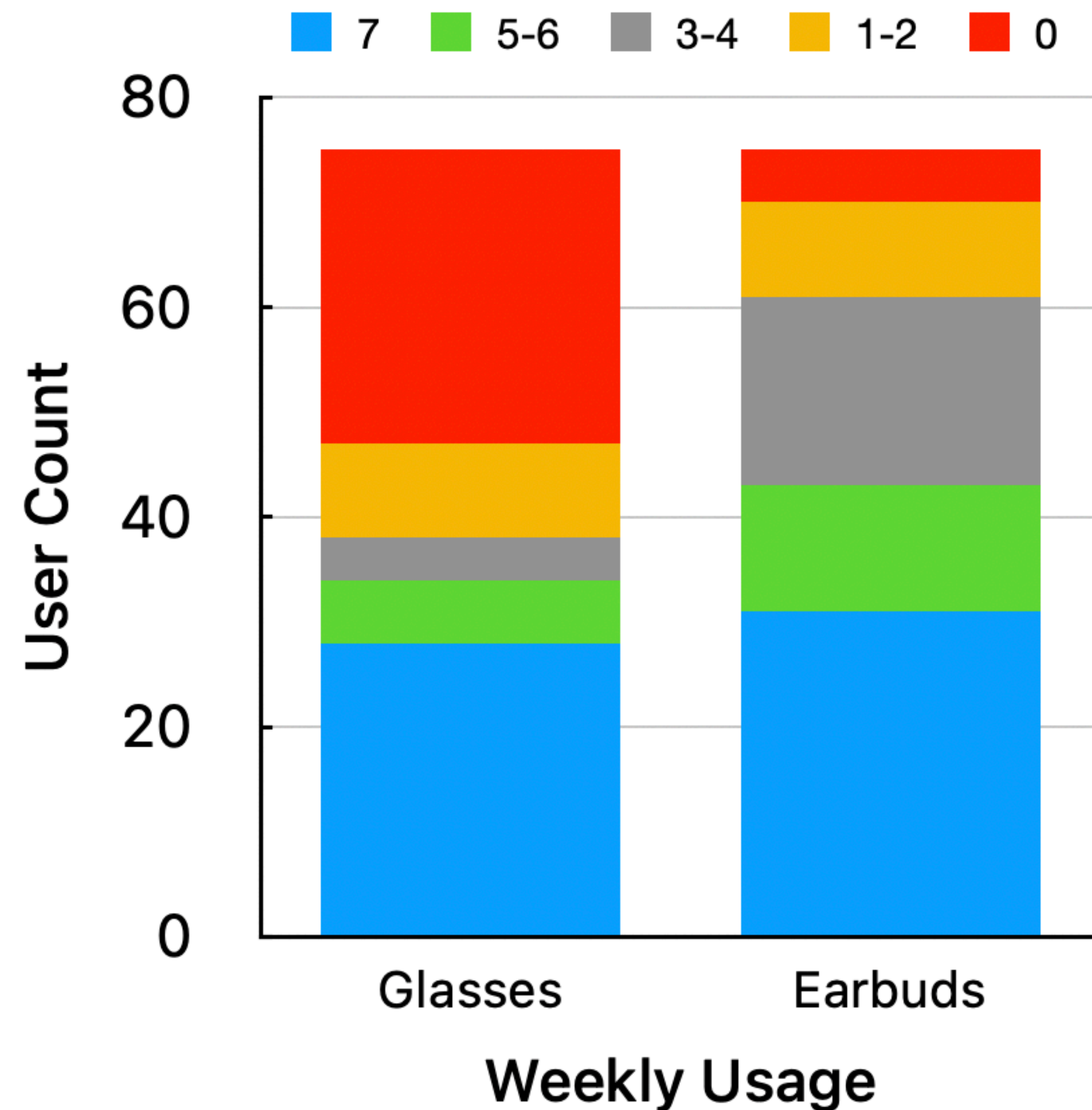
- Frequency of wearing ordinary glasses and earbuds
- Response quality evaluation between RayBan Meta and VueBuds

In-person Evaluation (n=16)

- Real world visual question answering (VQA) performance across different environments

Online Survey Results

VueBuds User Study | Platform Usage Analysis (n = 74)



Three usage patterns

- High (5+ days a week)
- Occasional (1-4 days a week)
- Non-users (0 days a week)

Glasses

- H: 45.3%, O: 17.3%, N: 37.3%

Earbuds




- H: 57.3%, O: 36%, N: 6.7%

Occasional + High Usage

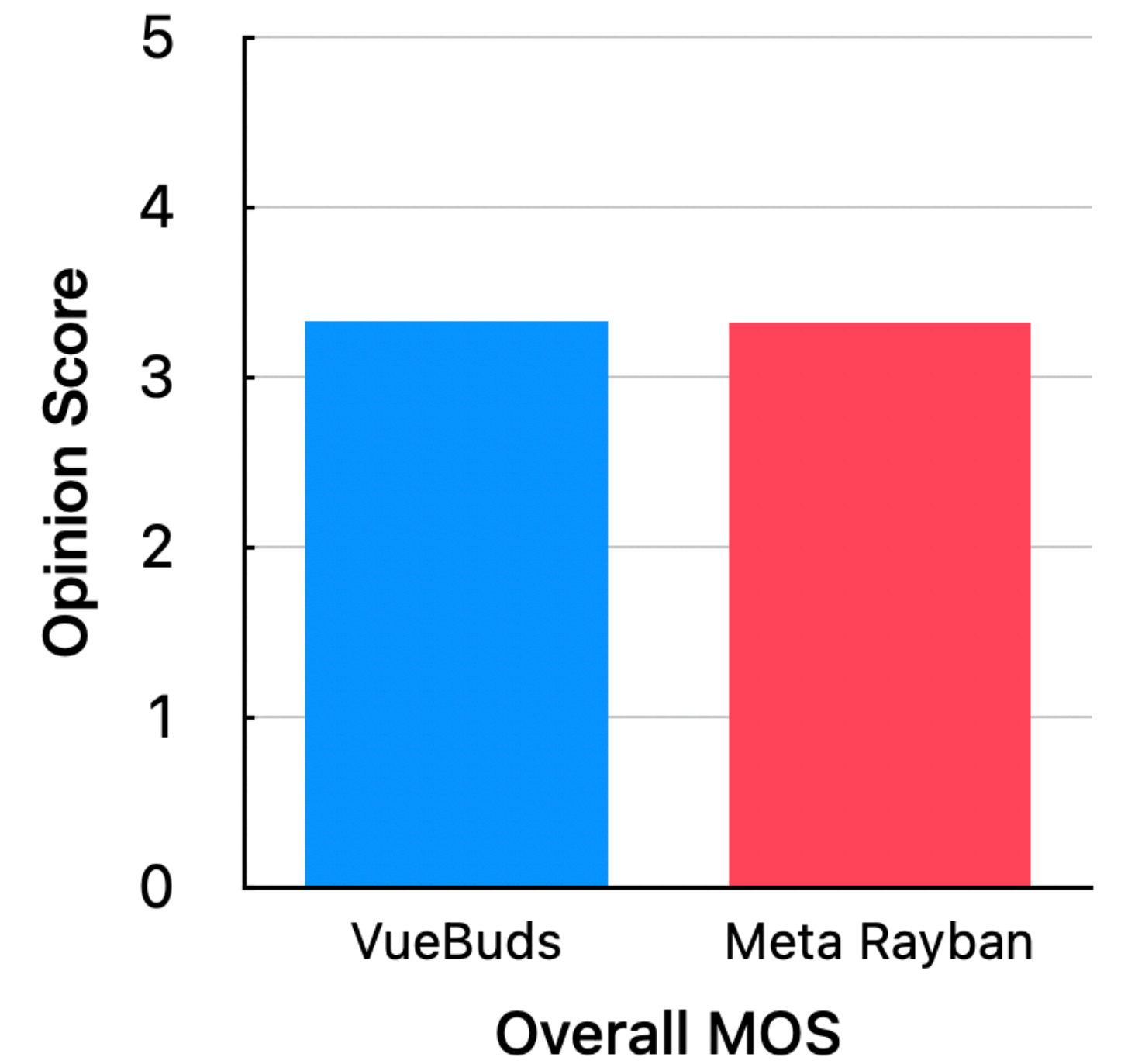
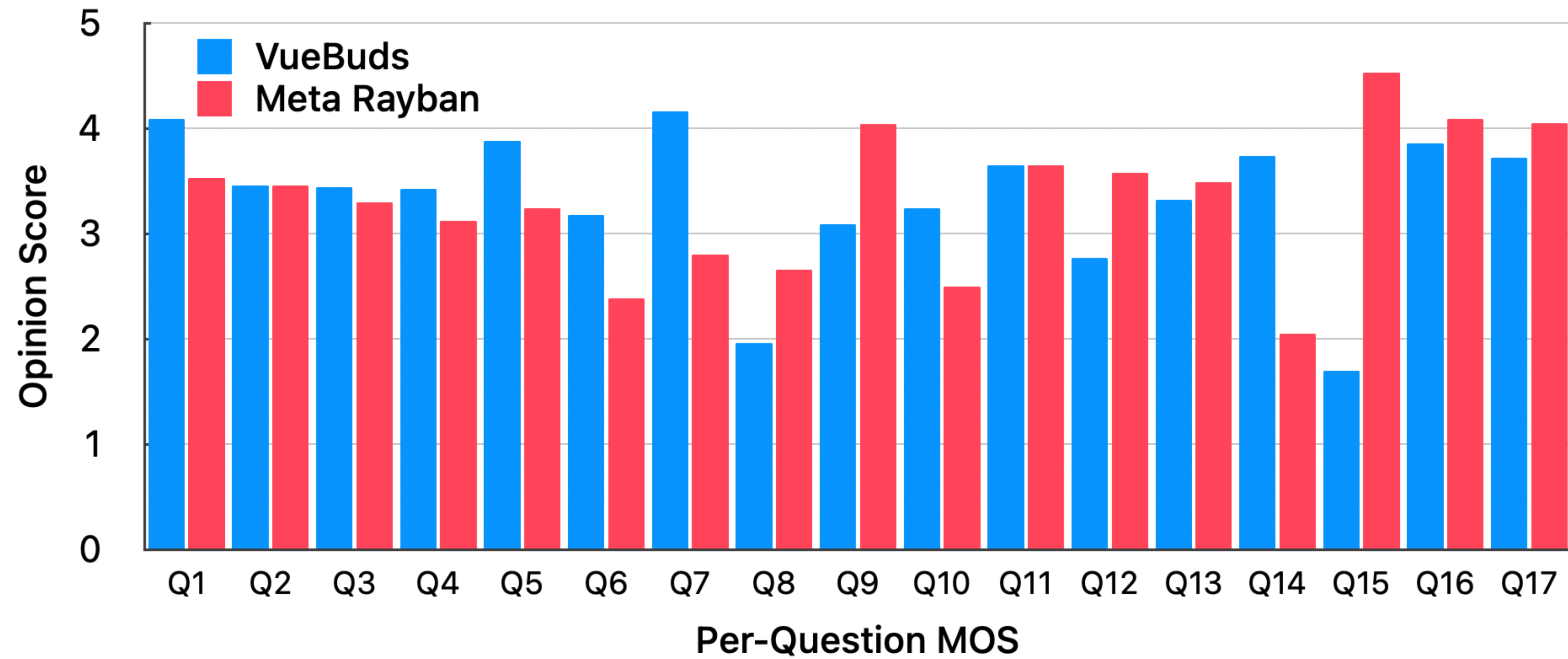
- Earbuds (93.3%)
- Glasses (62.7%)

VueBuds User Study | Blind comparative evaluation (n=74)

*Imagine you are wearing AI-enabled smart glasses.
You will see images with sample questions and two different AI responses.*

	<p>Q11 <i>"How many calories are in this?"</i></p>	<p>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.</p>	<p>The can of La Croix sparkling water has zero calories.</p>
	<p>Q7 <i>"translate this for me"</i></p>	<p>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.</p>	<p>The text appears to be a product labeled for a Korean cold noodle dish with ingredients and nutritional info listed.</p>
	<p>Q15 <i>"how many brushes do you see?"</i></p>	<p>In the image, there are six paintbrushes visible.</p>	<p>I see seven brushes.</p>

VueBuds User Study | Blind comparative evaluation (n=74)



Nearly all participants prioritized **factual correctness**

Participants consistently valued responses that **directly answer the query**

For translation tasks, they strongly preferred **actual translations over scene descriptions**

Prefer responses that matched "**everyday conversation**" style

Valued appropriate **uncertainty expression**.

Acknowledge limitations rather than provide overconfident incorrect responses.

"I really disliked when it got the count wrong"

"concise answers that actually answer the question correctly"







"way too much detail."

"explanations along with the answer" for complex scenarios.

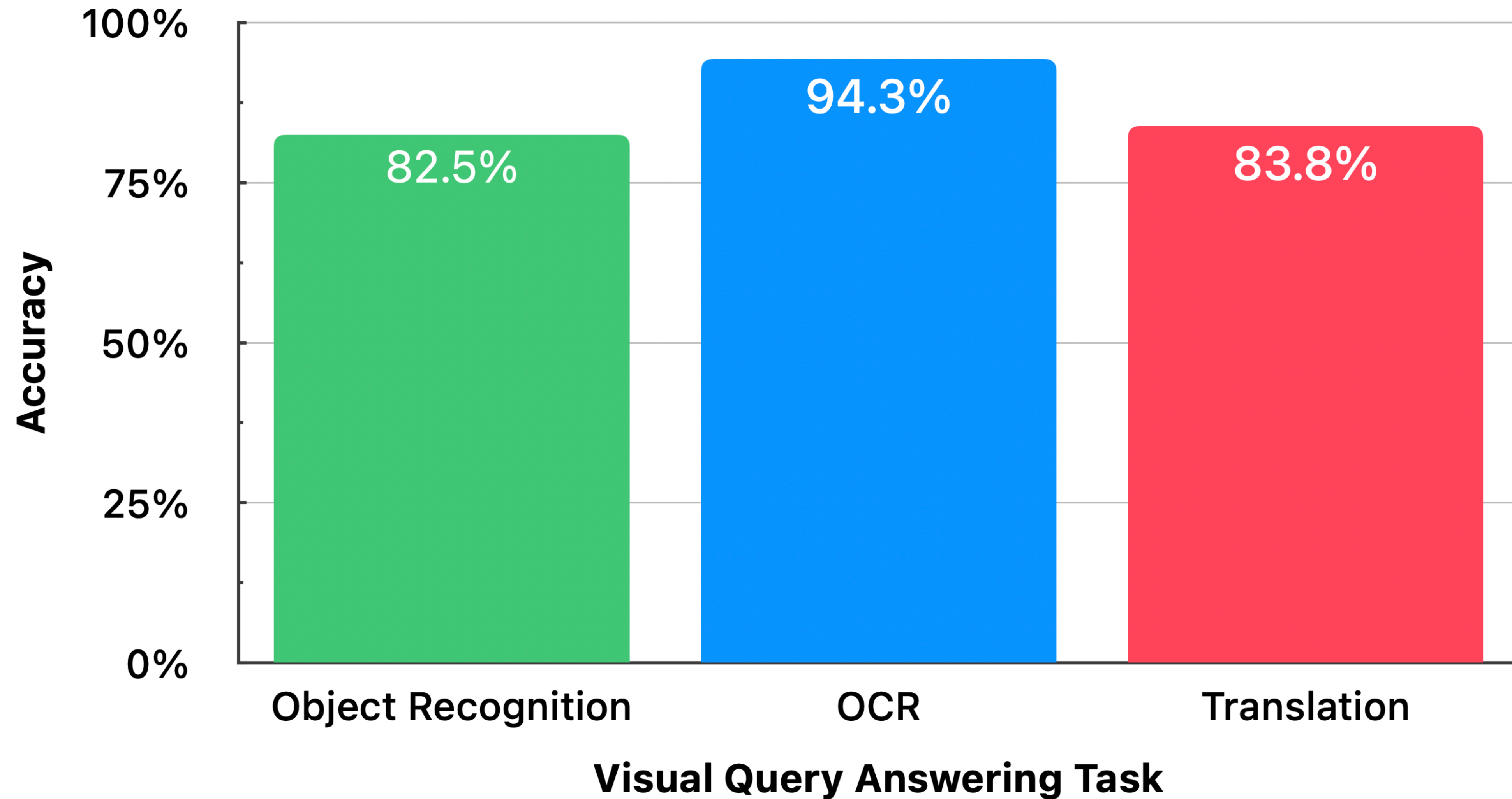
"answers where the AI guessed calorie content were rated poorly."

"prefer it to tell me what it's unconfident about rather than to be confidently wrong."

In-Person Evaluation

Stereo Images	Query	Response
	<p><i>"what are some songs off of this album?"</i></p>	<p>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."</p>
	<p><i>"what am I holding?"</i></p>	<p>I see you're holding a pair of scissors in your right hand.</p>
	<p><i>"what am I holding?"</i></p>	<p>I see a box of Cheez-It crackers in the foreground. It looks like you're holding it.</p>
	<p><i>"can you translate this for me?"</i></p>	<p>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.</p>
	<p><i>"what book is this?"</i></p>	<p>I see a book titled "Data Structures in C++ Using the Standard Template Library" by Timothy Budd.</p>
	<p><i>"who is this book by?"</i></p>	<p>I see the book in the image is titled "What's Our Problem?" and it's written by Tim Urban.</p>

VueBuds User Study | In-person evaluation (n=16, 130 total trials)



Most users reported that VueBuds **felt similar to regular earbuds**. Visual similarity with regular earbuds emerged as a common theme.

Predominantly **positive reactions** with several expressing enthusiasm.

Privacy concerns if streaming to the cloud.

Many favored VueBuds **over phone-ChatGPT workflows**, while some favored phone style workflows for more complex questions.

Translation and travel emerged as the prominent application.

"it feels like wearing normal earbuds"

"very fascinating idea ... similar to meta glasses but probably more convenient"

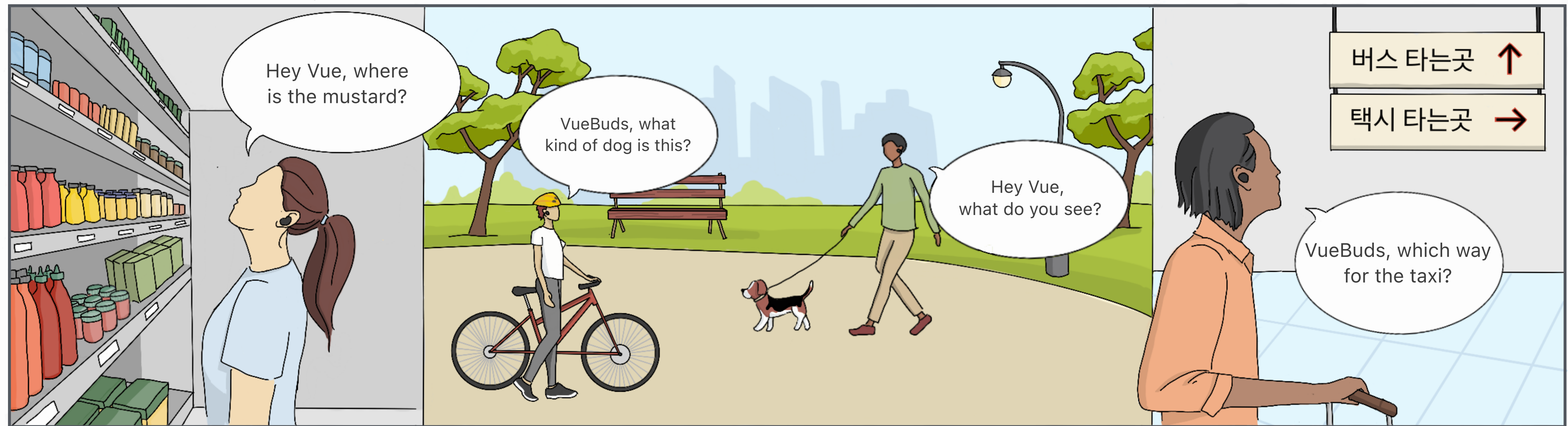
"could put my hair down and [to] stop it from recording any visual data ... can't be done with glasses."

"when I'm running or biking, it's a lot easier than taking out my phone, etc."

"favor phone for problems that require back and forth"

"traveling in a foreign country with language barriers"

VueBuds | Conclusion

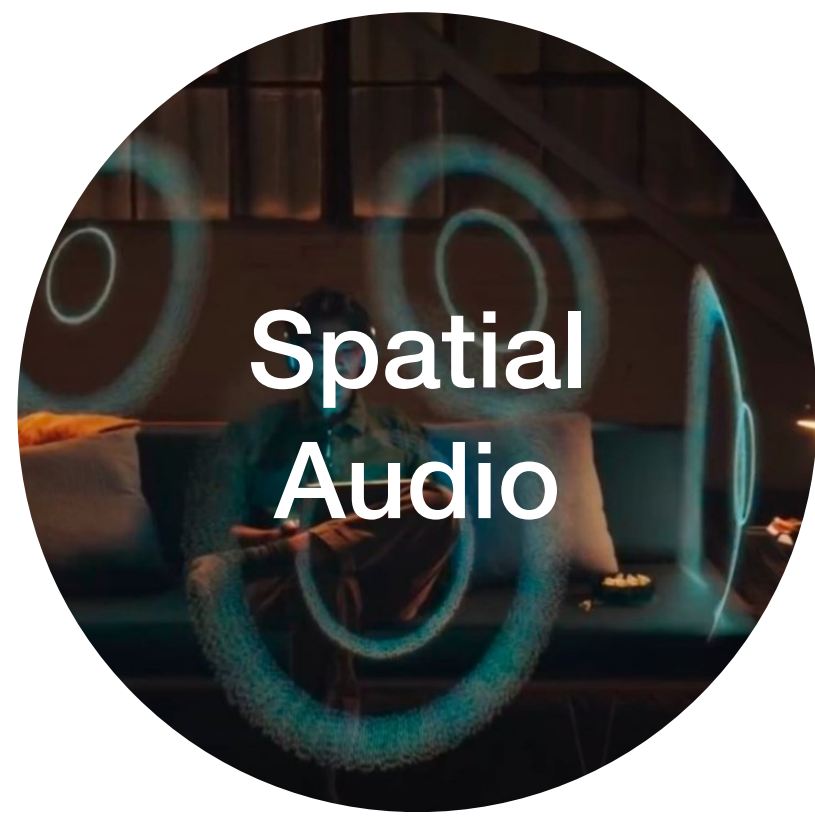


First wireless earbuds to integrate forward-facing cameras, without compromising size, weight, and power

With stereo vision, ear-level cameras are capable of egocentric vision applications, and can utilize stitching to reduce redundant information

Extensive user studies demonstrating comparable performance to RayBan Meta, with robust operation across varying users and environments

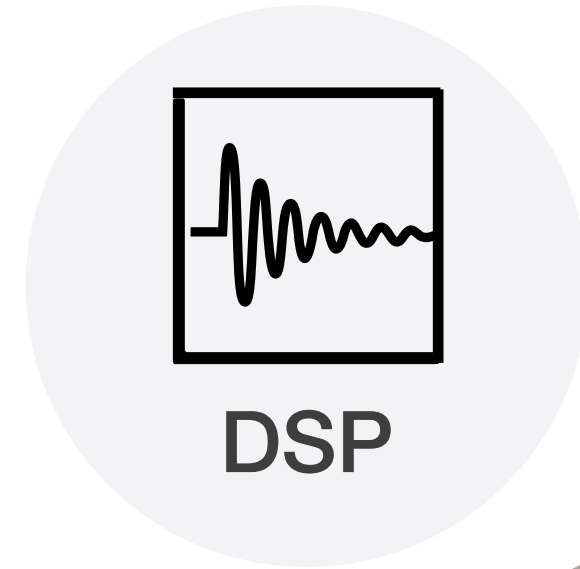
The Future of Intelligent Earbuds



Spatial
Audio



ANC and
Transparency



DSP



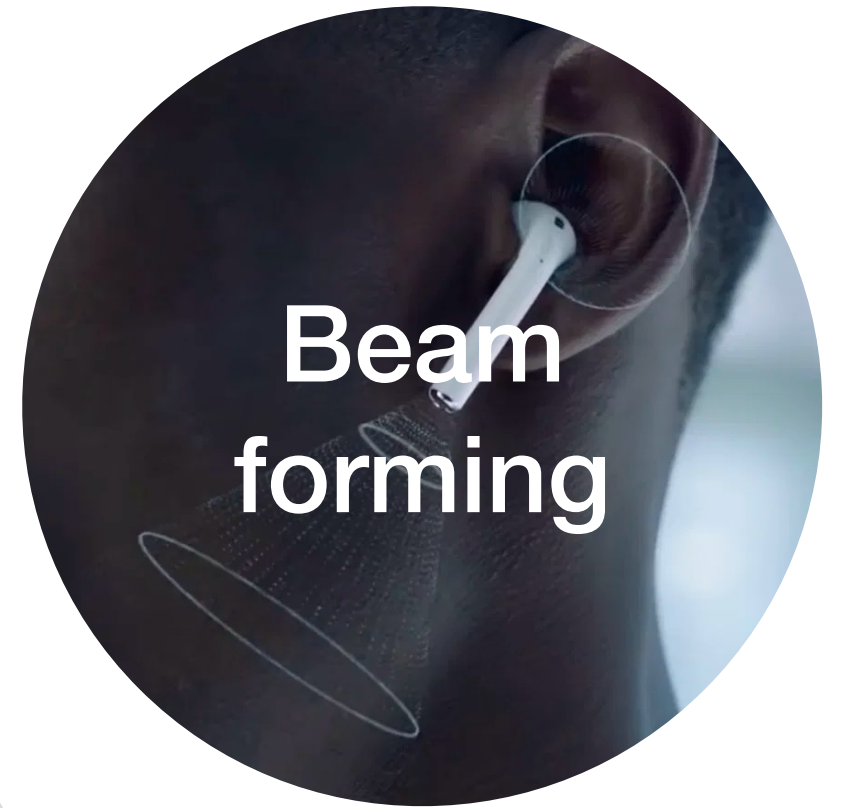
Speakers



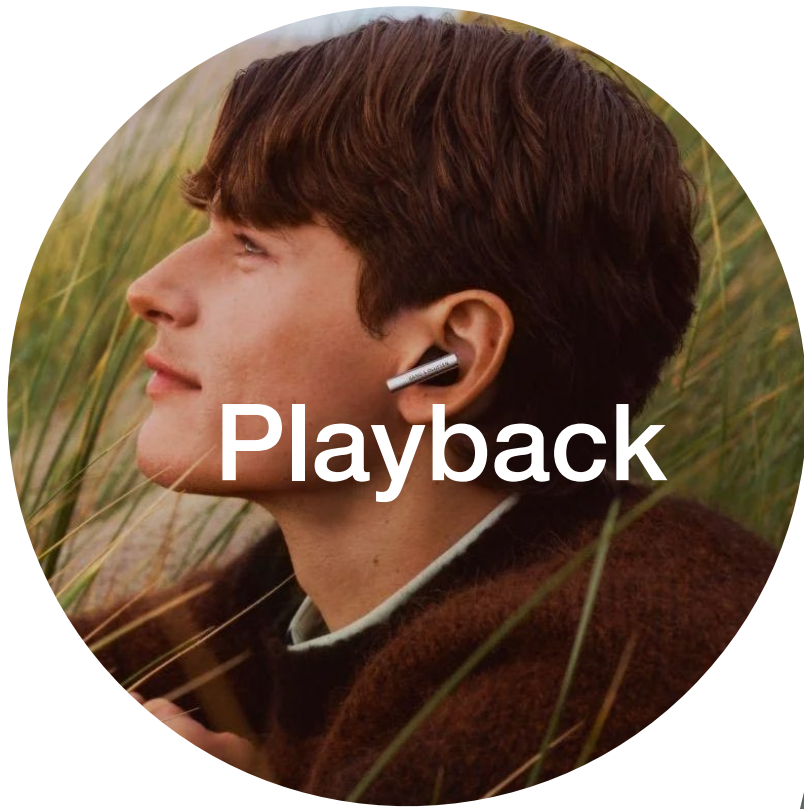
Head
Tracking



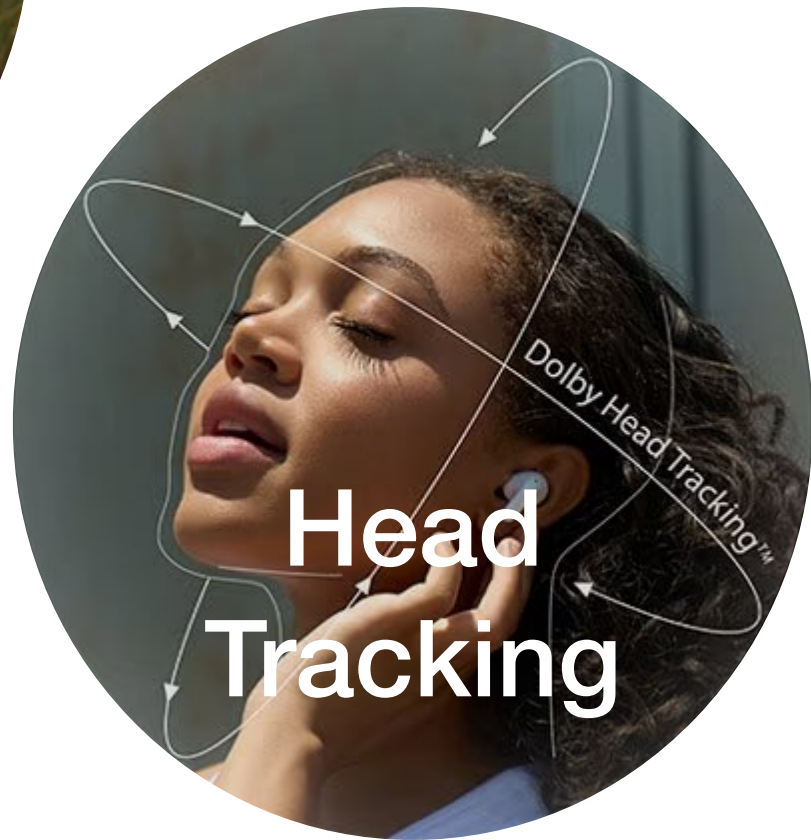
Telephony



Beam
forming



Playback



Head
Tracking



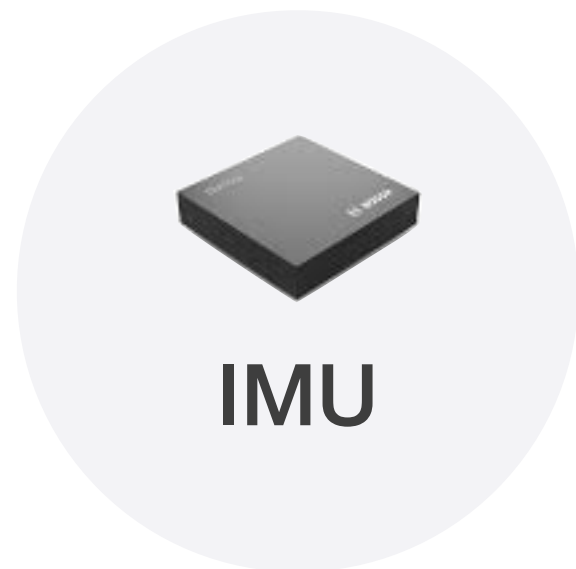
Mics



Voice
Isolation



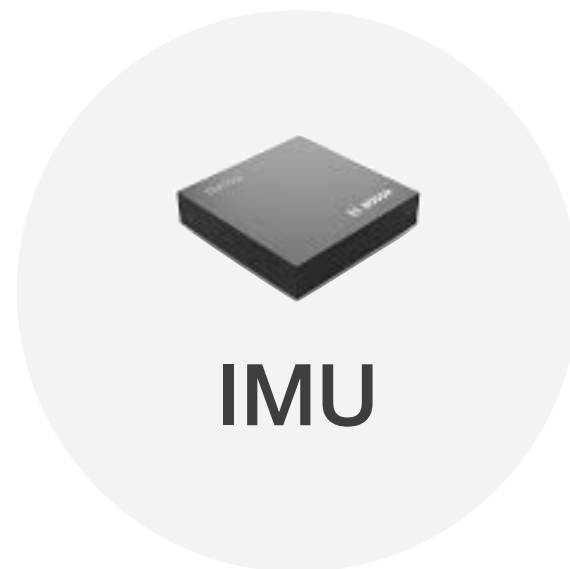
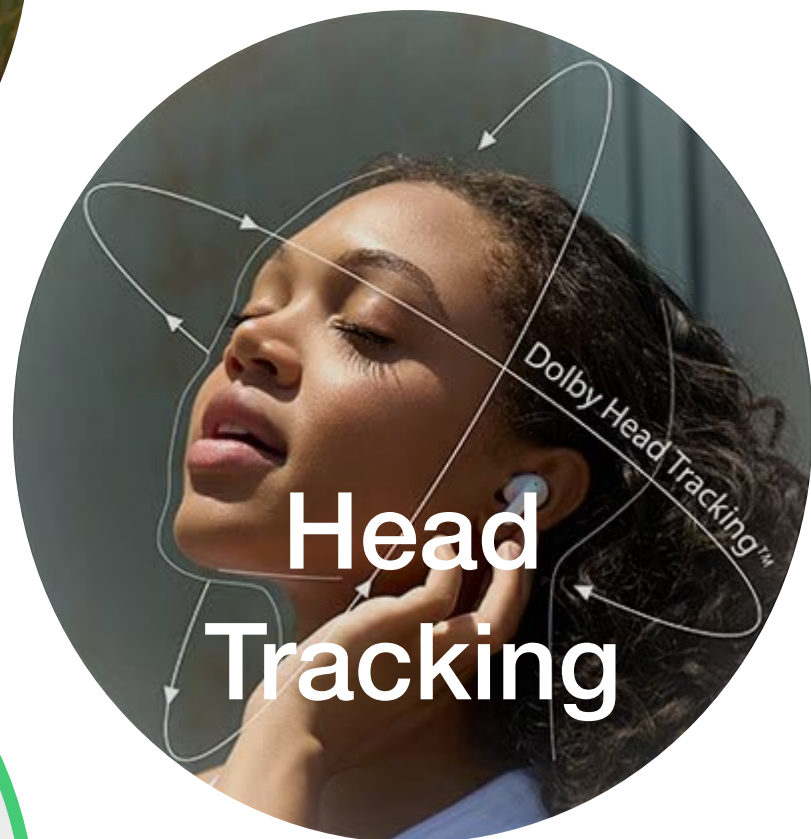
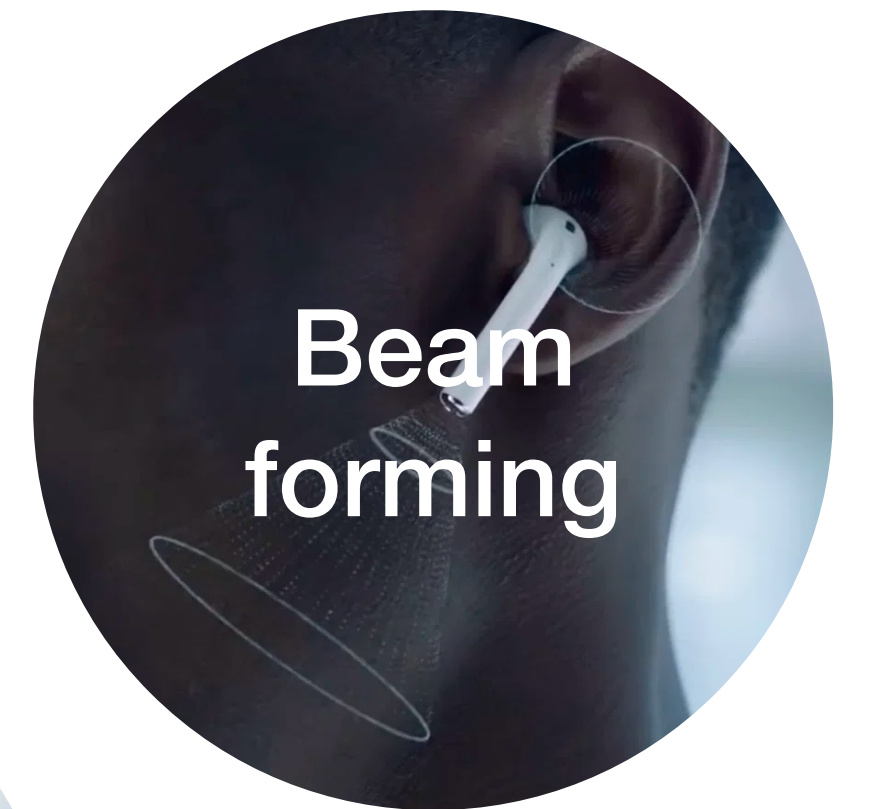
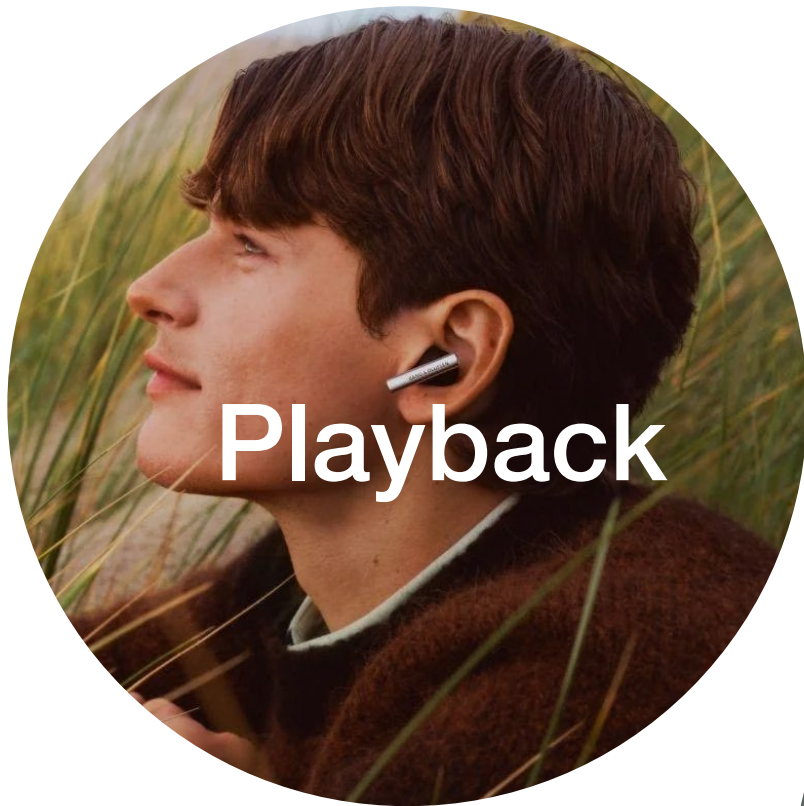
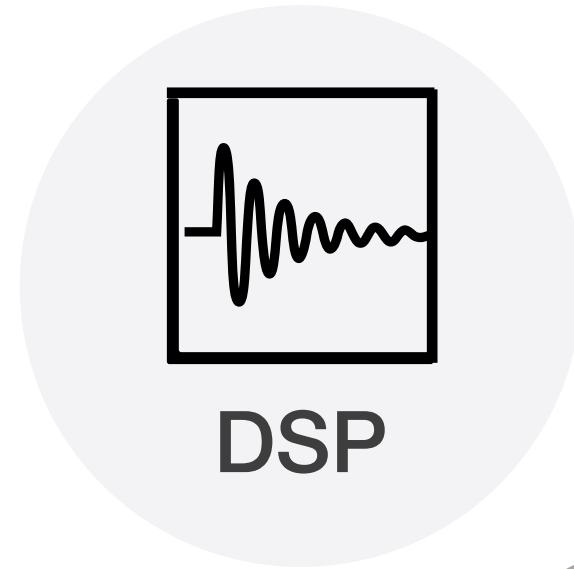
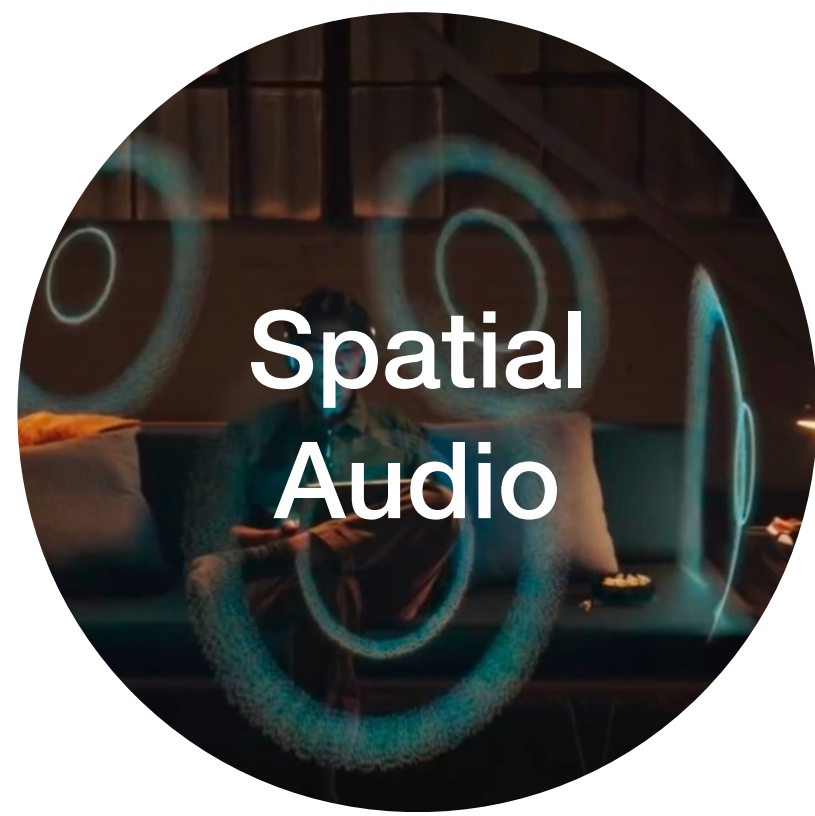
Input
Gestures



IMU



Voice
Controls



Spatial Video Capture

Audio Ray Tracing

Context Aware ANC

Proactive Intelligence

Turn-based Visual Intelligence

Depth Estimation

Spatial Audio

ANC and Transparency

DSP

Speakers

Telephony

Beam forming

Hello
Olá
Bonjour
Ciao
Guten Tag
привет
नमस्ते
ابج

Playback

Head Tracking

Mics

Voice Isolation

Cameras + Stereo Vision

Input Gestures

IMU

Voice Controls

Photo + Video Capture

Hand Tracking

Multimodal Augmented Audio

Cognitive Offloading

Thank you!



ClearBuds (MobiSys 2022)
Synchronized Wireless Earbuds
for Speech Enhancement



IRIS (UIST 2024)
Camera-Integrated Ring for
Vision-Based Interaction



VueBuds (CHI 2026)
Visual Intelligence with
Wireless Earbuds