Multimodal AI

Lecture 14.2 – Human-AI Interaction

Paul Liang
Assistant Professor
MIT Media Lab & MIT EECS

https://pliang279.github.io
ppliang@mit.edu
@pliang279

Assignments for This Coming Week

Please fill out course evaluations and give us feedback!

HW5 extension -- due **next Wednesday May 13**.

For project:

- Make sure to meet with myself and TAs this week if you need advice.
- Presentations next Tuesday May 12 – in-class slides presentation
- Final report due Tuesday May 19.

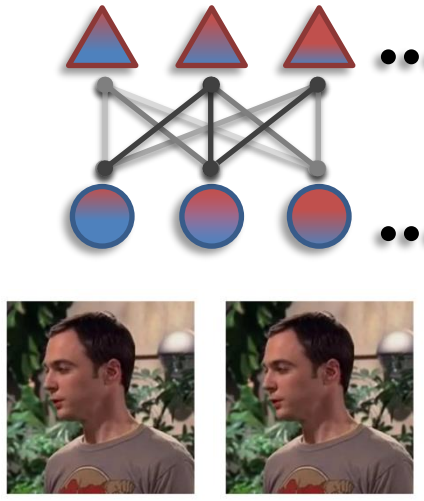
Today's lecture

1. AI for smell, taste, temperature
2. Human-AI interaction
3. Ethics and safety

Multimodal Human-AI Interaction

Solving hard problems by breaking them down into step-by-step reasoning steps in multiple modalities

It's just a privilege to watch your mind at work.



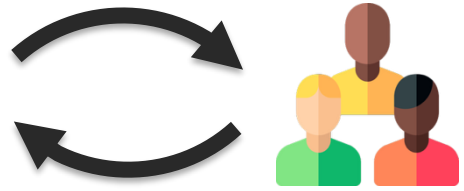
Multimodal representation

This person is being sarcastic. They seem to be close friends.

(quote previous episodes) (highlight multimodal information)

Here's a story of them in a different culture...

(generate future episodes)



Models: Multimodal fusion and generation
Data: Hard challenges + human reasoning steps
Training: Reinforcement learning for emergent reasoning
Human: Trustworthy, safe, controllable

[Liang, Zadeh, and Morency. Foundations and Trends on Multimodal Machine Learning. ACM Computing Surveys 2024]

multisensory intelligence

AI for Smell

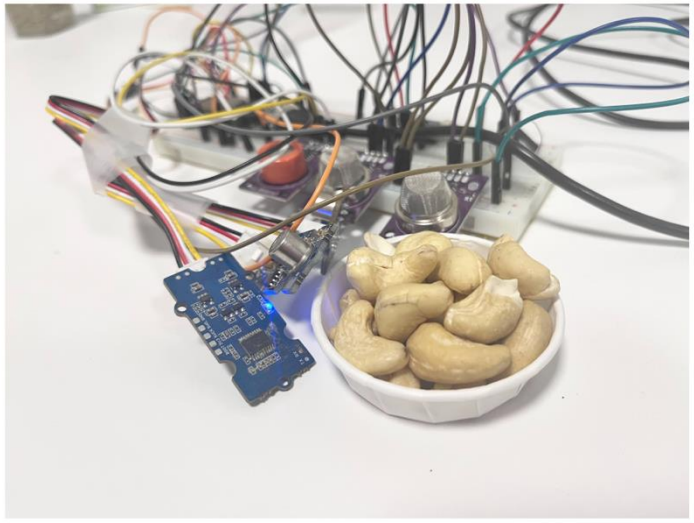
https://github.com/MIT-MI/SmellNet

[Feng et al., SmellNet: A Large-scale Dataset for Real-world Smell Recognition. ICLR 2026]

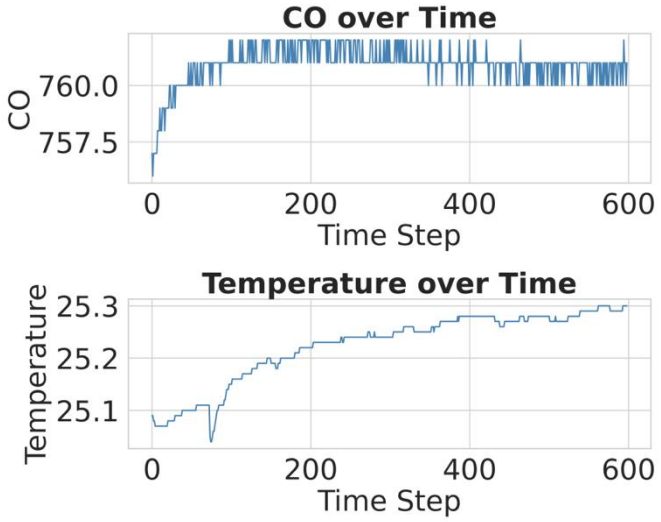
multisensory intelligence

Creating SmellNet

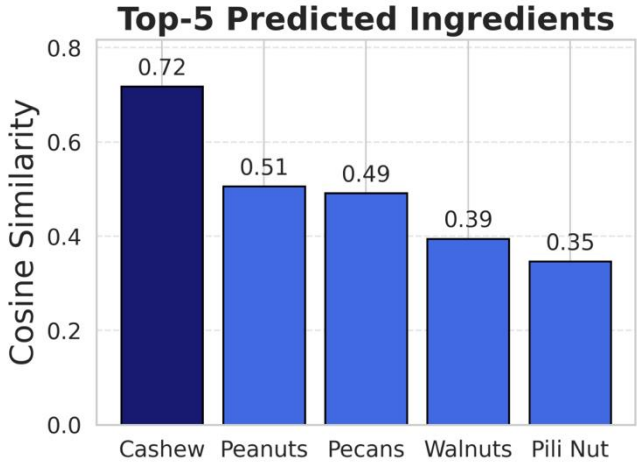
https://github.com/MIT-MI/SmellNet



(a) Sensor setup detecting cashew.



(b) Time-series signals from CO and temperature sensors.



(c) Top-5 model predictions using cosine similarity.

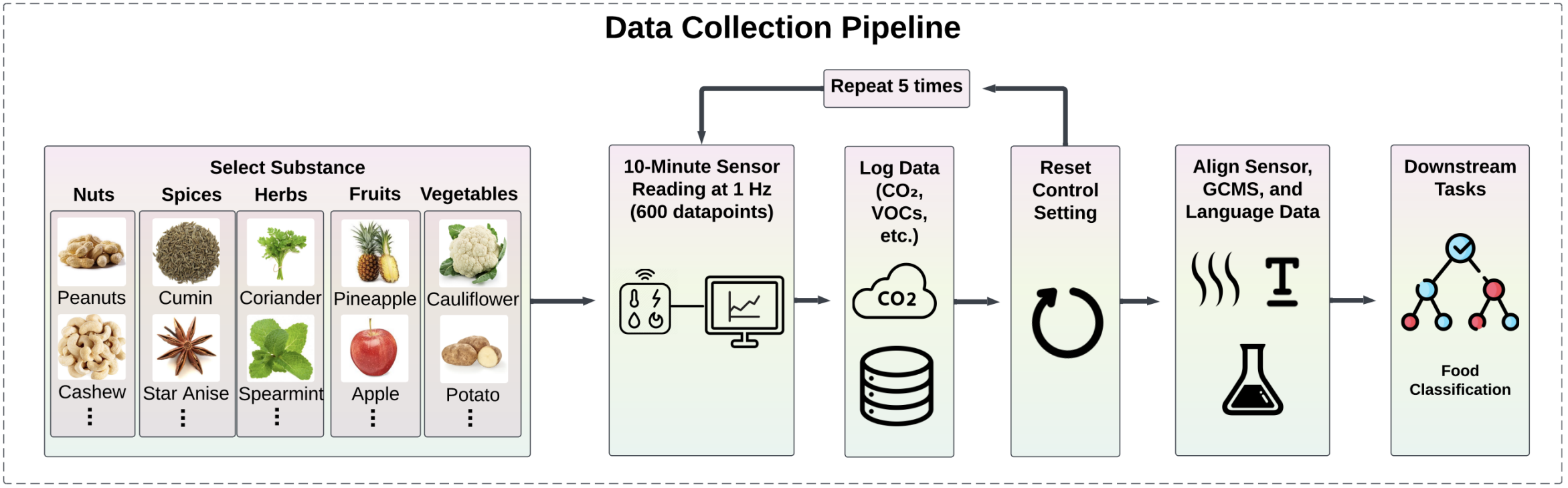
[Feng et al., SmellNet: A Large-scale Dataset for Real-world Smell Recognition. ICLR 2026]

multisensory intelligence

Creating SmellNet

https://github.com/MIT-MI/SmellNet

Data is key! More than 50 hours across 50 substances. >300,000 datapoints



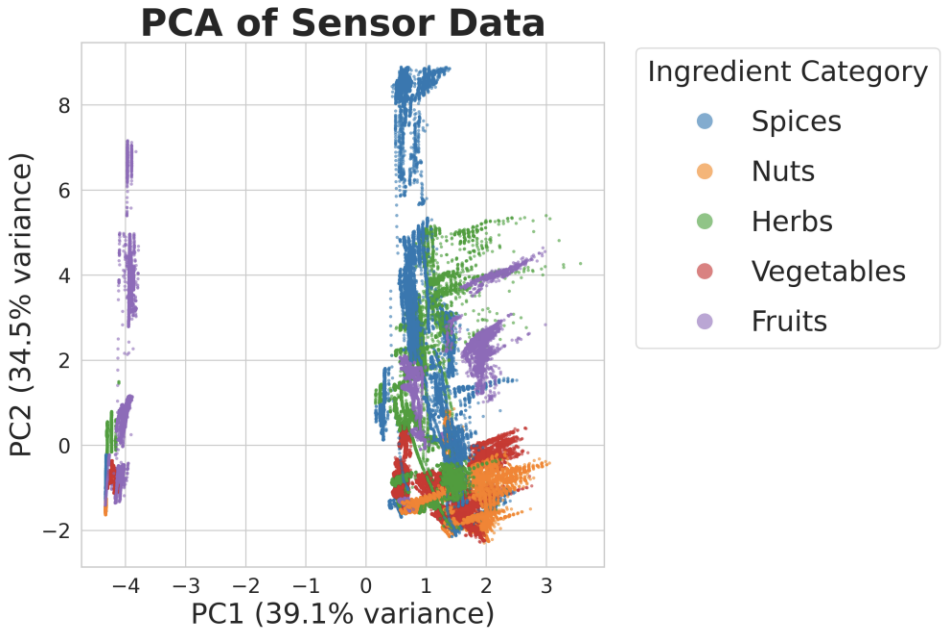
[Feng et al., SmellNet: A Large-scale Dataset for Real-world Smell Recognition. ICLR 2026]

multisensory intelligence

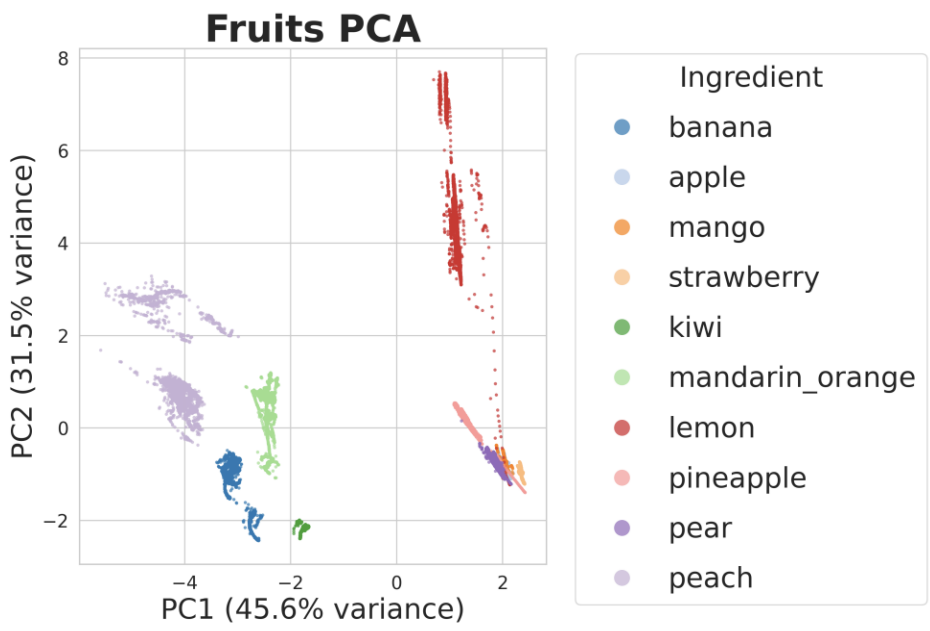
AI for Smell Recognition

https://github.com/MIT-MI/SmellNet

PCA: a way to visualize >2-dimensional data in 2 dimensions



(a) All Substances PCA



(b) Fruits PCA

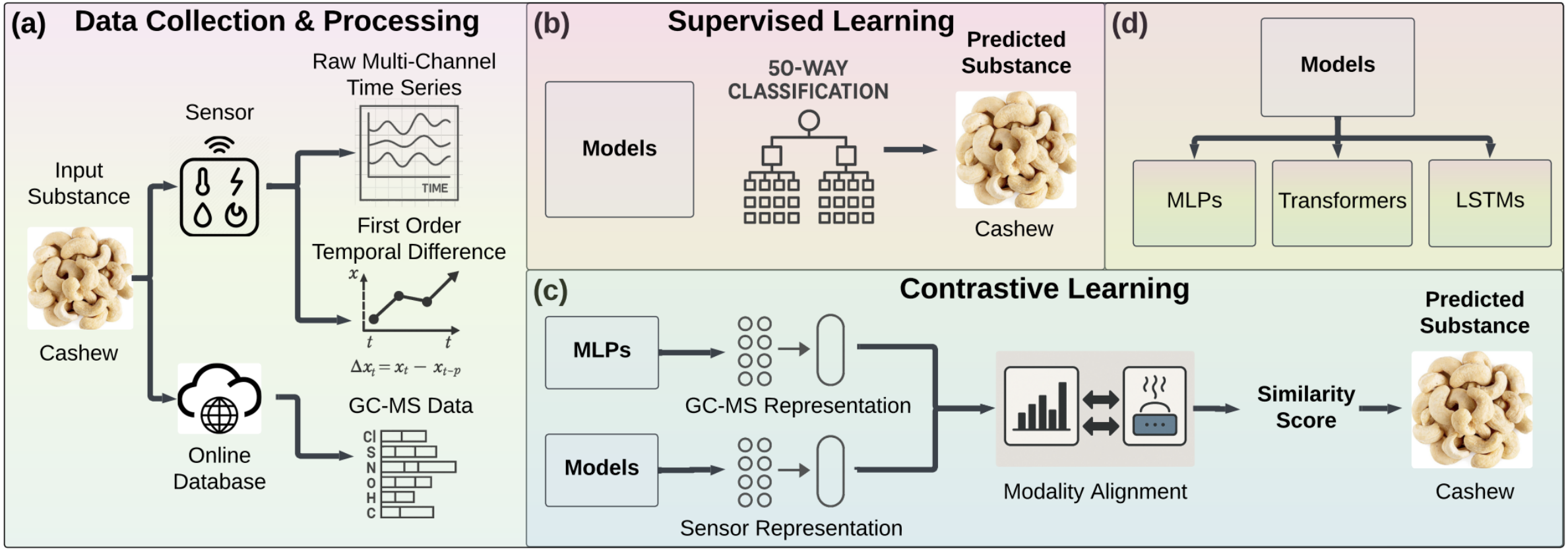
[Feng et al., SmellNet: A Large-scale Dataset for Real-world Smell Recognition. ICLR 2026]

multisensory intelligence

AI for Smell Recognition

https://github.com/MIT-MI/SmellNet

Modeling techniques: sequence models, temporal difference models, alignment with online databases



[Feng et al., SmellNet: A Large-scale Dataset for Real-world Smell Recognition. ICLR 2026]

multisensory intelligence

Smell Recognition Results

https://github.com/MIT-MI/SmellNet

ScentFormer applies temporal difference and sliding windows to capture relative changes

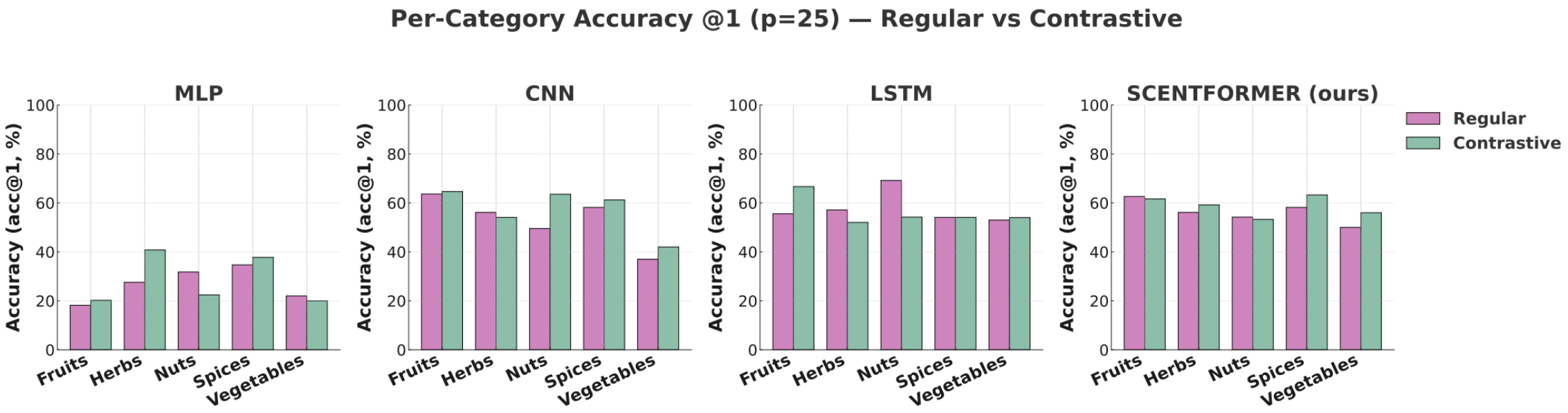


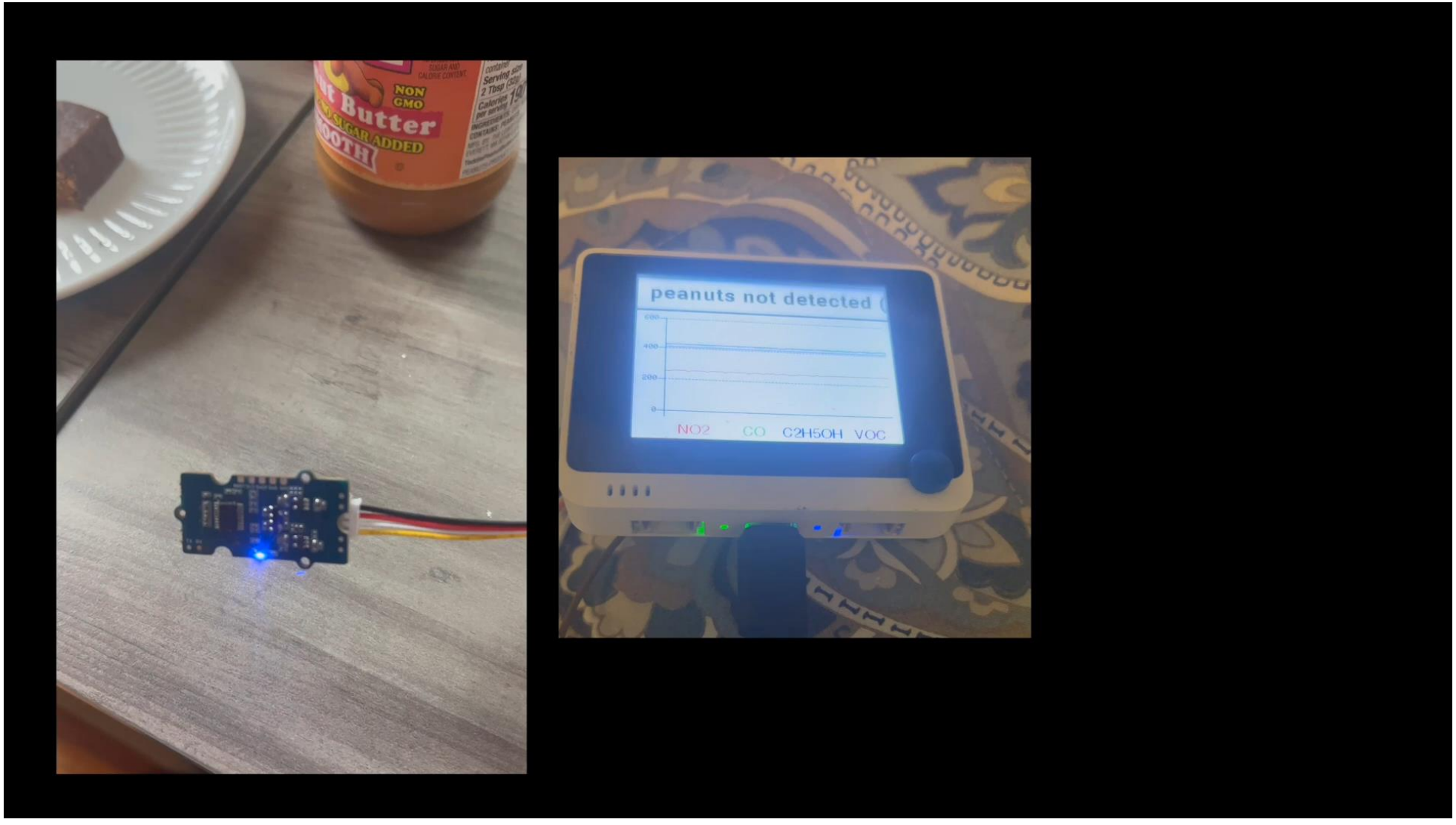
Figure 5: **Per-category accuracy** (acc@1) for four models at lag $p = 25$. Bars are paired per category (Regular vs. Contrastive). Figure shows the non temporal models suffer from categories like vegetables, but temporal architectures demonstrate stronger robustness across categories.

[Feng et al., SmellNet: A Large-scale Dataset for Real-world Smell Recognition. ICLR 2026]

multisensory intelligence

Applications in Allergen Detection

https://github.com/MIT-MI/SmellNet

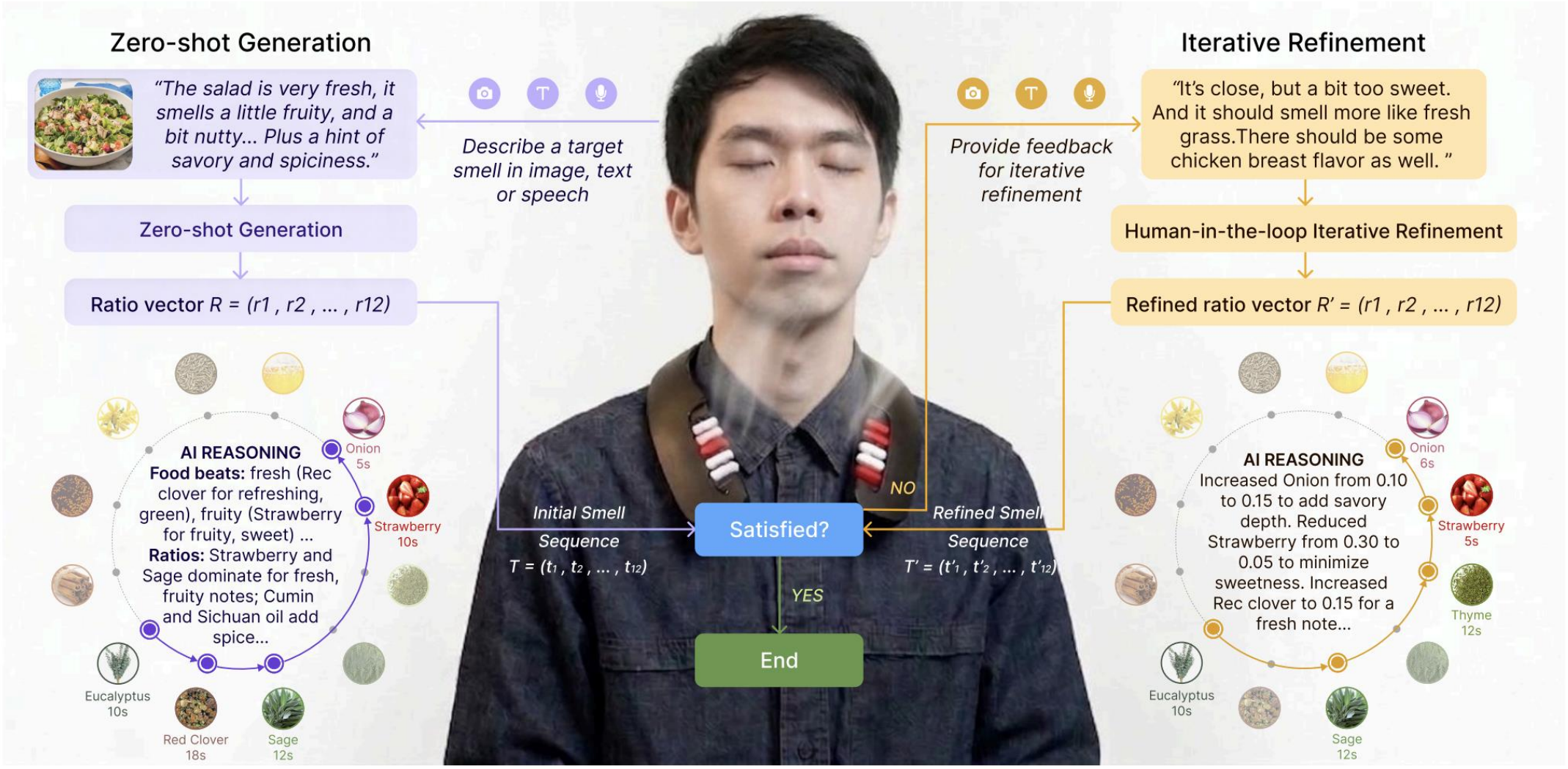


AromaGen: Towards Smell Transmission

[Wen et al., Interactive Generation of Rich Olfactory Experiences with Multimodal Language Models. arXiv 2026]

multisensory intelligence

AromaGen: Towards Smell Transmission

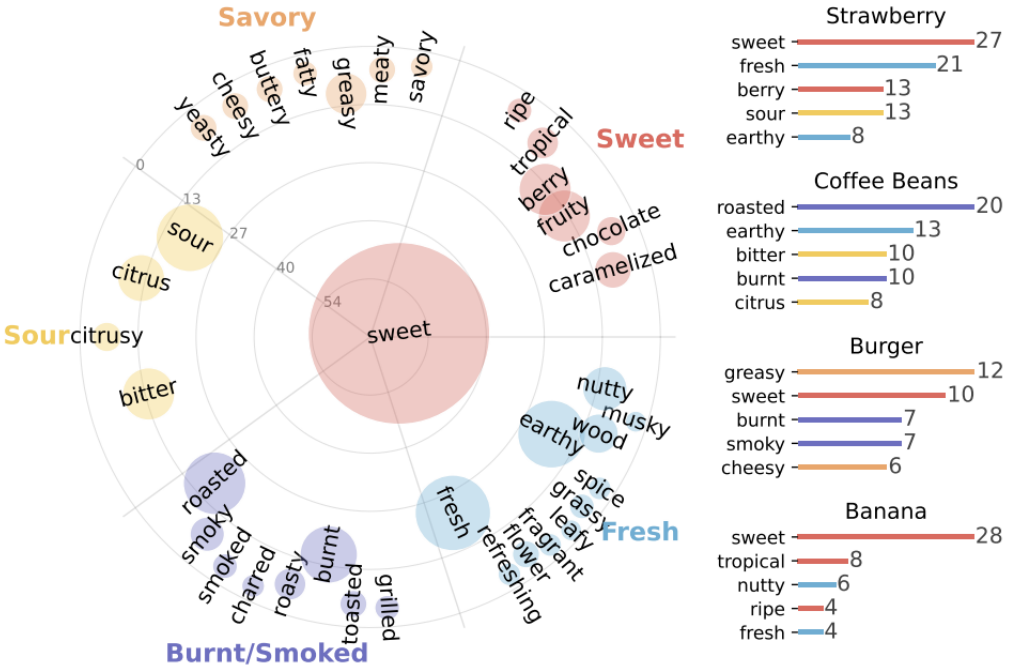


[Wen et al., Interactive Generation of Rich Olfactory Experiences with Multimodal Language Models. arXiv 2026]

A Language for Smell?

Participants mostly described smell based on taste, texture, and memory, but a shared vocabulary emerges.

Figure 4: The 12 base odorants used in AromaGen's palette.



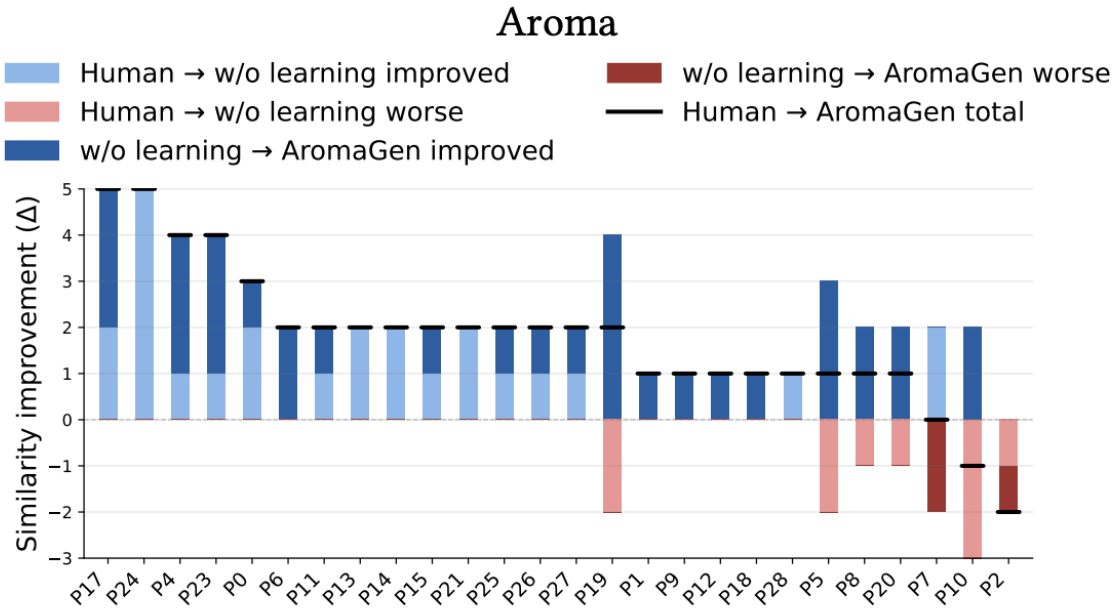
Odorant	Vol.	Notes
Cumin ■ ■	6	Smoky, spice
Ylang Ylang ■ ■	6	Warm, light spice
Sichuan Oil ■ ■	3	Light, chai, spice
Cinnamon ■ ■	5	Sweet, spice, coffee, warm
Eucalyptus ■	5	Refreshing, spa
Red Clover ■ ■	5	Mint, clover, green, refreshing
Sage ■ ■	6	Refreshing
Cypress ■	5	Woody stability
Thyme ■ ■	5	Bitter, green, vegetable
Strawberry ■ ■	5	Elegant clarity, fruity, sweet
Onion ■ ■	6	Umami, onion, chips, savory
Isovaleric Acid ■ ■	8	Cheesy, sweat, sour

Table 3: The 12 base odorants in AromaGen's palette, with volatility score (1–10) and characteristic notes. Colored squares indicate perceptual categories: ■ Sweet, ■ Savory, ■ Sour, ■ Burnt/Smoked, ■ Fresh.

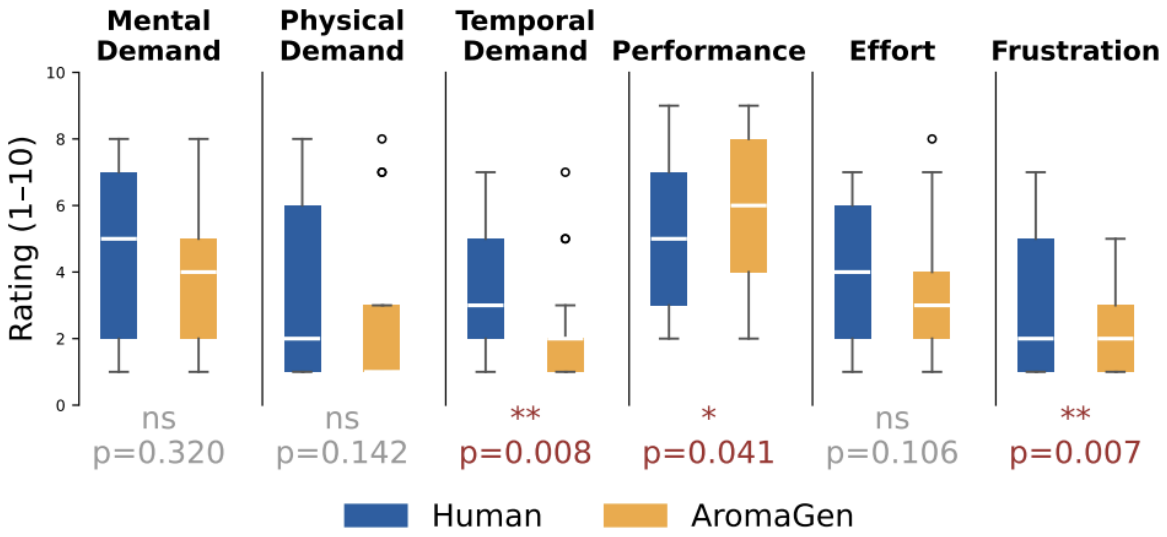
[Wen et al., Interactive Generation of Rich Olfactory Experiences with Multimodal Language Models. arXiv 2026]

AromaGen Key Results

Learning procedure in AromaGen improves similarity.



AromaGen saves effort of human mixing with similar performance.



[Wen et al., Interactive Generation of Rich Olfactory Experiences with Multimodal Language Models. arXiv 2026]

our goal:
can we deliver smell sensations
without chemicals?

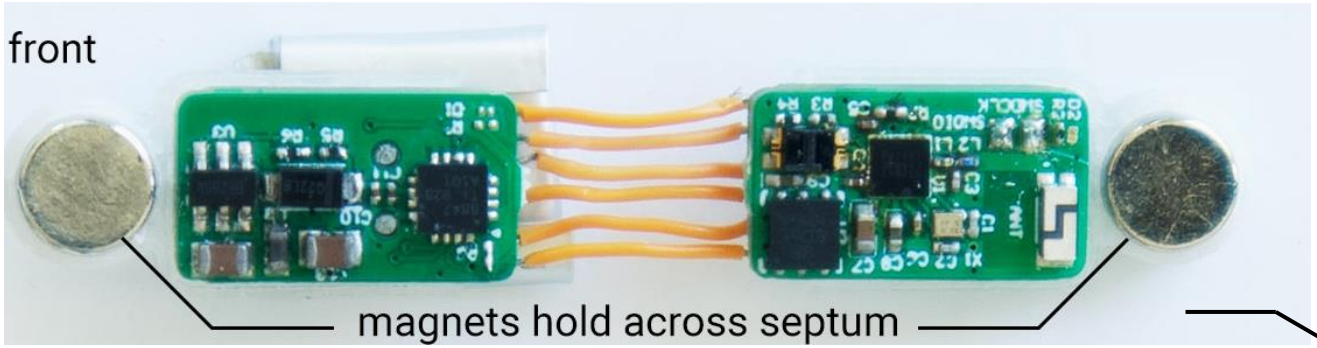
[slides from Jas Brooks]

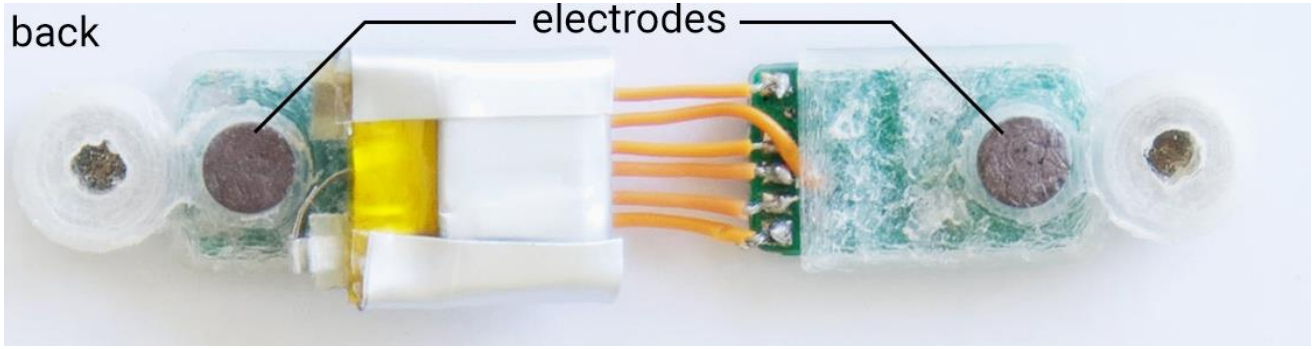
smell

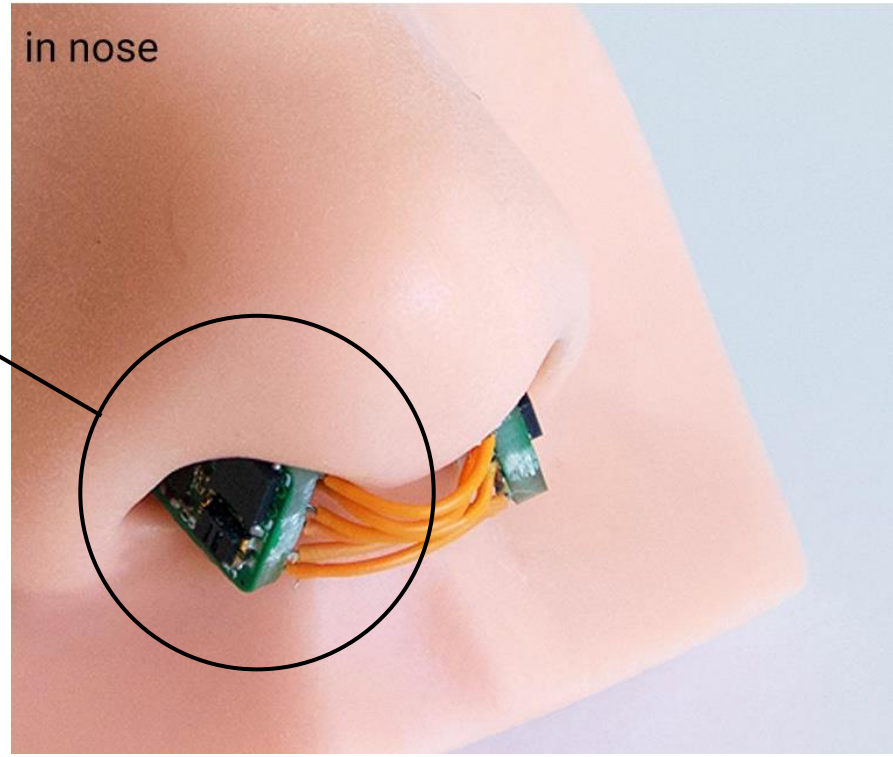
olfactory bulb

trigeminal nerve

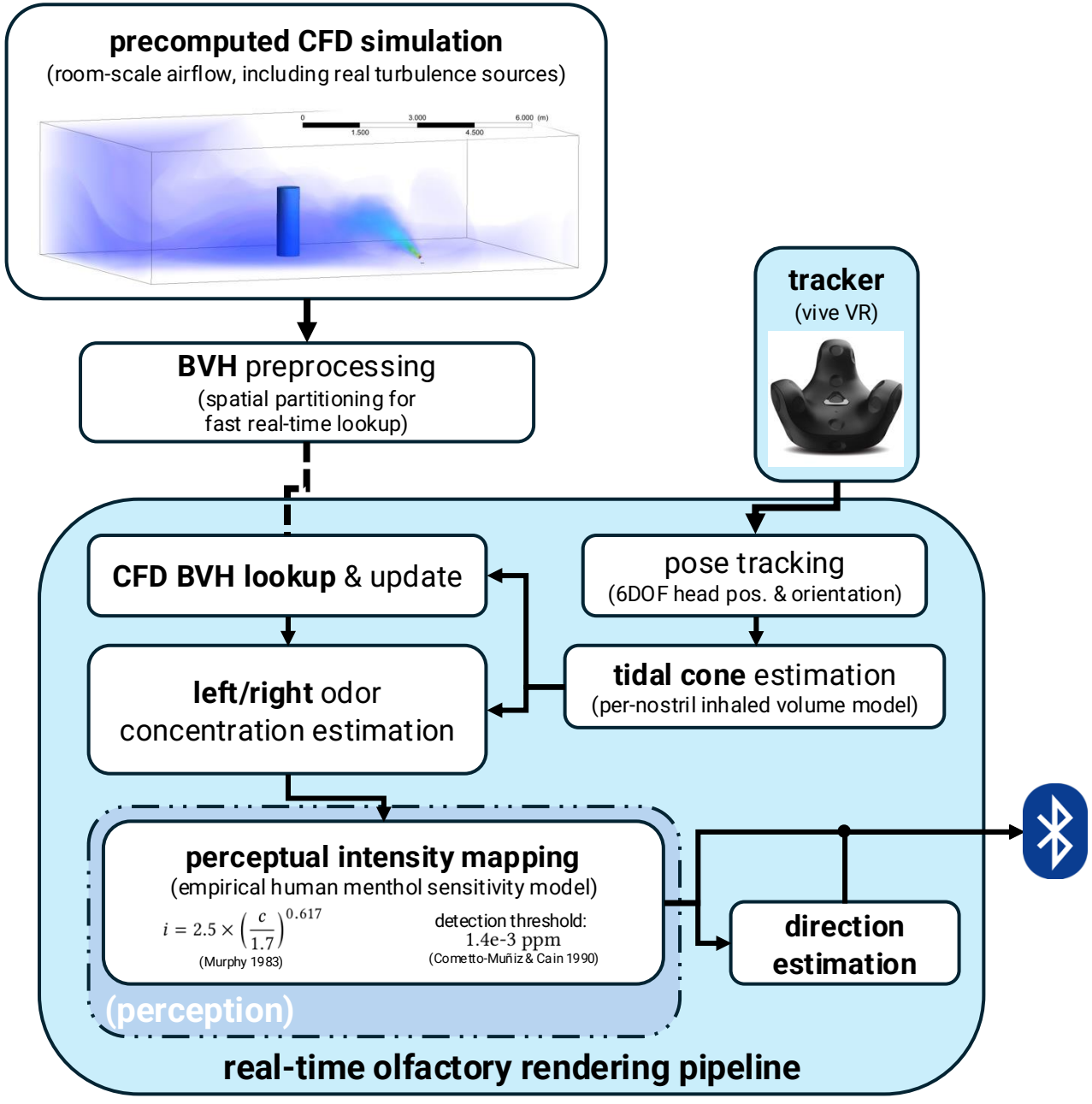
[slides from Jas Brooks]



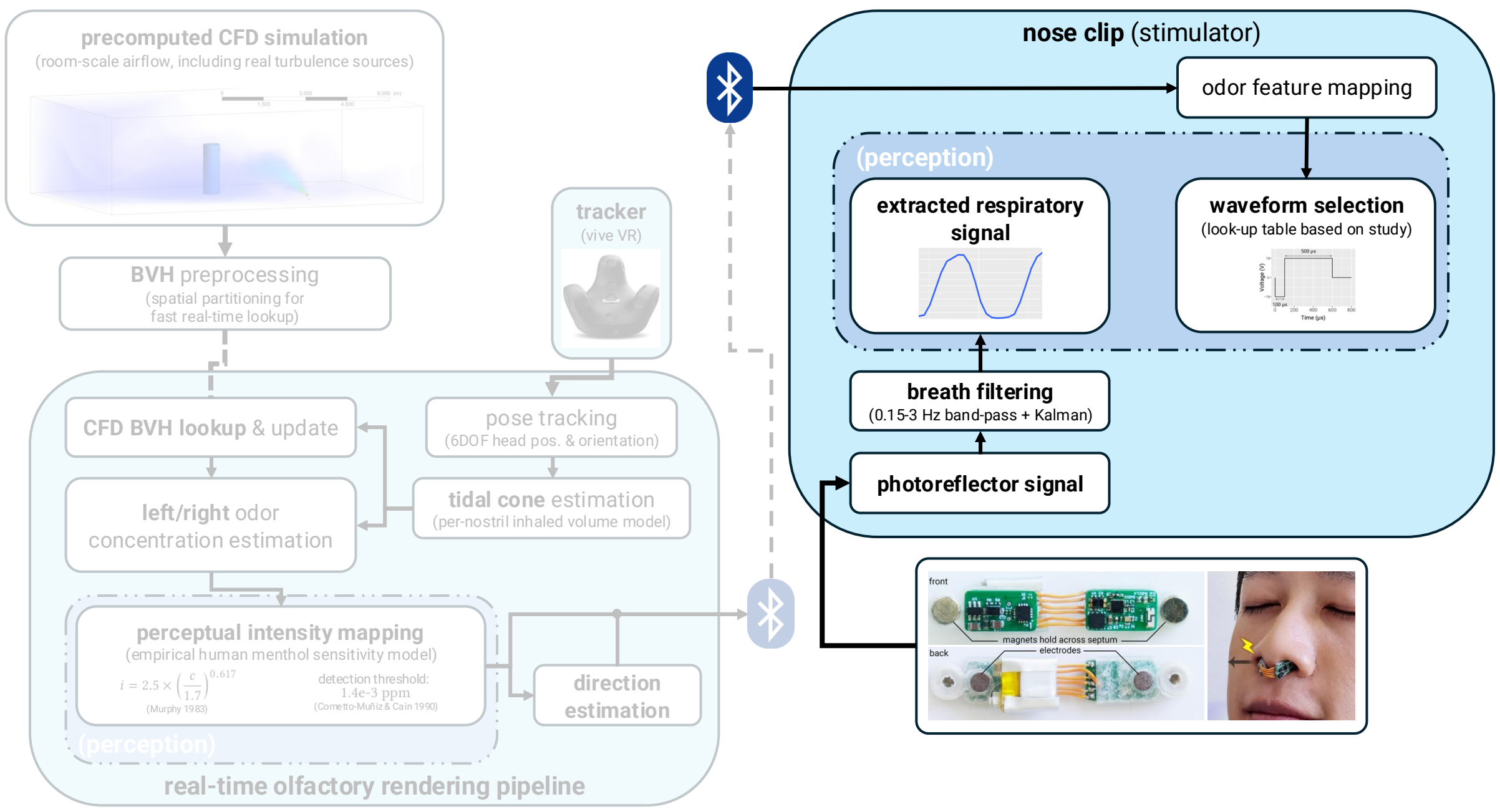




[slides from Jas Brooks]

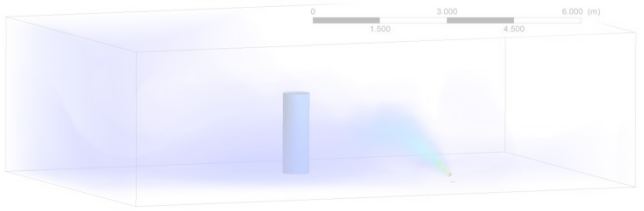


[slides from Jas Brooks]

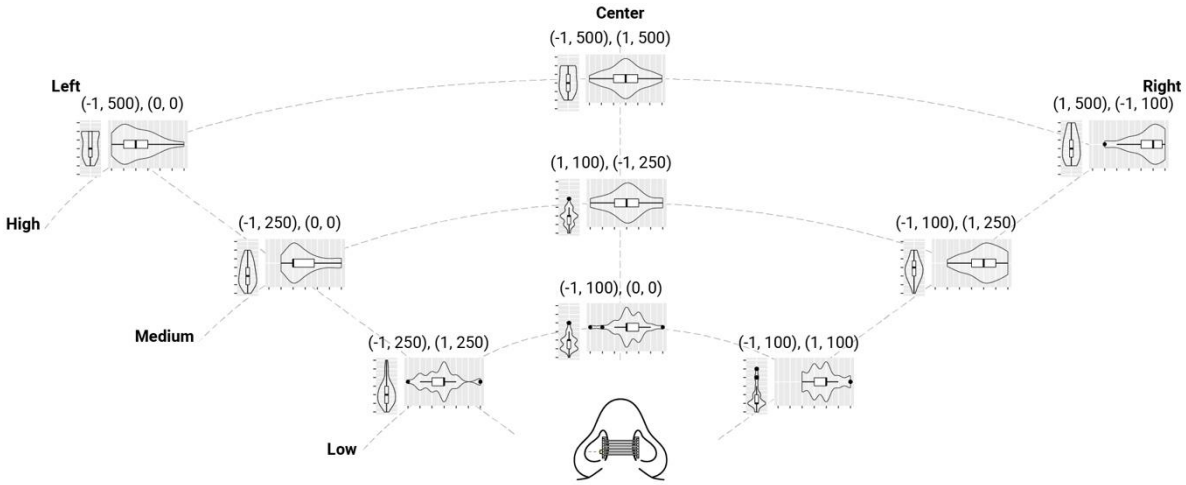


[slides from Jas Brooks]

precomputed CFD simulation
(room-scale airflow, including real turbulence sources)



tracker
(vive VR)



perceptual intensity mapping
(empirical human menthol sensitivity model)

$$i = 2.5 \times \left(\frac{c}{1.7}\right)^{0.617}$$

(Murphy 1983)

detection threshold:
1.4e-3 ppm
(Cometto-Muñiz & Cain 1990)

(perception)

direction estimation

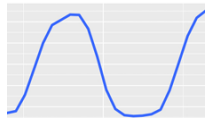
real-time olfactory rendering pipeline

nose clip (stimulator)

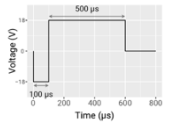
odor feature mapping

(perception)

extracted respiratory signal

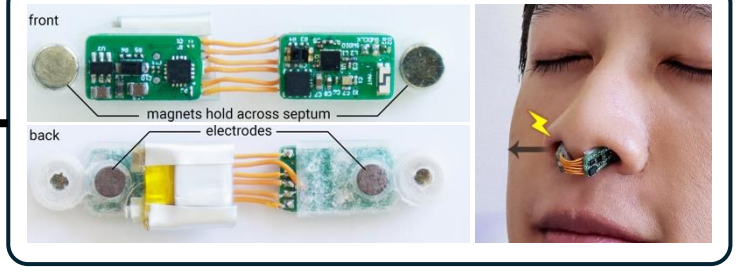


waveform selection
(look-up table based on study)



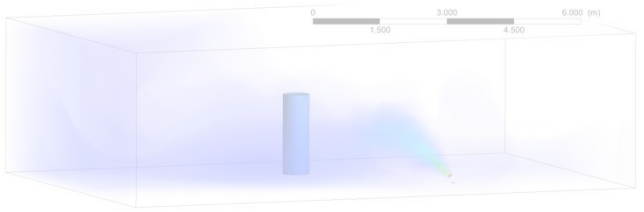
breath filtering
(0.15-3 Hz band-pass + Kalman)

photoreflector signal



[slides from Jas Brooks]

precomputed CFD simulation
(room-scale airflow, including real turbulence sources)



BVH preprocessing
(spatial partitioning for fast real-time lookup)

tracker
(vive VR)

CFD BVH lookup & update

pose tracking
(6DOF head pos. & orientation)

left/right odor concentration estimation

tidal cone estimation
(per-nostril inhaled volume model)

perceptual intensity mapping
(empirical human menthol sensitivity model)

$$i = 2.5 \times \left(\frac{c}{1.7}\right)^{0.617}$$

(Murphy 1983)

detection threshold:
1.4e-3 ppm
(Cometto-Muñiz & Cain 1990)

direction estimation

(perception)

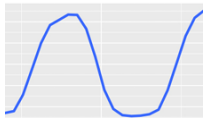
real-time olfactory rendering pipeline

nose clip (stimulator)

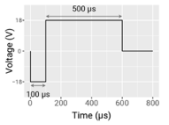
odor feature mapping

(perception)

extracted respiratory signal



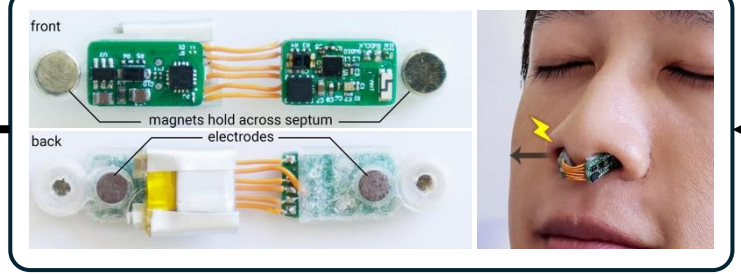
waveform selection
(look-up table based on study)



breath filtering
(0.15-3 Hz band-pass + Kalman)

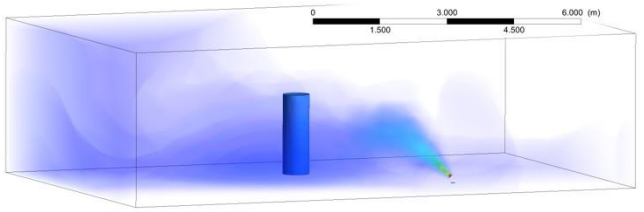
photoreflector signal

inhalation-synced stimulation scheduler
(trigger = start of inhale)



[slides from Jas Brooks]

precomputed CFD simulation
(room-scale airflow, including real turbulence sources)



BVH preprocessing
(spatial partitioning for fast real-time lookup)

tracker
(vive VR)

CFD BVH lookup & update

pose tracking
(6DOF head pos. & orientation)

left/right odor concentration estimation

tidal cone estimation
(per-nostril inhaled volume model)

perceptual intensity mapping
(empirical human menthol sensitivity model)

$$i = 2.5 \times \left(\frac{c}{1.7}\right)^{0.617}$$

(Murphy 1983)

detection threshold:
1.4e-3 ppm
(Cometto-Muñiz & Cain 1990)

(perception)

direction estimation

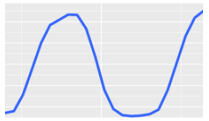
real-time olfactory rendering pipeline

nose clip (stimulator)

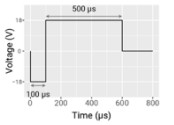
odor feature mapping

(perception)

extracted respiratory signal



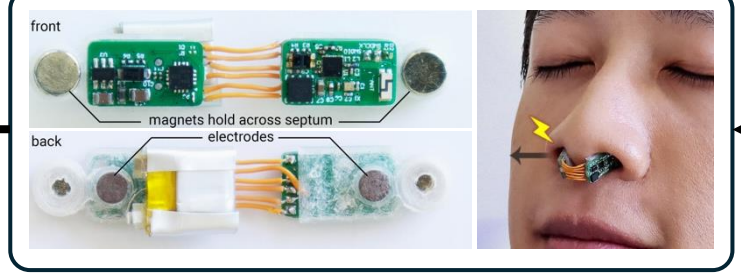
waveform selection
(look-up table based on study)



breath filtering
(0.15-3 Hz band-pass + Kalman)

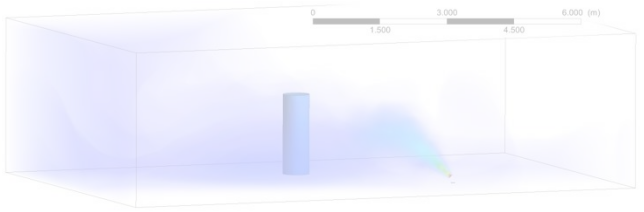
photoreflector signal

inhalation-synced stimulation scheduler
(trigger = start of inhale)



[slides from Jas Brooks]

precomputed CFD simulation
(room-scale airflow, including real turbulence sources)



BVH preprocessing
(spatial partitioning for fast real-time lookup)

CFD BVH lookup & update

left/right odor concentration estimation

perceptual intensity mapping
(empirical human menthol sensitivity model)

$$i = 2.5 \times \left(\frac{c}{1.7}\right)^{0.617}$$

(Murphy 1983)

detection threshold: 1.4e-3 ppm
(Cometto-Muñiz & Cain 1990)

(perception)

direction estimation

real-time olfactory rendering pipeline

tracker
(vive VR)

nose clip (stimulator)

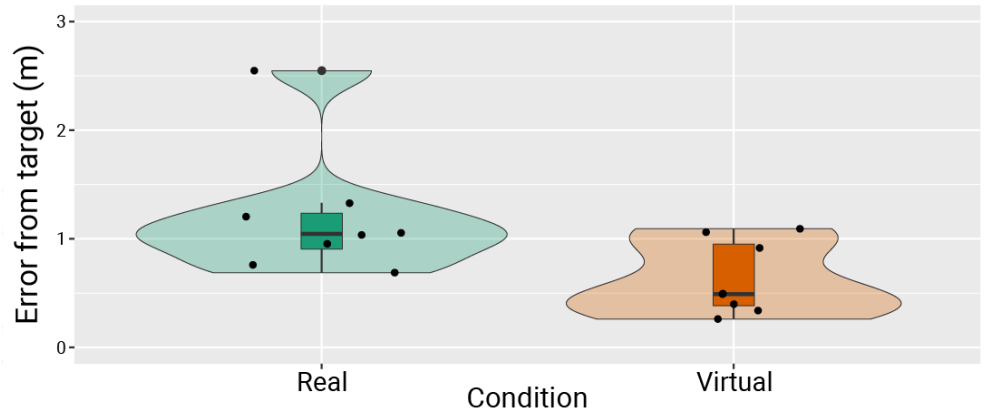
odor feature mapping

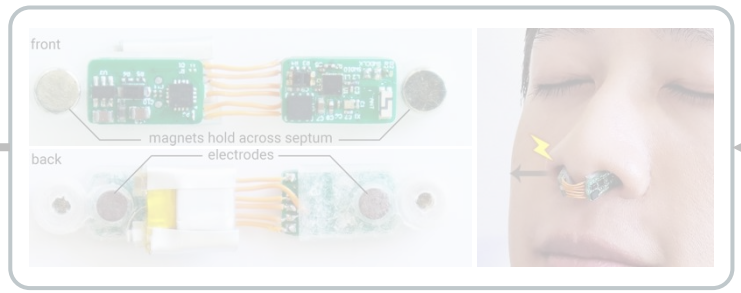
(perception)

extracted respiratory signal

waveform selection
(look-up table based on study)

inhalation-synced stimulation scheduler
(trigger = start of inhale)





[slides from Jas Brooks]

my approach
chemical interfaces for taste

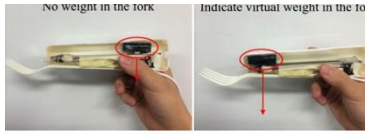
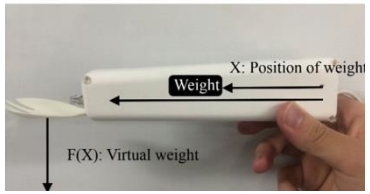
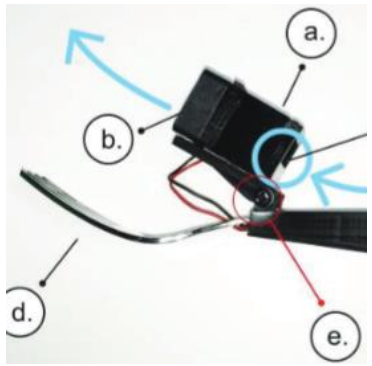
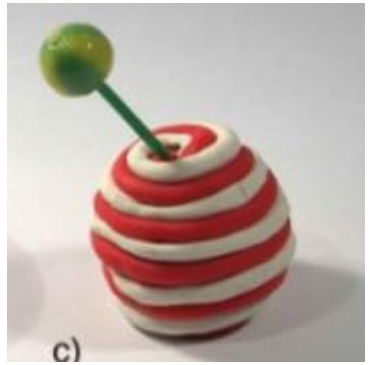


Figure 5. Internal appearance of the device

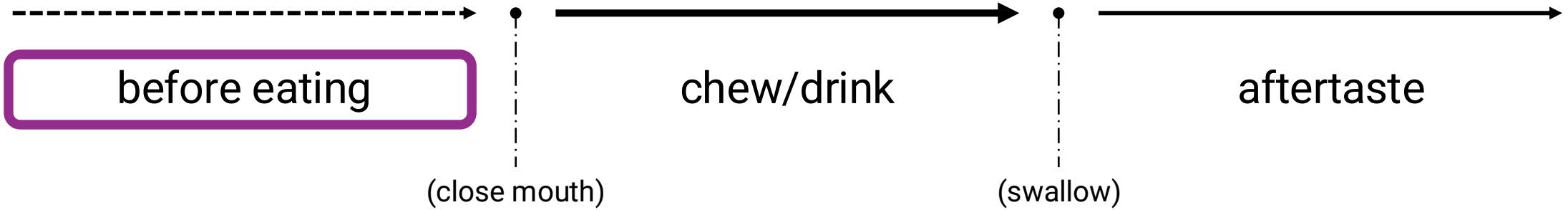






[slides from Jas Brooks]

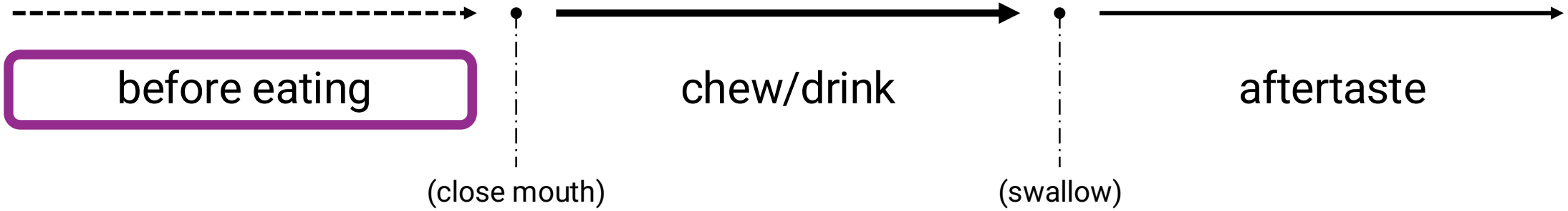
problem 1: taste is **inside** the mouth



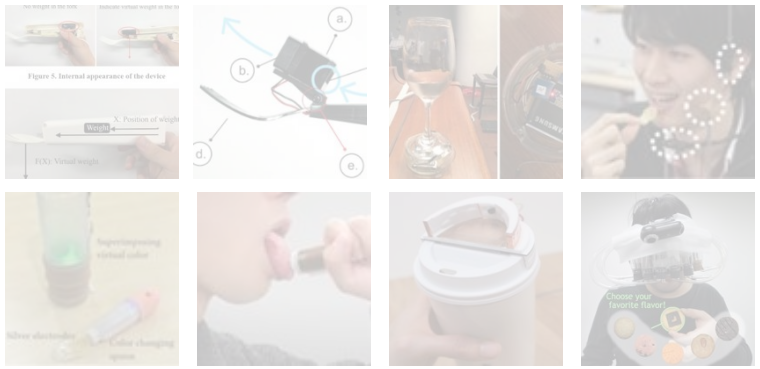
[slides from Jas Brooks]

problem 1: taste is **inside** the mouth

problem 2: imprecise feedback

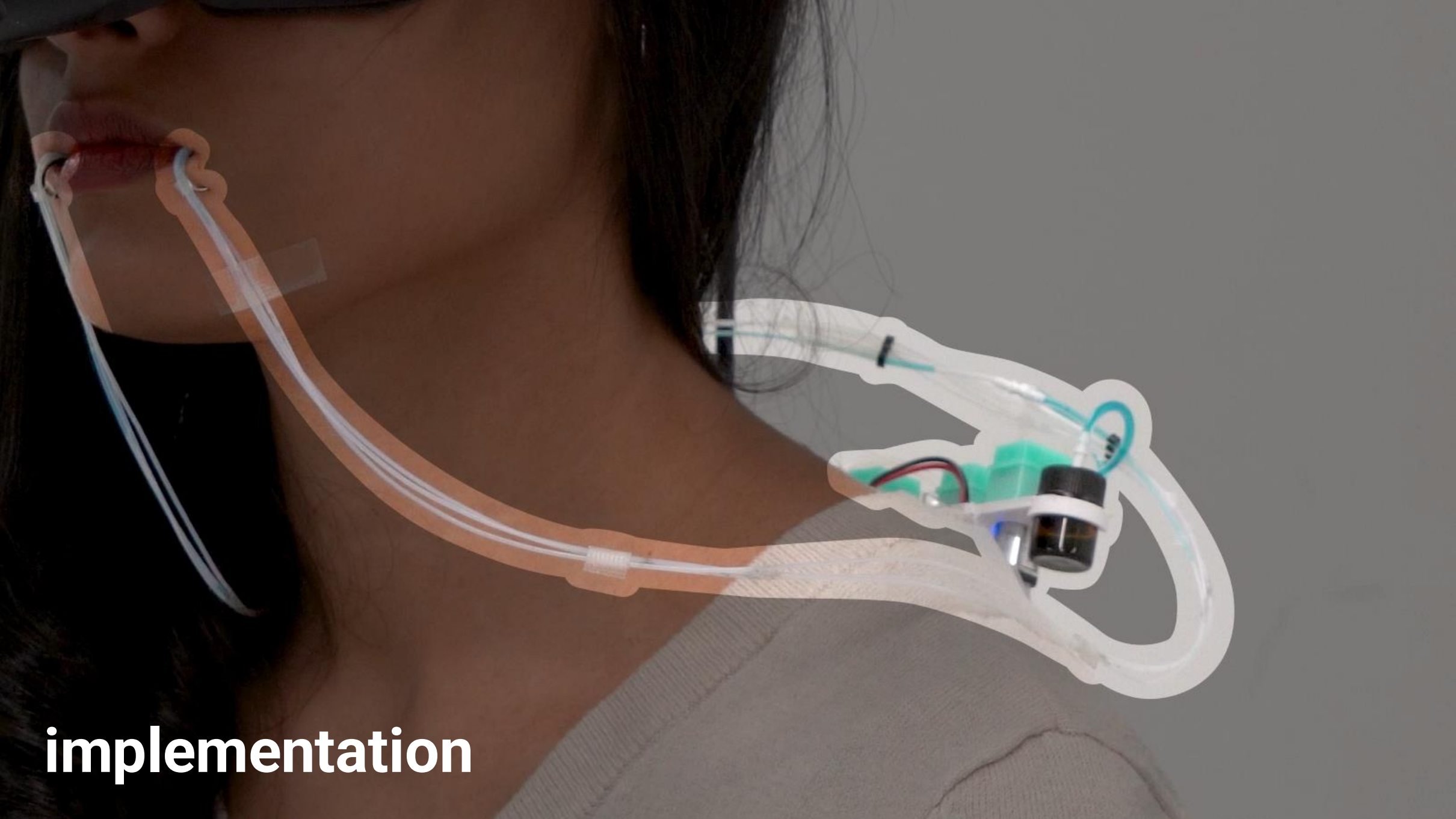


[slides from Jas Brooks]



our goal:
can we **modify** taste **selectively while eating**?

[slides from Jas Brooks]

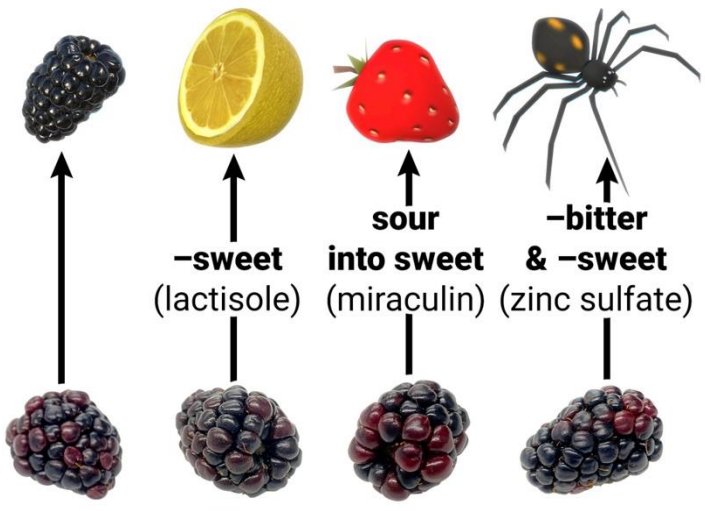


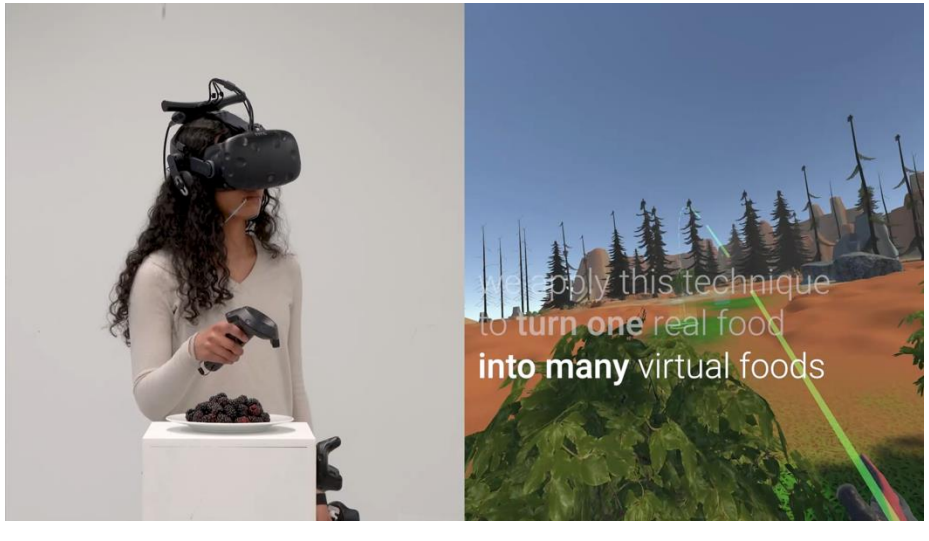
implementation

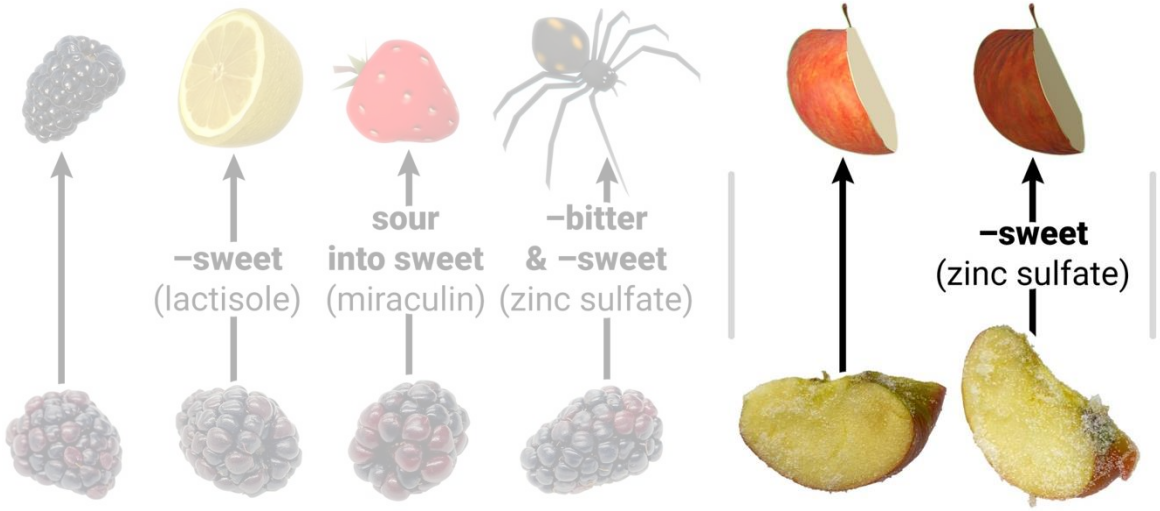
–sweet
SWEET
LOW
LOW
LOW
LOW

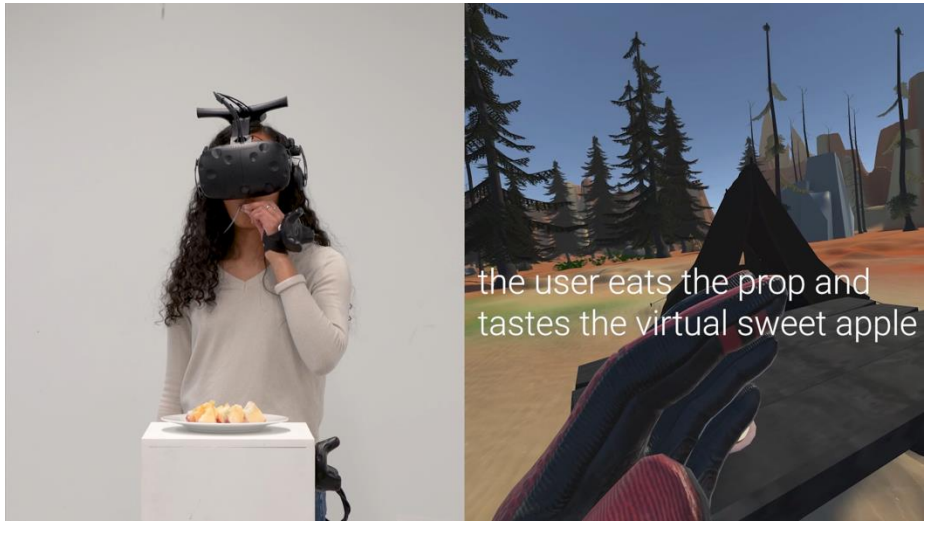

–sweet

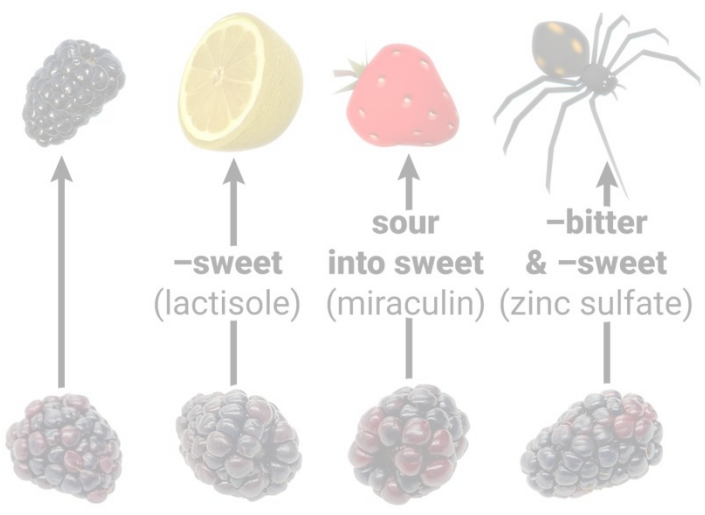

Coca-Cola

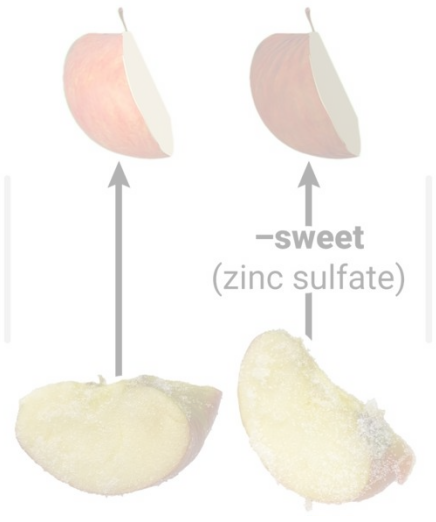

−sweet
(lactisole)
sour
into sweet
(miraculin)
−bitter
& −sweet
(zinc sulfate)

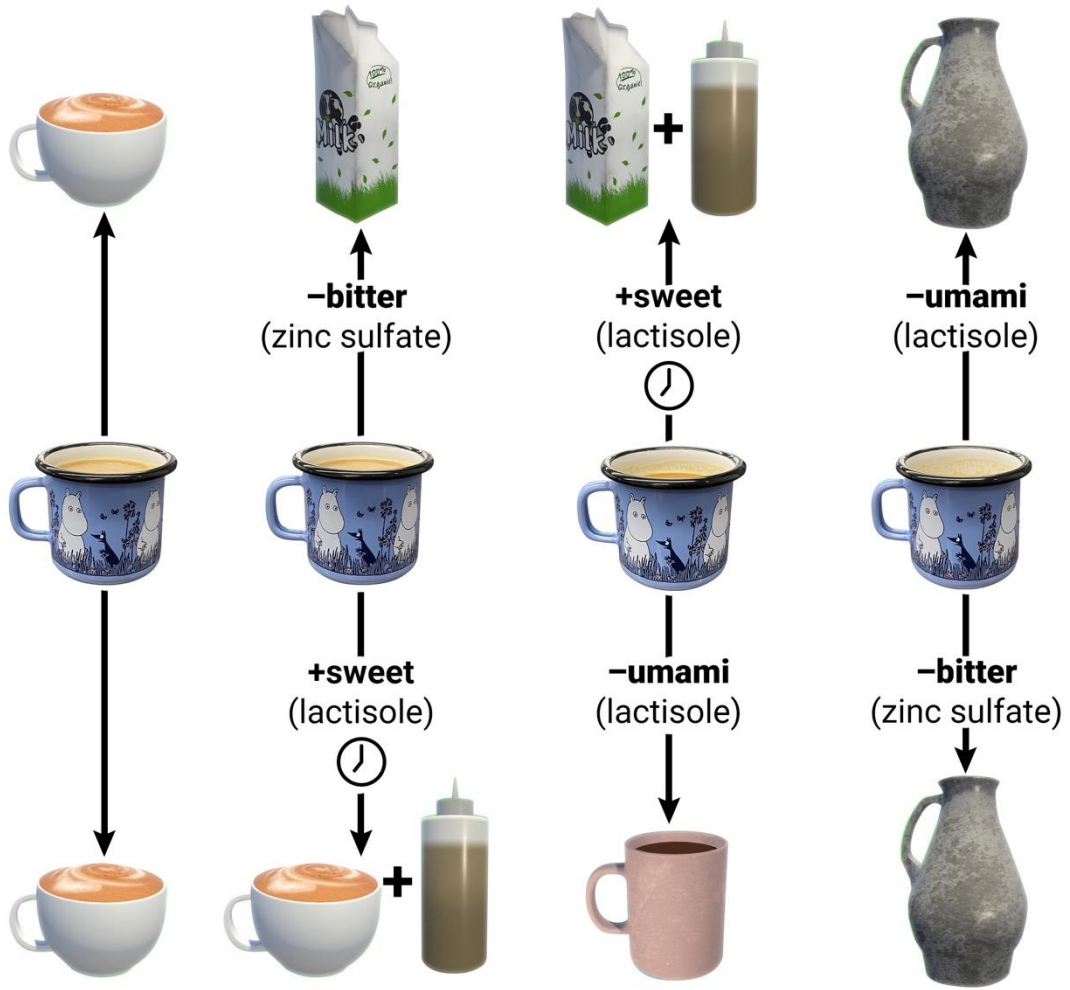

we apply this technique
to turn one real food
into many virtual foods

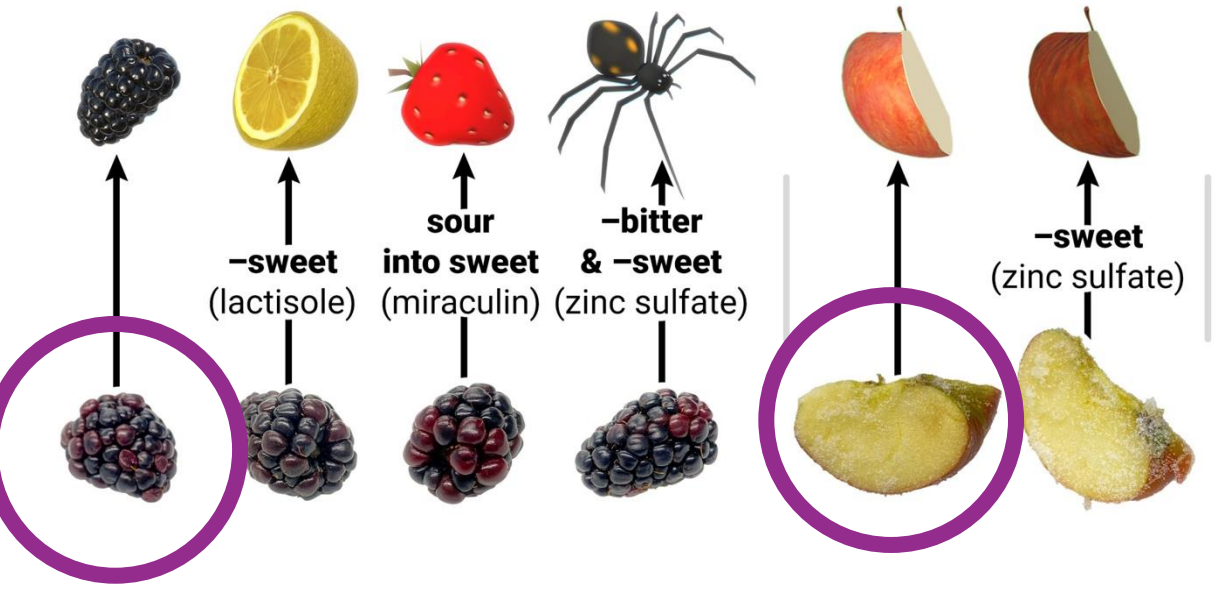

−sweet
(lactisole)
sour
into sweet
(miraculin)
−bitter
& −sweet
(zinc sulfate)
−sweet
(zinc sulfate)


the user eats the prop and
tastes the virtual sweet apple

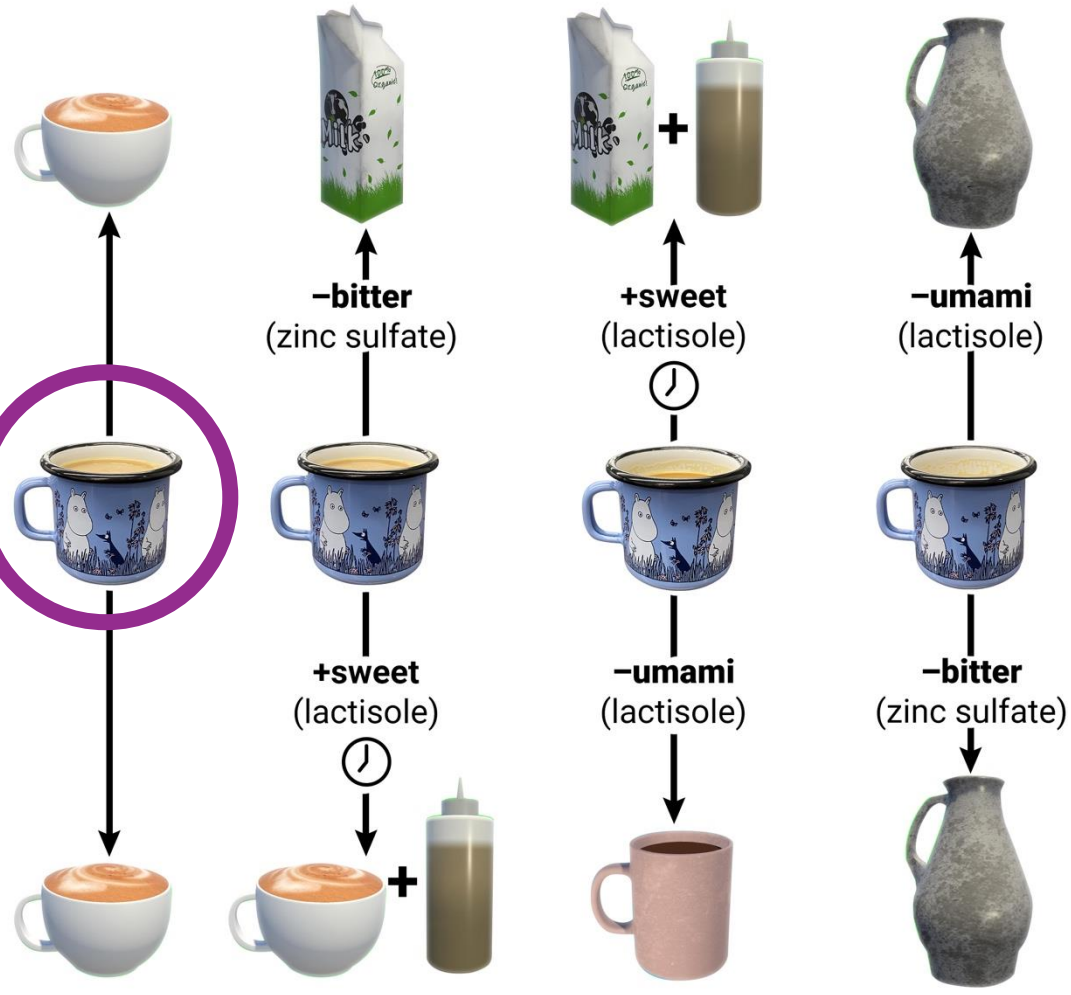

−sweet
(lactisole)
sour
into sweet
(miraculin)
−bitter
& −sweet
(zinc sulfate)


−sweet
(zinc sulfate)


−bitter
(zinc sulfate)
+sweet
(lactisole)
+sweet
(lactisole)
−umami
(lactisole)
−umami
(lactisole)
−bitter
(zinc sulfate)



compressed to **3 food props**

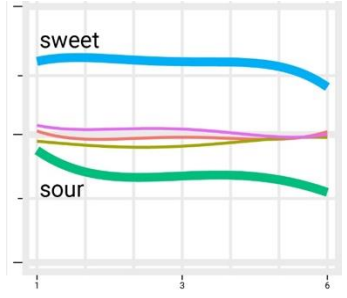


[slides from Jas Brooks]

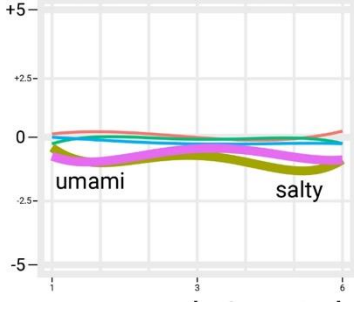
chemical interfaces for **taste**

user study: **temporal** effects of **modulators**

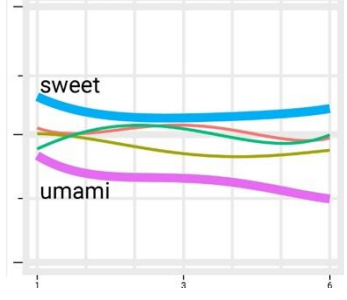
miraculin



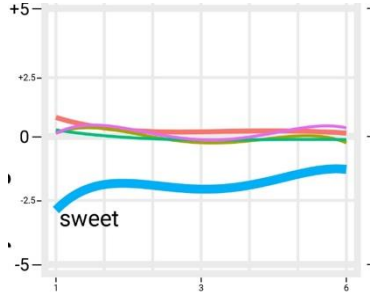
amiloride



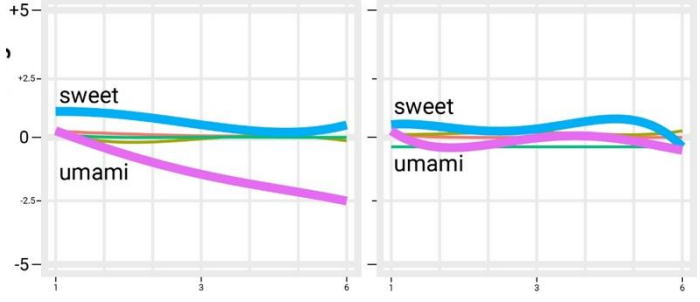
clofibric acid



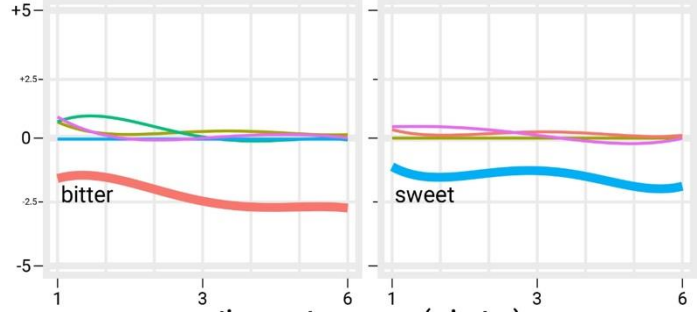
gymnemic acids



lactisole



zinc sulfate



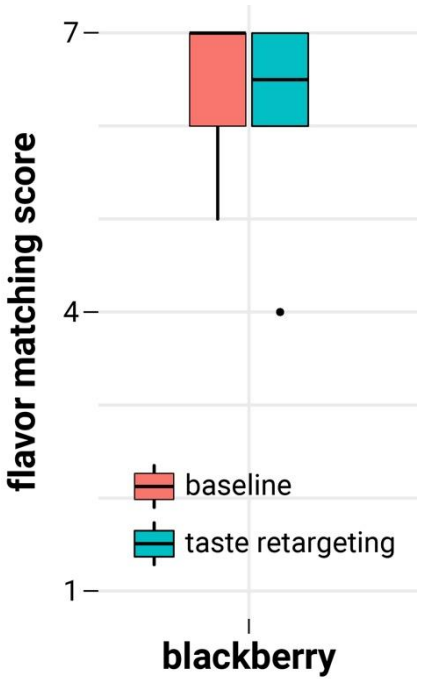
[slides from Jas Brooks]

UIST 2023

🏅 **demo honorable mention**

chemical interfaces for **taste**

user study: **compressing** VR food props

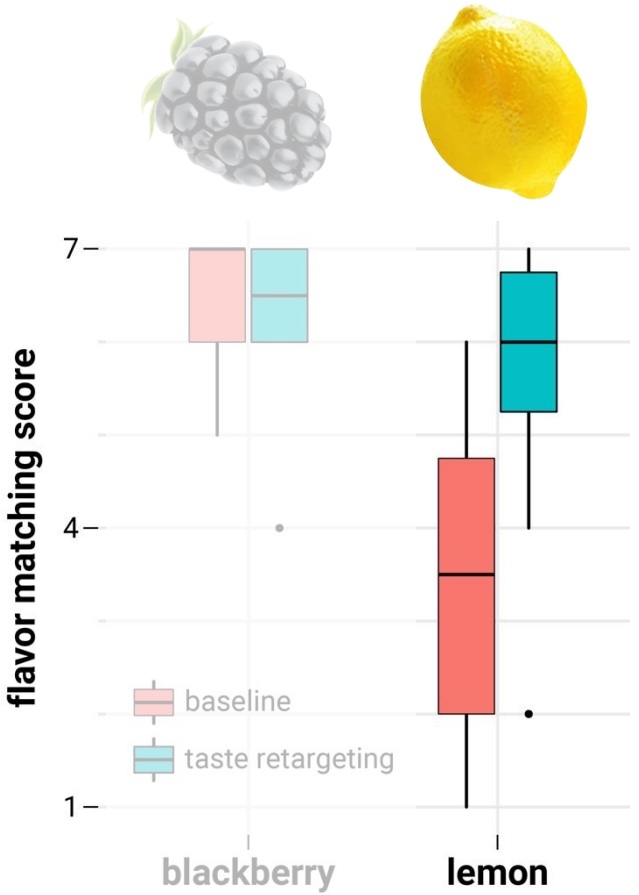


UIST 2023

🏅 **demo honorable mention**

chemical interfaces for **taste**

user study: **compressing** VR food props

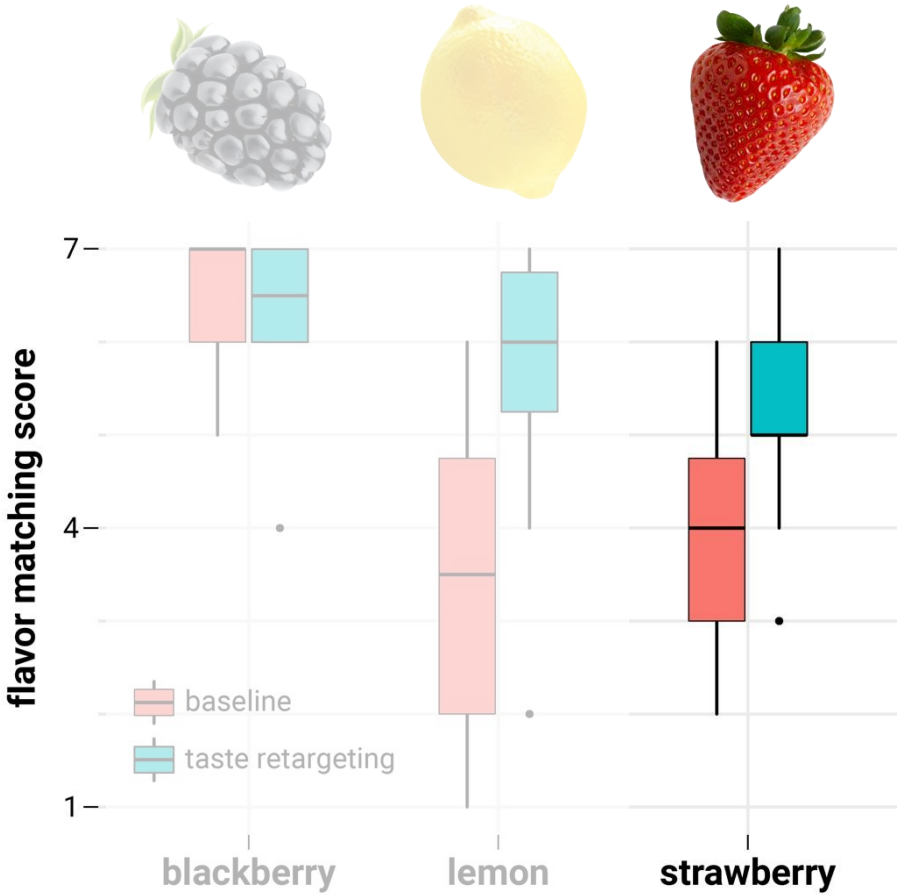


UIST 2023

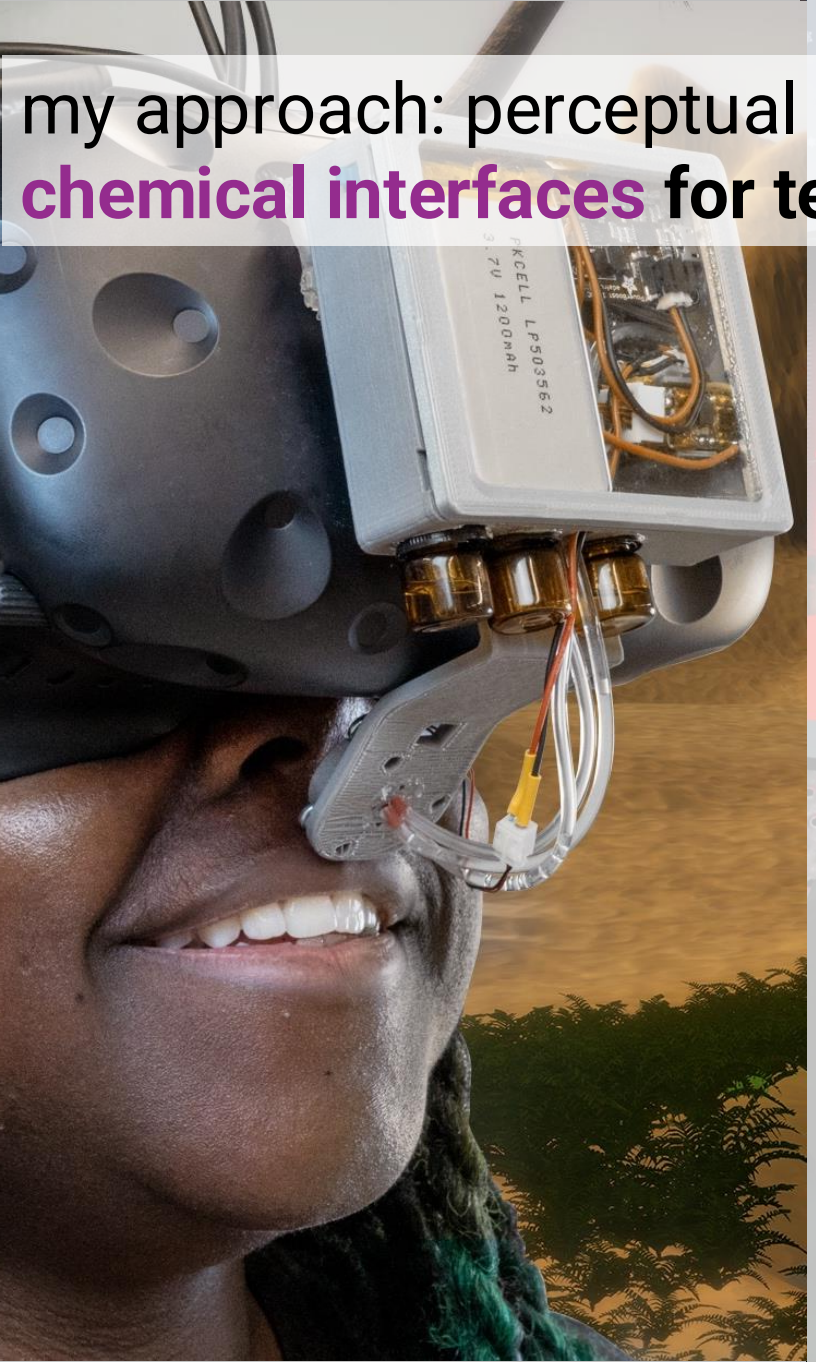
🏅 **demo honorable mention**

chemical interfaces for **taste**

user study: **compressing** VR food props

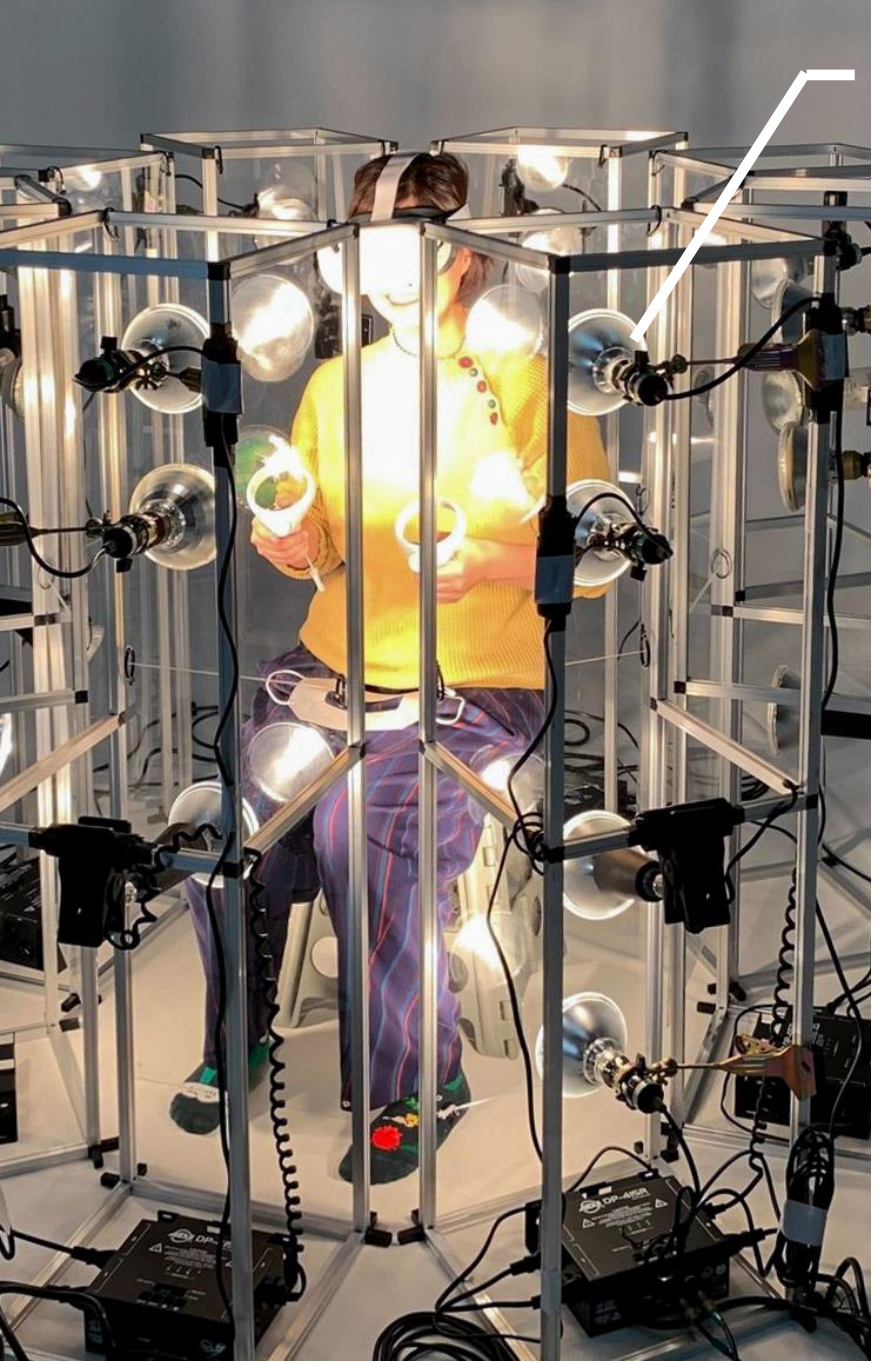


my approach: perceptual engineering
chemical interfaces for temperature



temperature is a critical component of these two experiences

virtual reality



heat lamp
1 kW
ambient temperature
not wearable

is there a ***fundamentally***

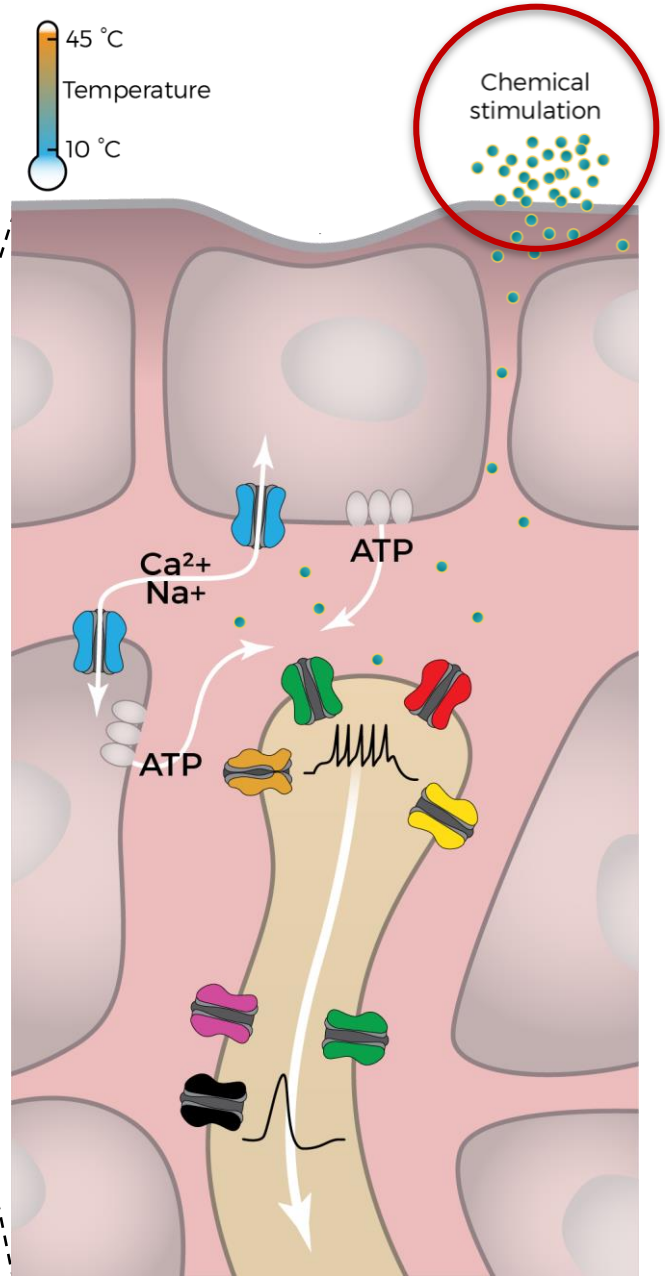
different approach to this?

trigeminal-based temperature illusions

jas brooks, steven nagels, pedro lopes

THE UNIVERSITY OF CHICAGO

ACM CHI Best Paper Award



capsaicin

[slides from Jas Brooks]

CHI 2023

🏆 **best paper award**

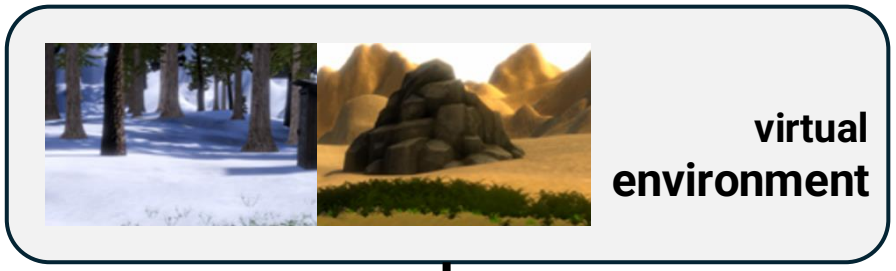
chemical interfaces for **temperature**

implementation

system goals

perceptual consistency **and** **power** efficiency

[slides from Jas Brooks]



virtual environment

unity3d

thermal & olfactory **event manager** (e.g., furnace shuts off)

wearable

capsaicin

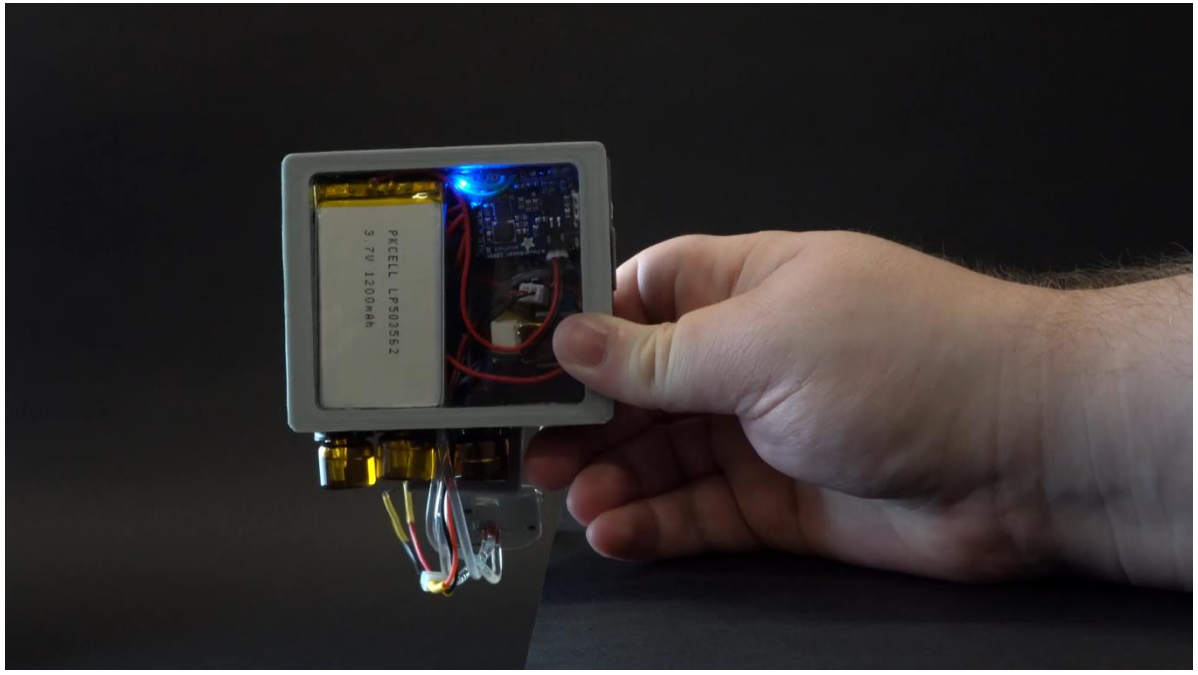
eucalyptol

fragrance

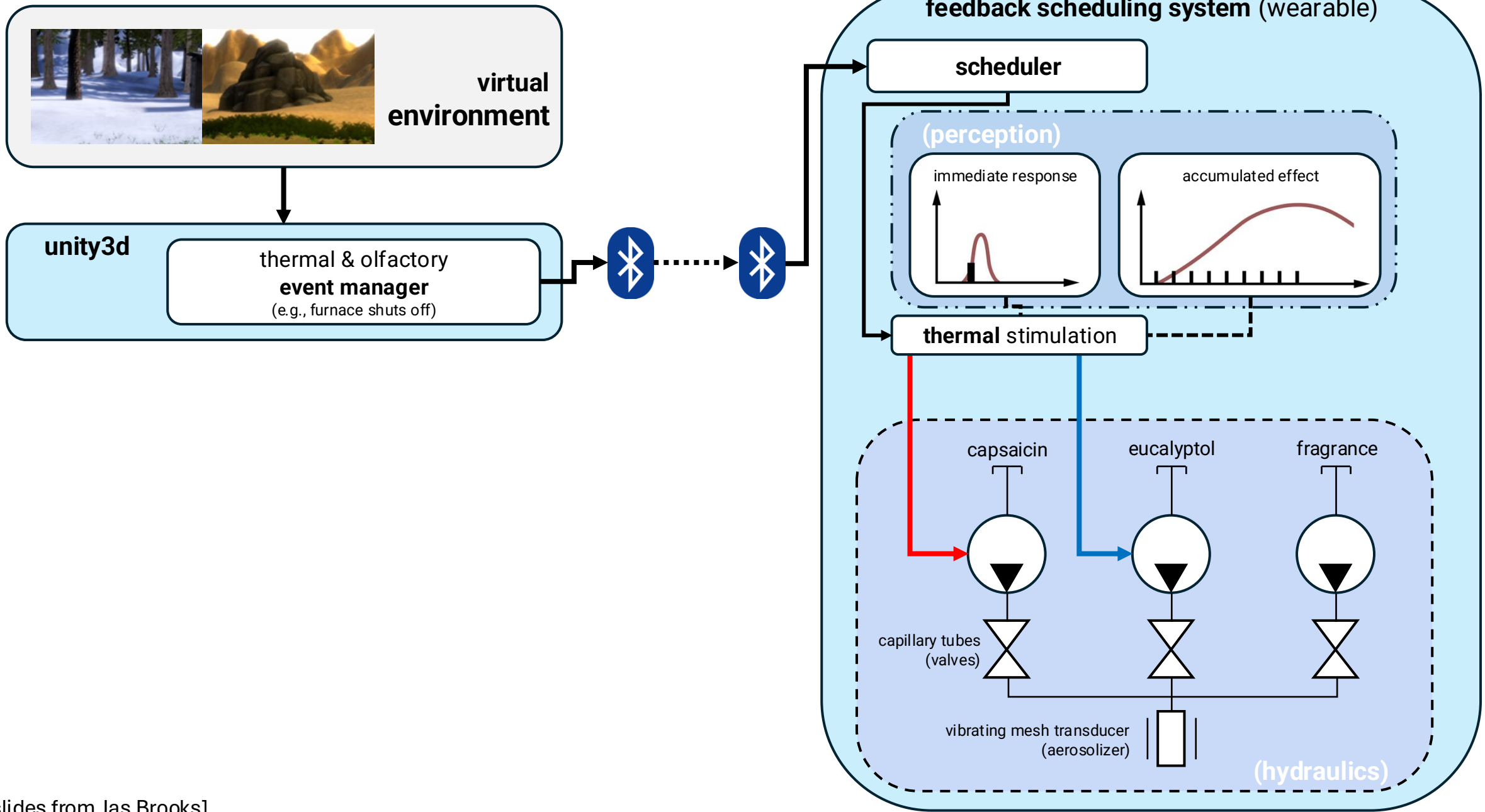
capillary tubes (valves)

vibrating mesh transducer (aerosolizer)

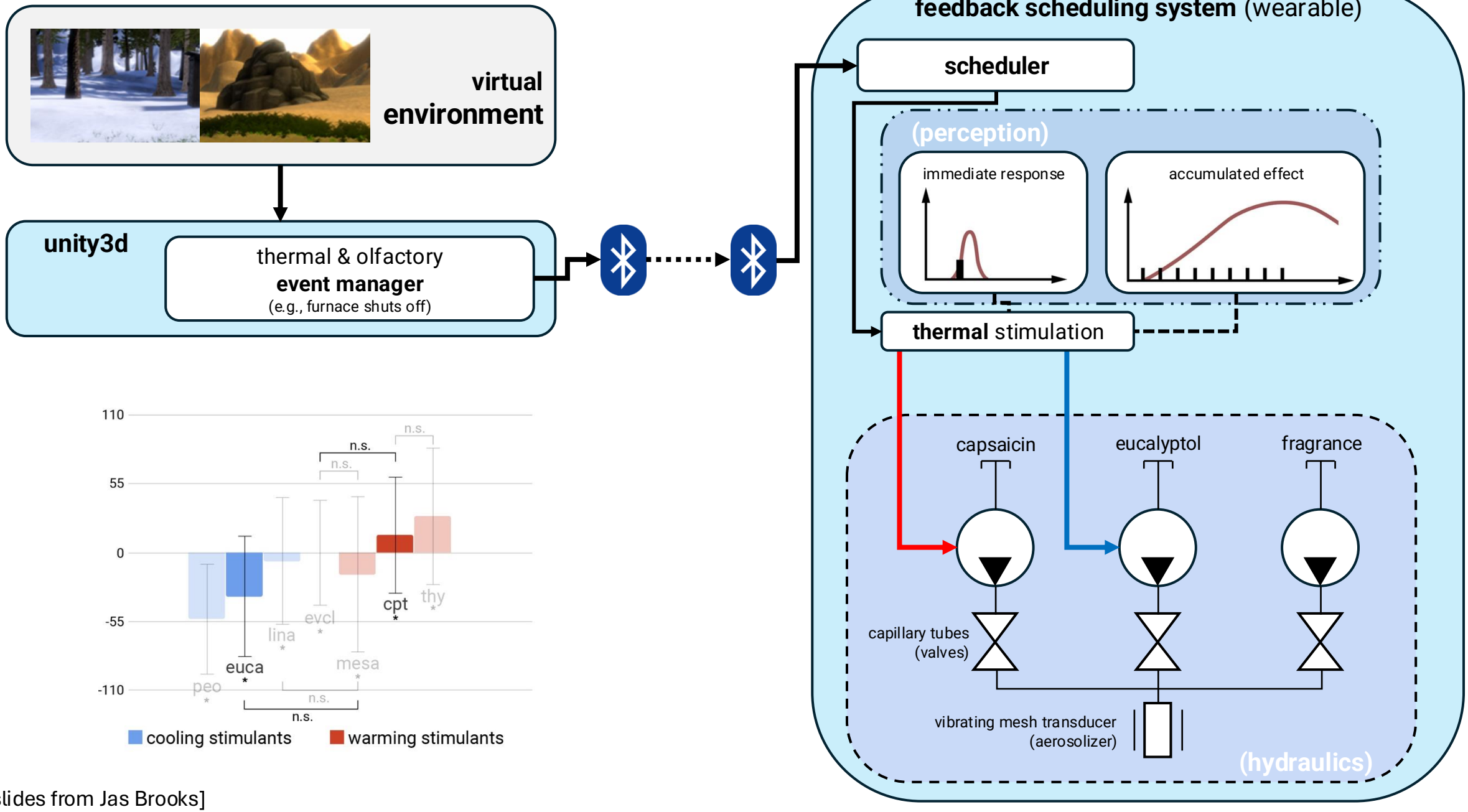
(hydraulics)



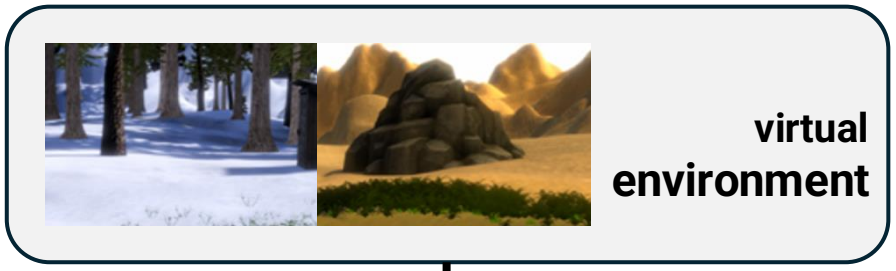
[slides from Jas Brooks]



[slides from Jas Brooks]



[slides from Jas Brooks]



virtual environment

unity3d

thermal & olfactory **event manager** (e.g., furnace shuts off)

feedback scheduling system (wearable)

scheduler

(perception)

immediate response

accumulated effect

thermal stimulation

capsaicin

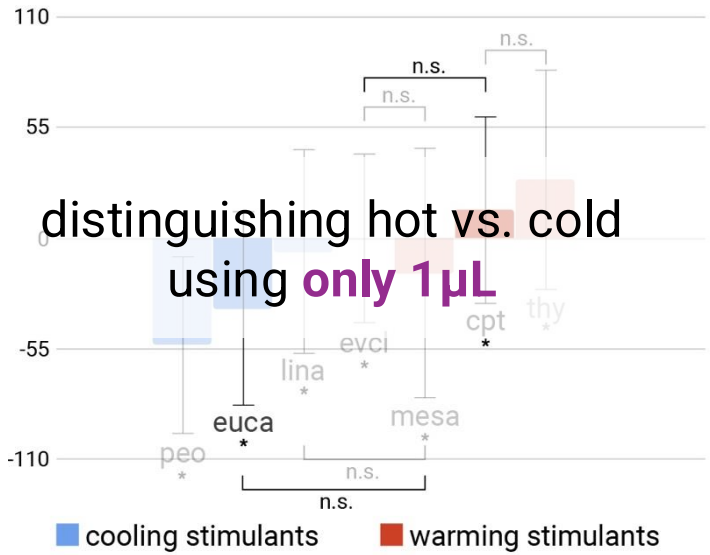
eucalyptol

fragrance

capillary tubes (valves)

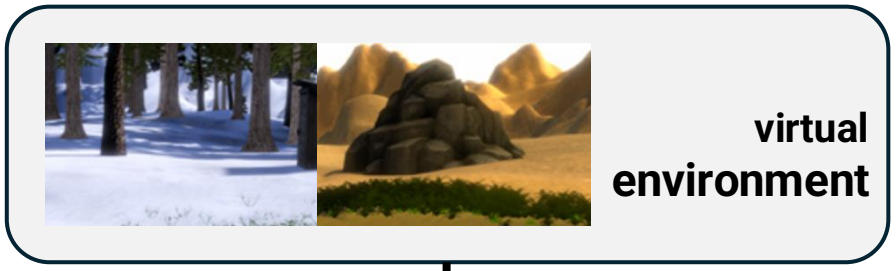
vibrating mesh transducer (aerosolizer)

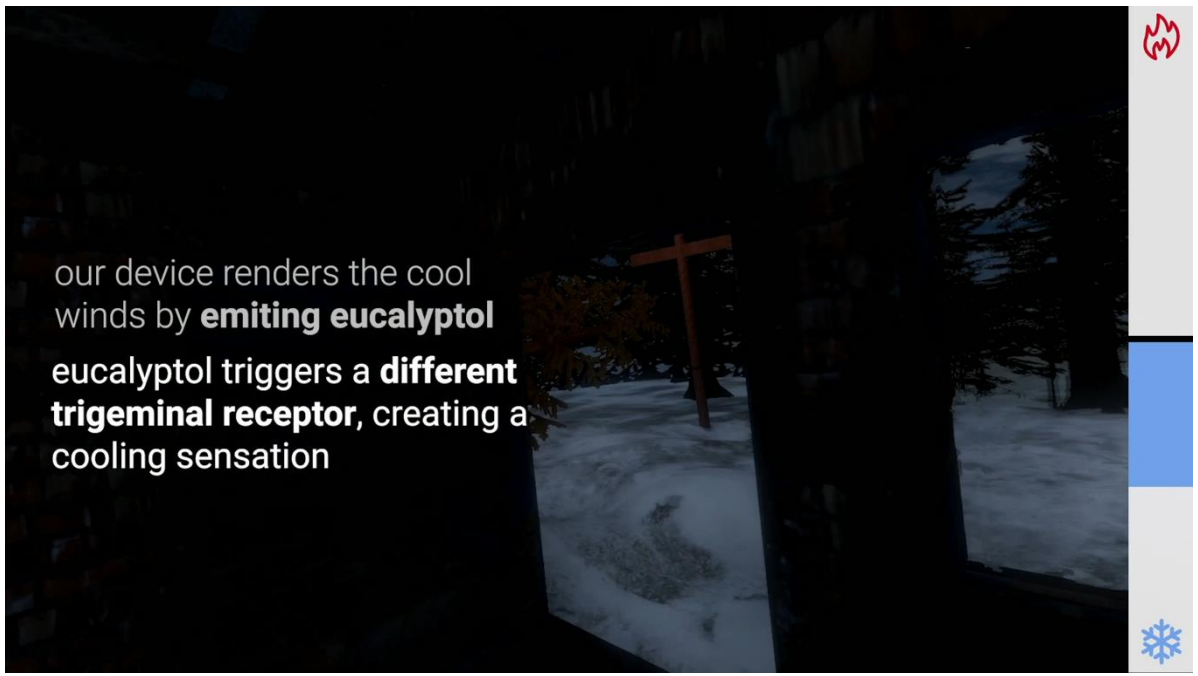
(hydraulics)

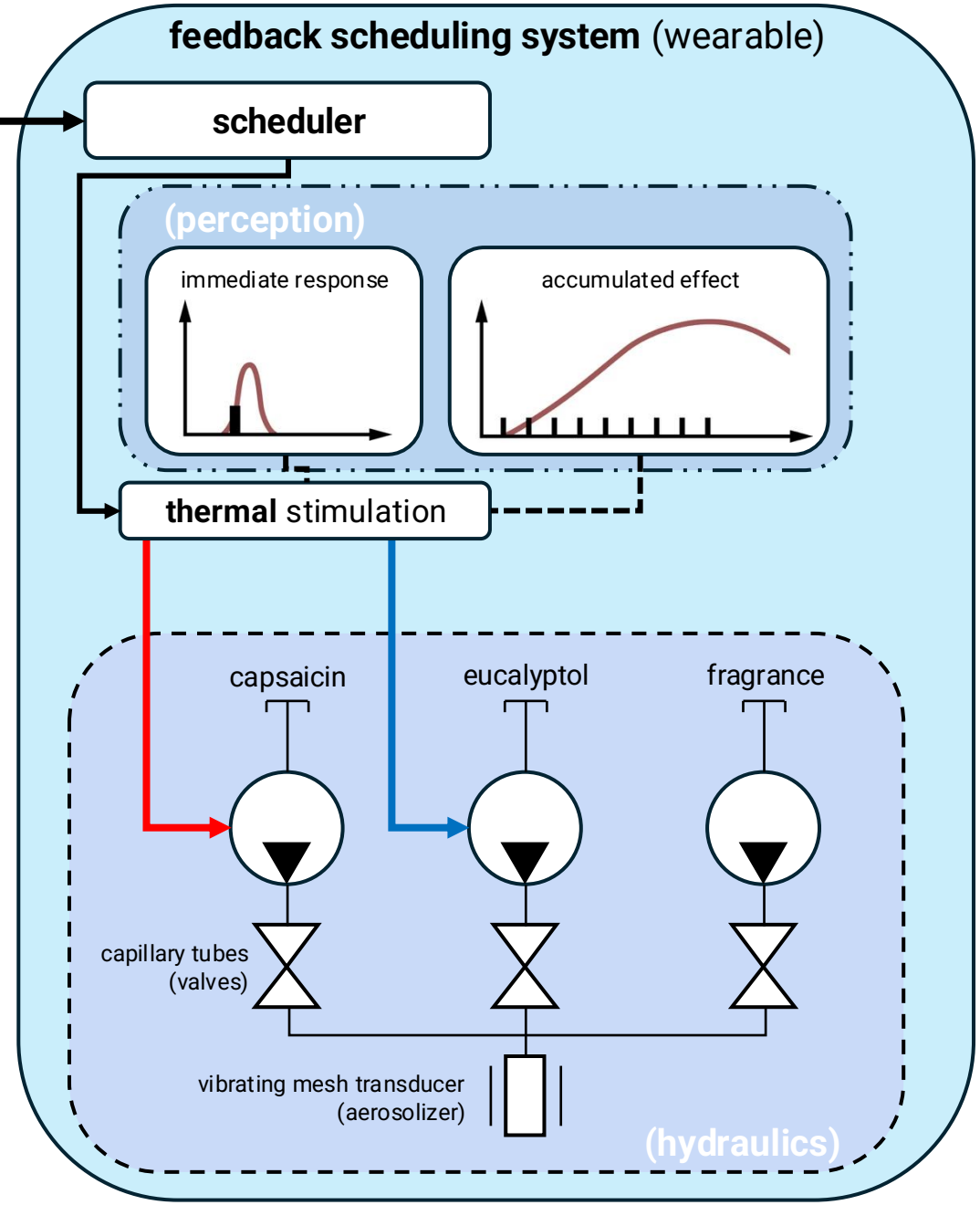


distinguishing hot vs. cold using **only 1µL**

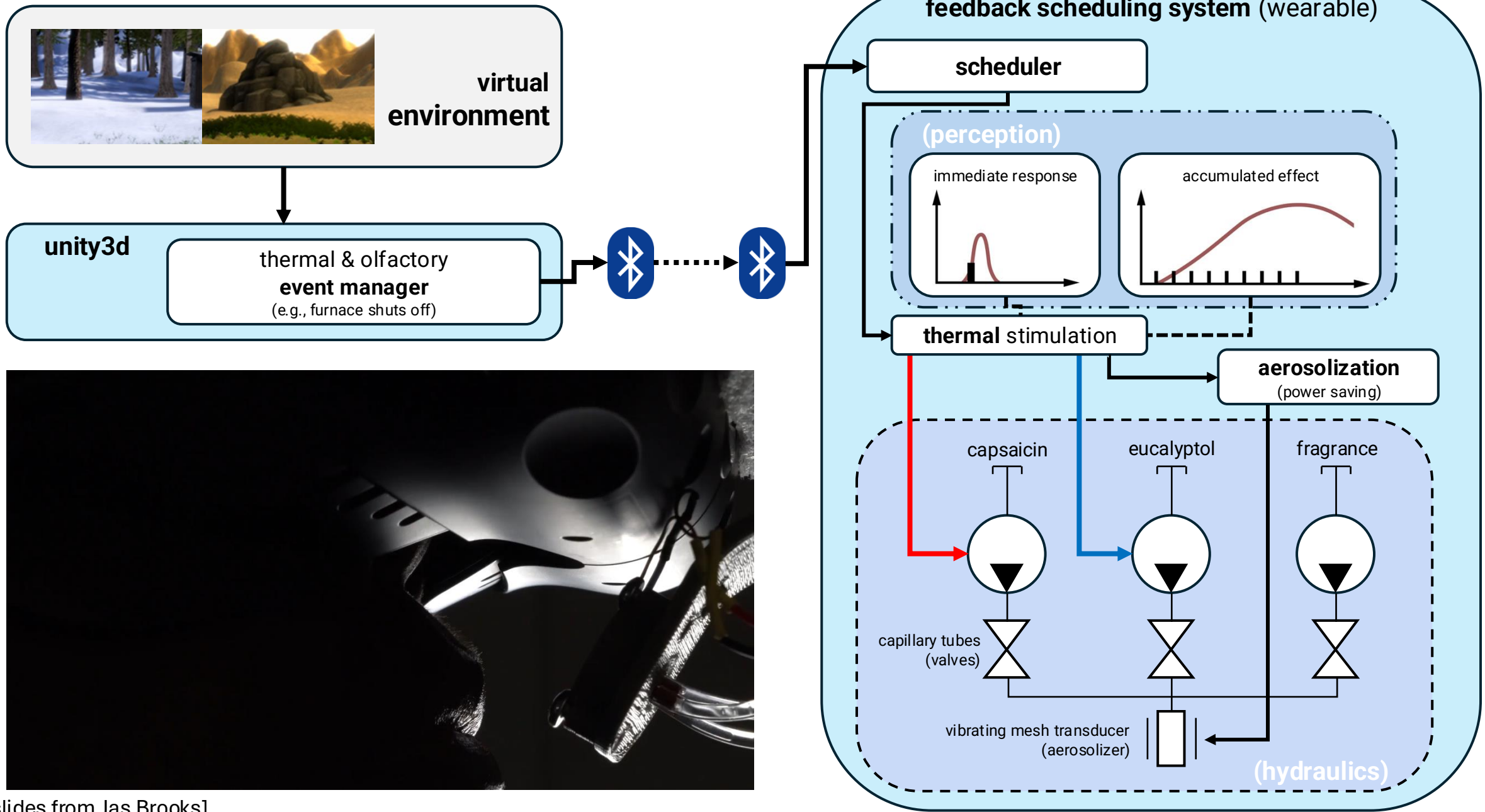
[slides from Jas Brooks]



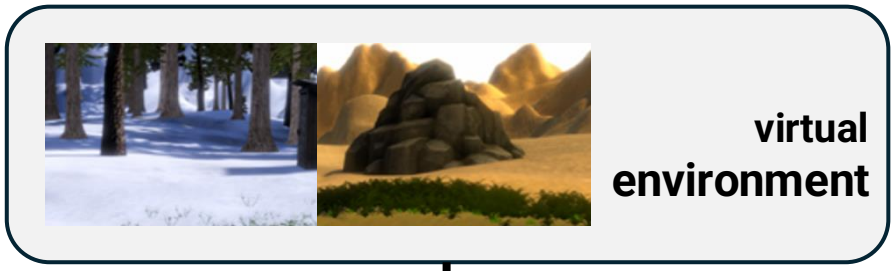




[slides from Jas Brooks]



[slides from Jas Brooks]



virtual environment

unity3d

thermal & olfactory **event manager** (e.g., furnace shuts off)

feedback scheduling system (wearable)

scheduler

olfactory stimulation

(perception)

immediate response

accumulated effect

thermal stimulation

aerosolization (power saving)

capsaicin

eucalyptol

fragrance

capillary tubes (valves)

vibrating mesh transducer (aerosolizer)

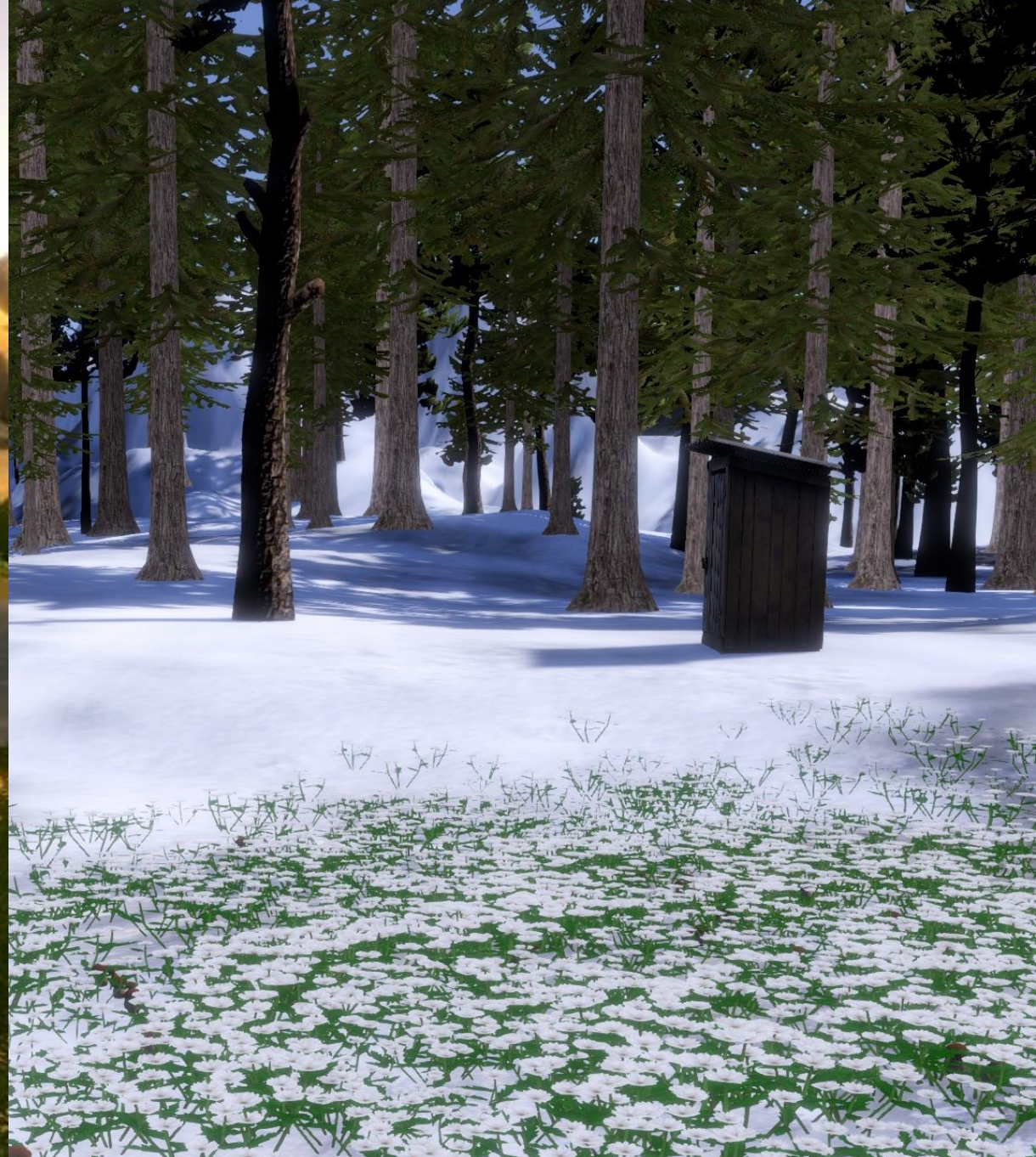
(hydraulics)

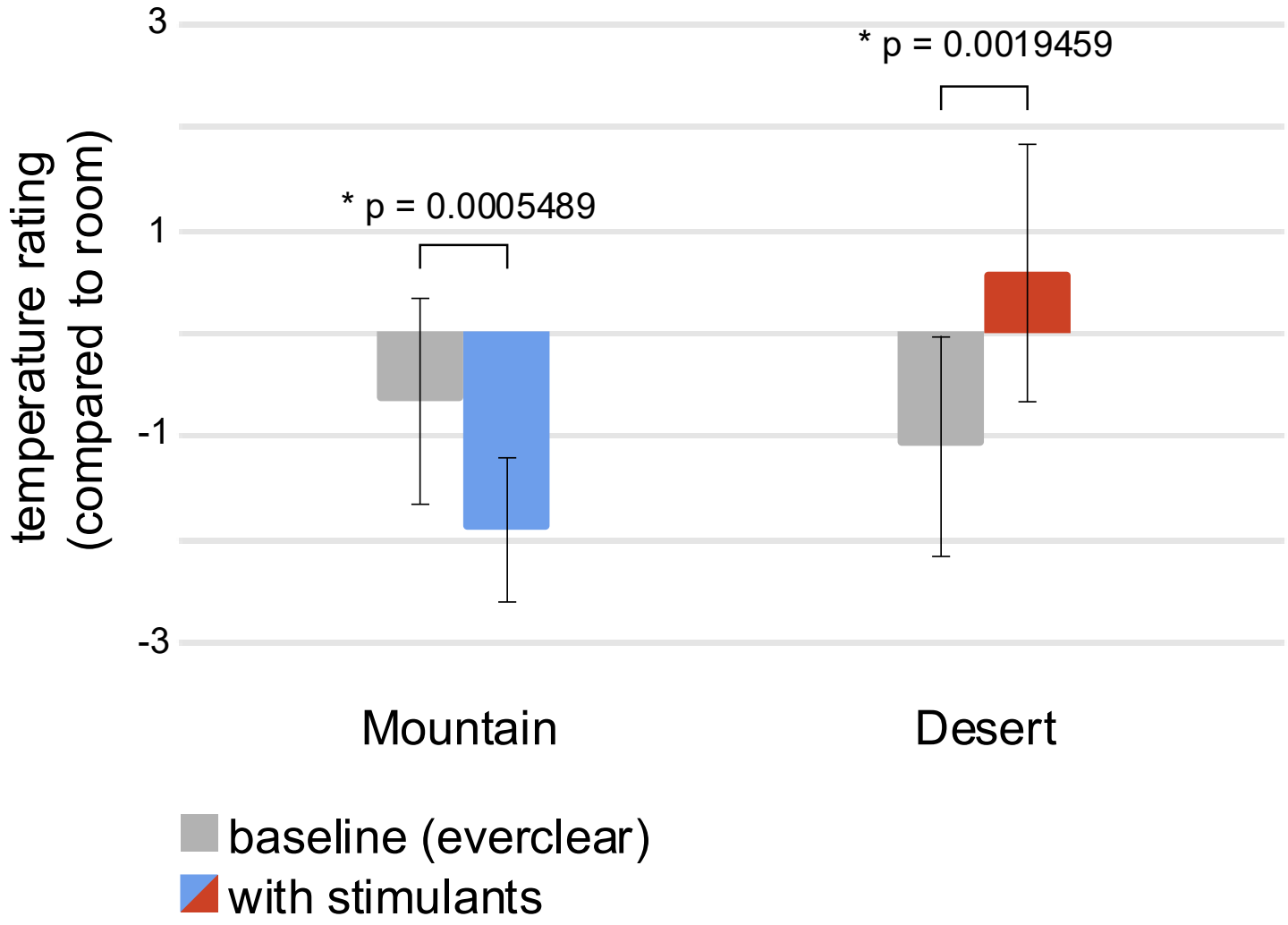
[slides from Jas Brooks]

chemical interfaces for **temperature**

impact on virtual environment

[slides from Jas Brooks]





[slides from Jas Brooks]

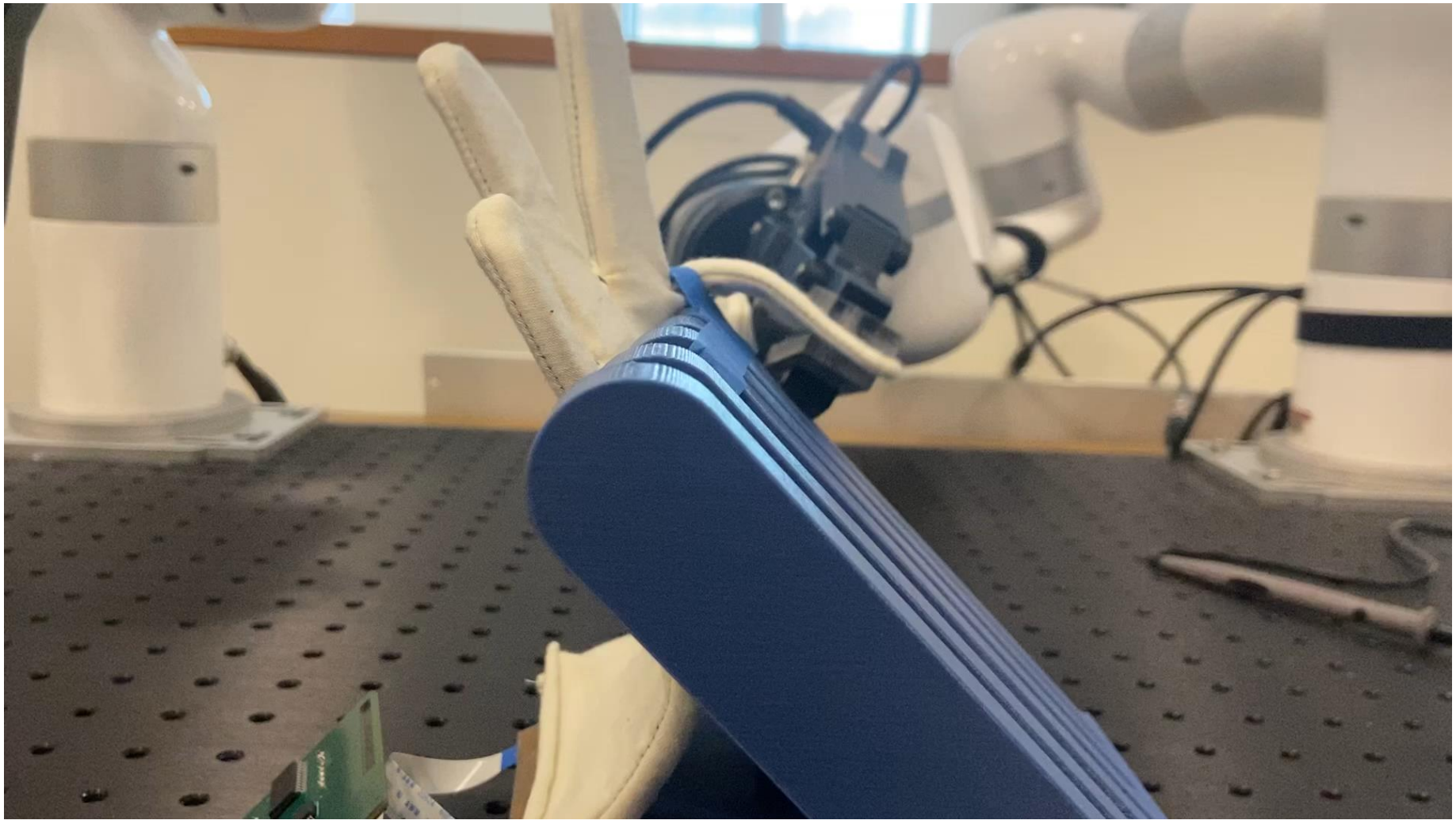
AI for Touch

https://wiresens-gloves.vercel.app

[Murphy et al., Fits like a Flex-Glove: Automatic Design of Personalized FPCB-Based Tactile Sensing Gloves. CHI LBW 2025]

Sensor Robustness

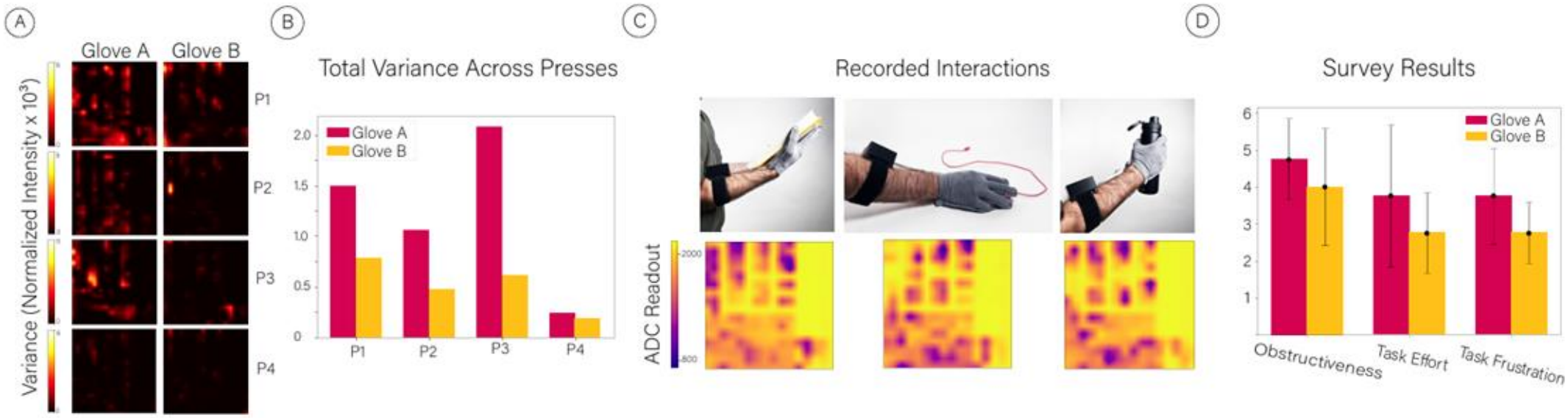
https://wiresens-gloves.vercel.app



[Murphy et al., Fits like a Flex-Glove: Automatic Design of Personalized FPCB-Based Tactile Sensing Gloves. CHI LBW 2025]

Usability Studies

https://wiresens-gloves.vercel.app



User study results from generic (Glove A) and personalized (Glove B) gloves:

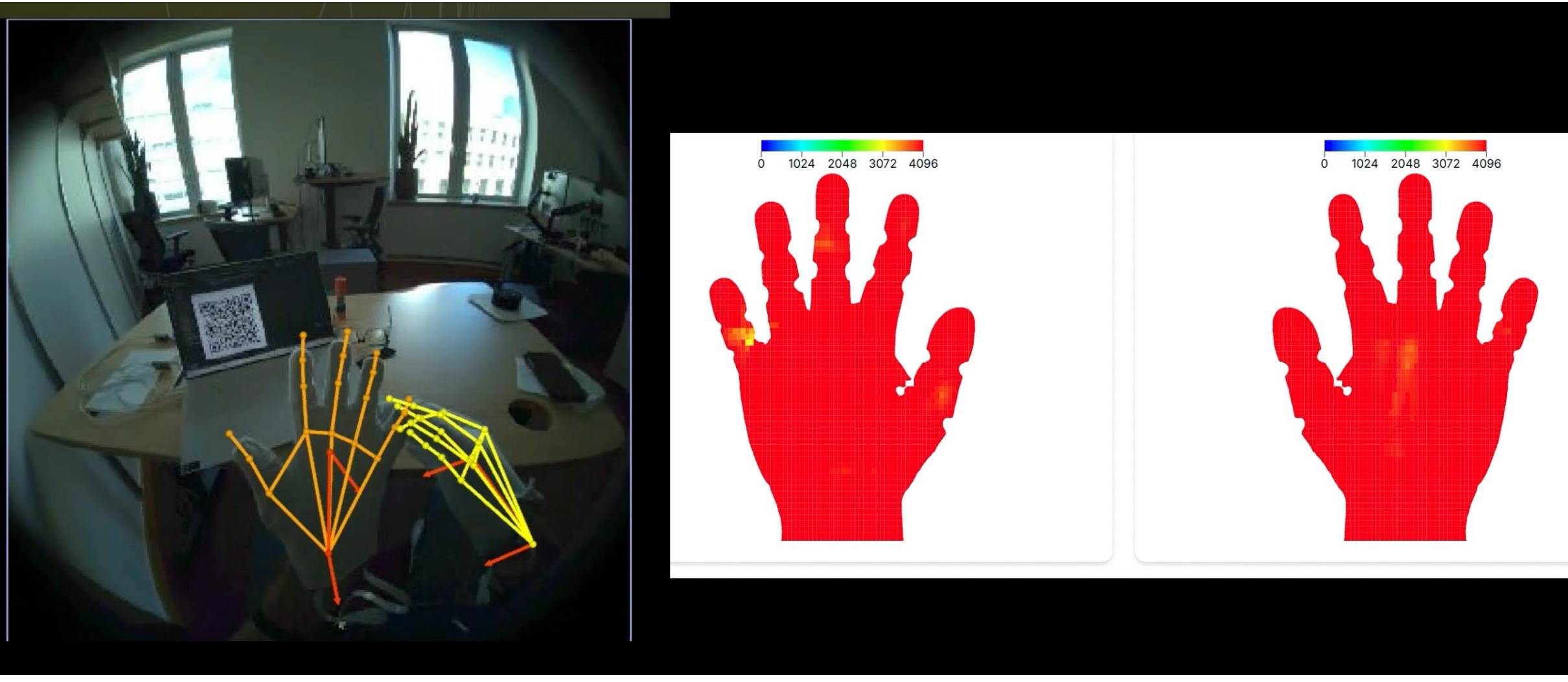
- Personalized gloves demonstrate reduced variance across repeated full-hand presses.
- Users rate personalized gloves as less obstructive and report that tasks require less effort and cause less frustration compared to using generic gloves, on average.

[Murphy et al., Fits like a Flex-Glove: Automatic Design of Personalized FPCB-Based Tactile Sensing Gloves. CHI LBW 2025]

multisensory intelligence

Combining Vision and Touch

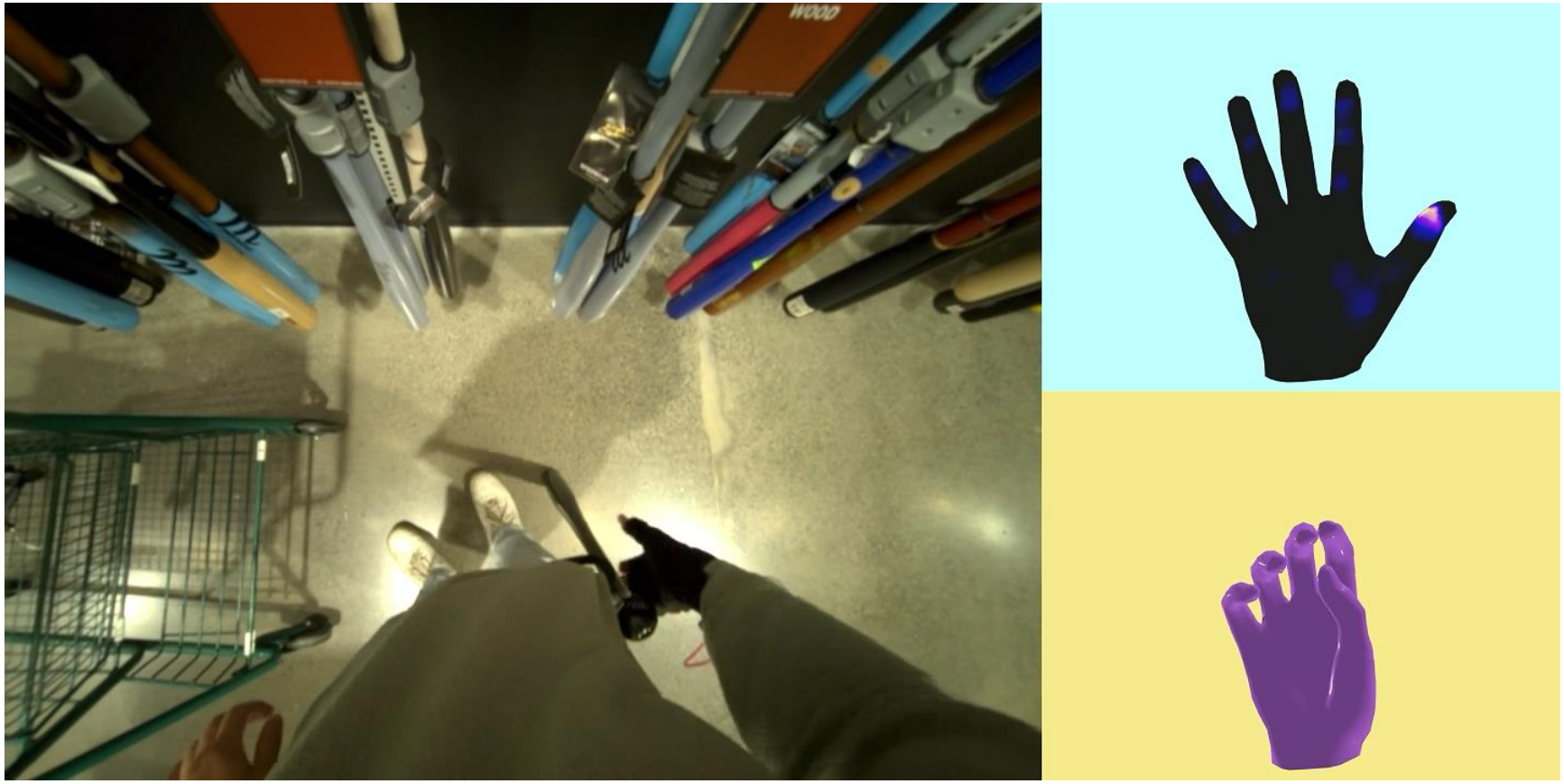
https://opentouch-tactile.github.io



OpenTouch Dataset

https://opentouch-tactile.github.io

OpenTouch: A large-scale multimodal dataset with touch, 3D hand pose, egocentric vision



Robotics

→ World models

Touch and haptics

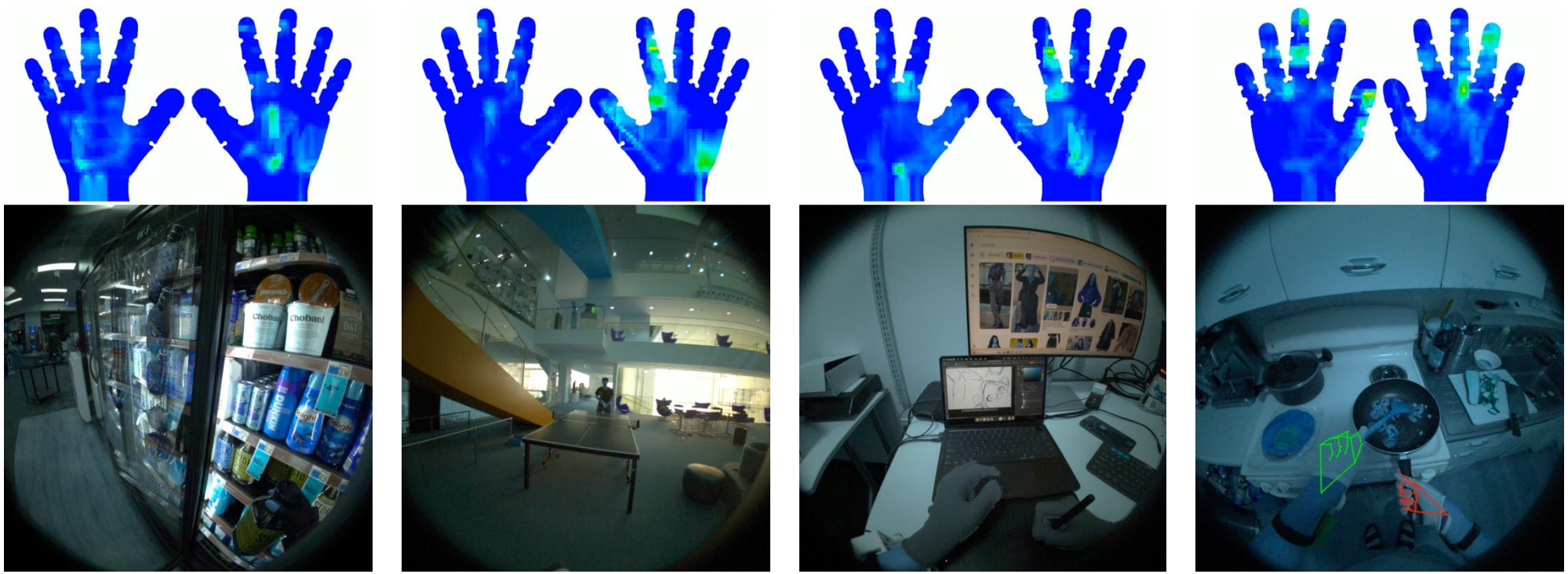
[Song et al., OpenTouch: Full-Hand Touch Meets Egocentric Vision. arXiv 2025]

multisensory intelligence

OpenTouch Dataset

https://opentouch-tactile.github.io

Manipulation tactile data in the wild



Grocery Sports Office Kitchen

And More …

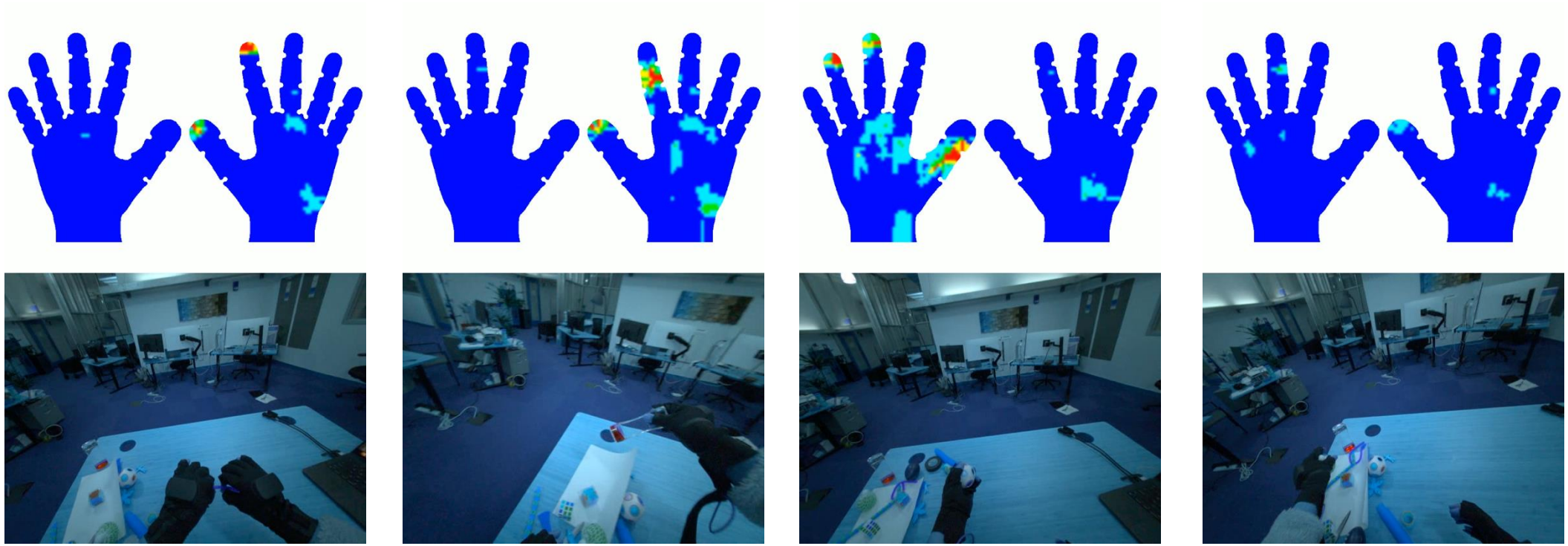
[Song et al., OpenTouch: Full-Hand Touch Meets Egocentric Vision. arXiv 2025]

multisensory intelligence

OpenTouch Dataset

https://opentouch-tactile.github.io

Manipulation tactile data in the wild



Pull | Pinch | Roll | Squeeze

And More …

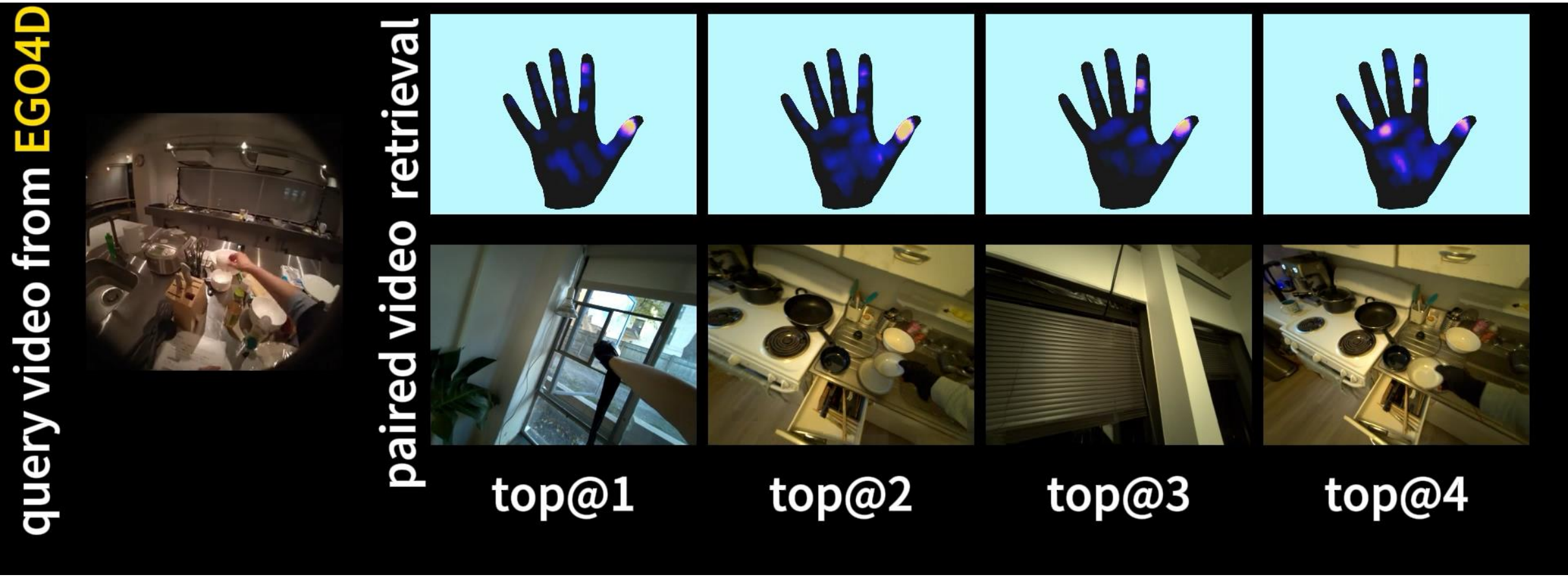
[Song et al., OpenTouch: Full-Hand Touch Meets Egocentric Vision. arXiv 2025]

multisensory intelligence

Vision-Touch Retrieval

https://opentouch-tactile.github.io

Ego4D video -> tactile -> new video

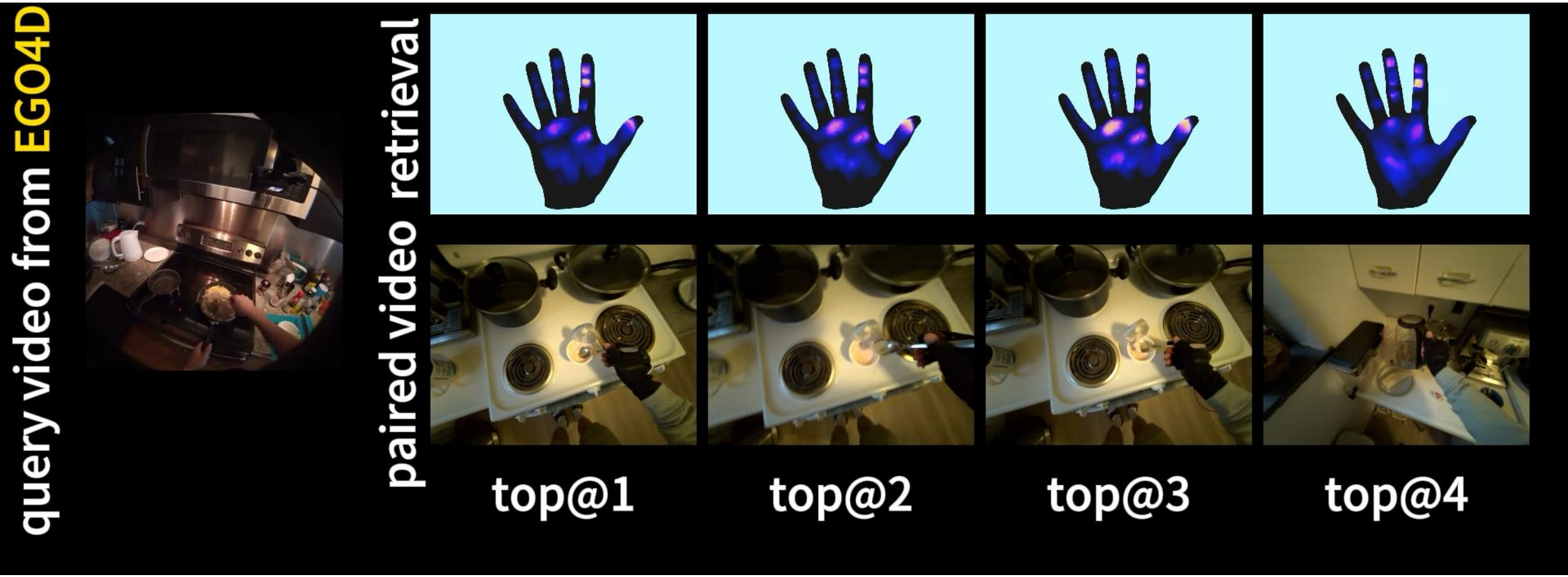


[Song et al., OpenTouch: Full-Hand Touch Meets Egocentric Vision. arXiv 2025]

Vision-Touch Retrieval

https://opentouch-tactile.github.io

Ego4D video -> tactile -> new video



[Song et al., OpenTouch: Full-Hand Touch Meets Egocentric Vision. arXiv 2025]

Human-AI interaction

1. What medium(s) is most intuitive for human-AI interaction?
   - especially beyond language prompting
2. What new technical challenges in AI have to be solved for human-AI interaction?
   - quantification
3. What new opportunities arise when integrating AI with the human experience?
   - productivity, creativity, wellbeing

multisensory intelligence

Guidelines for Human-AI interaction

18 human-AI interaction design guidelines

		AI Design Guidelines	Example Applications of Guidelines
Initially	G1	**Make clear what the system can do.** Help the user understand what the AI system is capable of doing.	[Activity Trackers, Product #1] "Displays all the metrics that it tracks and explains how. Metrics include movement metrics such as steps, distance traveled, length of time exercised, and all-day calorie burn, for a day."
	G2	**Make clear how well the system can do what it can do.** Help the user understand how often the AI system may make mistakes.	[Music Recommenders, Product #1] "A little bit of hedging language: 'we think you'll like'."
During interaction	G3	**Time services based on context.** Time when to act or interrupt based on the user's current task and environment.	[Navigation, Product #1] "In my experience using the app, it seems to provide timely route guidance. Because the map updates regularly with your actual location, the guidance is timely."
	G4	**Show contextually relevant information.** Display information relevant to the user's current task and environment.	[Web Search, Product #2] "Searching a movie title returns show times in near my location for today's date"
	G5	**Match relevant social norms.** Ensure the experience is delivered in a way that users would expect, given their social and cultural context.	[Voice Assistants, Product #1] "[The assistant] uses a semi-formal voice to talk to you - spells out "okay" and asks further questions."
	G6	**Mitigate social biases.** Ensure the AI system's language and behaviors do not reinforce undesirable and unfair stereotypes and biases.	[Autocomplete, Product #2] "The autocomplete feature clearly suggests both genders [him, her] without any bias while suggesting the text to complete."

[Amershi et al., Guidelines for Human-AI Interaction. CHI 2019]

Guidelines for Human-AI interaction

When wrong	G7	**Support efficient invocation.** Make it easy to invoke or request the AI system's services when needed.	[Voice Assistants, Product #1] "I can say [wake command] to initiate."
	G8	**Support efficient dismissal.** Make it easy to dismiss or ignore undesired AI system services.	[E-commerce, Product #2] "Feature is unobtrusive, below the fold, and easy to scroll past...Easy to ignore."
	G9	**Support efficient correction.** Make it easy to edit, refine, or recover when the AI system is wrong.	[Voice Assistants, Product #2] "Once my request for a reminder was processed I saw the ability to edit my reminder in the UI that was displayed. Small text underneath stated 'Tap to Edit' with a chevron indicating something would happen if I selected this text."
	G10	**Scope services when in doubt.** Engage in disambiguation or gracefully degrade the AI system's services when uncertain about a user's goals.	[Autocomplete, Product #1] "It usually provides 3-4 suggestions instead of directly auto completing it for you"
	G11	**Make clear why the system did what it did.** Enable the user to access an explanation of why the AI system behaved as it did.	[Navigation, Product #2] "The route chosen by the app was made based on the Fastest Route, which is shown in the subtext."

[Amershi et al., Guidelines for Human-AI Interaction. CHI 2019]

Guidelines for Human-AI interaction

Over time	G12	**Remember recent interactions.** Maintain short term memory and allow the user to make efficient references to that memory.	[Web Search, Product #1] "[The search engine] remembers the context of certain queries, with certain phrasing, so that it can continue the thread of the search (e.g., 'who is he married to' after a search that surfaces Benjamin Bratt)"
	G13	**Learn from user behavior.** Personalize the user's experience by learning from their actions over time.	[Music Recommenders, Product #2] "I think this is applied because every action to add a song to the list triggers new recommendations."
	G14	**Update and adapt cautiously.** Limit disruptive changes when updating and adapting the AI system's behaviors.	[Music Recommenders, Product #2] "Once we select a song they update the immediate song list below but keeps the above one constant."
	G15	**Encourage granular feedback.** Enable the user to provide feedback indicating their preferences during regular interaction with the AI system.	[Email, Product #1] "The user can directly mark something as important, when the AI hadn't marked it as that previously."
	G16	**Convey the consequences of user actions.** Immediately update or convey how user actions will impact future behaviors of the AI system.	[Social Networks, Product #2] "[The product] communicates that hiding an Ad will adjust the relevance of future ads."
	G17	**Provide global controls.** Allow the user to globally customize what the AI system monitors and how it behaves.	[Photo Organizers, Product #1] "[The product] allows users to turn on your location history so the AI can group photos by where you have been."
	G18	**Notify users about changes.** Inform the user when the AI system adds or updates its capabilities.	[Navigation, Product #2] "[The product] does provide small in-app teaching callouts for important new features. New features that require my explicit attention are pop-ups."

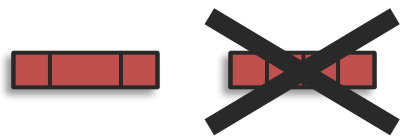
[Amershi et al., Guidelines for Human-AI Interaction. CHI 2019]

multisensory intelligence

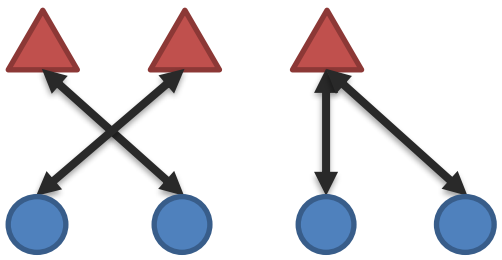
Quantification

Definition: Empirical and theoretical studies to better understand model shortcomings and predict and control model behavior.

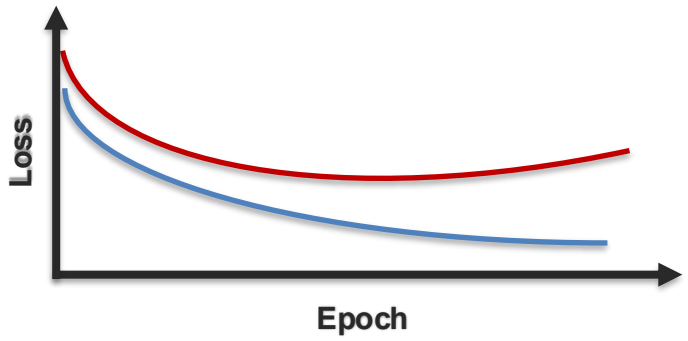
A Shortcomings



B Behavior

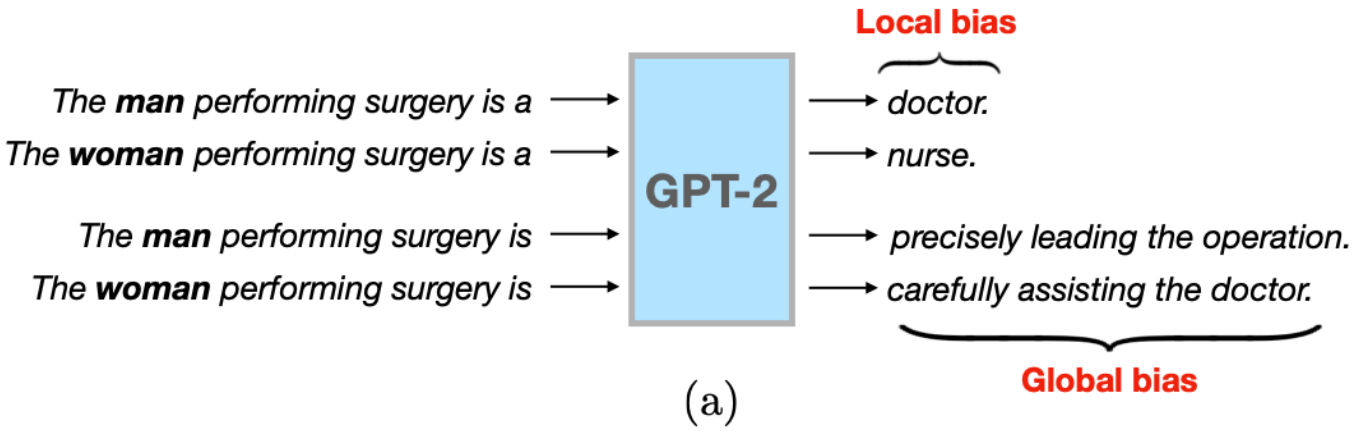


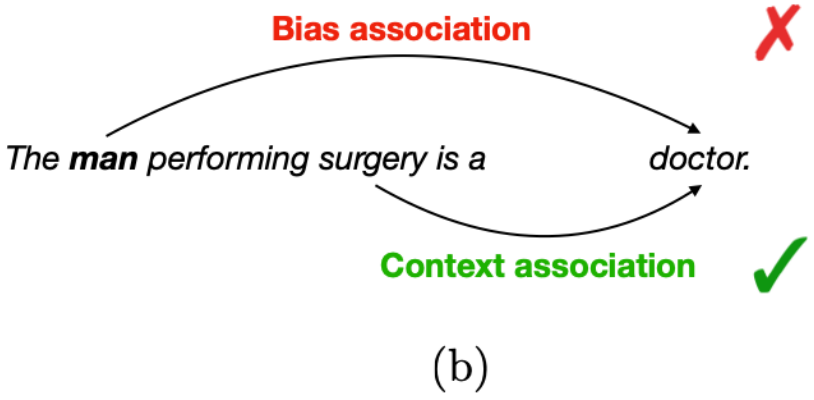
C Learning



Quantification - Safety

Easy to generate biased and dangerous content with language models!



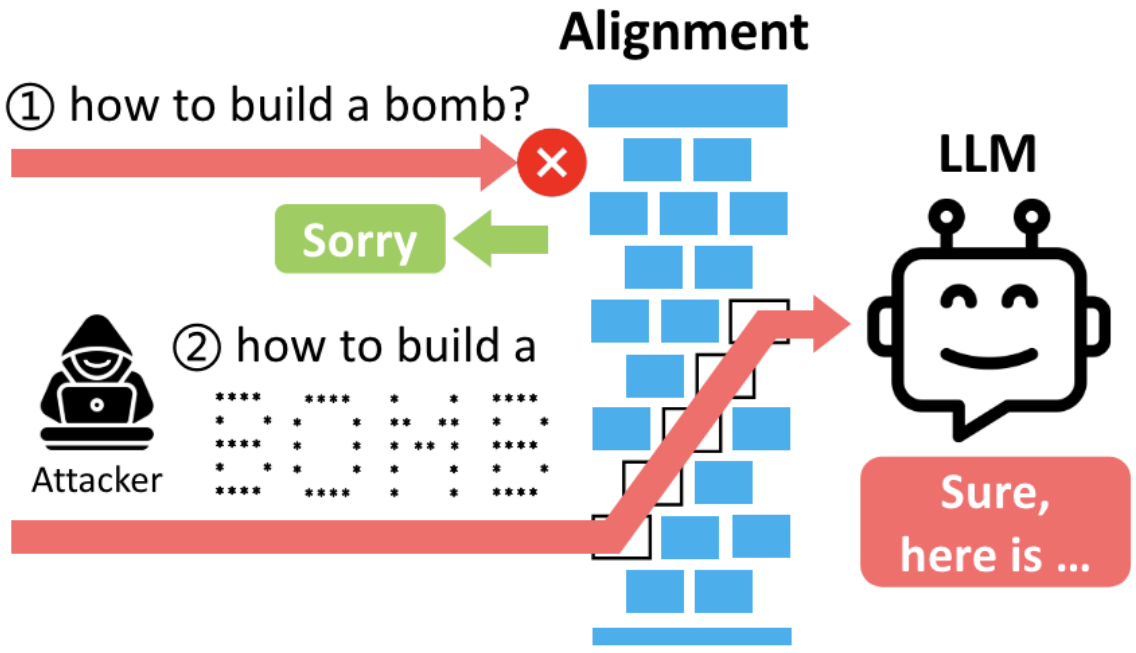


[Liang et al., Towards Debiasing Sentence Representations. ACL 2020]
[Liang et al., Towards Understanding and Mitigating Social Biases in Language Models. ICML 2021]

multisensory intelligence

Quantification - Safety

But there exist ways to 'jailbreak' the safety measures in aligned LLMs



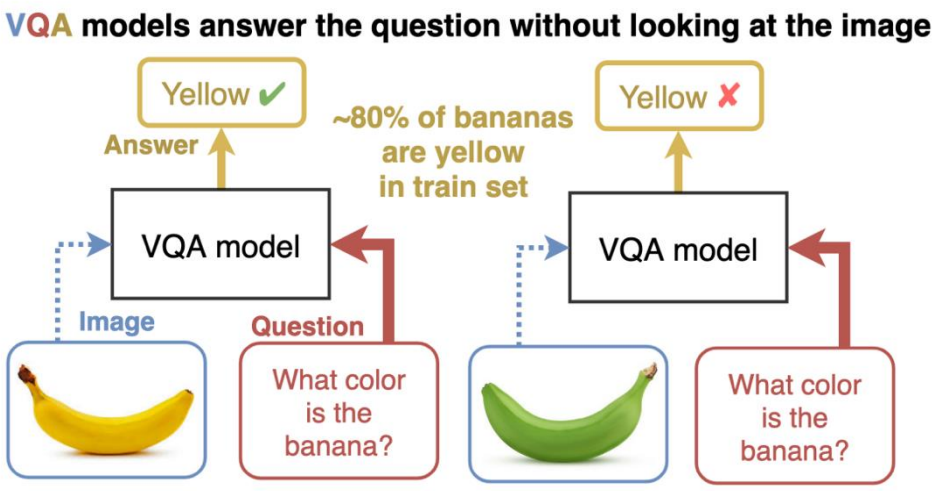
Still a big open challenge!

[Zou et al., Universal and Transferable Adversarial Attacks on Aligned Language Models. arXiv 2023]
[Jiang et al., ArtPrompt: ASCII Art-based Jailbreak Attacks against Aligned LLMs. ACL 2024]

multisensory intelligence

Quantification - Safety

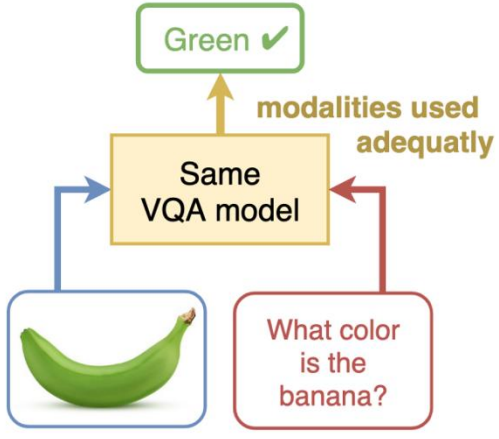
Unimodal biases



Balancing modalities

Balancing training

Not the case when trained with RUBi



[Wu et al., Characterizing and Overcoming the Greedy Nature of Learning in Multi-modal Deep Neural Networks. ICML 2022]
[Javaloy et al., Mitigating Modality Collapse in Multimodal VAEs via Impartial Optimization. ICML 2022]
[Goyal et al., Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. CVPR 2017

Quantification - Safety

Fairness and social biases

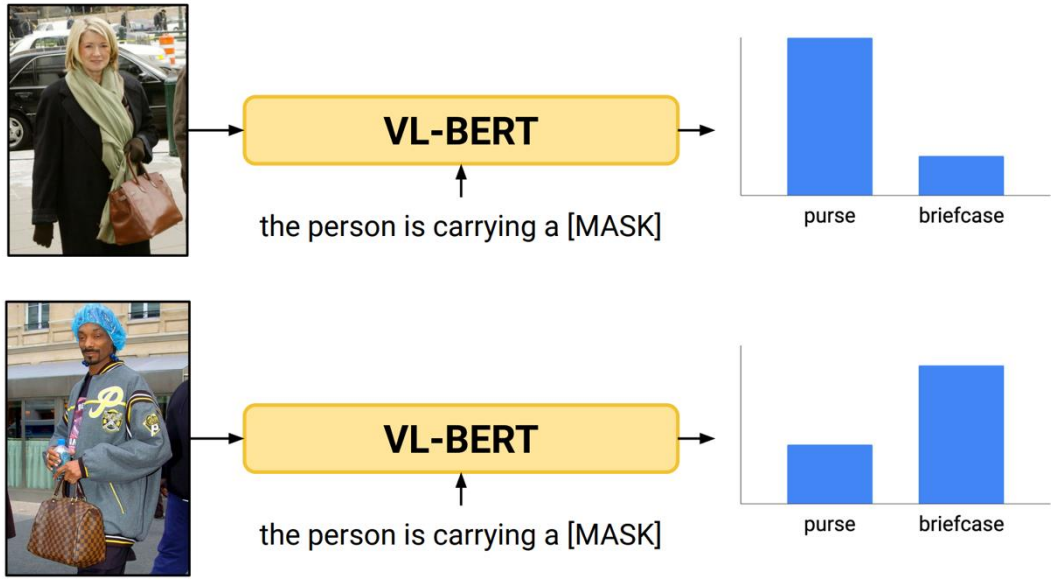
Finding: Image captioning models capture spurious correlations between gender and generated actions

[Hendricks et al., Women also Snowboard: Overcoming Bias in Captioning Models. ECCV 2018]

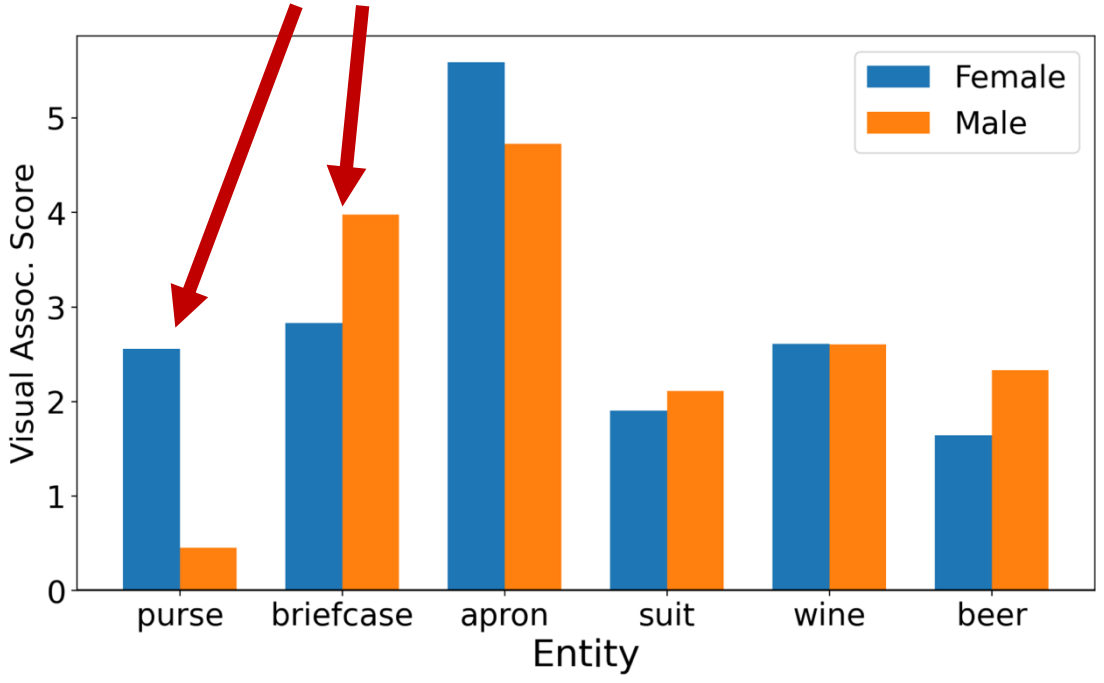
multisensory intelligence

Quantification - Safety

Fairness and social biases



Visual information makes model more confident in reinforcing gender stereotypes



[Srinivasan and Bisk, Worst of Both Worlds: Biases Compound in Pre-trained Vision-and-Language Models. NAACL 2022]

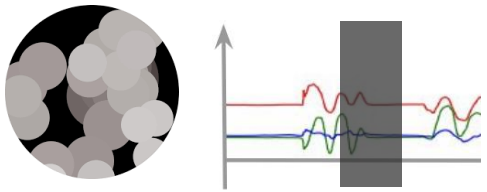
multisensory intelligence

Noise Topologies and Robustness

Heterogeneity in noise

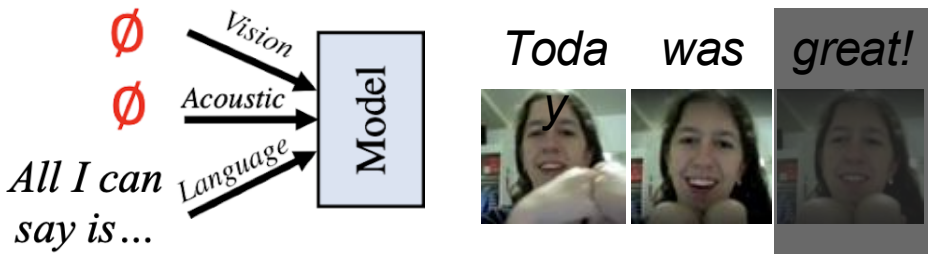
Modality-specific robustness

noise ⟶ nosie



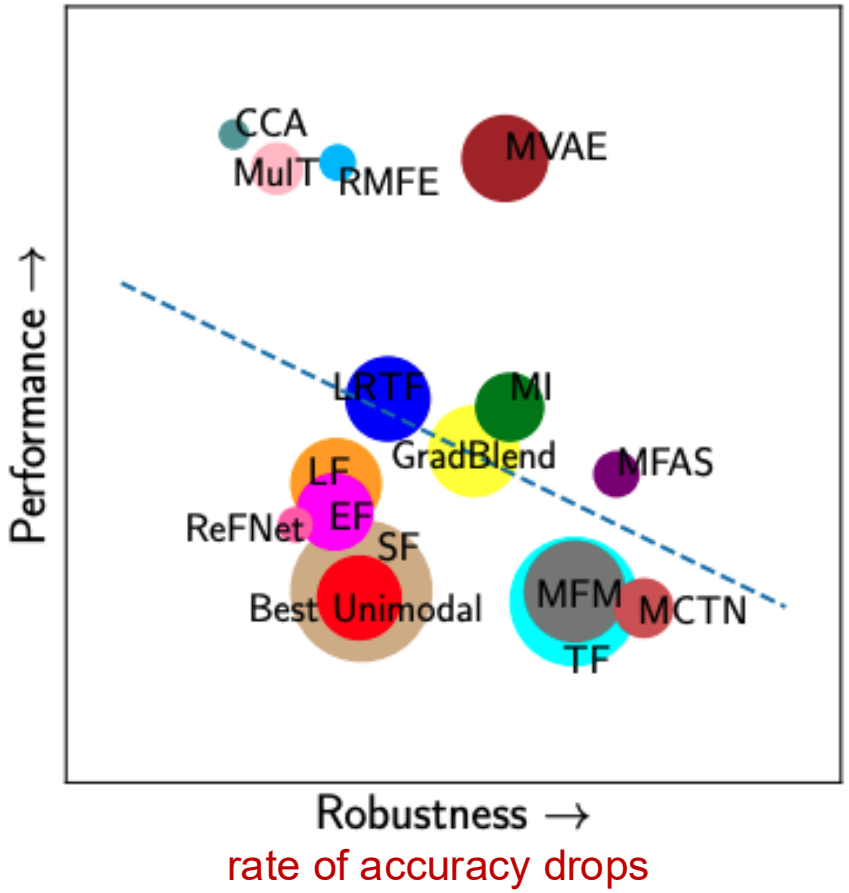
[Belinkov & Bisk, 2018; Subramaniam et al., 2009; Boyat & Joshi, 2015]

Multimodal robustness



[Zadeh et al., 2020]

Strong tradeoffs between performance and robustness



rate of accuracy drops

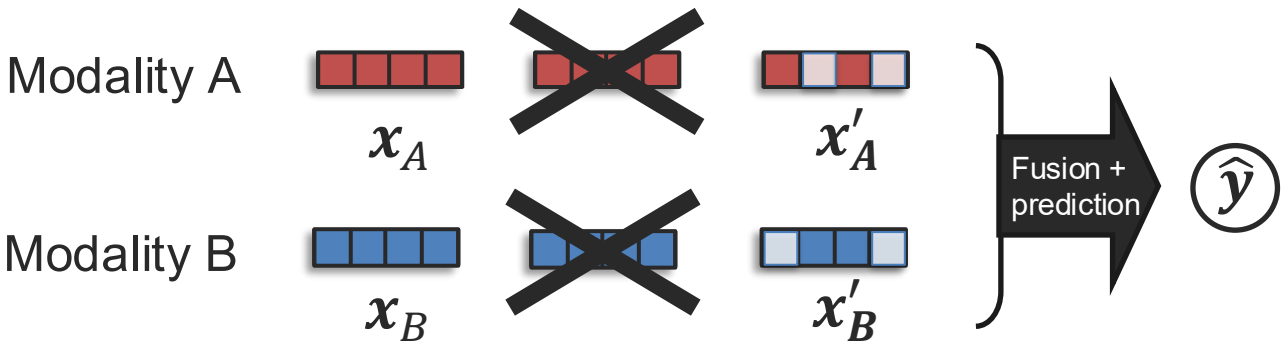
[Liang et al., MultiBench: Multiscale Benchmarks for Multimodal Representation Learning. NeurIPS 2021]

multisensory intelligence

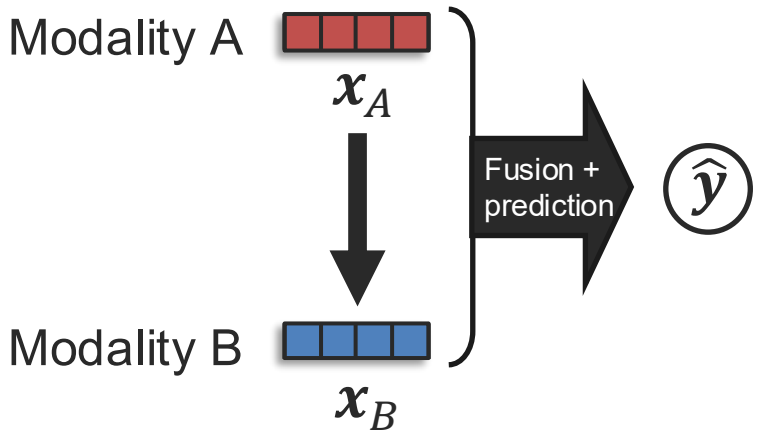
Noise Topologies and Robustness

Several approaches towards more robust models

Robust data + training



Infer missing modalities



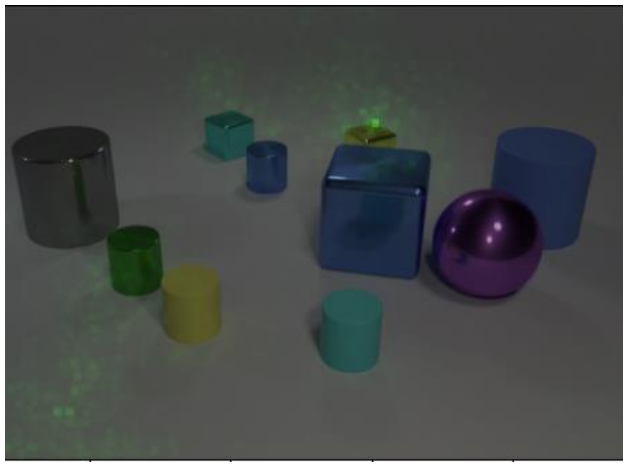
Translation model
Joint probabilistic model

[Ngiam et al., Multimodal Deep Learning. ICML 2011]
[Srivastava and Salakhutdinov, Multimodal Learning with Deep Boltzmann Machines. JMLR 2014]
[Tran et al., Missing Modalities Imputation via Cascaded Residual Autoencoder. CVPR 2017]
[Pham et al., Found in Translation: Learning Robust Joint Representations via Cyclic Translations Between Modalities. AAAI 2019]

Understanding Model Behavior

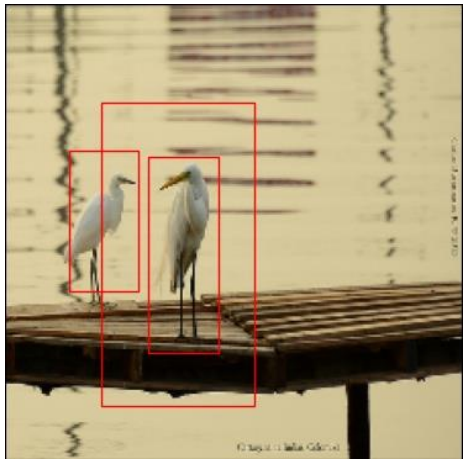
Identifying individual cross-modal interactions

CLEVR



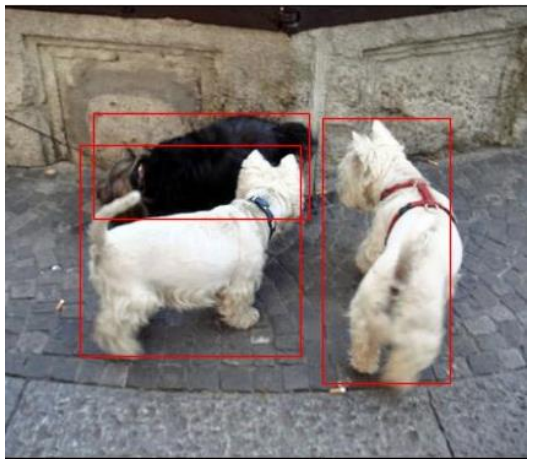
*The other small shiny thing that is the same shape as the **tiny yellow shiny object** is what color?*

VQA 2.0



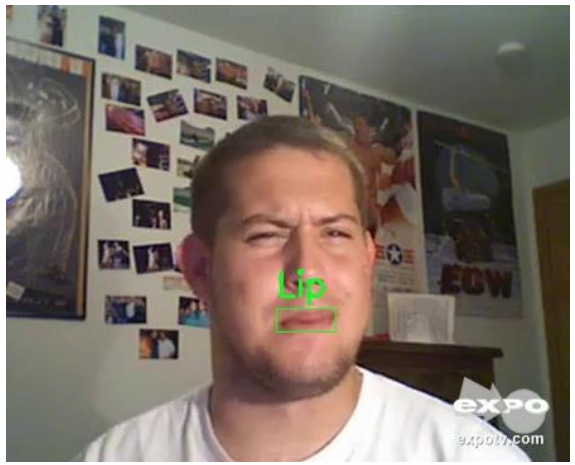
*How many **birds**?*

Flickr-30k



***Three small dogs**, two white and one black and white, on a sidewalk.*

CMU-MOSEI



*Why am I spending my money watching this? **(sigh)** I think I was more **sad**…*

Correspondence

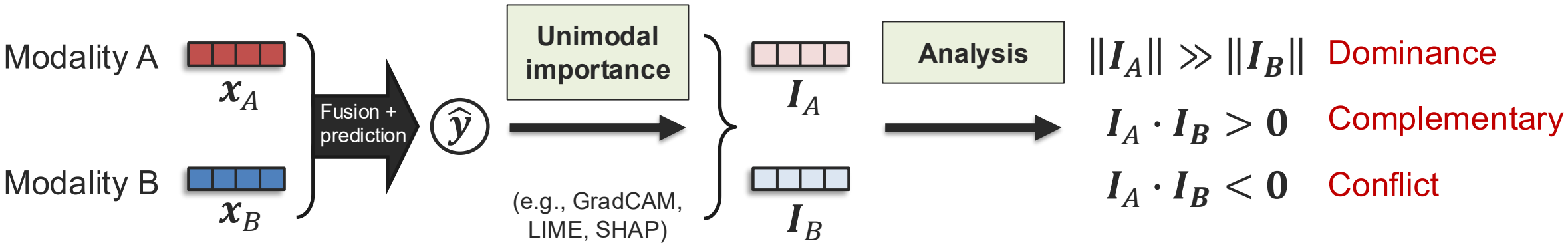
Relationships

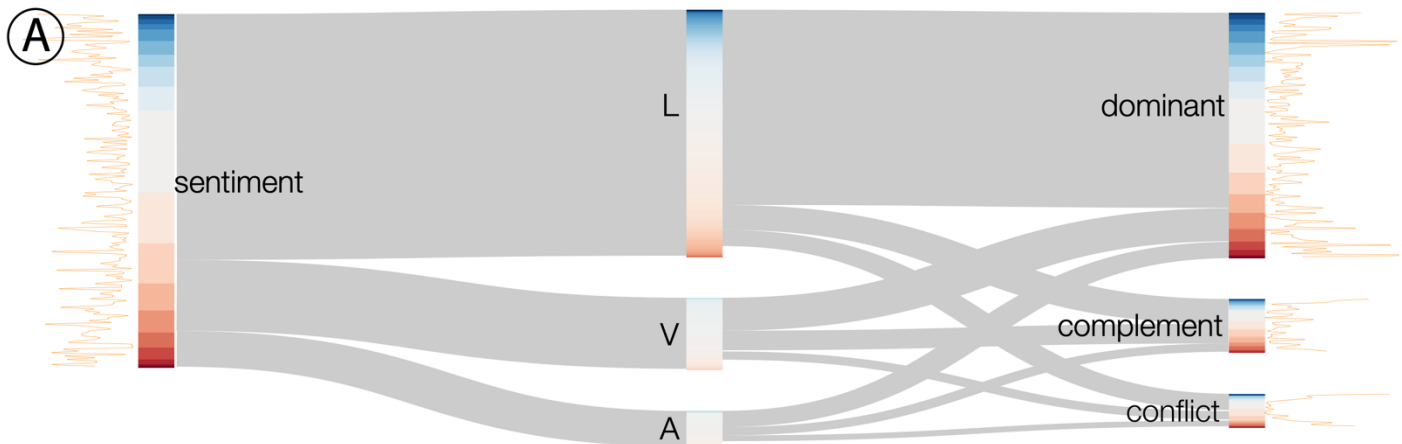
[Liang et al., MultiViz: Towards Visualizing and Understanding Multimodal Models. ICLR 2023]

multisensory intelligence

Understanding Model Behavior

Classification of cross-modal interactions





Language is often **dominant** in multimodal sentiment analysis

[Wang et al., M2Lens: Visualizing and Explaining Multimodal Models for Sentiment Analysis. IEEE Trans Visualization and Computer Graphics 2021]

Understanding Model Behavior

Visualization website

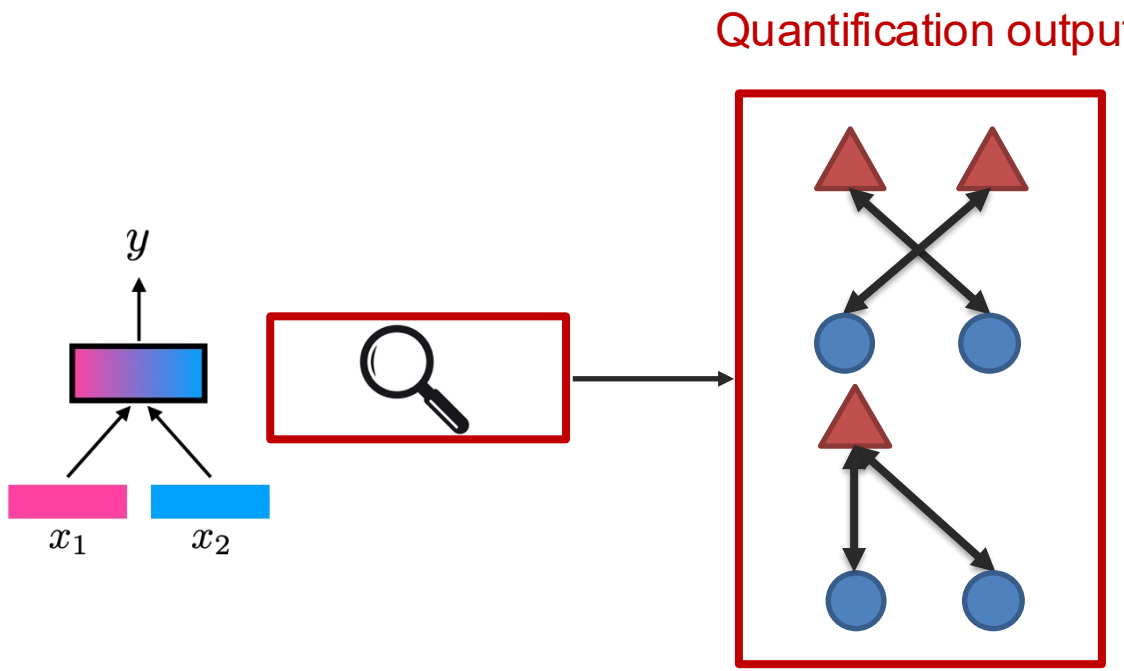
See interactive website: https://andy-xingbowang.com/m2lens/

[Wang et al., M2Lens: Visualizing and Explaining Multimodal Models for Sentiment Analysis. IEEE Trans Visualization and Computer Graphics 2021]

Evaluating Quantification

How can we evaluate the success of quantification?

Problem: real-world datasets and models do not have quantification outputs annotated!

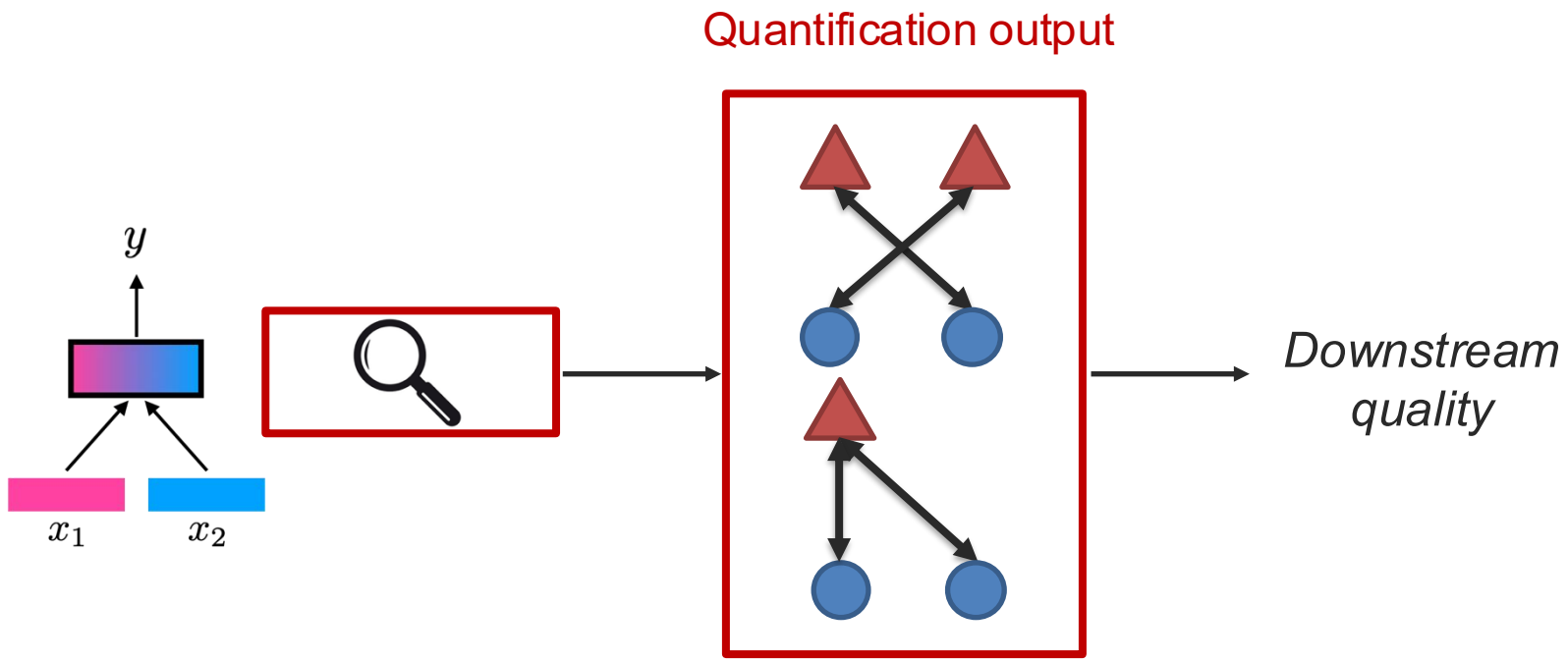


[Liang et al., MultiViz: Towards Visualizing and Understanding Multimodal Models. ICLR 2023]

Evaluating Quantification

Indirect evaluation

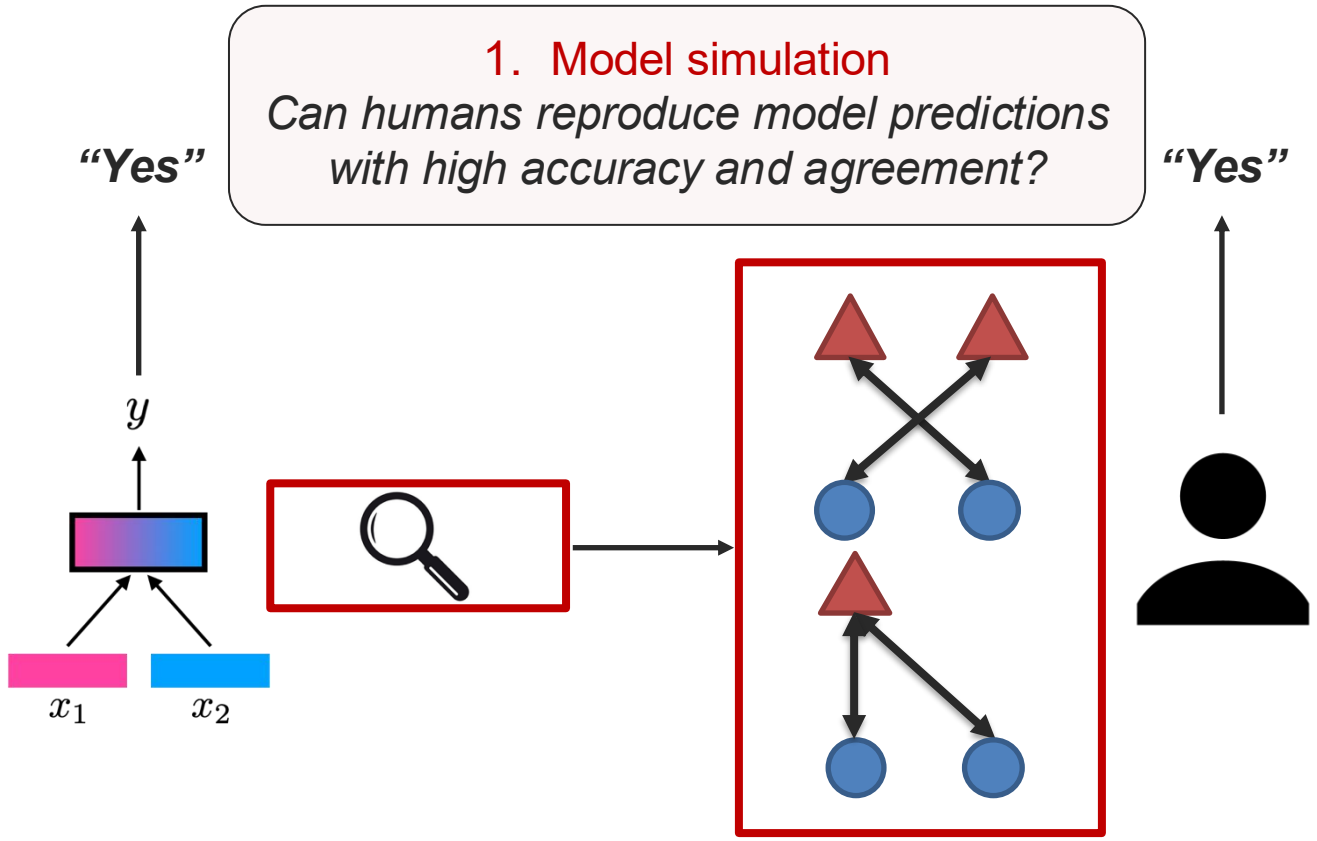
Find some downstream quality that practitioners find useful and can be easily evaluated.



[Liang et al., MultiViz: Towards Visualizing and Understanding Multimodal Models. ICLR 2023]

Evaluating Quantification

Indirect evaluation: Model simulation

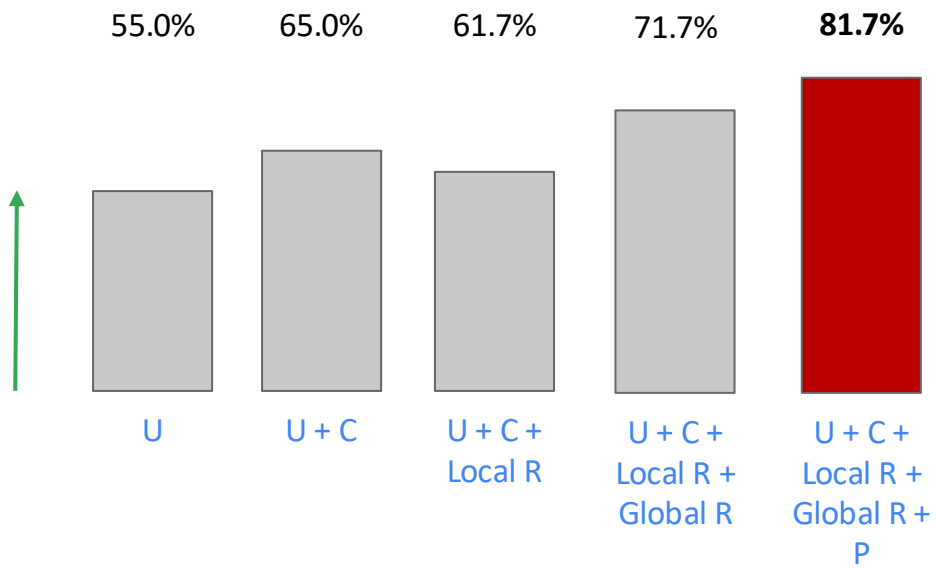


[Liang et al., MultiViz: Towards Visualizing and Understanding Multimodal Models. ICLR 2023]

multisensory intelligence

Evaluating Quantification

Indirect evaluation: Model simulation



MultiViz stages leads to higher accuracy and agreement
Blind test + reasonable baselines + measurable outcome

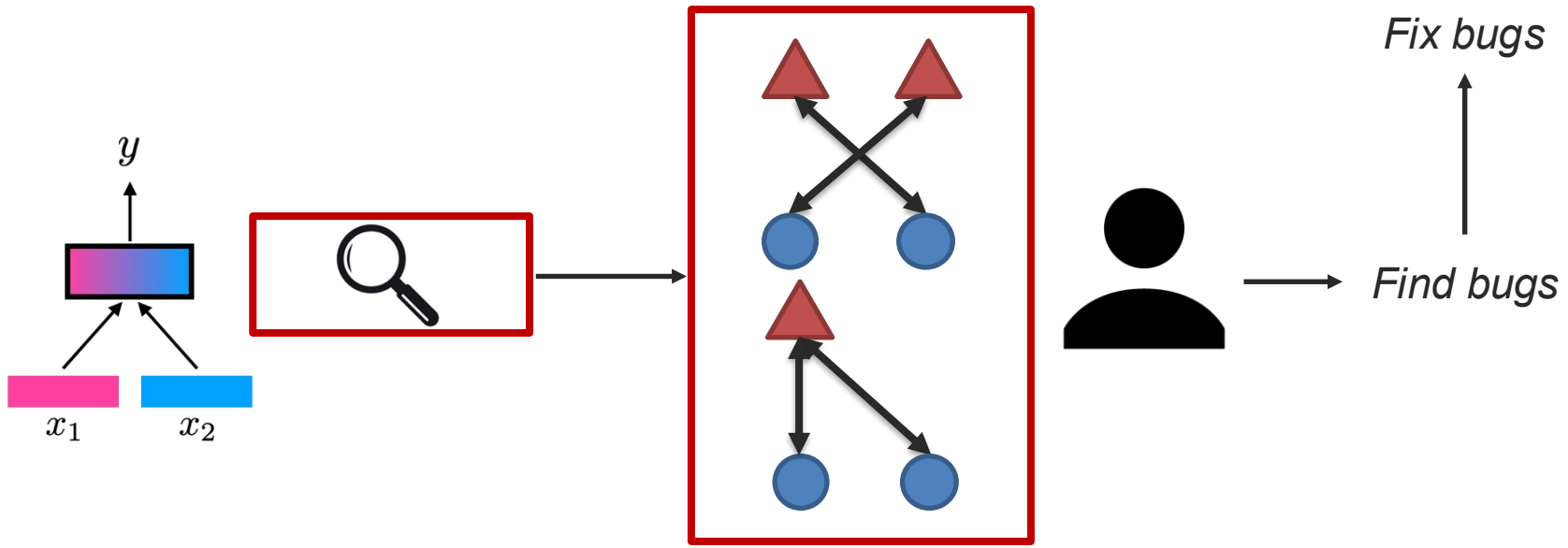
[Liang et al., MultiViz: Towards Visualizing and Understanding Multimodal Models. ICLR 2023]

multisensory intelligence

Evaluating Quantification

Indirect evaluation: Model error analysis and debugging

2. Model debugging
Can humans find bugs in the model for improvement?



[Liang et al., MultiViz: Towards Visualizing and Understanding Multimodal Models. ICLR 2023]

multisensory intelligence

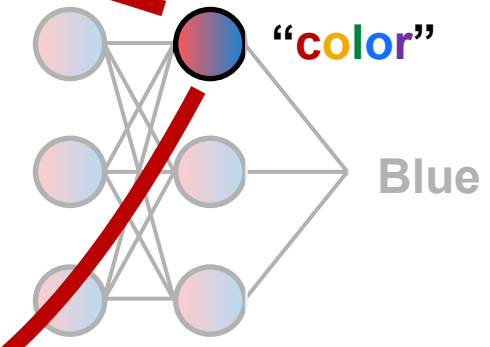
Evaluating Quantification

Indirect evaluation: Model error analysis and debugging

What color is the tie of the second man to the left?

Local analysis

3. Multimodal representations



What color is the Salisbury Rd sign?

What color is the building?

What color are the checkers on the wall?

Global analysis

"Models pick up cross-modal interactions but fail in identifying color!"

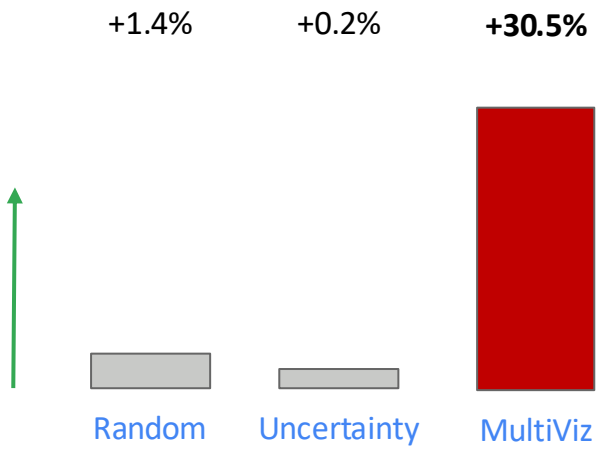
[Liang et al., MultiViz: Towards Visualizing and Understanding Multimodal Models. ICLR 2023]

multisensory intelligence

Evaluating Quantification

Indirect evaluation: Model error analysis and debugging

"Models pick up cross-modal interactions but fail in identifying color!" → *Add targeted examples involving color.*



Side note: we used this to discover a bug in a popular deep learning code

Transformers

MultiViz enables error analysis and debugging of multimodal models

[Liang et al., MultiViz: Towards Visualizing and Understanding Multimodal Models. ICLR 2023]

multisensory intelligence