GazePointAR

A Context-Aware Multimodal Voice Assistant for Pronoun Disambiguation in Wearable Augmented Reality

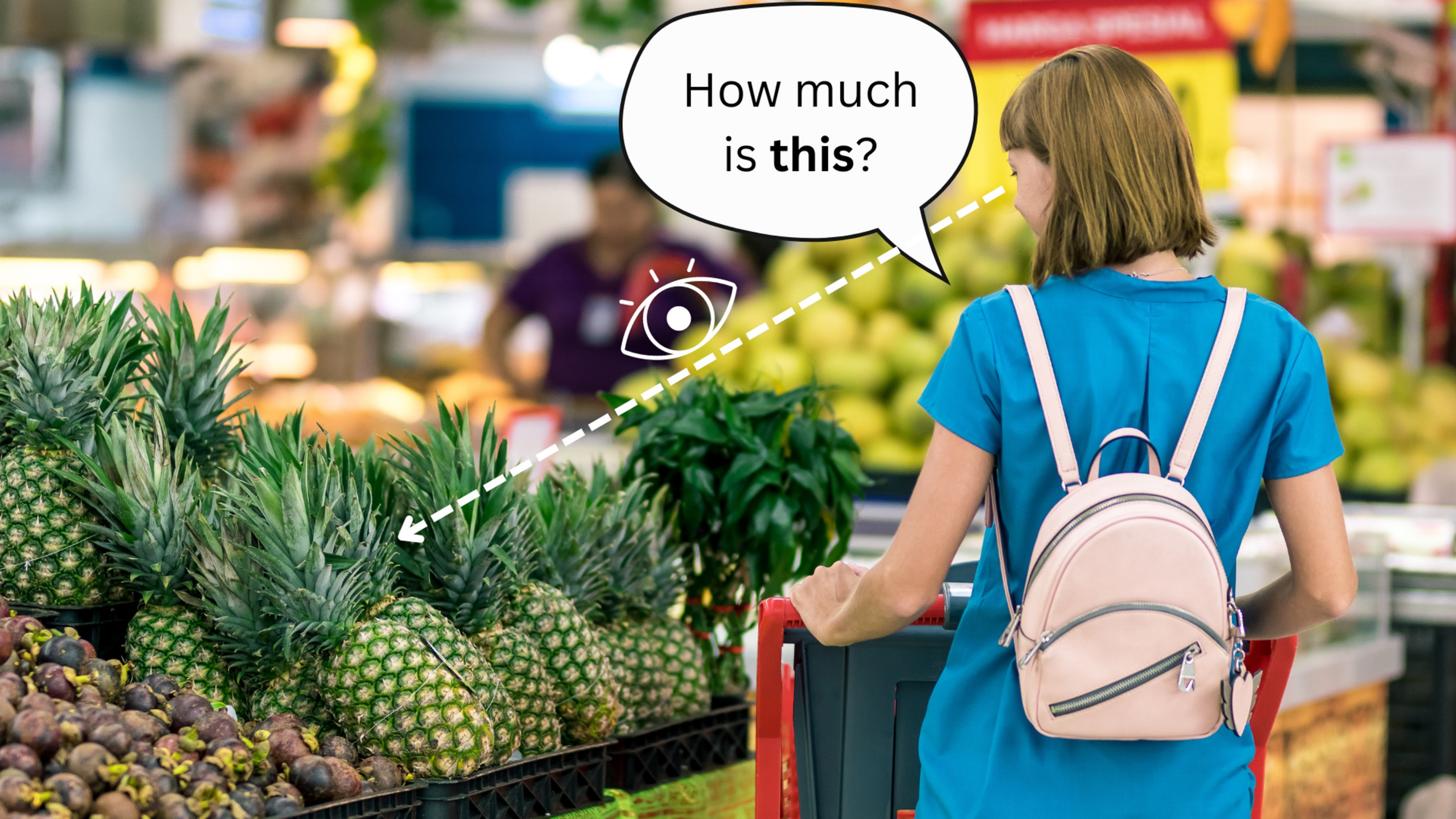
Jaewook Lee[1], Jun Wang[1], Elizabeth Brown[1], Liam Chu[1],
Sebastian S. Rodriguez[2], and Jon E. Froehlich[1]
University of Washington[1], University of Illinois at Urbana-Champaign[2]

MAKEABILITY LAB

PAUL G. ALLEN SCHOOL
OF COMPUTER SCIENCE & ENGINEERING

UNIVERSITY OF
ILLINOIS
URBANA-CHAMPAIGN

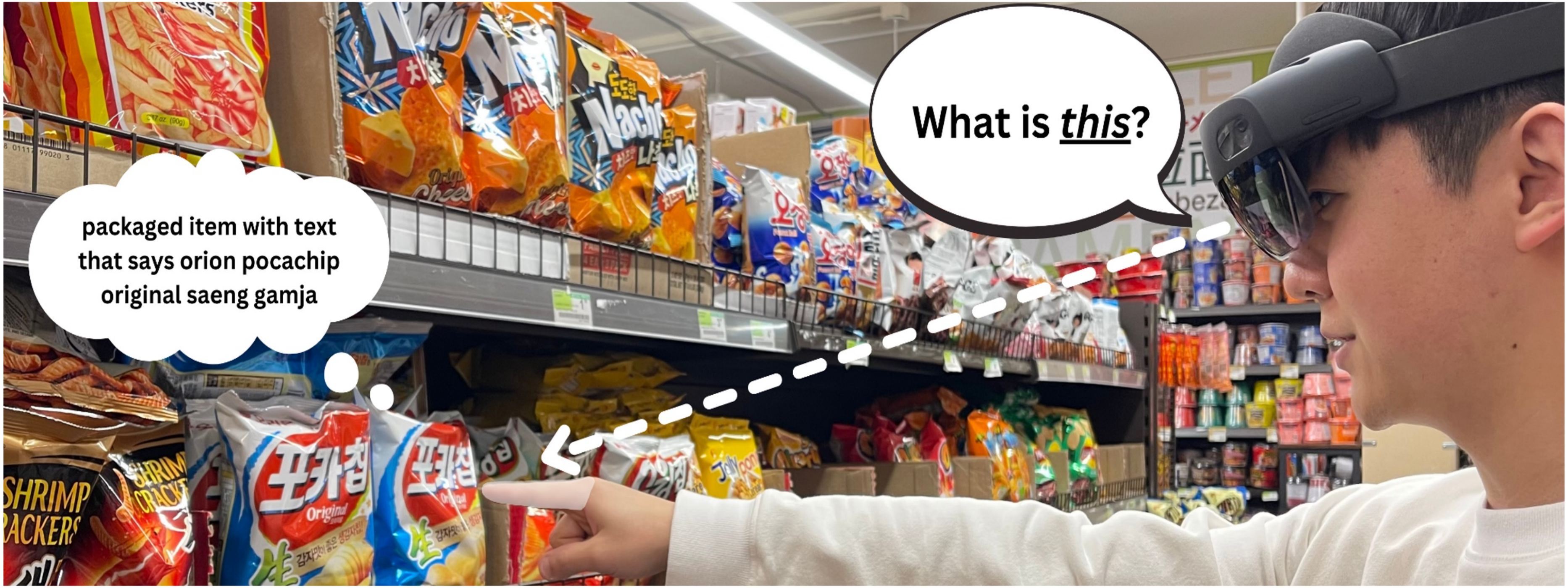
How much is **this**?

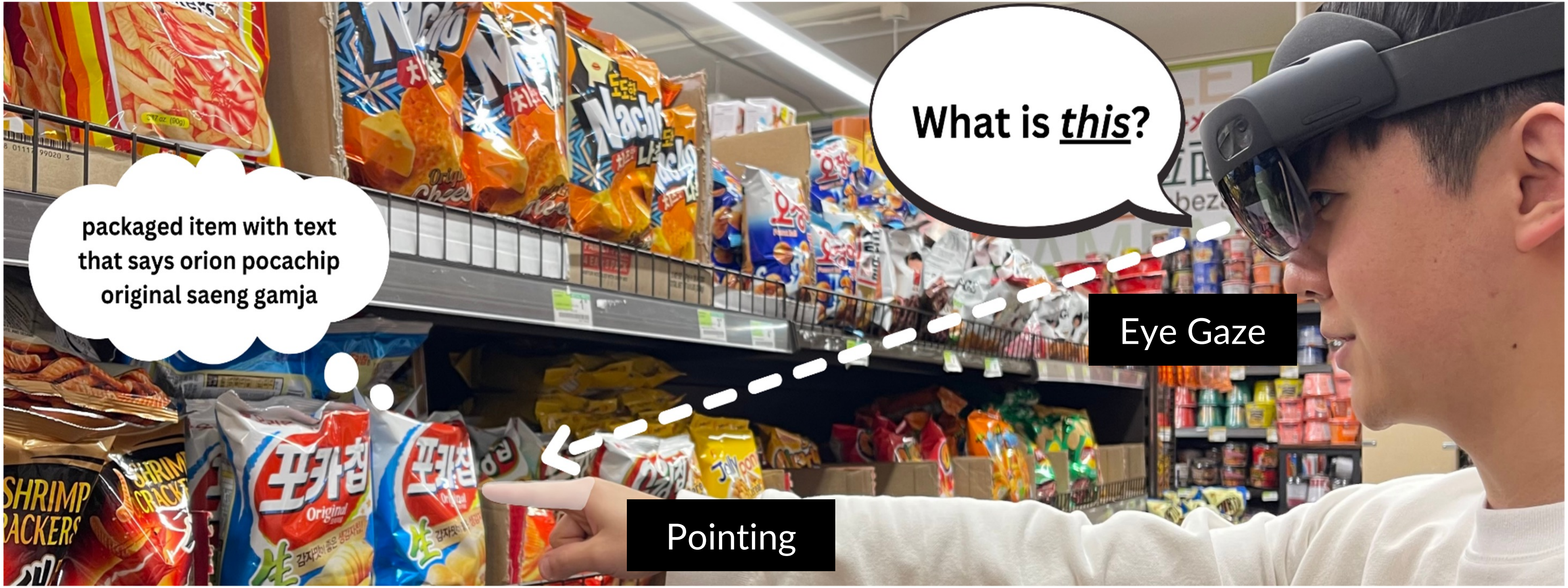


How much is **this**?

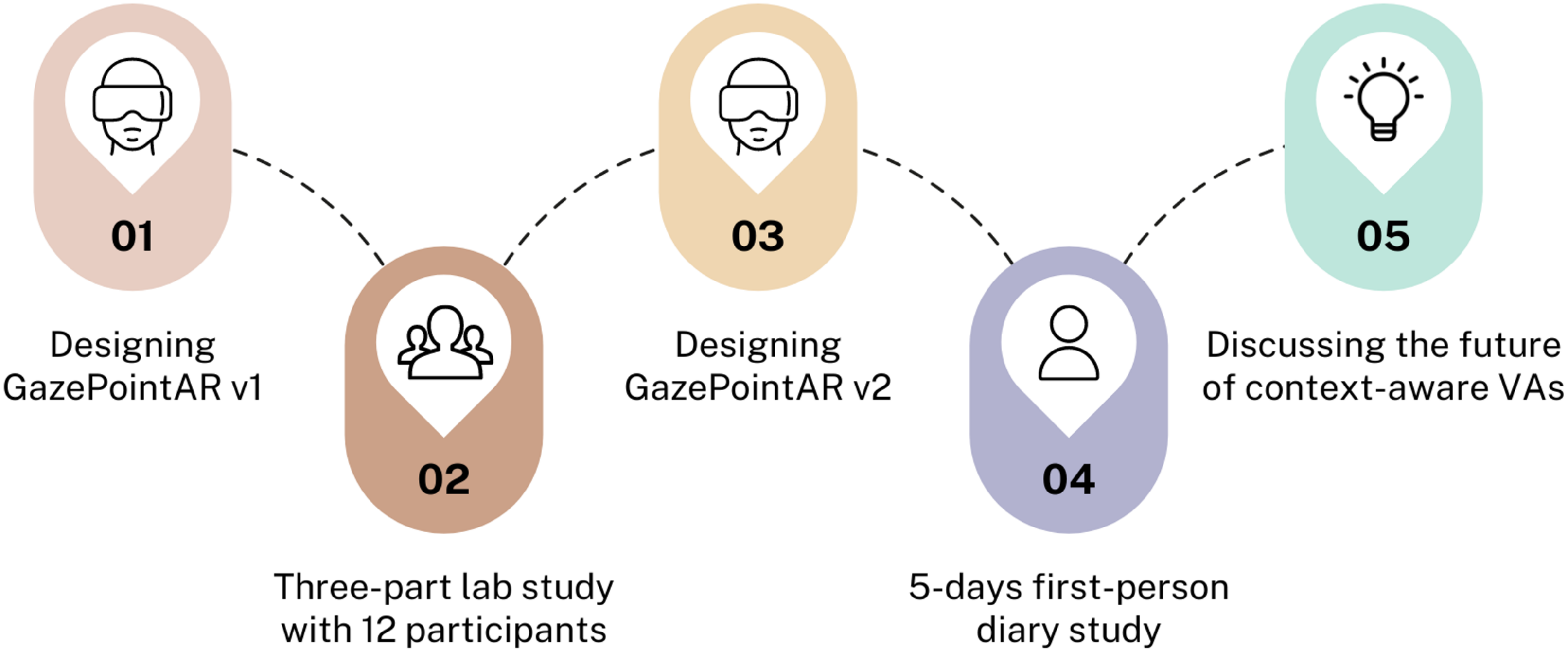
How much is **this**?

But voice assistants don’t understand “**this**”


What is *this*?
packaged item with text that says orion pocachip original saeng gamja


What is *this*?
packaged item with text that says orion pocachip original saeng gamja
Eye Gaze
Pointing

GazePointAR Timeline



01

Designing GazePointAR v1

02

Three-part lab study with 12 participants

03

Designing GazePointAR v2

04

5-days first-person diary study

05

Discussing the future of context-aware VAs

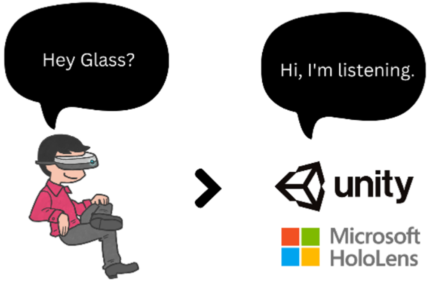
GazePointAR Timeline

01

Designing GazePointAR v1

GazePointAR v1

GazePointAR v1



GazePointAR v1

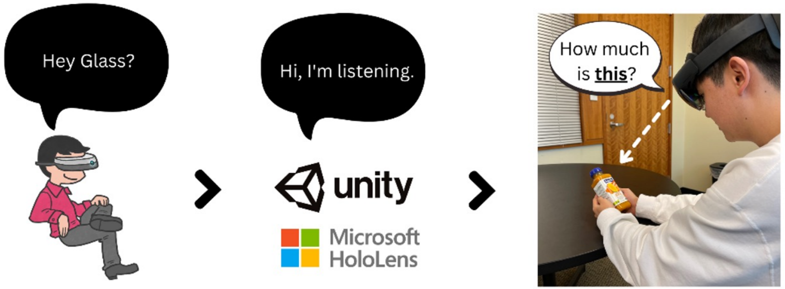
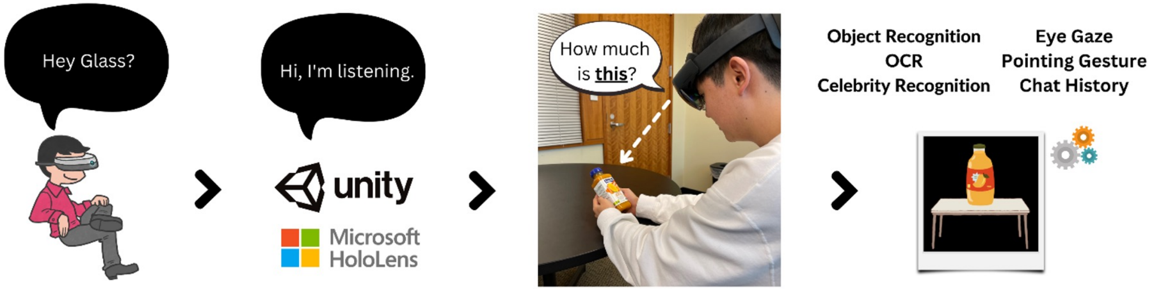


Image Capture: 2.27s ± 0.16

GazePointAR v1



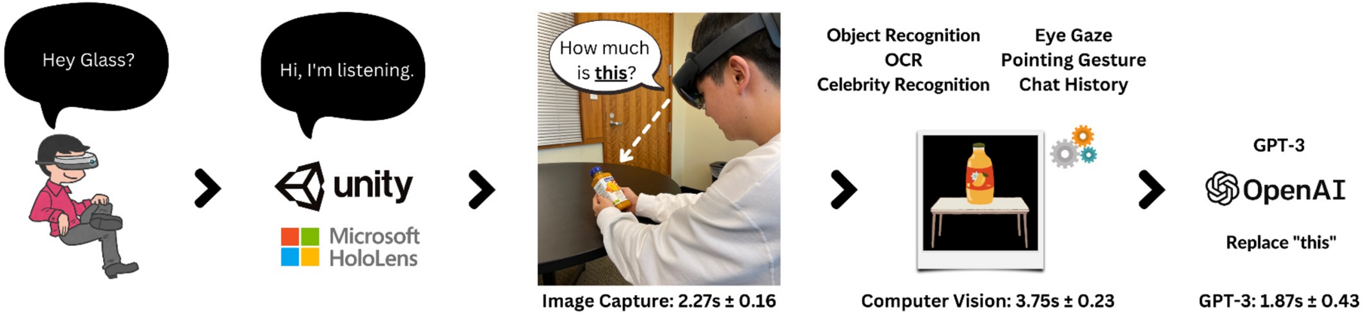
Object Recognition
OCR
Celebrity Recognition

Eye Gaze
Pointing Gesture
Chat History

Image Capture: 2.27s ± 0.16

Computer Vision: 3.75s ± 0.23

GazePointAR v1



Object Recognition
OCR
Celebrity Recognition

Eye Gaze
Pointing Gesture
Chat History

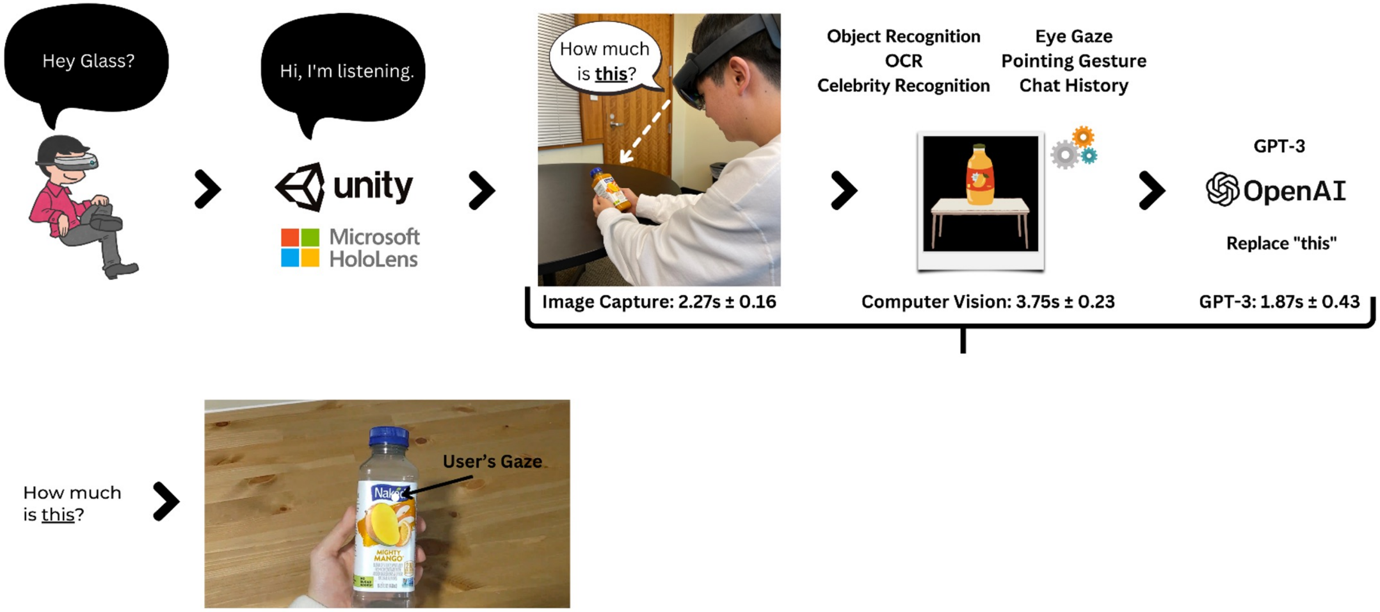
Replace "this"

Image Capture: 2.27s ± 0.16

Computer Vision: 3.75s ± 0.23

GPT-3: 1.87s ± 0.43

GazePointAR v1



Hey Glass?

Hi, I'm listening.

unity

Microsoft HoloLens

How much is this?

Object Recognition
OCR
Celebrity Recognition

Eye Gaze
Pointing Gesture
Chat History

GPT-3

OpenAI

Replace "this"

Image Capture: 2.27s ± 0.16

Computer Vision: 3.75s ± 0.23

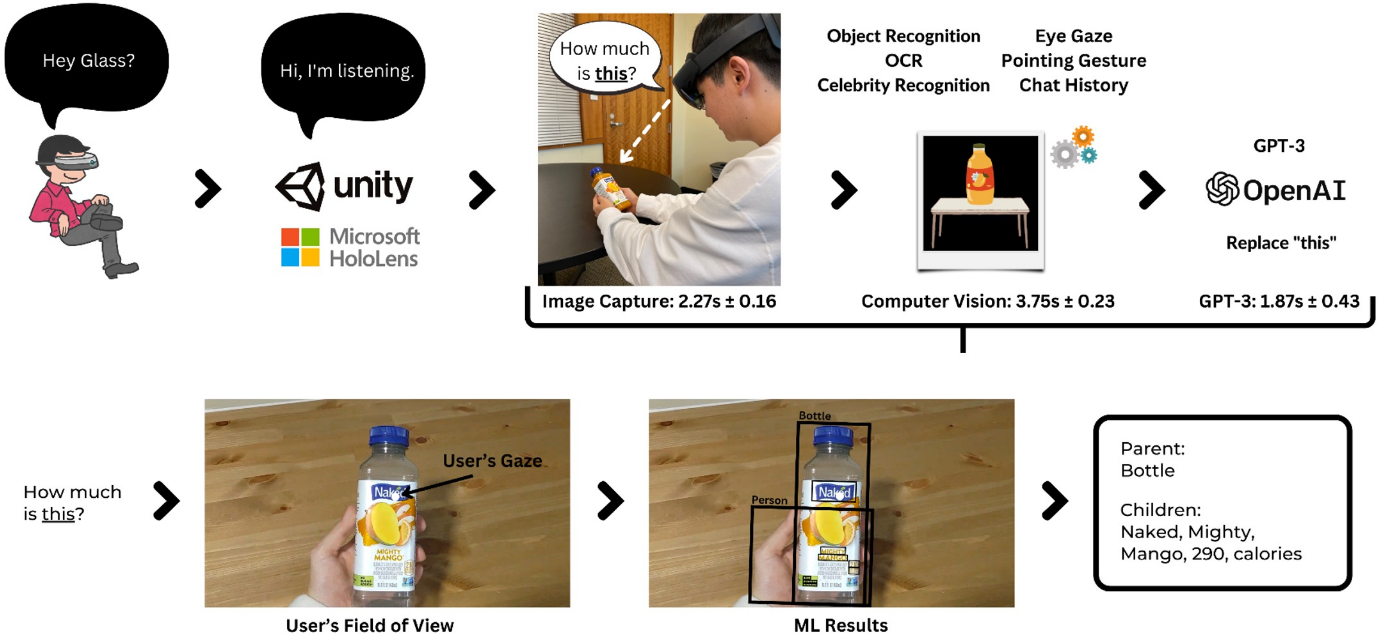
GPT-3: 1.87s ± 0.43

How much is this?

User's Gaze

User's Field of View

GazePointAR v1



Hey Glass?

Hi, I'm listening.

unity

Microsoft HoloLens

How much is this?

Object Recognition
OCR
Celebrity Recognition

Eye Gaze
Pointing Gesture
Chat History

GPT-3
OpenAI
Replace "this"

Image Capture: 2.27s ± 0.16

Computer Vision: 3.75s ± 0.23

GPT-3: 1.87s ± 0.43

How much is this?

User's Gaze

User's Field of View

Bottle

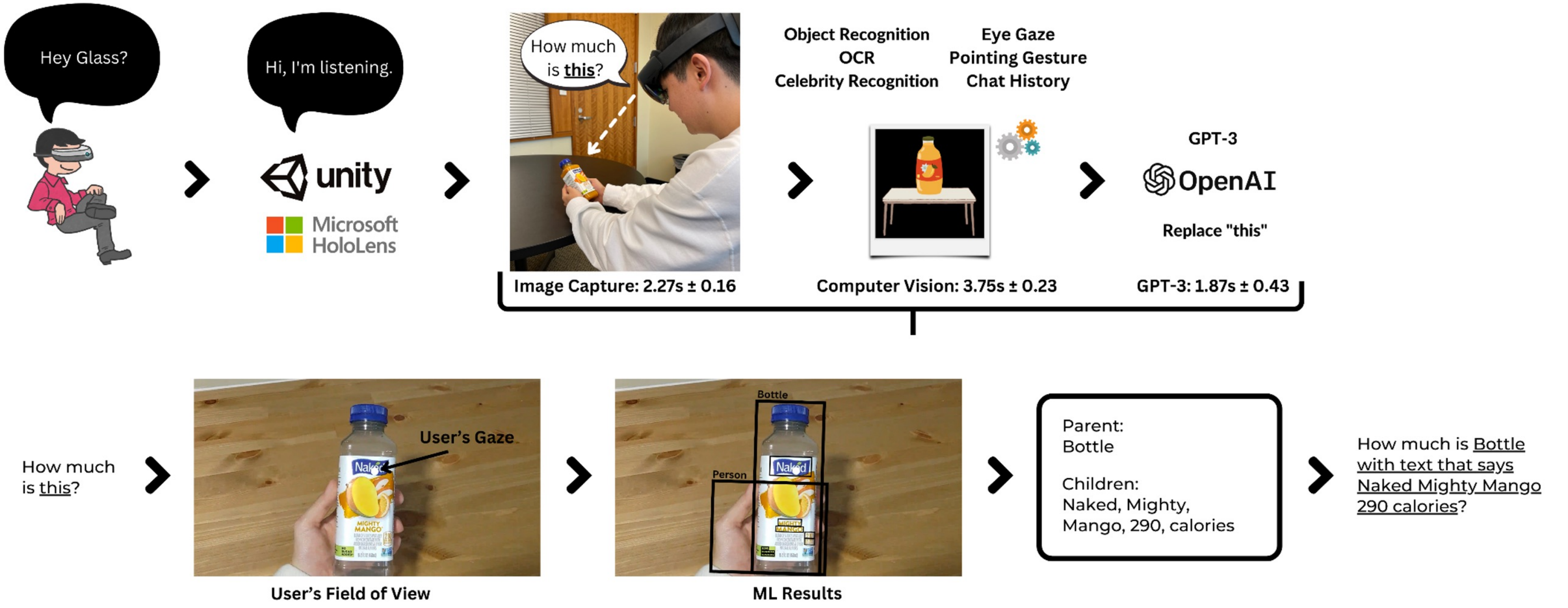
Person

ML Results

Parent:
Bottle

Children:
Naked, Mighty,
Mango, 290, calories

GazePointAR v1



Object Recognition
OCR
Celebrity Recognition

Eye Gaze
Pointing Gesture
Chat History

GPT-3

Replace "this"

Image Capture: 2.27s ± 0.16

Computer Vision: 3.75s ± 0.23

GPT-3: 1.87s ± 0.43

How much is this?

User's Field of View

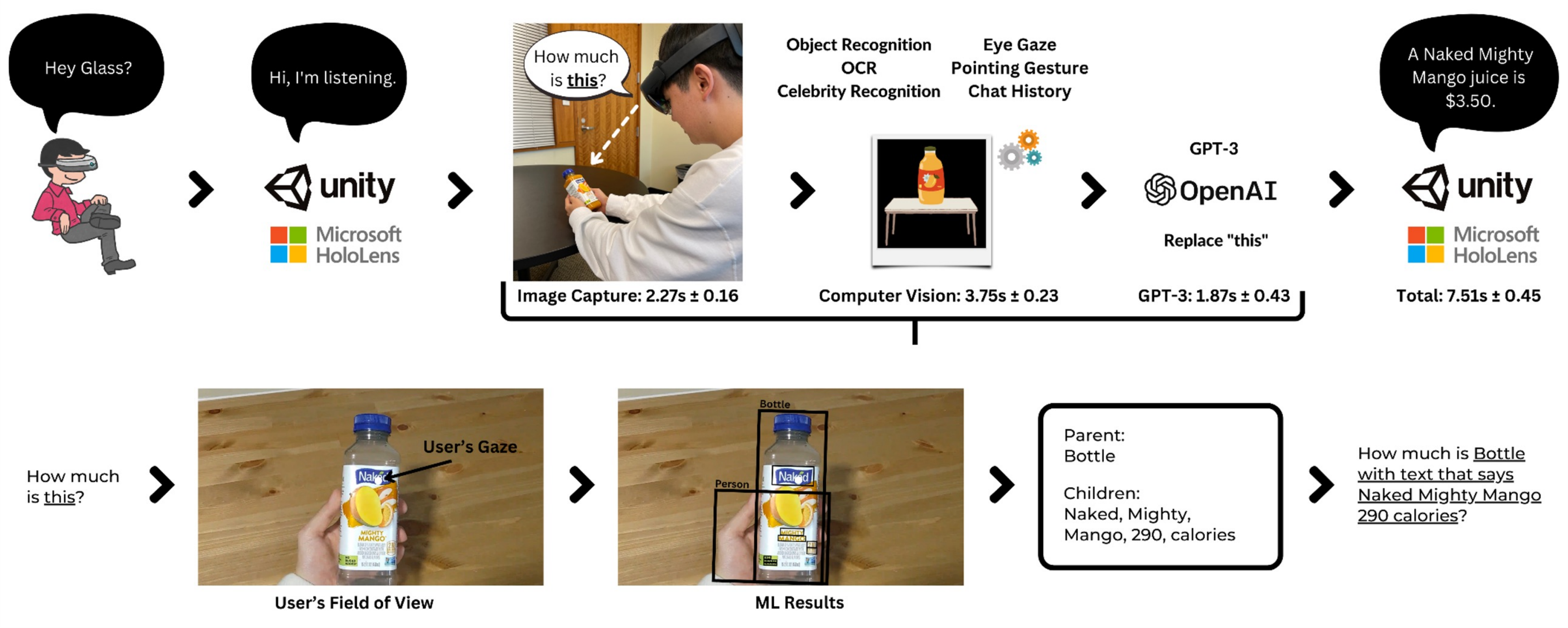
ML Results

Parent:
Bottle

Children:
Naked, Mighty,
Mango, 290, calories

How much is Bottle with text that says Naked Mighty Mango 290 calories?

GazePointAR v1



Hey Glass?

Hi, I'm listening.

How much is this?

Object Recognition
OCR
Celebrity Recognition

Eye Gaze
Pointing Gesture
Chat History

GPT-3

Replace "this"

A Naked Mighty Mango juice is $3.50.

Image Capture: 2.27s ± 0.16

Computer Vision: 3.75s ± 0.23

GPT-3: 1.87s ± 0.43

Total: 7.51s ± 0.45

How much is this?

User's Gaze

User's Field of View

Bottle

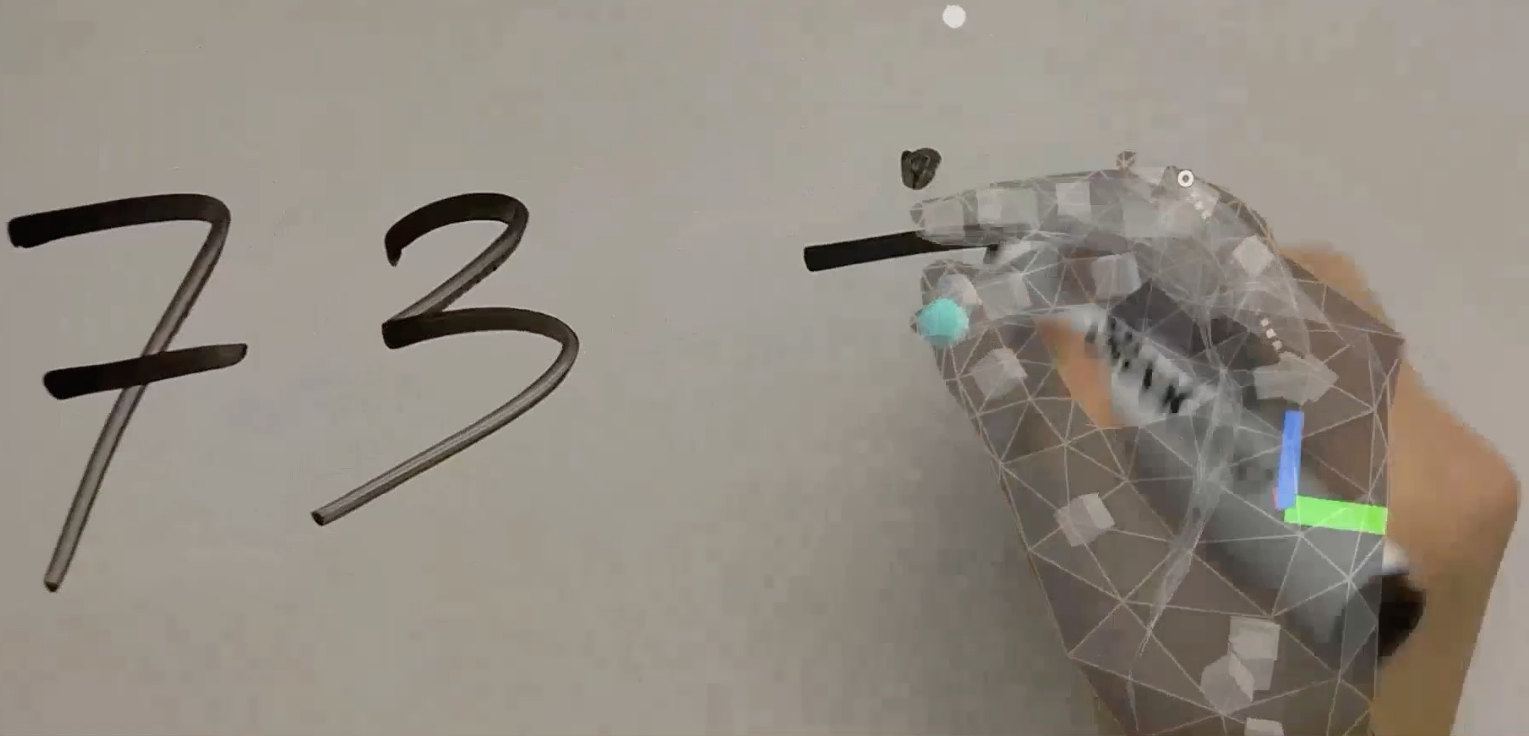
Person

ML Results

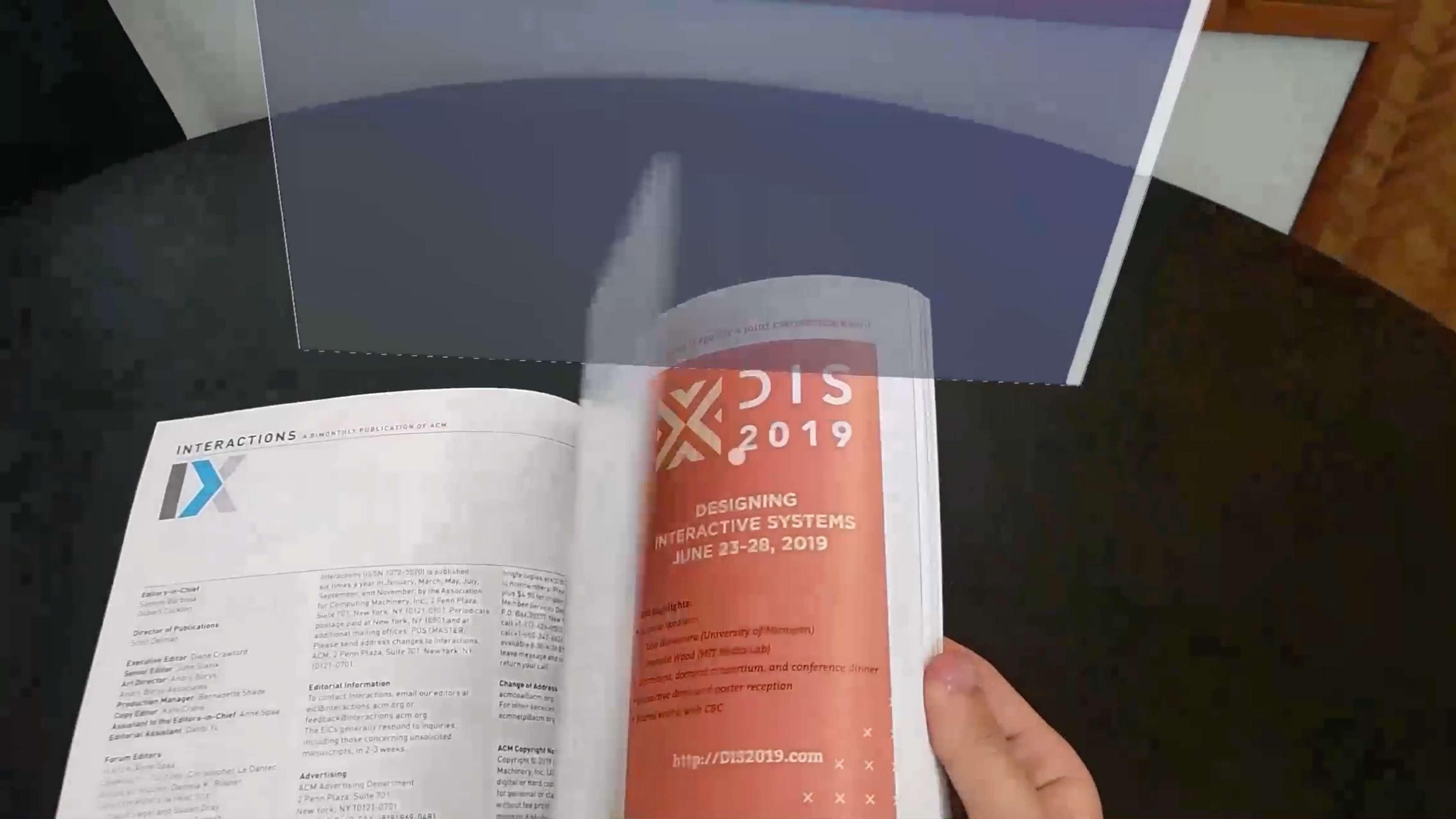
Parent:
Bottle

Children:
Naked, Mighty, Mango, 290, calories

How much is Bottle with text that says Naked Mighty Mango 290 calories?


73 -

Voice Assistant Question and Answer
LANDFILL
RECYCLING
COMPOST



INTERACTIONS A BIMONTHLY PUBLICATION OF ACM

Interactions (ISSN 1072-5520) is published six times a year in January, March, May, July, September, and November, by the Association for Computing Machinery, Inc., 2 Penn Plaza, Suite 701, New York, NY 10121-0701. Periodicals postage paid at New York, NY 10001 and at additional mailing offices. POSTMASTER: Please send address changes to Interactions, ACM, 2 Penn Plaza, Suite 701, New York, NY 10121-0701.

Editorial Information
To contact *Interactions*, email our editors at eic@interactions.acm.org or feedback@interactions.acm.org
The EICs generally respond to inquiries, including those concerning unsolicited manuscripts, in 2-3 weeks.

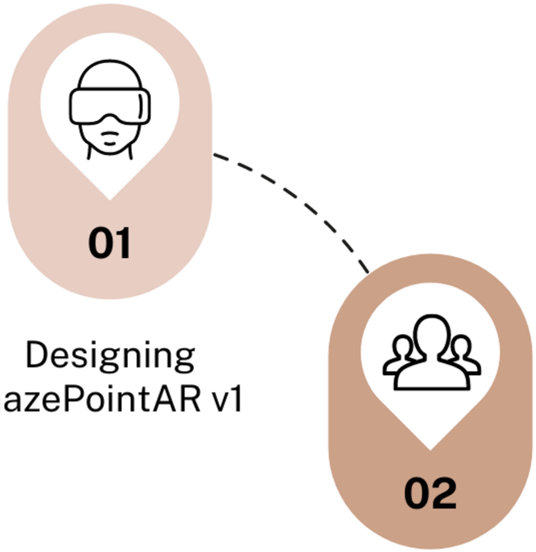
Advertising
ACM Advertising Department
2 Penn Plaza, Suite 701
New York, NY 10121-0701

Single copies are [illegible] to nonmembers [illegible] plus $4.95 for [illegible] Member Services [illegible] P.O. Box [illegible] New [illegible] call +1-212-626-0500 [illegible] call +1-800-342-6626 [illegible] available 8:30-4:30 [illegible] leave message and [illegible] return your call

Change of Address
acmcoa@acm.org
For other services [illegible]
acmhelp@acm.org

ACM Copyright No[illegible]
Copyright © 2019 [illegible] Machinery, Inc. [illegible] digital or hard cop[illegible] for personal or cla[illegible] without fee prov[illegible]

GazePointAR Timeline



01

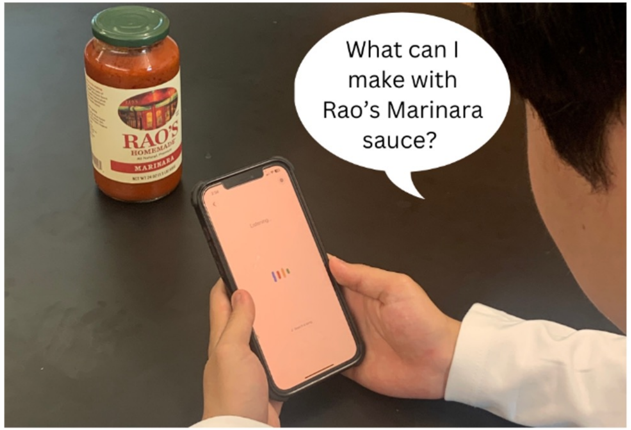
Designing GazePointAR v1

02

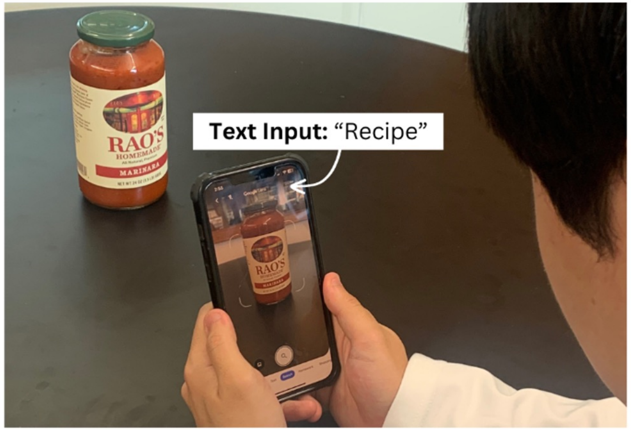
Three-part lab study with 12 participants

Part 1 - Comparing VAs

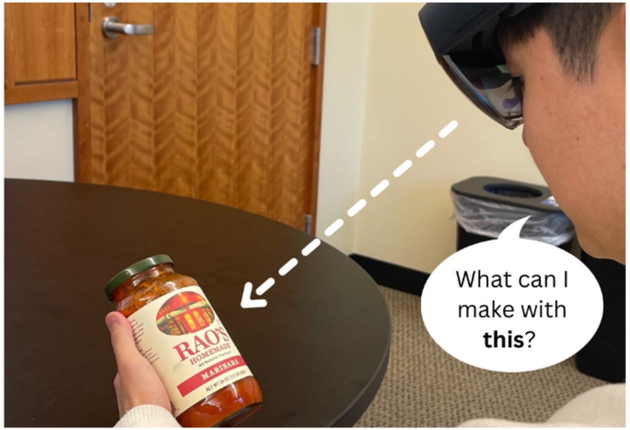
In Part 1, participants used Google VA, Google Lens, and GazePointAR to find a recipe that uses a specific pasta sauce.



Google VA



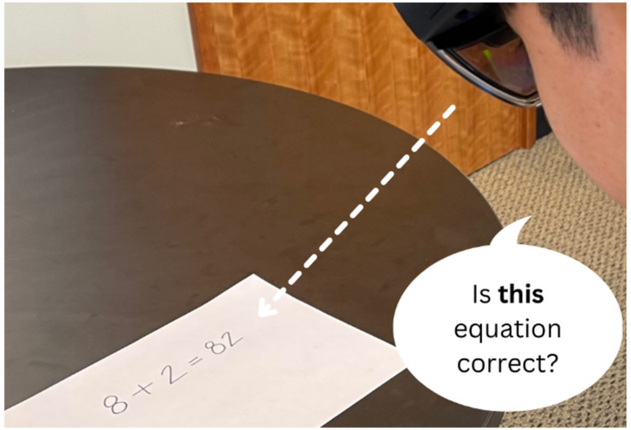
Google Lens



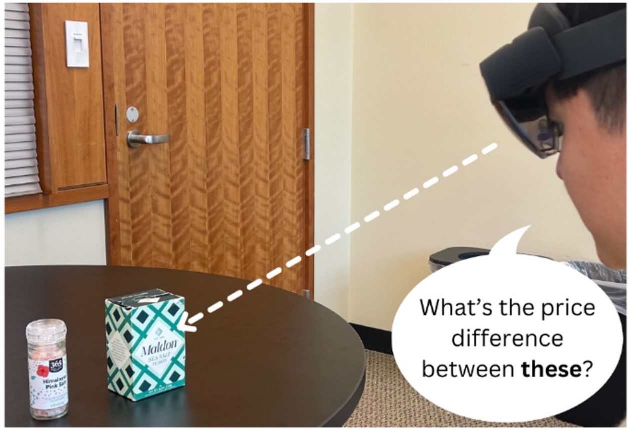
GazePointAR

Part 2 - Ambiguous Queries with GazePointAR

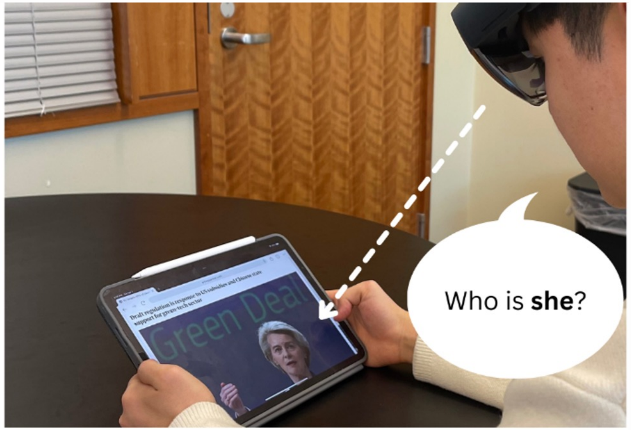
In Part 2, participants interacted with GazePointAR to complete three additional query tasks: math, price comparison, and celebrity search.



Math Task



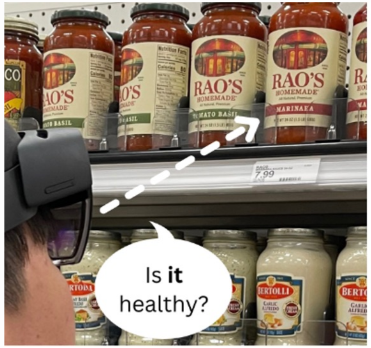
Price Comparison Task



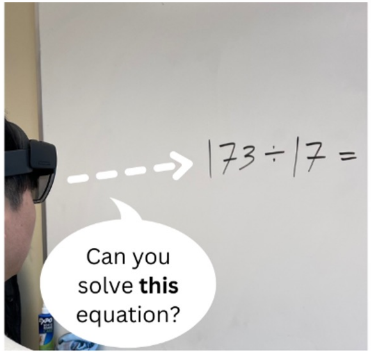
Celebrity Task

Part 3 - Design Probe & Co-Design

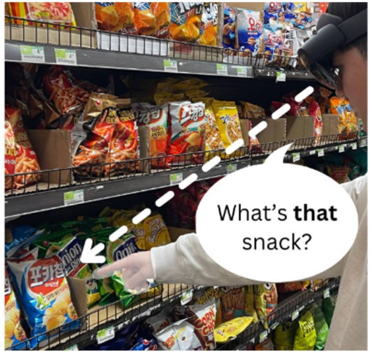
In Part 3, participants first watched five design probe videos, then brainstormed and tried their own context-sensitive queries.



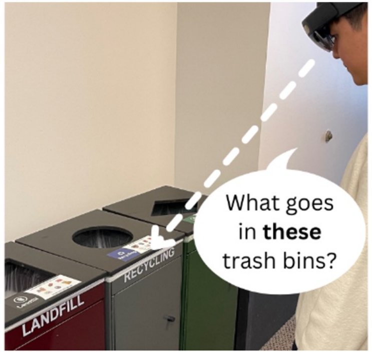
Cooking



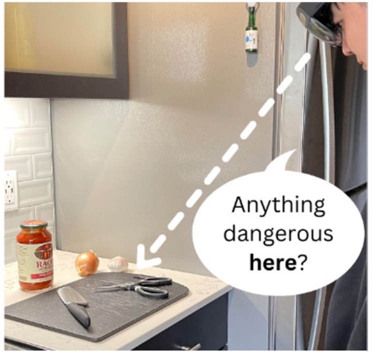
Math



Language Translation



Recycling

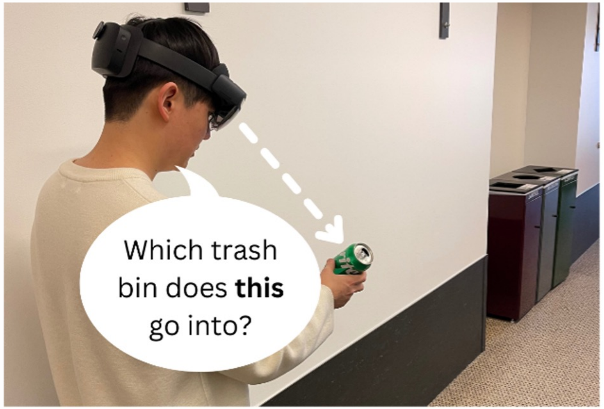


Accessibility

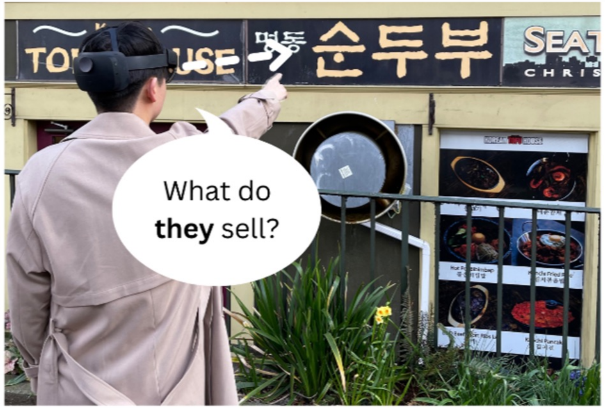
Key Findings

- GazePointAR is simple, fast, natural, and human-like.
- Participants preferred to speak pronouns, but not always.
- Pronouns are often for difficult-to-pronounce, long, or unknown object names.
- Gaze-only is preferred to keep interactions hands-free.

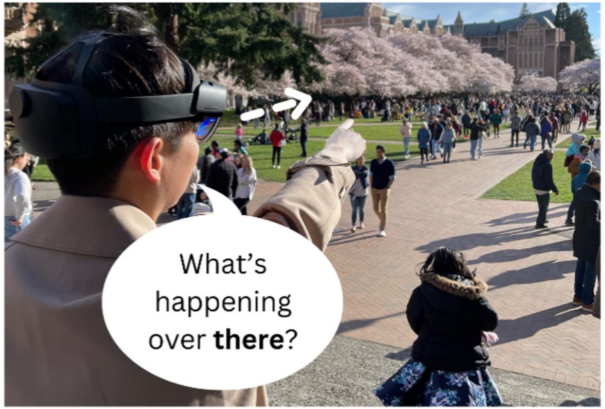
Answerable Queries



A: A Sprite can goes into recycling.

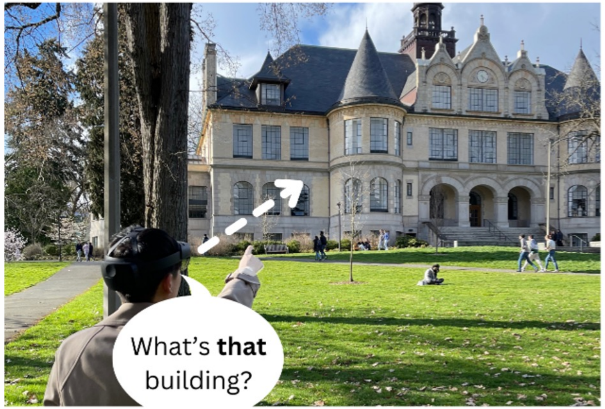


A: 명동순두부 sells sundubu jjigae...

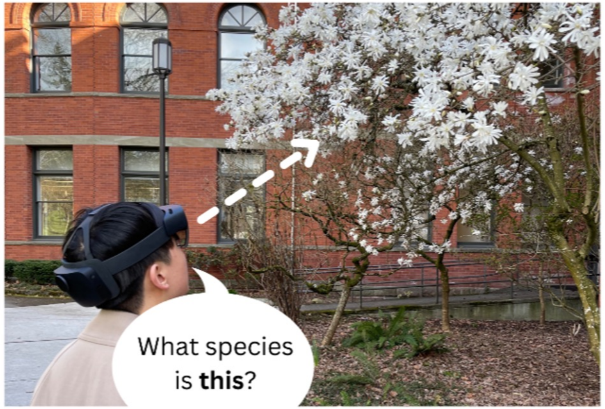


A: There are people around trees...

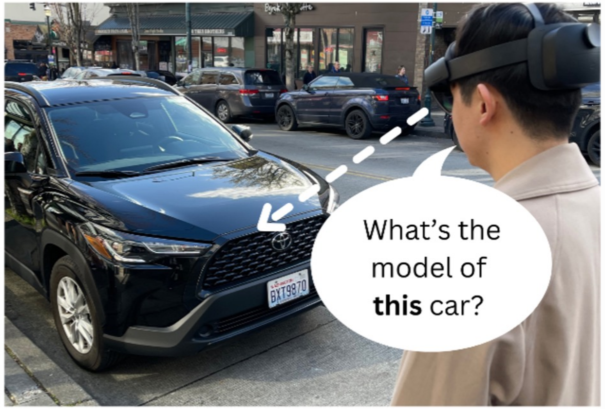
Unanswerable Queries



A: I didn't understand your question.



A: This is a tree...

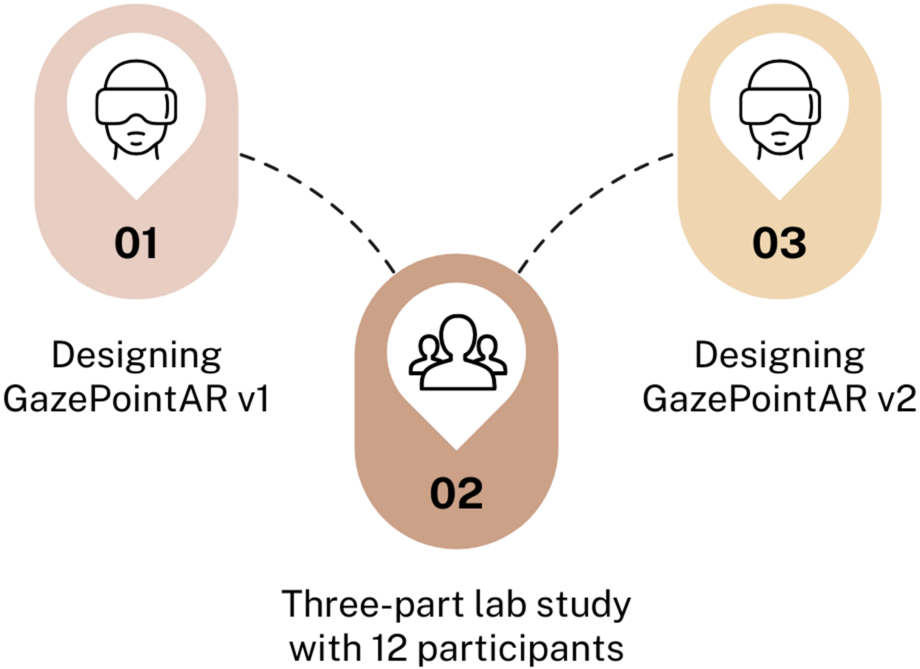


A: That depends on the car...

Key Limitations

- Support multiple pronouns (e.g., “Which is healthier, **this** or **that**?”).
- Support queries with no pronoun (e.g., “What would be good for dinner?”).
- Provide explanations to its answers.
- Reduce having to dwell.

GazePointAR Timeline



01

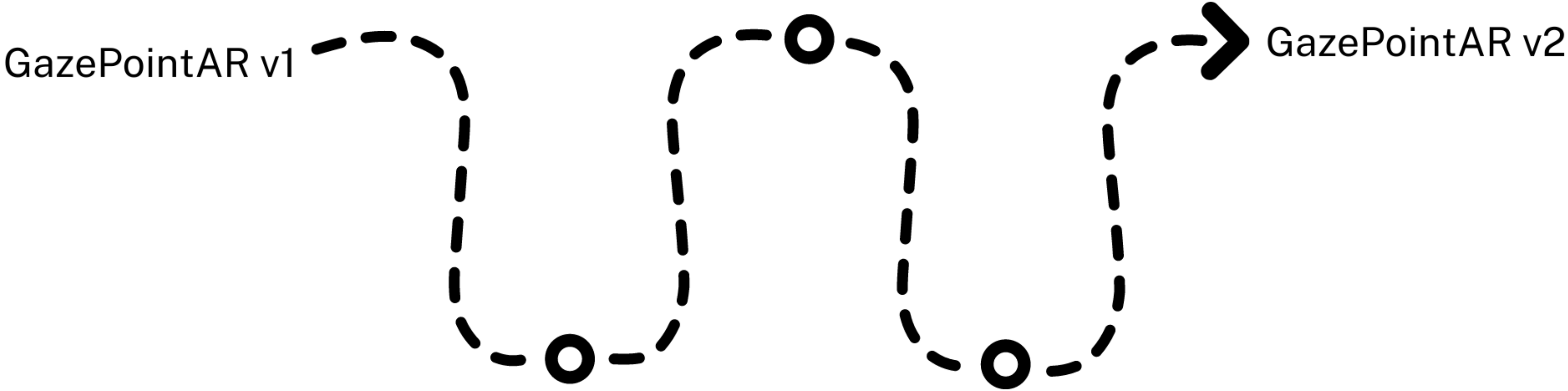
Designing GazePointAR v1

02

Three-part lab study with 12 participants

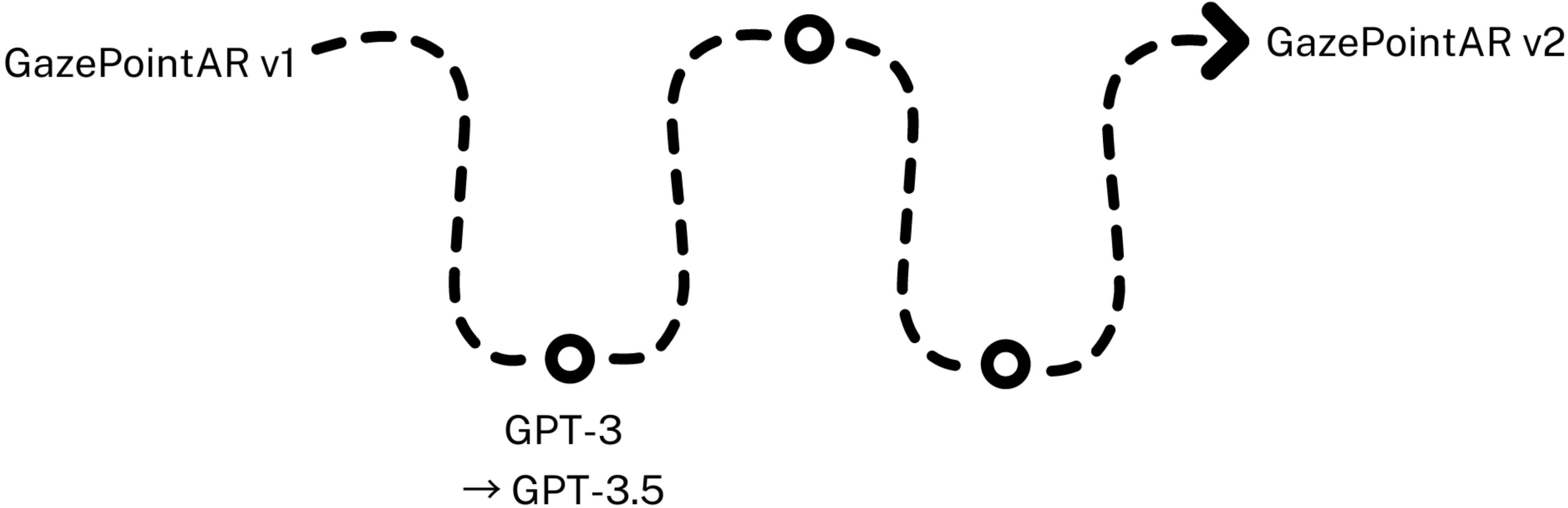
03

Designing GazePointAR v2



GazePointAR v1

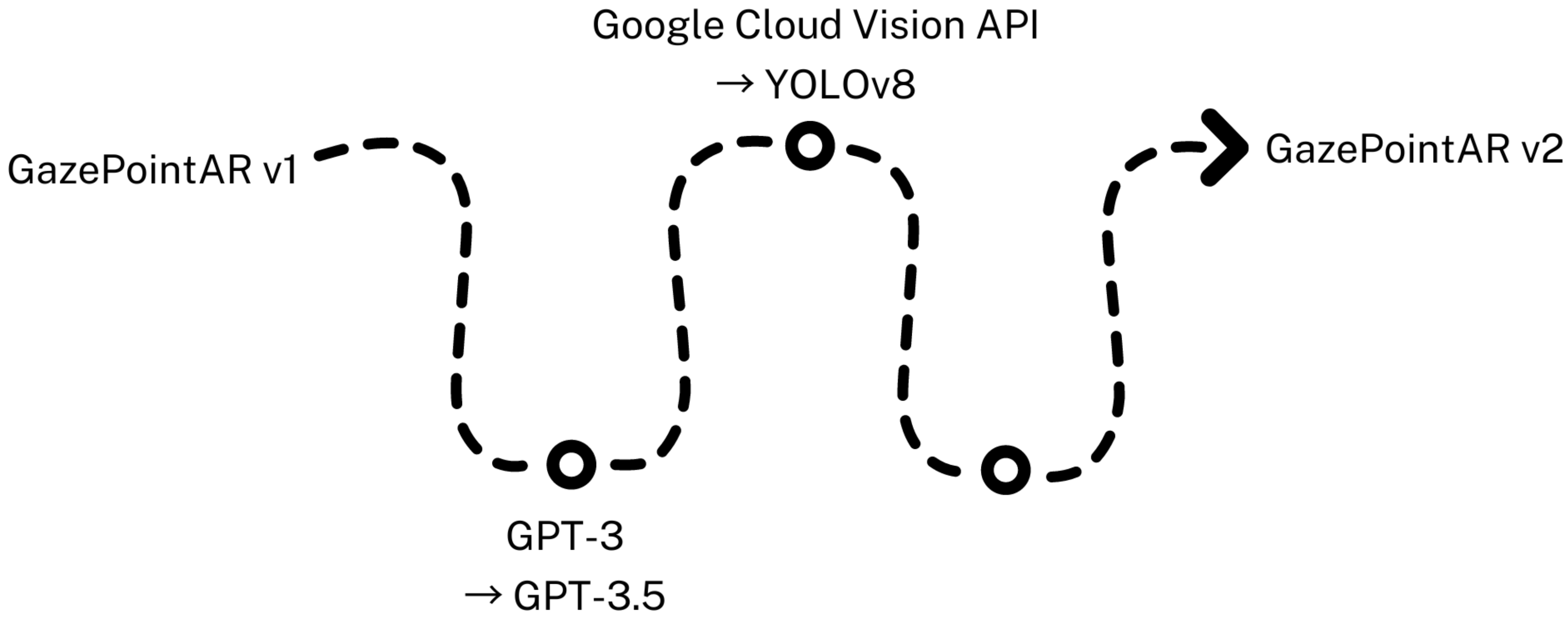
GazePointAR v2



GazePointAR v1

GPT-3
→ GPT-3.5

GazePointAR v2



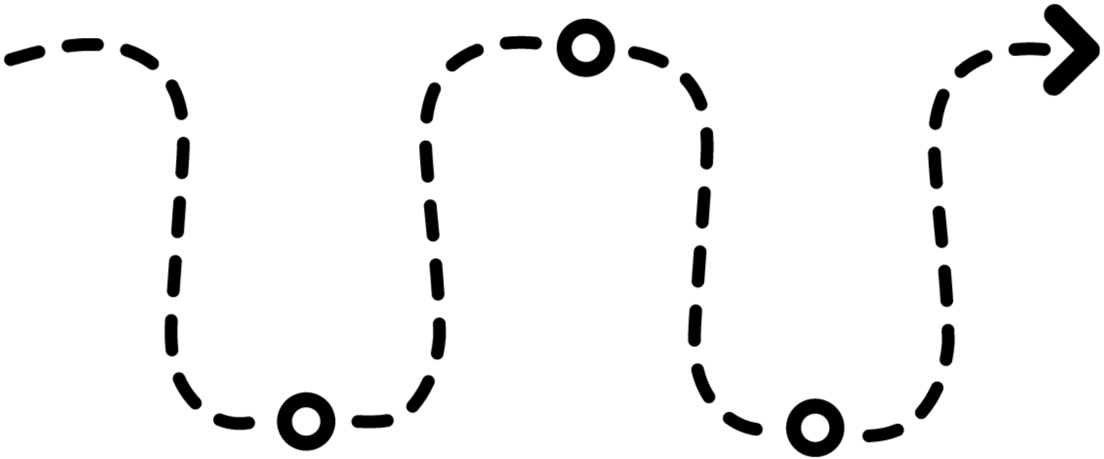
GazePointAR v1

GPT-3
→ GPT-3.5

Google Cloud Vision API
→ YOLOv8

GazePointAR v2

GazePointAR v1



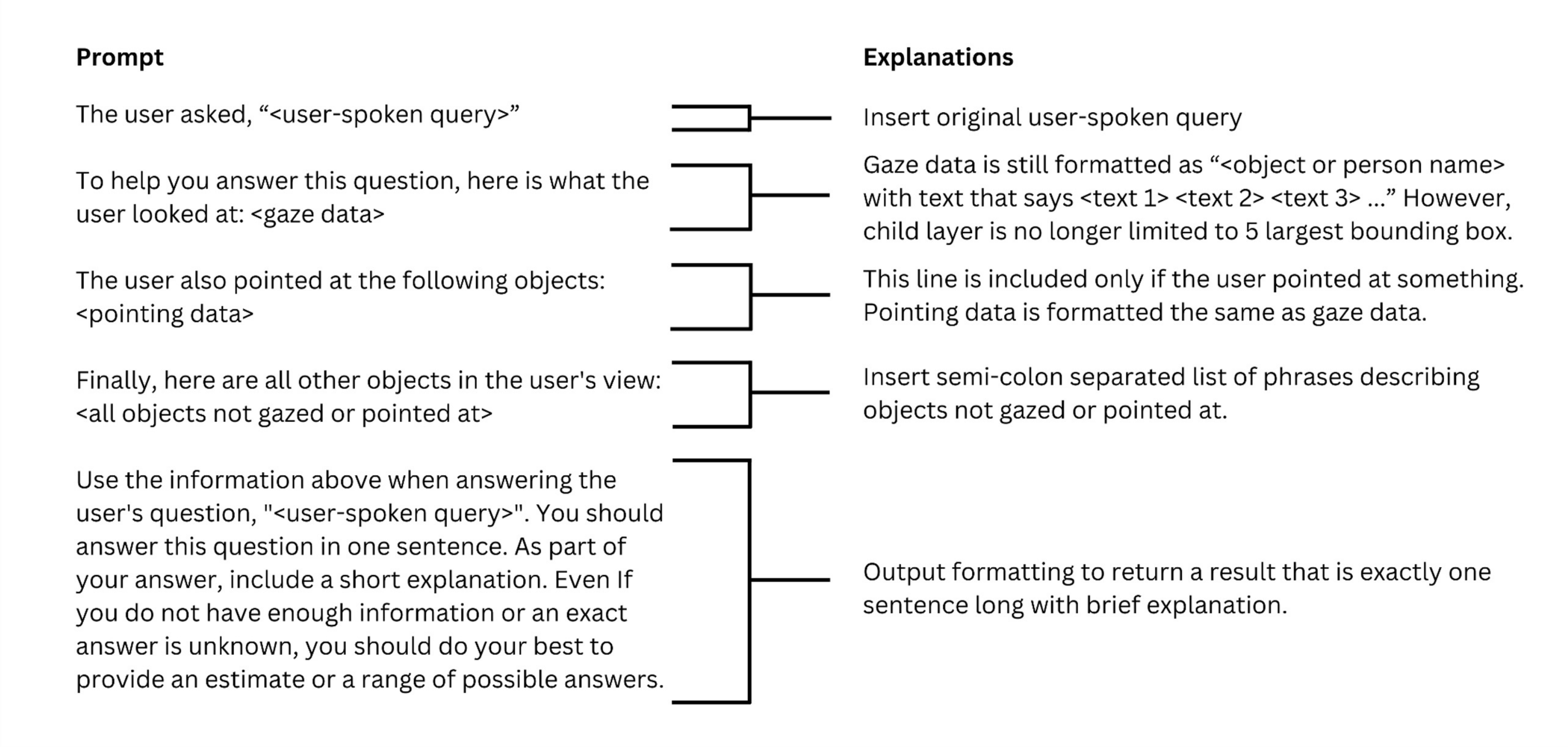
GPT-3
→ GPT-3.5

Google Cloud Vision API
→ YOLOv8

Manual Pronoun Replacement
→ Prompt Engineering

GazePointAR v2

GazePointAR v2’s Prompt Method

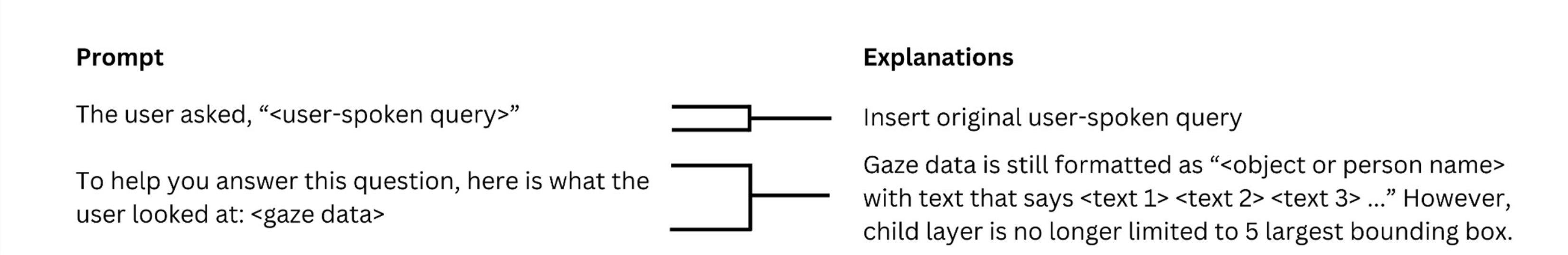


Prompt	Explanations
The user asked, “<user-spoken query>”	Insert original user-spoken query
To help you answer this question, here is what the user looked at: <gaze data>	Gaze data is still formatted as “<object or person name> with text that says <text 1> <text 2> <text 3> ...” However, child layer is no longer limited to 5 largest bounding box.
The user also pointed at the following objects: <pointing data>	This line is included only if the user pointed at something. Pointing data is formatted the same as gaze data.
Finally, here are all other objects in the user's view: <all objects not gazed or pointed at>	Insert semi-colon separated list of phrases describing objects not gazed or pointed at.
Use the information above when answering the user's question, "<user-spoken query>". You should answer this question in one sentence. As part of your answer, include a short explanation. Even If you do not have enough information or an exact answer is unknown, you should do your best to provide an estimate or a range of possible answers.	Output formatting to return a result that is exactly one sentence long with brief explanation.

GazePointAR v2’s Prompt Method

Prompt	Explanations
The user asked, “<user-spoken query>”	Insert original user-spoken query

GazePointAR v2’s Prompt Method



Prompt

The user asked, “<user-spoken query>”

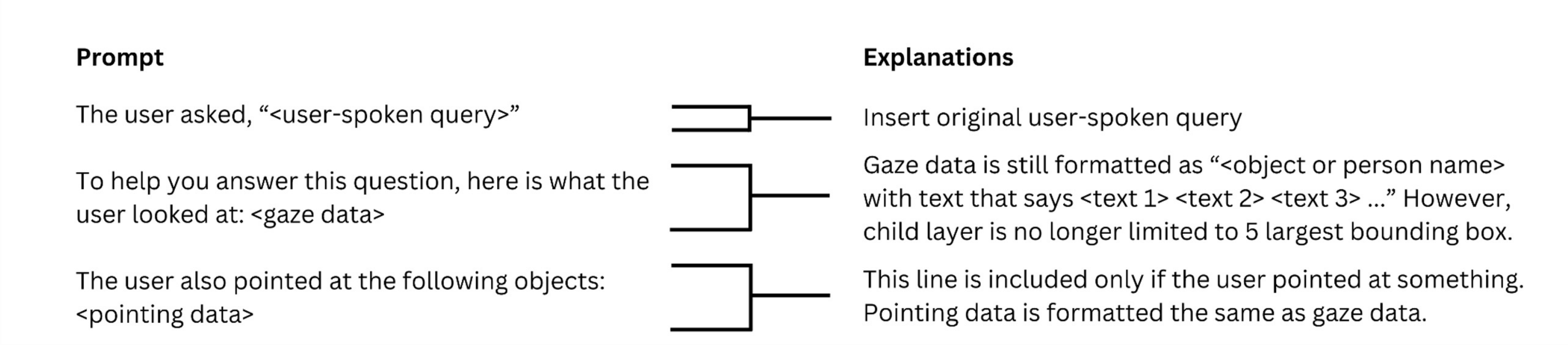
To help you answer this question, here is what the user looked at: <gaze data>

Explanations

Insert original user-spoken query

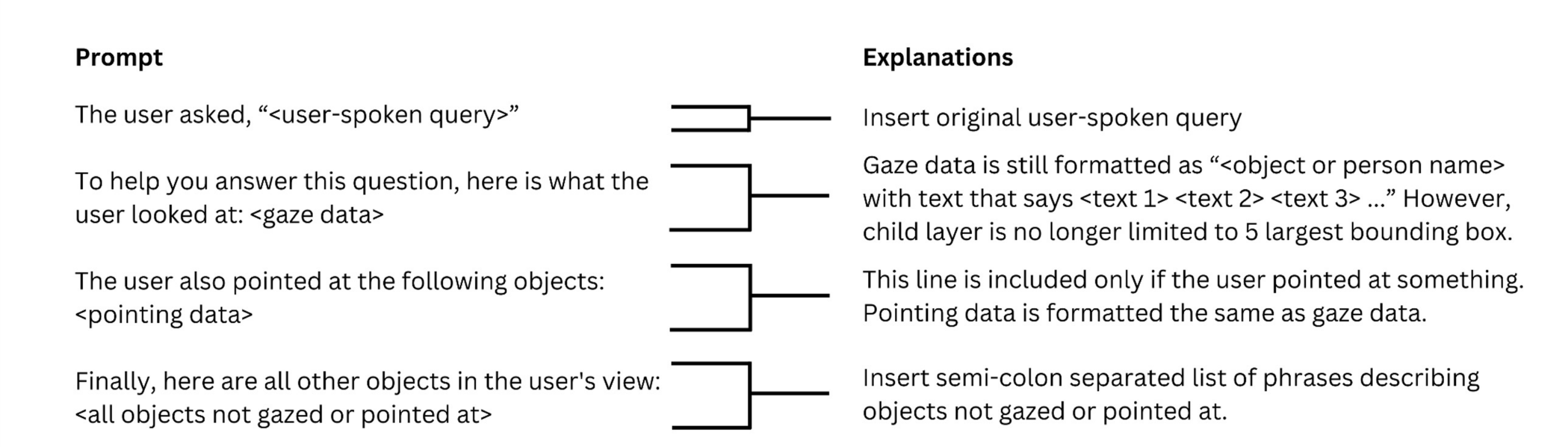
Gaze data is still formatted as “<object or person name> with text that says <text 1> <text 2> <text 3> …” However, child layer is no longer limited to 5 largest bounding box.

GazePointAR v2’s Prompt Method



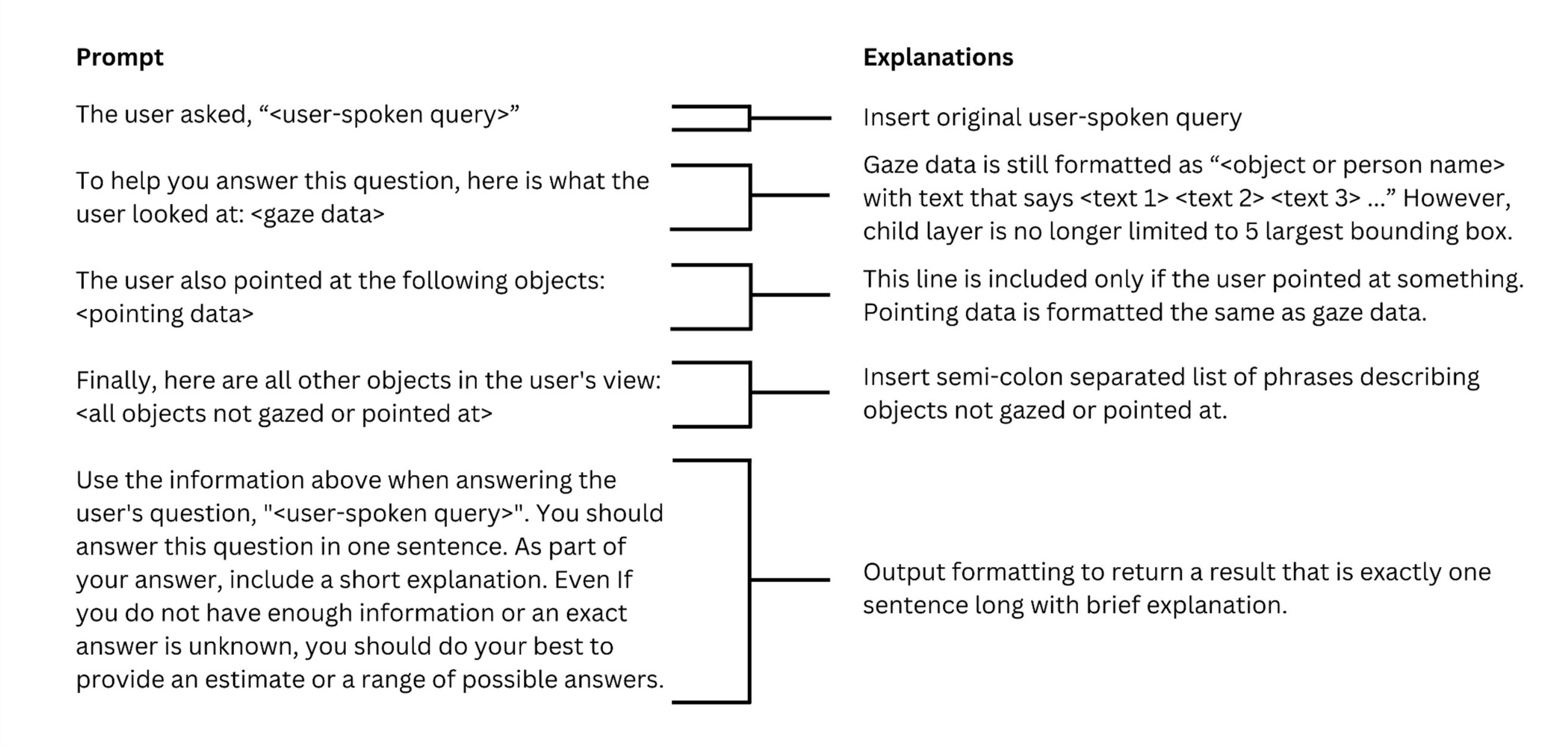
Prompt	Explanations
The user asked, “<user-spoken query>”	Insert original user-spoken query
To help you answer this question, here is what the user looked at: <gaze data>	Gaze data is still formatted as “<object or person name> with text that says <text 1> <text 2> <text 3> …” However, child layer is no longer limited to 5 largest bounding box.
The user also pointed at the following objects: <pointing data>	This line is included only if the user pointed at something. Pointing data is formatted the same as gaze data.

GazePointAR v2’s Prompt Method



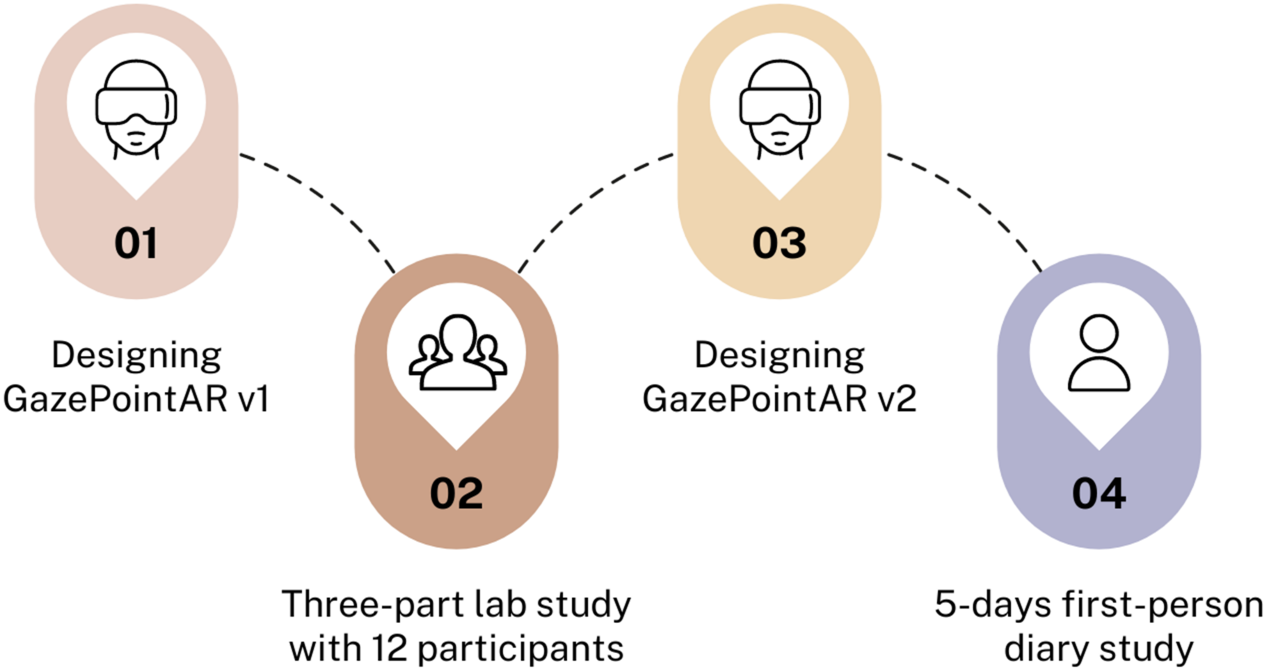
Prompt	Explanations
The user asked, “<user-spoken query>”	Insert original user-spoken query
To help you answer this question, here is what the user looked at: <gaze data>	Gaze data is still formatted as “<object or person name> with text that says <text 1> <text 2> <text 3> ...” However, child layer is no longer limited to 5 largest bounding box.
The user also pointed at the following objects: <pointing data>	This line is included only if the user pointed at something. Pointing data is formatted the same as gaze data.
Finally, here are all other objects in the user's view: <all objects not gazed or pointed at>	Insert semi-colon separated list of phrases describing objects not gazed or pointed at.

GazePointAR v2’s Prompt Method



Prompt	Explanations
The user asked, “<user-spoken query>”	Insert original user-spoken query
To help you answer this question, here is what the user looked at: <gaze data>	Gaze data is still formatted as “<object or person name> with text that says <text 1> <text 2> <text 3> ...” However, child layer is no longer limited to 5 largest bounding box.
The user also pointed at the following objects: <pointing data>	This line is included only if the user pointed at something. Pointing data is formatted the same as gaze data.
Finally, here are all other objects in the user's view: <all objects not gazed or pointed at>	Insert semi-colon separated list of phrases describing objects not gazed or pointed at.
Use the information above when answering the user's question, "<user-spoken query>". You should answer this question in one sentence. As part of your answer, include a short explanation. Even If you do not have enough information or an exact answer is unknown, you should do your best to provide an estimate or a range of possible answers.	Output formatting to return a result that is exactly one sentence long with brief explanation.

GazePointAR Timeline



01

Designing GazePointAR v1

02

Three-part lab study with 12 participants

03

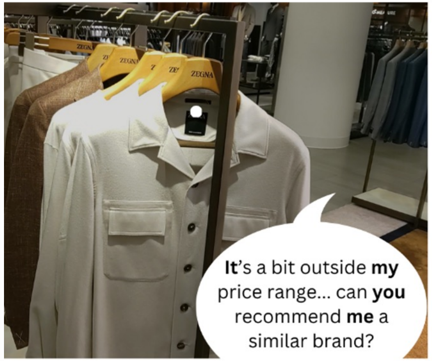
Designing GazePointAR v2

04

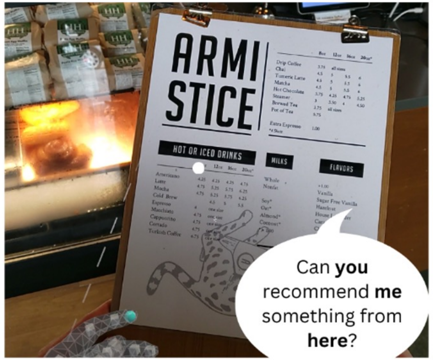
5-days first-person diary study

Study 2 - First-Person Diary Study

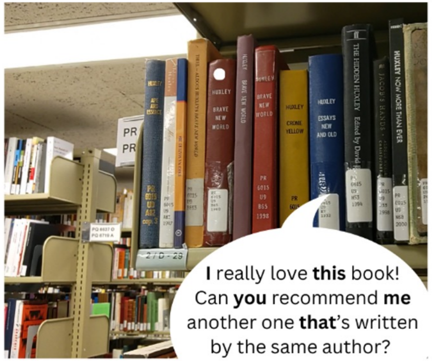
With GazePointAR v2, the first author used GazePointAR in their day-to-day activities for **five** days, recording their observations.


It's a bit outside my price range... can you recommend me a similar brand?

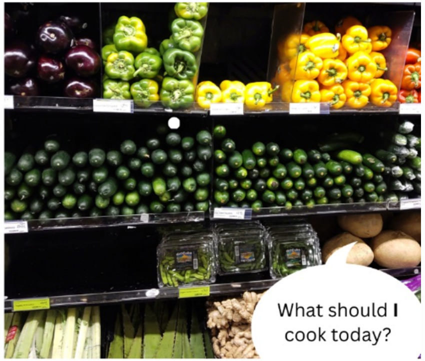
A: ...you may consider brand like Hugo Boss, Brooks Brothers, or J.Crew, which offer quality men's cloth at relatively lower costs.



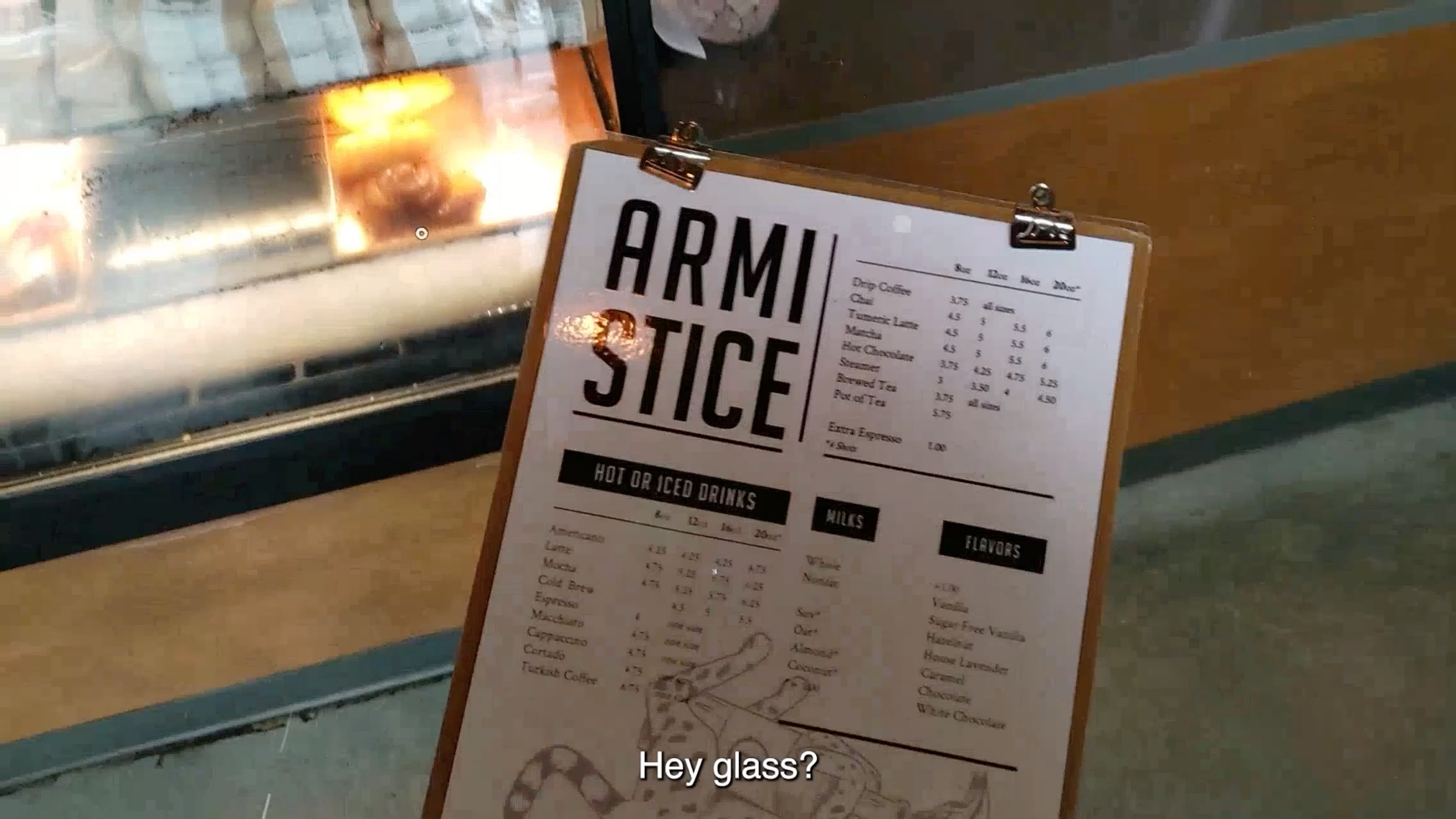
A: Based on your gaze, you are interested in the cappuccino, so I would recommend you to try the cappuccino.


I really love this book! Can you recommend me another one that's written by the same author?

A: ...The Brave New World Revisited by Aldous Huxley, as it is written by the same author and is often considered a companion piece.



A: ...consider making stuffed zucchini boats, as they are a delicious and healthy way to use both zucchini and peppers.



ARMI STICE

	8oz	12oz	16oz	20oz*
Drip Coffee	3.75	all sizes		
Chai	4.5	5	5.5	6
Tumeric Latte	4.5	5	5.5	6
Matcha	4.5	5	5.5	6
Hot Chocolate	3.75	4.25	4.75	5.25
Steamer	3	3.50	4	4.50
Brewed Tea	3.75	all sizes		
Pot of Tea	5.75			

Extra Espresso *4 Shots	1.00

HOT OR ICED DRINKS

	8oz	12oz	16oz	20oz*
Americano	4.25	4.25	4.25	4.75
Latte	4.75	5.25	5.75	6.25
Mocha	4.75	5.25	5.75	6.25
Cold Brew		4.5	5	5.5
Espresso	4	one size		
Macchiato	4.75	one size		
Cappuccino	4.75	one size		
Cortado	4.75			
Turkish Coffee	4.75			

MILKS

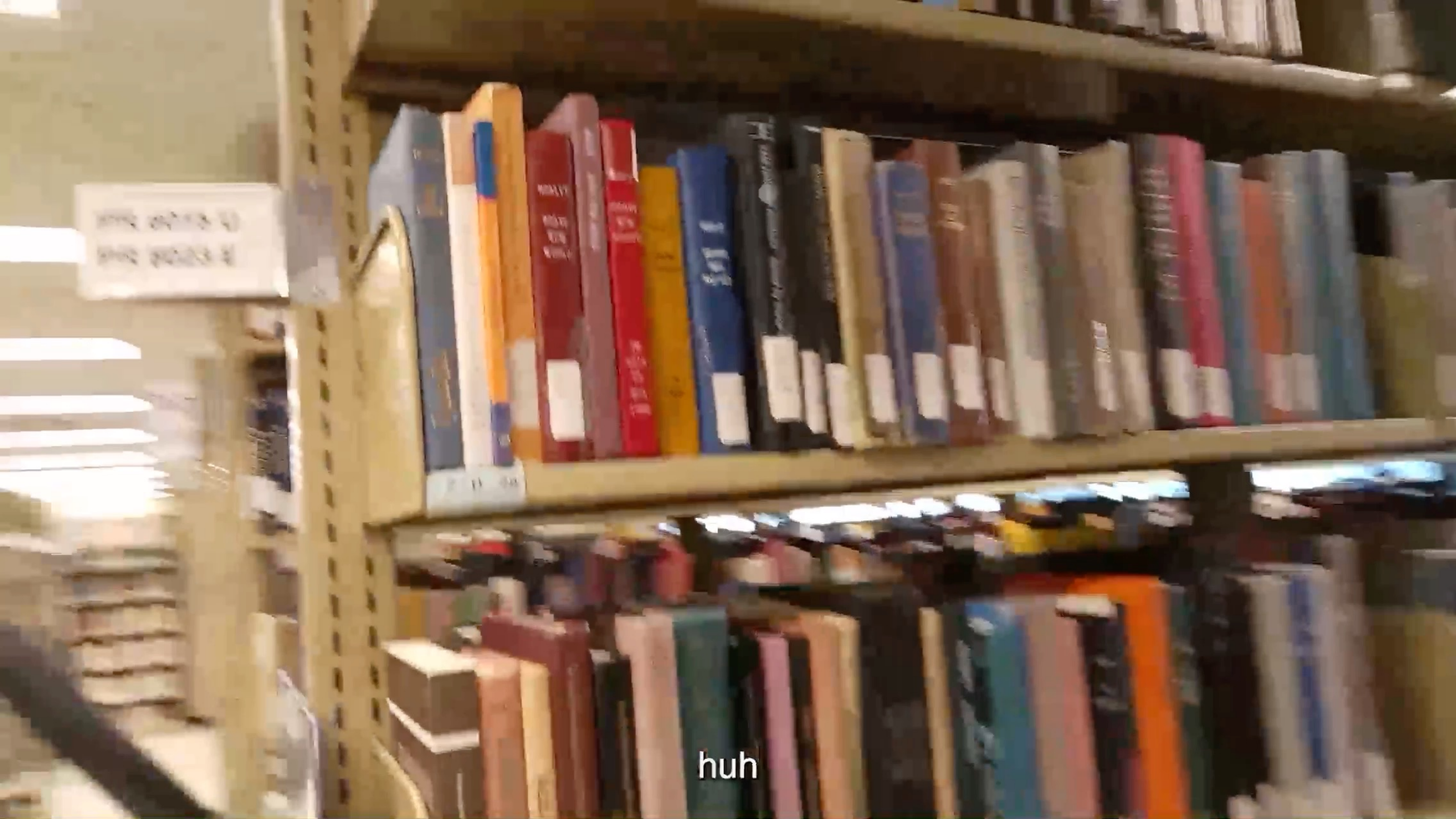
Whole
Nonfat

Soy*
Oat*
Almond*
Coconut*
*.50

FLAVORS

+1.00
Vanilla
Sugar Free Vanilla
Hazelnut
House Lavender
Caramel
Chocolate
White Chocolate

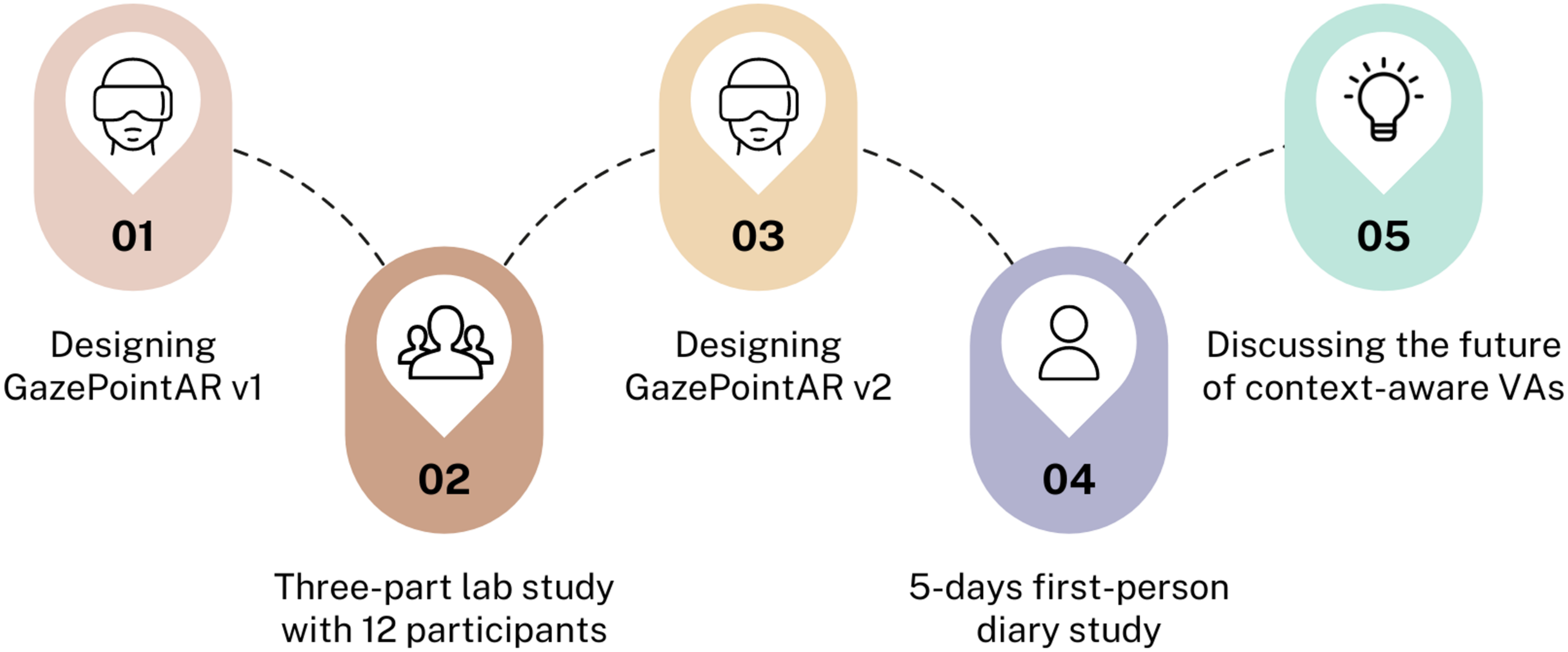
Hey glass?


huh

Study 2 Findings

- GazePointAR is natural and companion-like.
- GazePointAR still has several limitations:
  - Referents in the past require gaze history (e.g., “Where did I leave my keys?”).
  - Multiple referents require gaze shift (e.g., “Which is the healthier, **this** or **that**?”) .
  - Pointing is impractical in public.
  - Extended dwelling causes eye fatigue.

GazePointAR Timeline



01 Designing GazePointAR v1

02 Three-part lab study with 12 participants

03 Designing GazePointAR v2

04 5-days first-person diary study

05 Discussing the future of context-aware VAs

So, What's Next?

- GazePointAR is simple, natural, and human-like to speak to.
- Future context-aware VAs should leverage longitudinal natural eye gaze input.
- Future research should further study LLM-driven query disambiguation.

“I want to say queries with and without pronouns, because whichever comes to mind first, that’s the one I want to say. GazePointAR should adapt to me and my natural eye gaze” [P12].

Thanks for listening!

Let’s Connect!

Email	jaewook4@cs.washington.edu
Website	https://jaewook-lee.com
Twitter	https://twitter.com/jaewook_jae