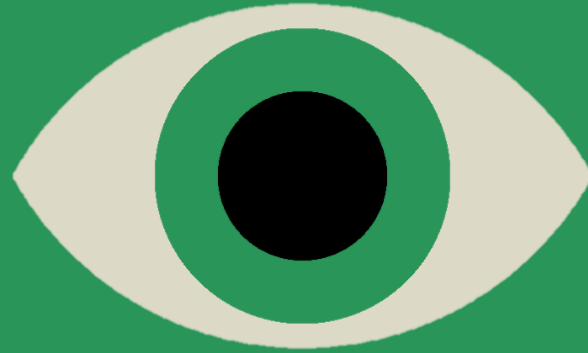




Multimodal User Interfaces

Table of Contents

- Introduction
- Software Architecture
- Evaluation
- Limitations and Future Work
- Conclusion



Introduction

Concept & Aim

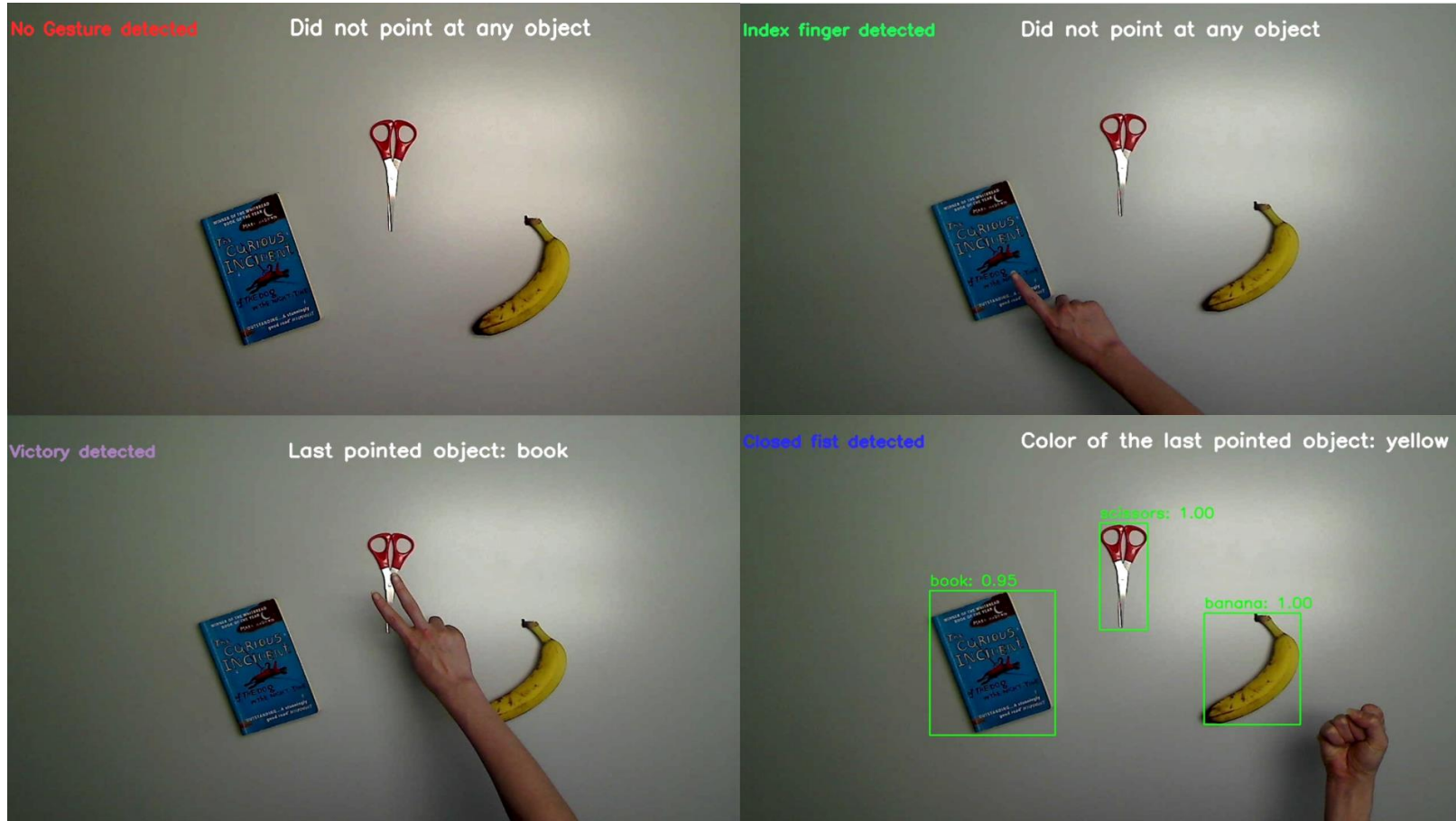
- Clinical application
 - Associative visual agnosia
 - Intact vision
 - Object recognition deficits
 - Can involve colors
- Develop visual assistant



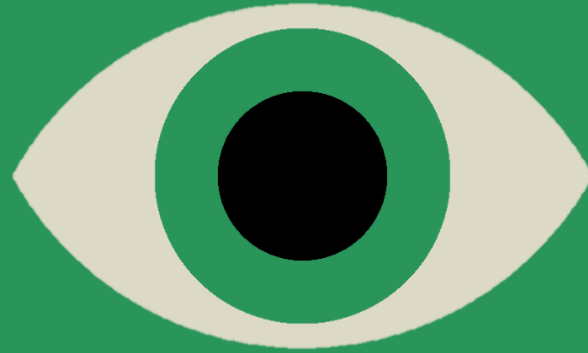
Modalities overview



Recognized Gestures

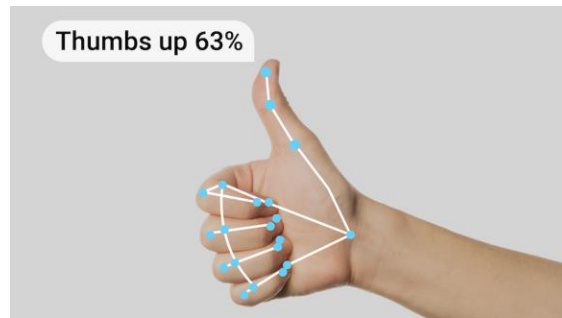


VuBot

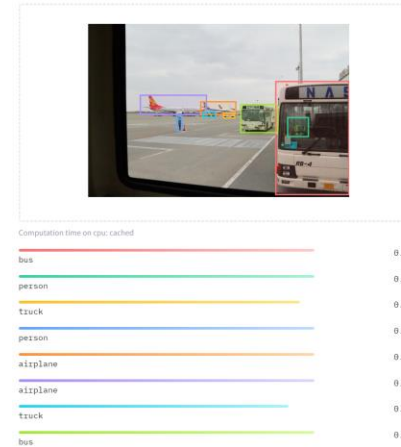


Software Architecture

Libraries



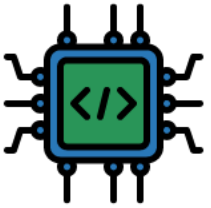
[facebook/detr-resnet-50](#)



[openai/whisper-large-v3](#)

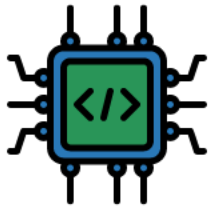


Modality Fusion

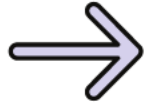


Thread 1

Modality Fusion

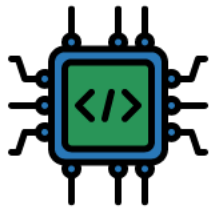


Thread 1

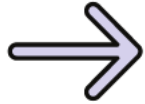


Frame Capture
(Open CV)

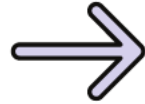
Modality Fusion



Thread 1

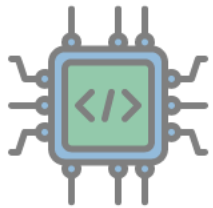


Frame Capture
(Open CV)

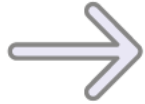


Gesture Recognition
(Media Pipe)

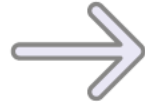
Modality Fusion



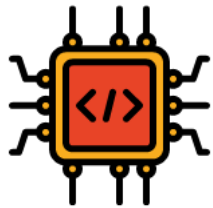
Thread 1



Frame Capture
(Open CV)

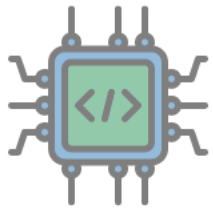


Gesture Recognition
(Media Pipe)

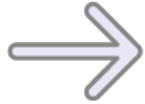


Thread 2

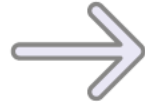
Modality Fusion



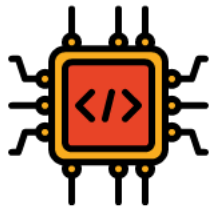
Thread 1



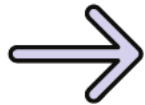
Frame Capture
(Open CV)



Gesture Recognition
(Media Pipe)

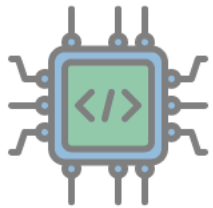


Thread 2

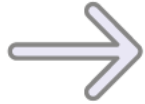


User Speaks

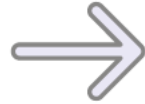
Modality Fusion



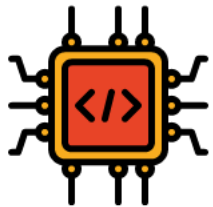
Thread 1



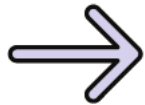
Frame Capture
(Open CV)



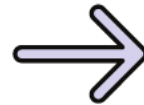
Gesture Recognition
(Media Pipe)



Thread 2

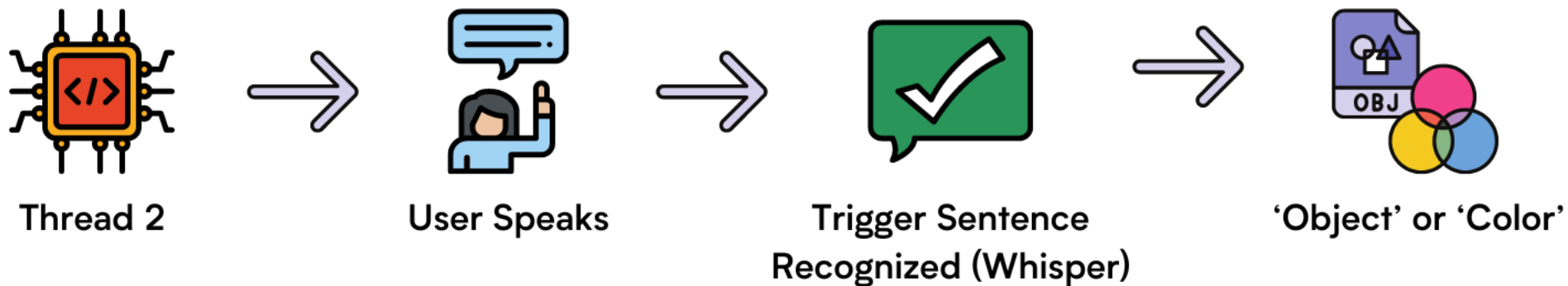
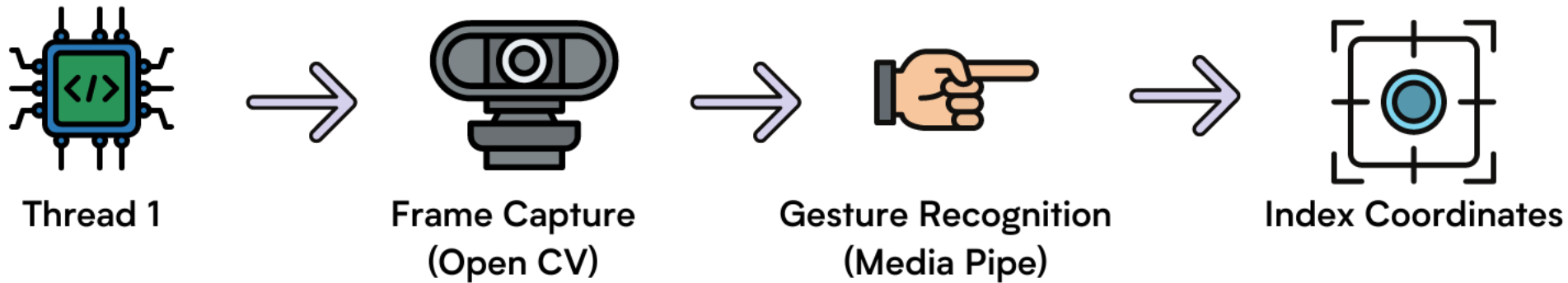


User Speaks

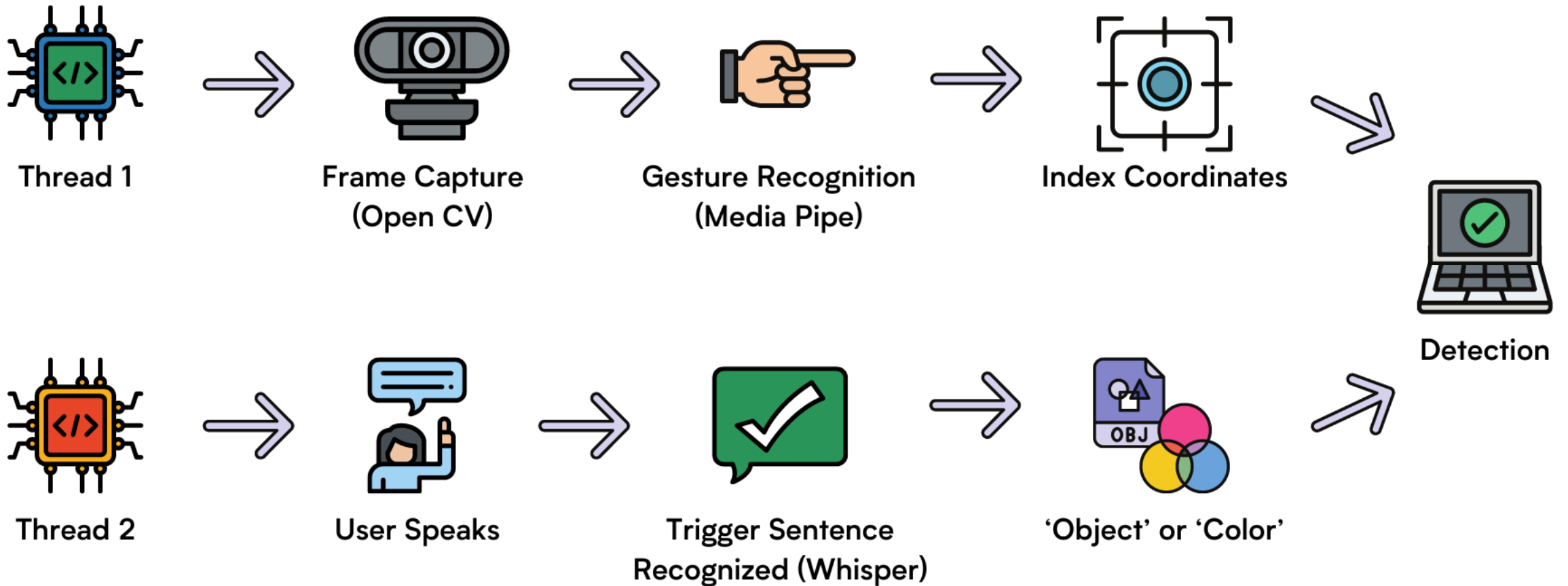


Trigger Sentence
Recognized (Whisper)

Modality Fusion

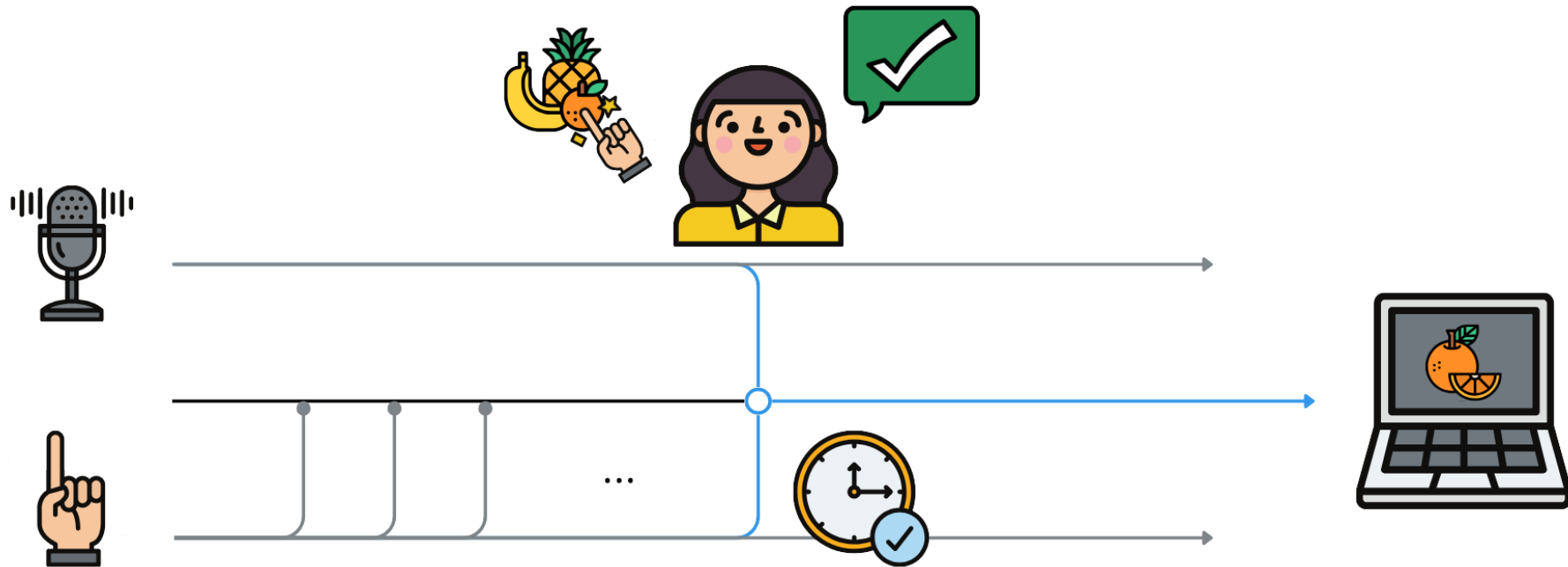


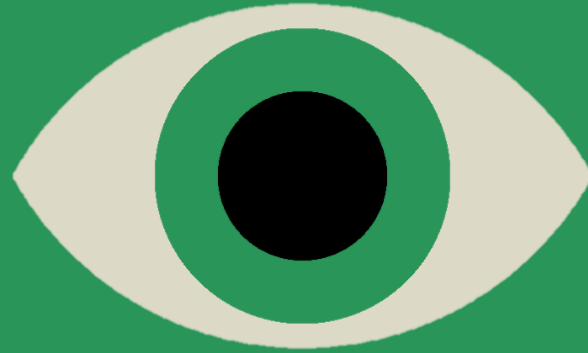
Modality Fusion



CARE/CASE Models

- CARE: Complimentary
- CASE: Synergistic

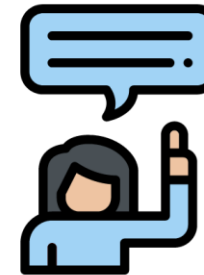




Evaluation

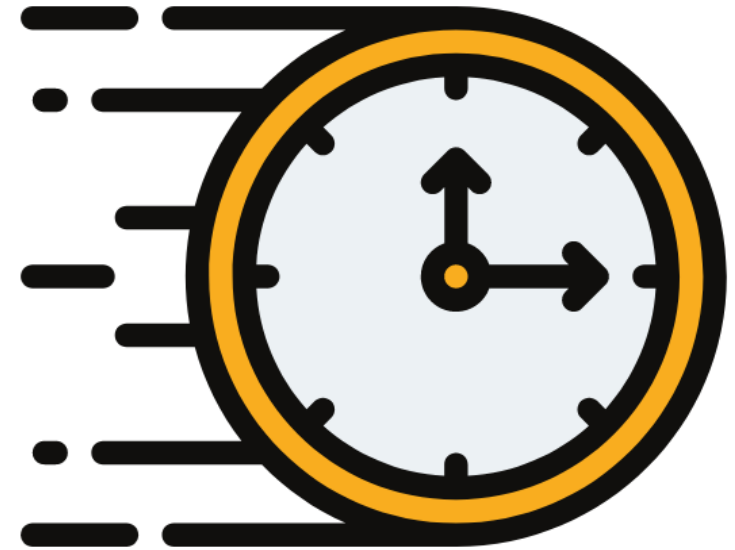
Experiment

- Task: find target objects and colors
 - Vary input modality
 - Speech input (original version)
 - Keyboard input (alternate version)
- Healthy participants
 - Manipulate labels to simulate recognition deficits



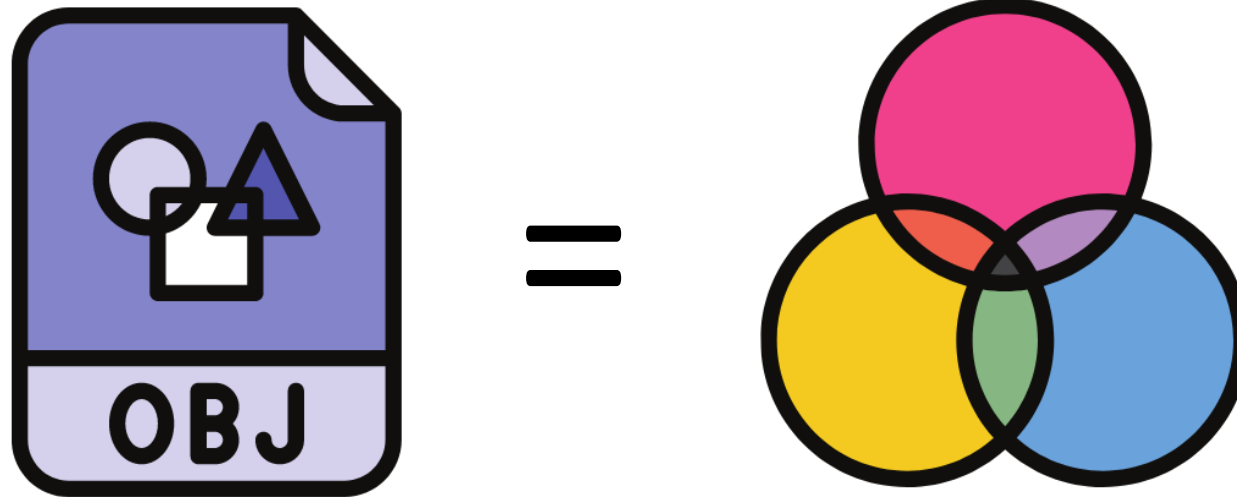
Statistical Analysis

- Time
 - Query time (object vs. color)
 - Task time (keys vs. speech)
- Accuracy
 - Task versions (keys vs. speech)
 - Models



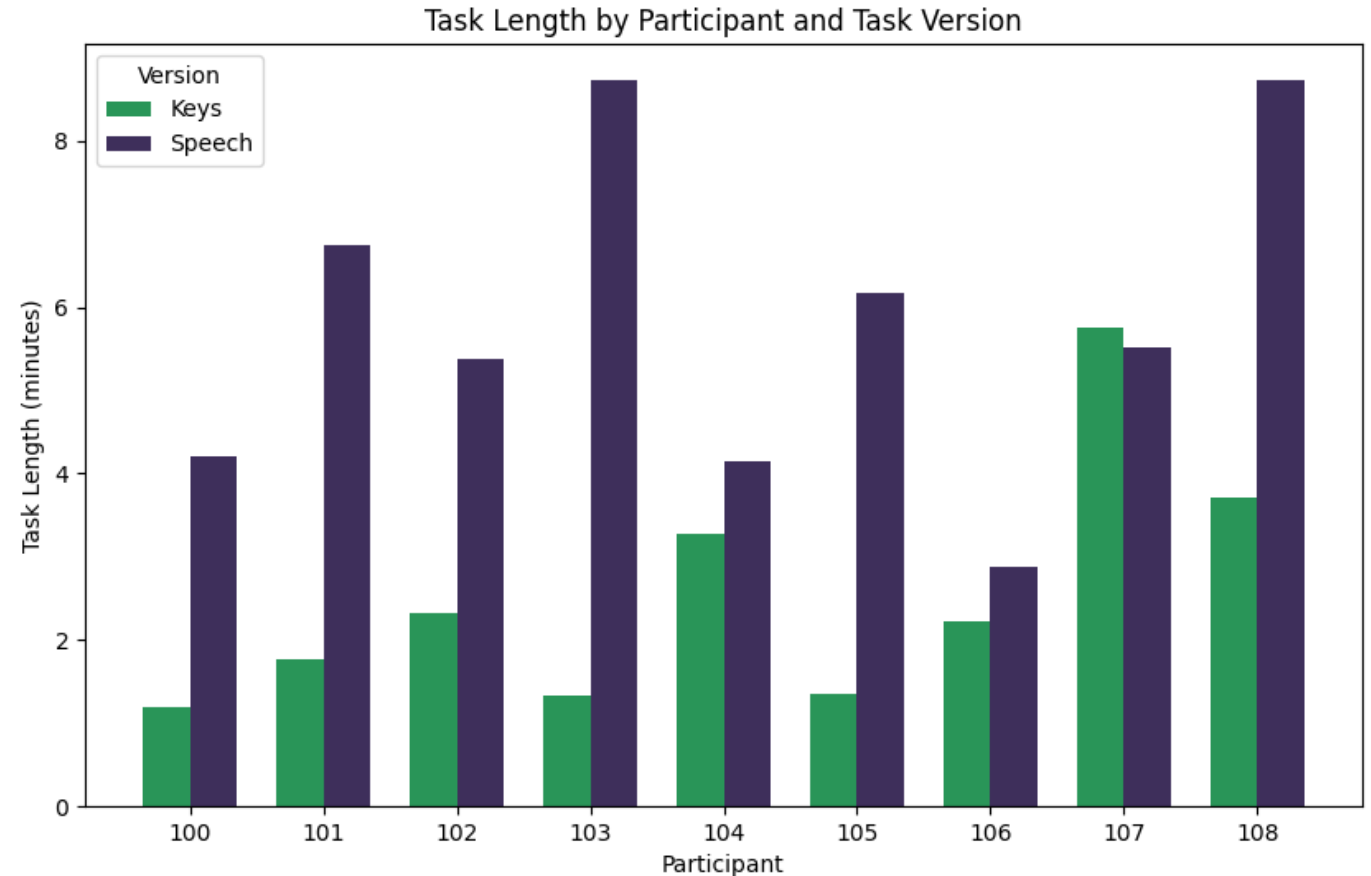
Query Runtimes

- Object and color queries take the same amount of time to compute



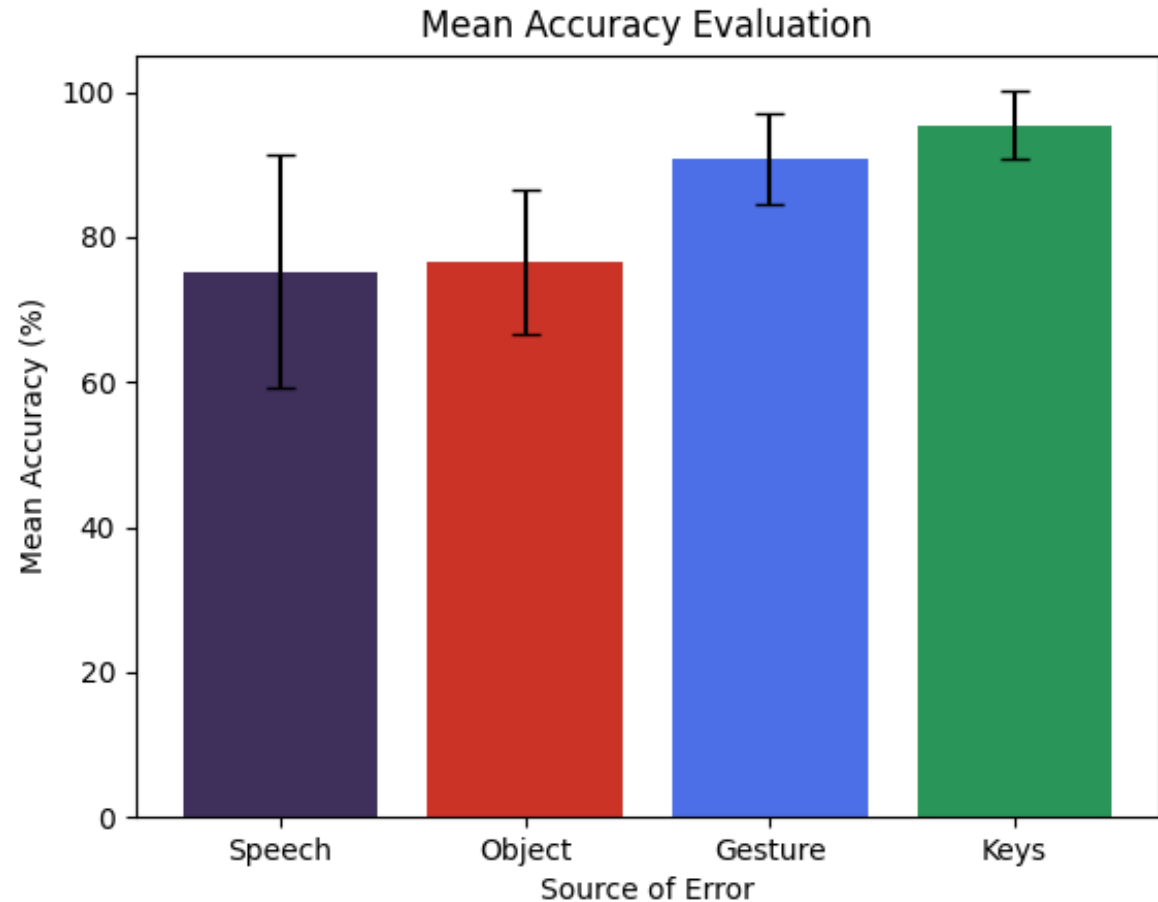
Runtime Performance

- Task is shorter when querying with keys than with speech
- $M_{keys} = 2.55min$
- $M_{speech} = 5.83min$



Accuracy

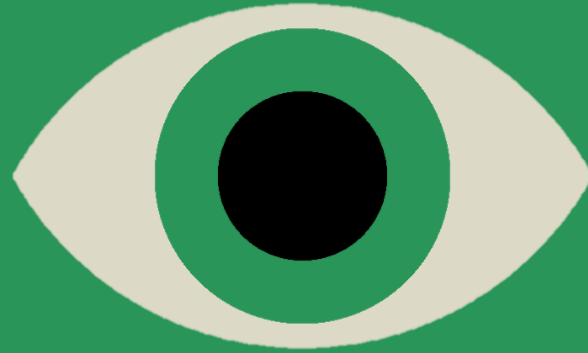
- Tracking of errors via screen recordings
- More speech errors than key errors
- Generally high accuracy across recognition models (80%)



Interpretation of Results

- Keys outperformed speech
- Speech remains the best modality for the future development of VuBot
 - Natural
 - Hands-free
 - Simpler scalability





Limitations and Future Work

Limitations

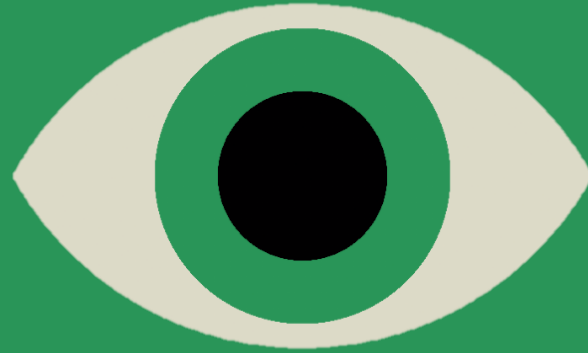
- General Recognition Errors:
 - Mistakes due to lighting or camera angles
- Color Recognition
 - Background and hand influence
- Speech Recognition:
 - Not meant for live transcription
 - Pronunciation of trigger words



Future Works

- Mobile Version
- Large Language Model (LLM) Integration

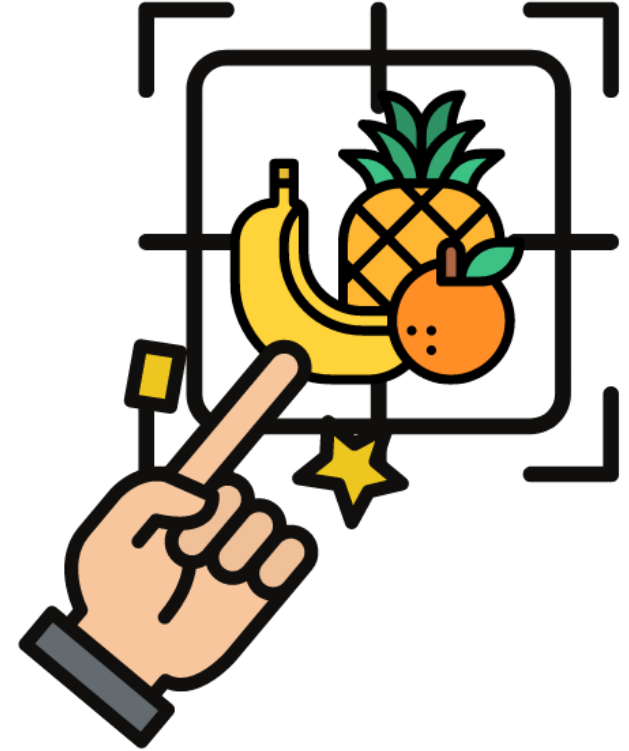




Conclusion

Conclusion

- VuBot is a visual assistant
- Complementary and synergistic fusion
- Recognizes objects and colors
- Many limitations
- This combination of modalities is best for future development
- Potential to empower individuals to have more independence



Thanks for listening

VuBot

References

- Icons : Canva – iconixar
- Lee, J., Wang, J., Brown, E., Chu, L., Rodriguez, S. S., & Froehlich, J. E. (2024). *GazePointAR: A Context-Aware Multimodal Voice Assistant for Pronoun Disambiguation in Wearable Augmented Reality* (arXiv:2404.08213). arXiv. <https://doi.org/10.48550/arXiv.2404.08213>
- *Openai/whisper-large-v3 · Discussions*. (n.d.). Retrieved 14 May 2024, from <https://huggingface.co/openai/whisper-large-v3/discussions>
- Hugging Face, facebook/detr-resnet-50, <https://huggingface.co/facebook/detr-resnet-50>
- Google AI, Gesture recognition task guide, https://ai.google.dev/edge/mediapipe/solutions/vision/gesture_recognizer?hl=fr
- OpenCV, <https://opencv.org/>
- Github, Whisper repo, <https://github.com/openai/whisper>