Multimodal AI

Lecture 1.2 – Multimodal Research Tasks

Paul Liang
Assistant Professor
MIT Media Lab & MIT EECS

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

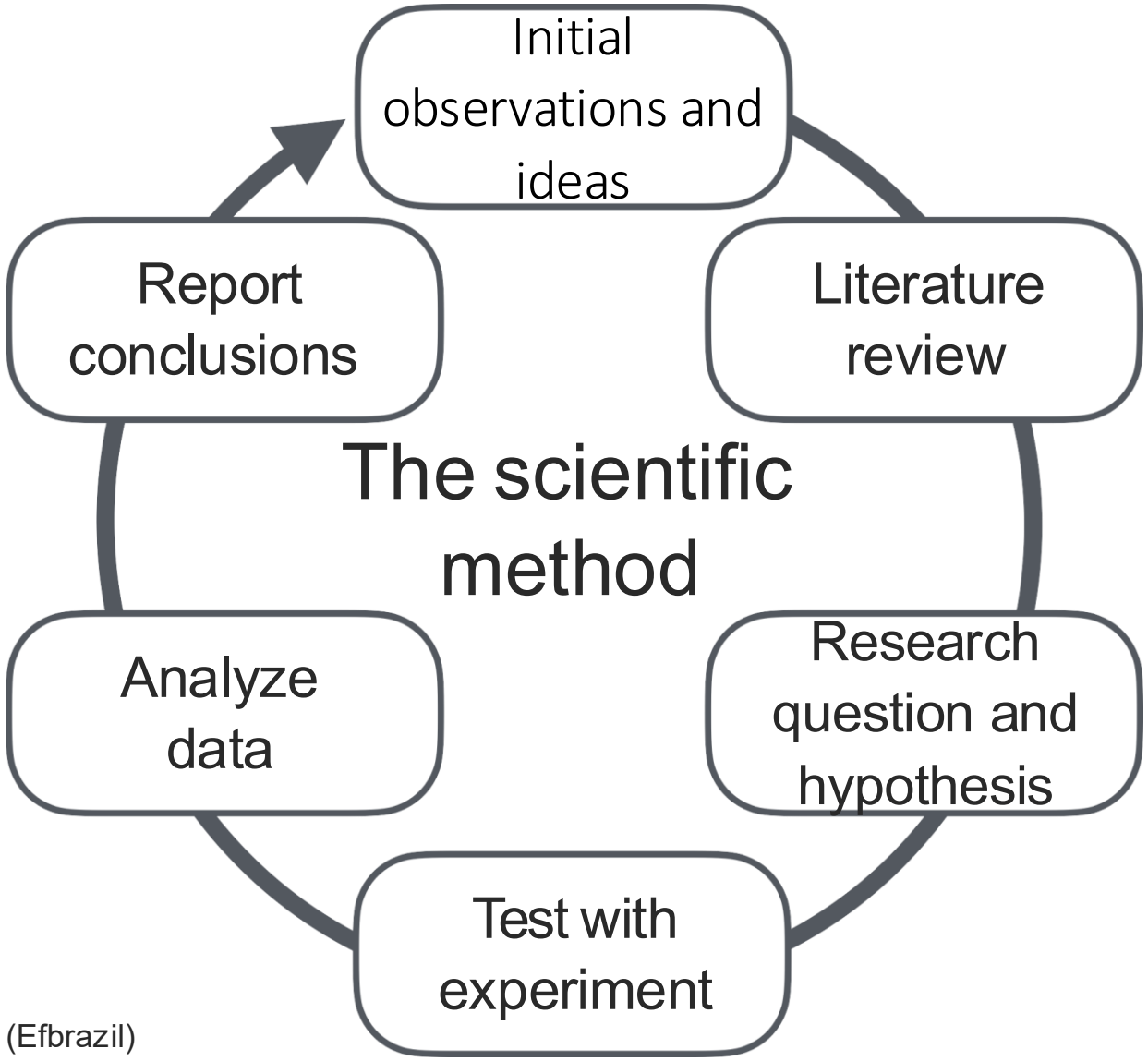
mit media lab

MIT EECS

Today's lecture

1. How to do AI research
2. Multimodal tasks and datasets
3. Advice on research projects
4. Mingle and find teammates

The Research Process



Credit: Adapted From Wikipedia (Efbrazil)

multisensory intelligence

How Do We Get Research Ideas?

Turn a concrete understanding of existing research's failings to a higher-level experimental question.

- **Bottom-up discovery** of research ideas
- Great tool for incremental progress, but may preclude larger leaps

Move from a higher-level question to a lower-level concrete testing of that question.

- **Top-down design** of research ideas
- Favors bigger ideas, but can be disconnected from reality

Bottom-Up Discovery

The proposal report should identify existing methods

The homeworks and midterm assignment will enable this bottom-up discovery:
1. Experiment with state-of-the-art models
2. Analyze successes and failures of these models
3. Identify ways you could improve on these failure cases

The final report should propose new ideas that do better

Your research ideas will evolve during the semester!

Top-down Design

Brainstorming: Take the time to brainstorm with your teammates, with TAs and with instructors.

- Office hours with TAs these coming 2 weeks
- Project hours with instructors in the next month

Literature review: The first assignment will allow you to review recent work related to your dataset and your initial research ideas

- When exploring the dataset, you should also expand your research ideas

Midterm and final report focused on incrementally achieving proposed capabilities

Scientific Research Questions and Hypotheses

Research Questions

- One or several explicit questions regarding the thing that you want to know
- Hypotheses are easier to draft with “Yes-no" questions than "how to" questions

Hypothesis:

- What you think the answer to the question may be a-priori
- Should be *falsifiable*: if you get a certain result the hypothesis will be validated, otherwise disproved

multisensory intelligence

Scientific Research Questions and Hypotheses

Good examples of research questions

Multimodal representation learning is fundamentally about transforming incomparable modalities into comparable representations. While prior research primarily focused on *explicitly* aligning these representations through targeted learning objectives and model architectures, a recent line of work has found that independently trained unimodal models of increasing scale and performance can become *implicitly* aligned with each other. These findings raise fundamental questions regarding the emergence of aligned representations in multimodal learning. Specifically: (1) when and why does alignment emerge implicitly? and (2) is alignment a reliable indicator of performance? Through a comprehensive em-

While fact-checking systems have shown mixed results on helping people detect misinformation and correct their beliefs about it [13, 67], recent work on AI chatbots has demonstrated that dialogues with an AI chatbot can reduce belief in conspiracy theories by 22% [15], suggesting potential for AI-assisted belief correction. Furthermore, AI chatbots for education have shown promise in various domains [9–11, 58], particularly when acting as collaborative partners rather than authoritative sources. This makes the AI chatbot a potential tool to teach people how to detect and update their beliefs about misinformation. Yet emerging research on Human-AI interaction suggests that AI chatbots can foster overreliance [18, 39], with users deferring to system outputs even when incorrect [18], potentially undermining the development of independent reasoning skills [7, 25]. This raises an important question: **Do dialogues with AI chatbots actually build lasting misinformation detection skills, or do they create dependency that undermines humans' independent judgment?**

Scientific Research Questions and Hypotheses

Not the best examples of research questions

4 Experiments

We design experiments to answer the following research questions. Details are included in App. D.

RQ1: How does DRPO compare with other critic-free RL methods and models? As detailed in Sec. 3.2, we train and evaluate QoQ-Med on a combination of 30 clinical diagnosis datasets across 9 clinical domains. A description of each dataset is included in App. C. The models are evaluated with balanced accuracy and macro-F1. We compare our training method DRPO against supervised fine-tuning (SFT), PPO [74] and four popular critic-free RL training methods: GRPO [75], RLOO [2], Reinforce++ [33], and ReMax [50]. We further compare our trained model QoQ-Med against medical VLMs (Llava-Med [48], Med-R1 [46]) and closed source VLMs (GPT-4o [37], o4-mini [61]).

RQ2: How well does DRPO handle mixed multimodal inputs? We repeat the comparison on MIMIC-IV, where samples contain a chest X-ray, a 12-lead ECG trace, and an accompanying clinical record. We train and evaluate the models on two tasks: length of stay (LOS) prediction, binned into a 4-day interval, and 48-hour in-hospital mortality (48-IHM). We evaluate the model with accuracy and F1 score in the same way as RQ1.

RQ3: How is the quality of the reasoning traces and bounding boxes learned by DRPO? We did both a qualitative and a quantitative analysis on QoQ-Med's reasoning and bounding box outputs. We evaluate the bounding box quality via the intersection over union (IoU) against the ground truth segmentation available in the dataset. We further collaborated with clinicians to annotate the reasoning traces on the validation dataset, grading the traces by their relevance to the final diagnosis.

Beware "Does X Make Y Better?" "Yes"

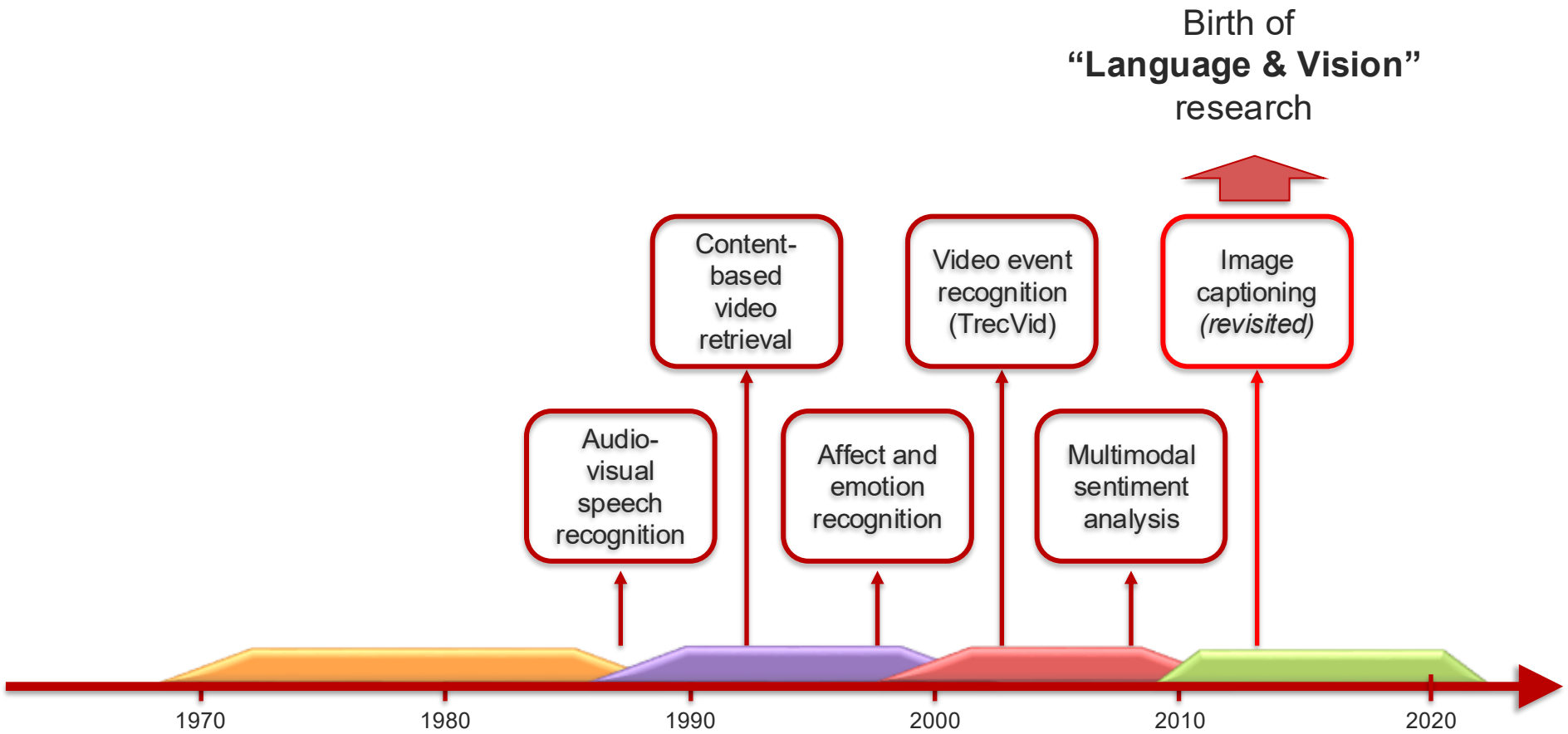
The above question/hypothesis is natural, but indirect

- If the answer is "no" after your experiments, how do you tell what's going wrong?

Usually you have an intuition about *why* X will make Y better (not just random)

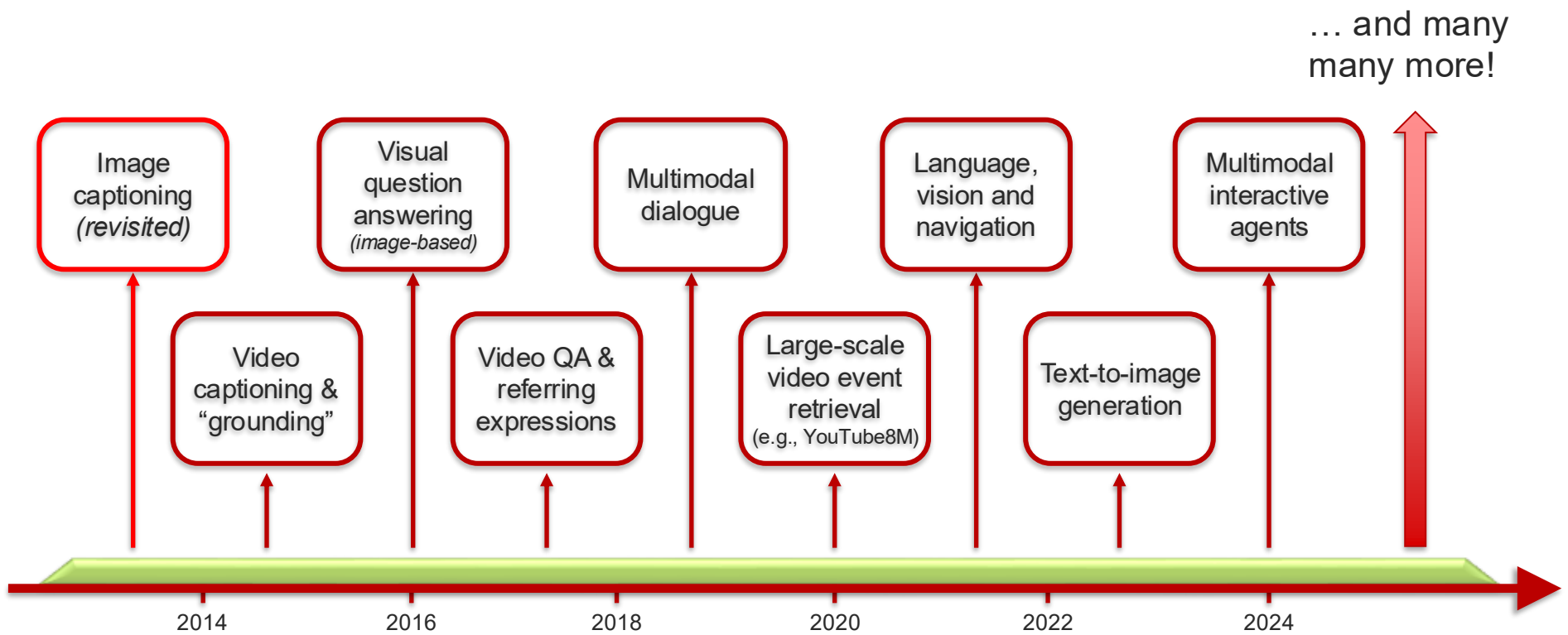
Can you think of other research questions/ hypotheses that confirm/falsify these assumptions

Multimodal Research Tasks



[original slide co-developed with Louis-Philippe Morency for CMU course 11-777]

Multimodal Research Tasks



[original slide co-developed with Louis-Philippe Morency for CMU course 11-777]

Multimodal Research Tasks

New modalities and applications

Core multimodal fusion, alignment, foundation modeling

Multimodal reasoning

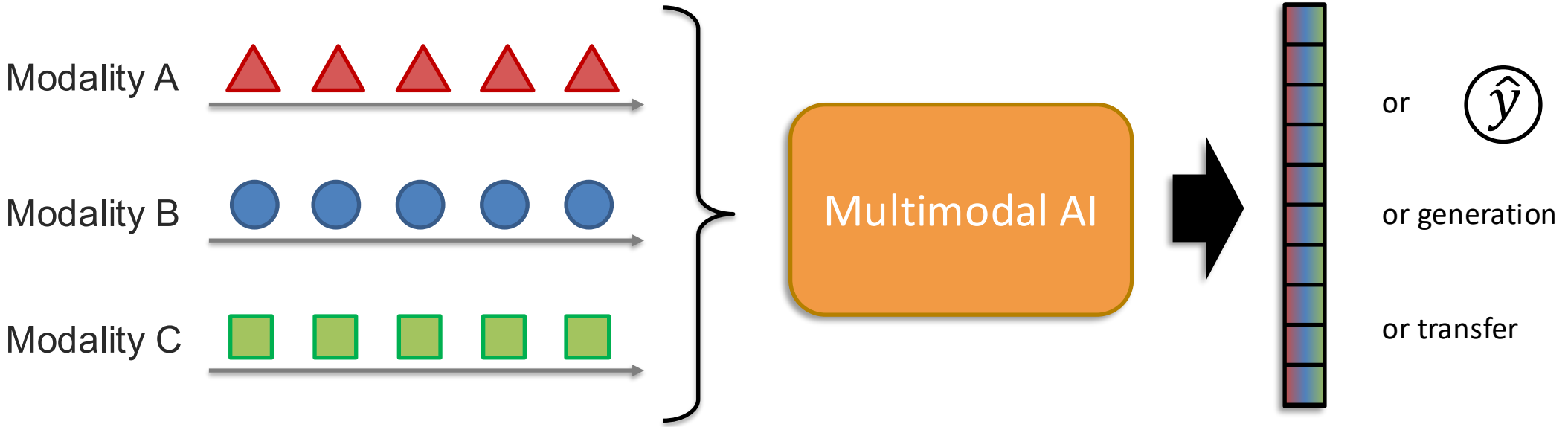
Interactive agents

Socially intelligent AI

Embodied AI

Human-AI interaction, ethics, safety

Multimodal Tasks



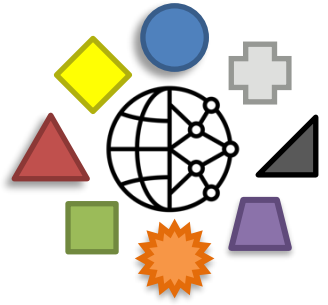
multisensory intelligence

Research Projects on New Modalities

Motivation: Many ways of sensing the world, each with complementary information, privacy concerns, invasiveness & efficiency tradeoffs.

Challenges:

- AI for physiological sensing, IoT sensing in cities, climate, environment, engineering
- Smell, taste, art, music, creative domains, tangible and embodied systems
- Multimodal with limited unimodal and multimodal data



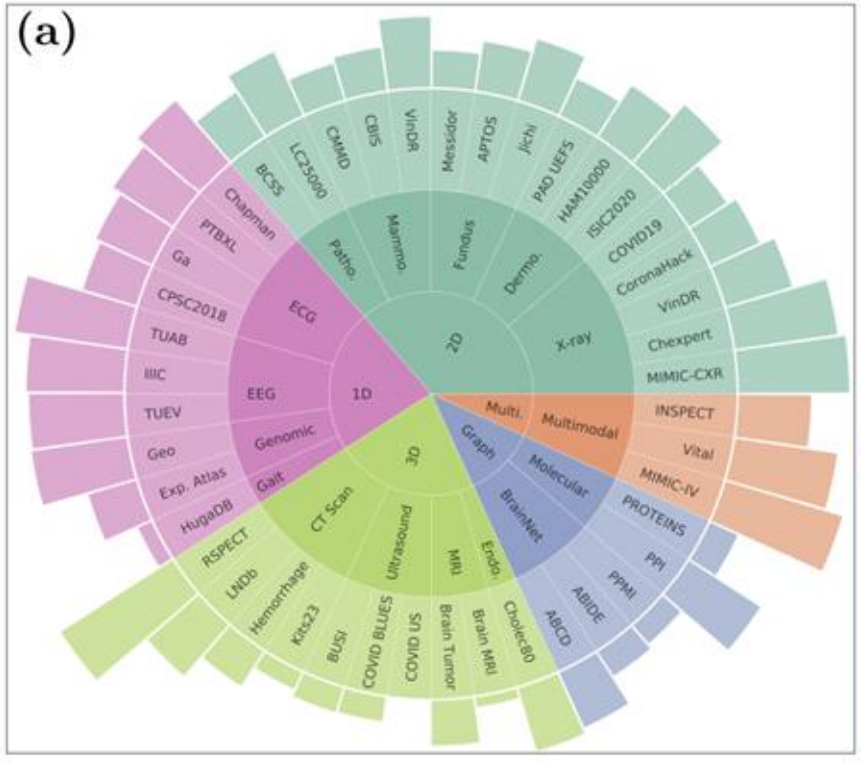
Potential models and dataset to start with

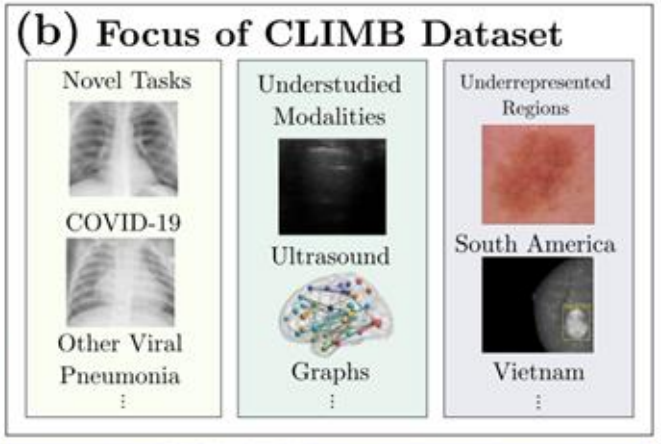
- Brain EEG Signal: https://arxiv.org/abs/2306.16934
- Speech: https://arxiv.org/pdf/2310.02050.pdf
- Facial Motion: https://arxiv.org/abs/2308.10897
- Tactile: https://arxiv.org/pdf/2204.00117.pdf
- SensorLM: https://research.google/blog/sensorlm-learning-the-language-of-wearable-sensors/

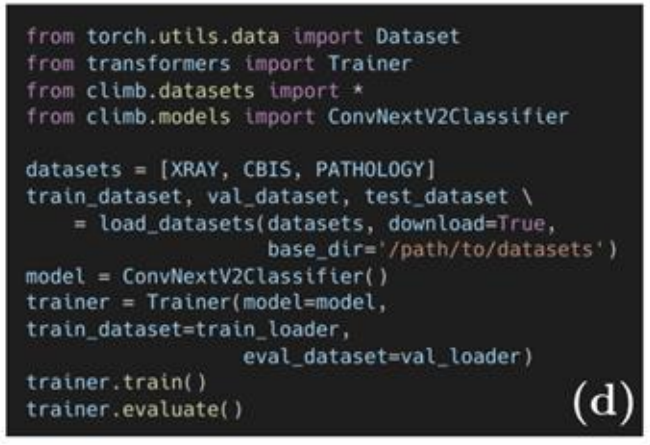
CLIMB: Multimodal Clinical Dataset

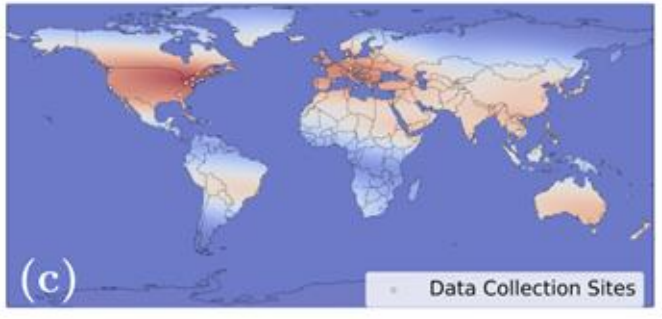
https://github.com/DDVD233/climb

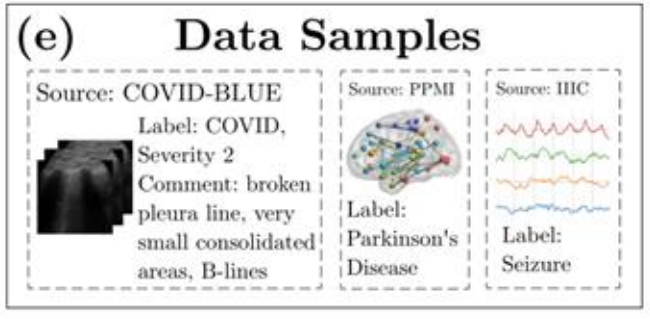
4.57M annotated samples from 44 datasets, totaling 19.01T in size







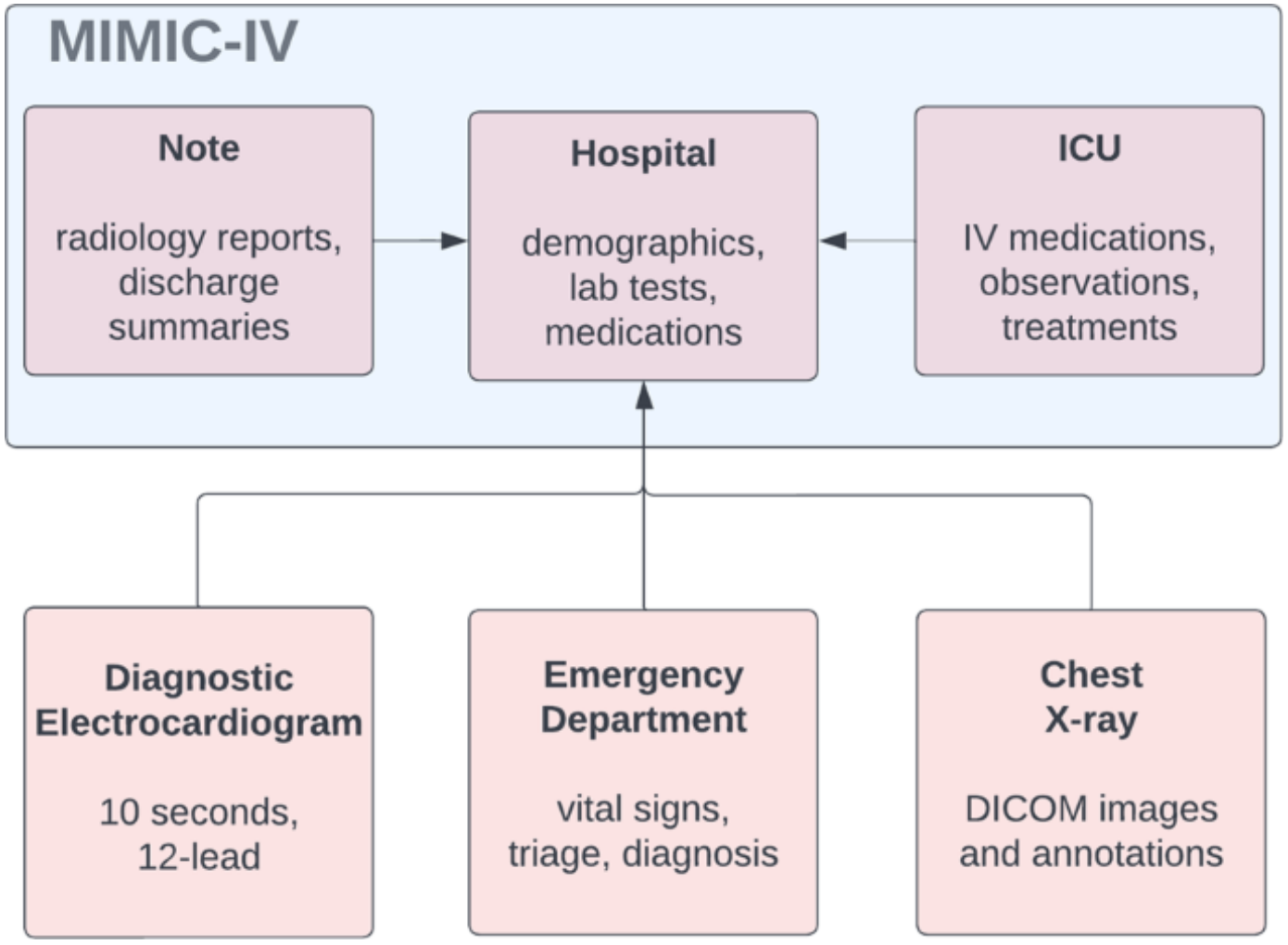




[Dai et al., CLIMB: Data Foundations for Large Scale Multimodal Clinical Foundation Models. ICML 2025]

multisensory intelligence

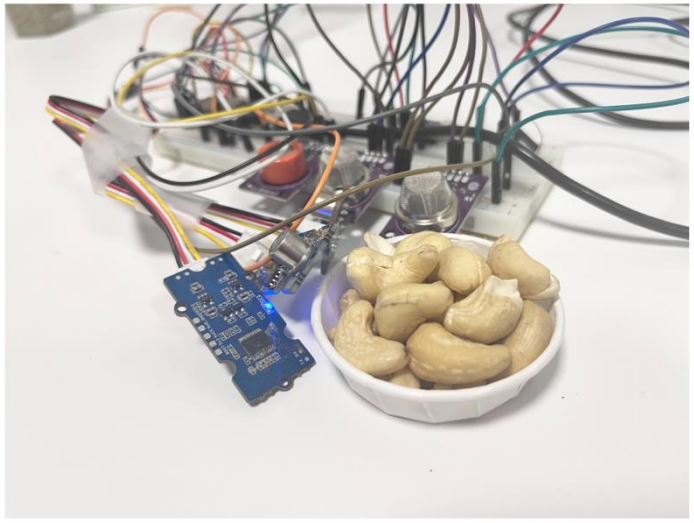
MIMIC



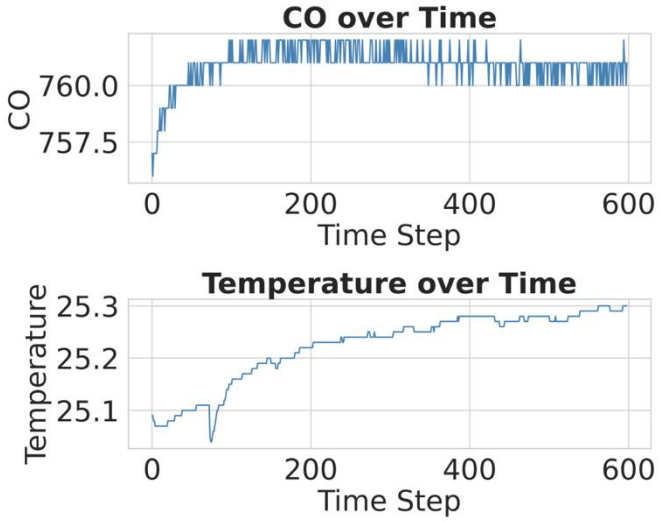
multisensory intelligence

SmellNet

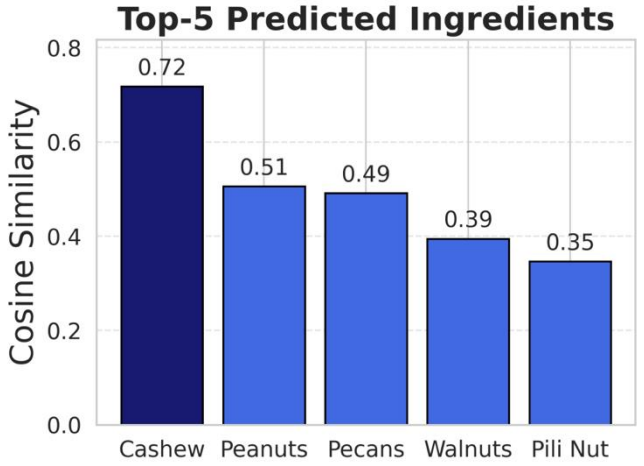
https://github.com/MIT-MI/SmellNet



(a) Sensor setup detecting cashew.



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



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

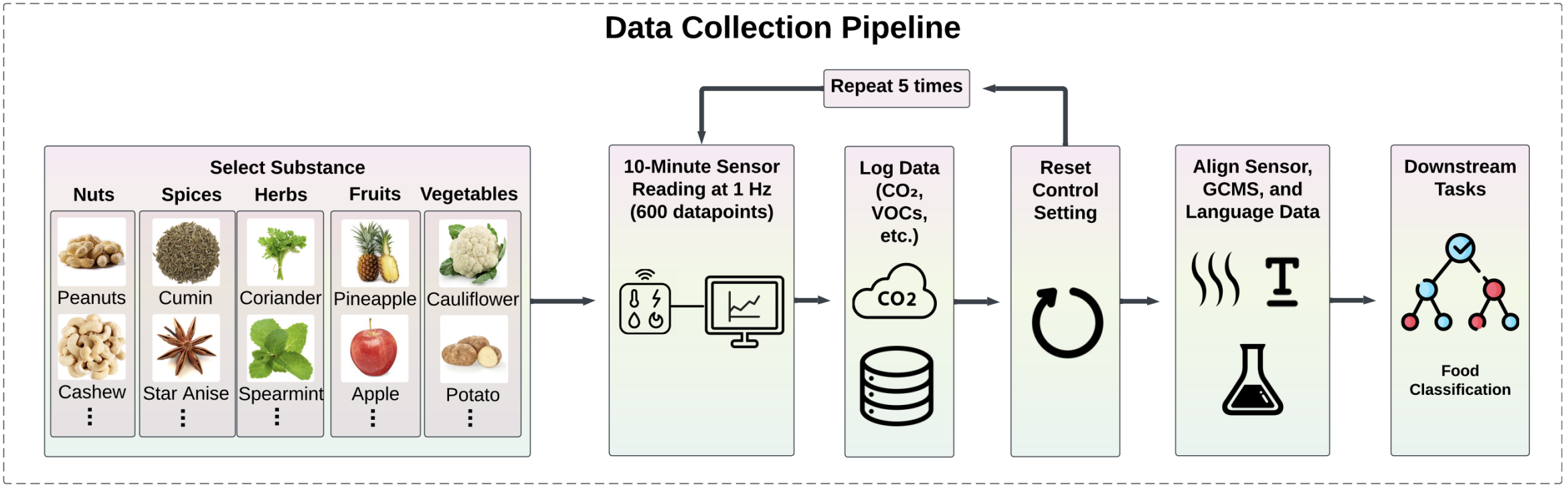
[Feng et al., SmellNet: A Large-scale Dataset for Real-world Smell Recognition. ICLR 2026]

multisensory intelligence

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

OpenTouch Dataset

https://opentouch-tactile.github.io

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

Robotics

→ World models

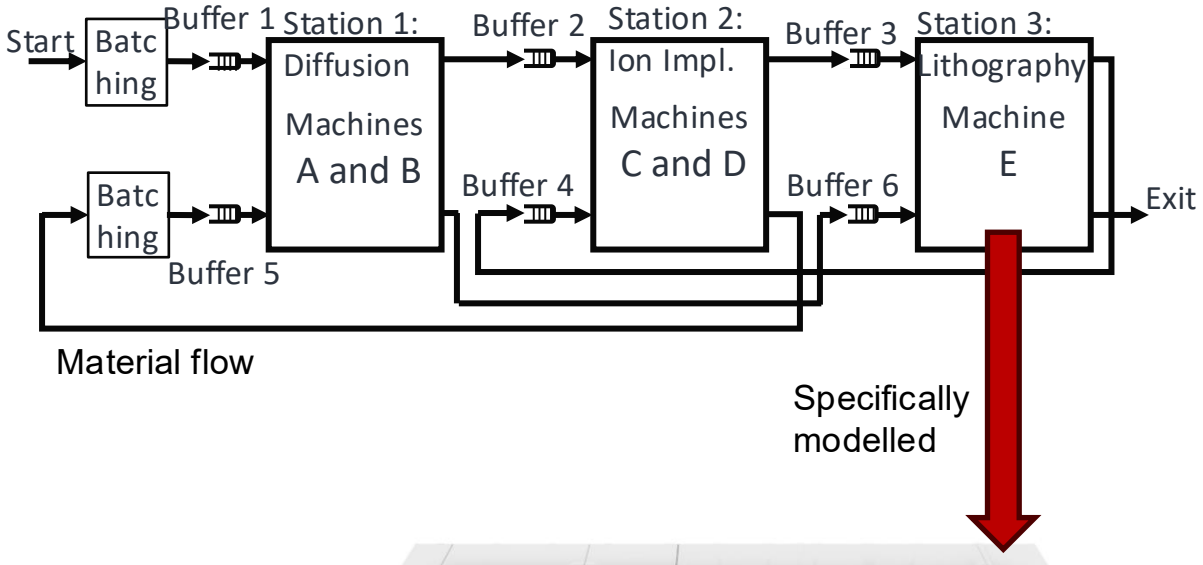
Touch and haptics

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

multisensory intelligence

IntelMini Fab - Manufacturing

System visualization (Simulation)



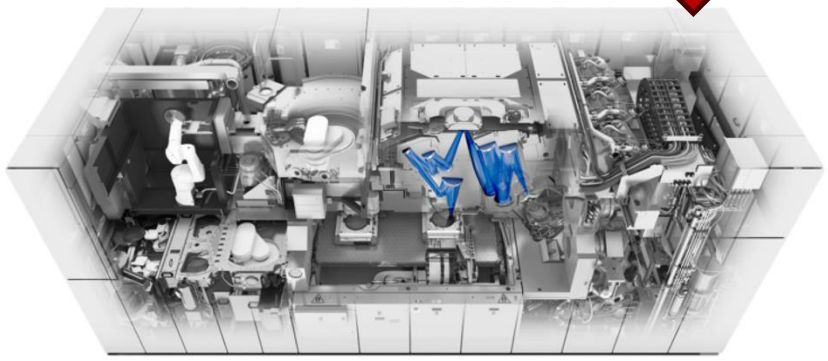
Subsystems,

Energy usage,

Failure rates,

Industrial Data

For instantiation



ASML EUV System. Source: Zeiss 2022

Description:

Discrete Event Simulation (Python, SimPy based), targets:

Factory performance (how many semiconductor wafers per hour), Quality performance (quality rate), Costs (processing, maintenance, …)

Main control levers:
- Material Flow (which jobs=wafer set to prioritize)
- Maintenance (when to perform which maintenance)

Focus on contingencies (non steady states):
- Contingency: Examples:

1. Machine Breakdown (Maintenance required) -> how to repair, how long, who, implications, … + how to route material flow in the meantime
2. Supply Chain (Input not available) -> how to react
3. Power/Utilities or workforce issues -> how to run the factory
4. Quality issues -> how to resolve for individual lot and how to restore quality in factory
5. …..

IntelMini Fab - Manufacturing

System visualization (Real-Data)

Turning/Milling Data

Real industrial large scale machine, retrofitted with various sensors

Quality evaluation

Grinding Data

Real industrial large scale machine

Quality evaluation available

Description:

Real machine process data (from partner):

Data is preprocessed and analyzed traditionally

Hundreds (?) of hours of grinding data:

- Acoustic emission signals
- Force measurements

Hundreds of hours of turning/milling data:

- Accelerometer, microphone, dynamometer

Discussion underway to get more industrial data as well

Goal: Build a turning/milling and/or grinding foundation model

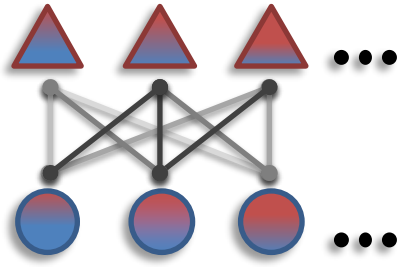
Extension possible with cross-attention and publicly available knowledge about the processes

Research Projects: Multimodal Fusion & Alignment

Motivation: Designing new ways to integrate and connect highly heterogeneous signals, bridging spatial & temporal, discrete & continuous.

Challenges:

- Multimodal deep learning + time-series analysis + tabular models
- Multimodal fusion of discrete and continuous data
- Discretizing and tokenizing continuous data for multimodal alignment
- Multimodal fusion and alignment without tokenization



Potential models and dataset to start with

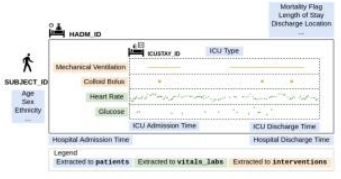
- Omni-modal models
- Holistic benchmarks

MultiBench

Domains

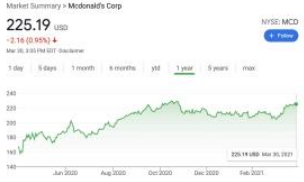
Affective computing

Healthcare



Robotics

Finance



HCI

Multimedia

Modalities

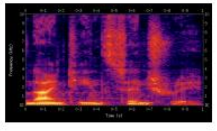
Language

All I can say is he's a pretty average guy.

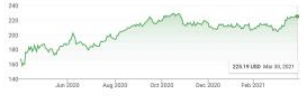
Image

Video

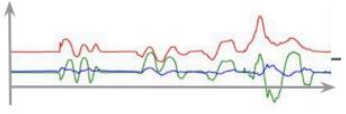
Audio



Time-series



Force sensors



Proprioception

Set

Table

SUBJECT_ID
Age
Sex
Ethnicity
...

Optical flow

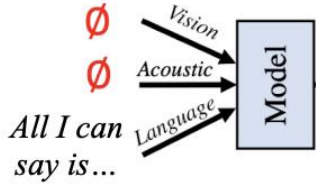
Evaluation

Performance

Rank	Method	Test Accuracy	Validation Accuracy
1	SAGN+SLE	0.8428 ± 0.0014	0.9287 ± 0.0003
2	MLP + C&S	0.8418 ± 0.0007	0.9147 ± 0.0009

Complexity

Robustness



multisensory intelligence

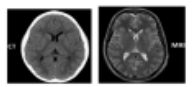
MultiBench++

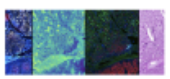
Datasets

Specialized Domains

Affective computing

Health care





Remote sensing

Others

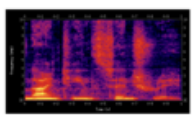
15+ Modalities

Text

He is a smart guy.

Image

Audio



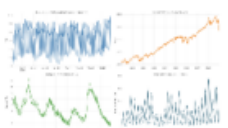
Table

SUBJECT_ID

Age
Sex
Ethnicity
...

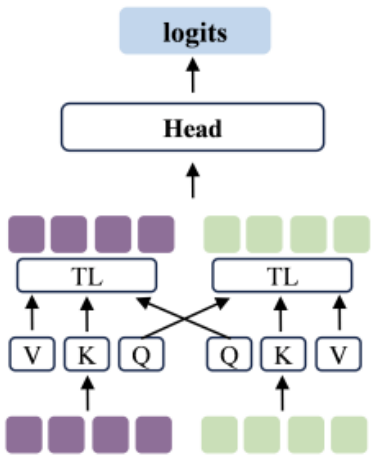
Video

Time-series

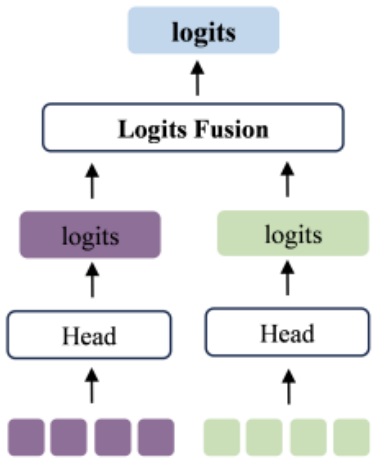


Broader & Deeper Datasets

Fusion Paradigms



Feature-level Fusion



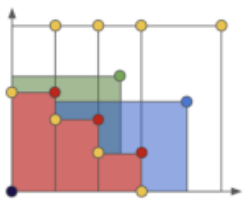
Decision-level Fusion

More Advanced Fusion Paradigms

Tuning

Define Objective Function

Create study

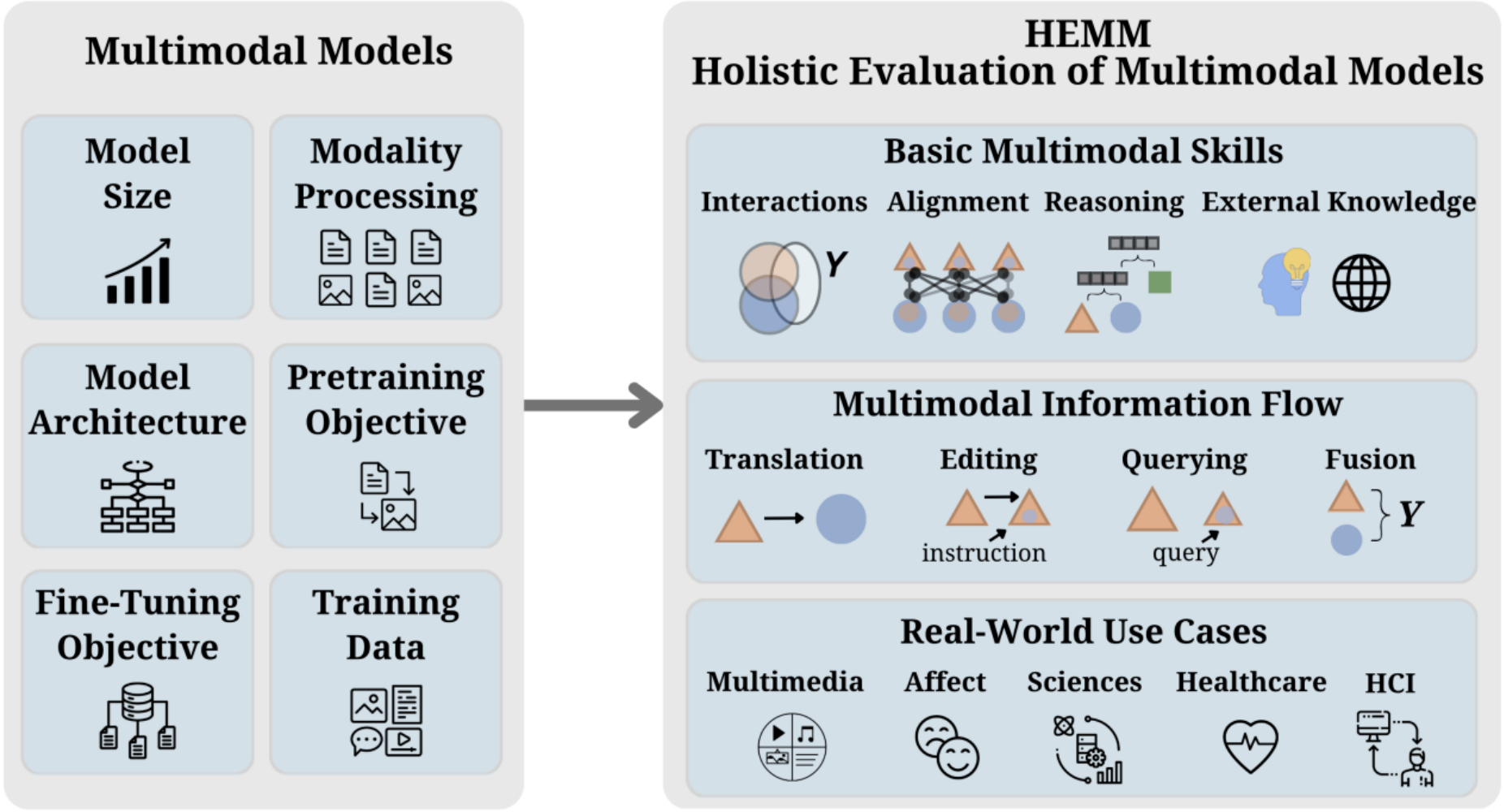


Hyper-Parameter Tuning

Automated Hyper-Parameter Tuning

multisensory intelligence

HEMM

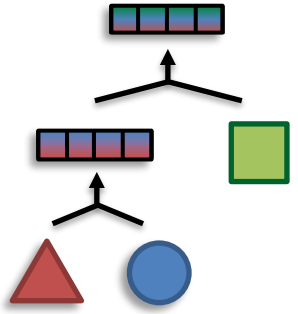

Multimodal Models
Model Size
Modality Processing
Model Architecture
Pretraining Objective
Fine-Tuning Objective
Training Data
HEMM
Holistic Evaluation of Multimodal Models
Basic Multimodal Skills
Interactions
Alignment
Reasoning
External Knowledge
Y
Multimodal Information Flow
Translation
Editing
Querying
Fusion
instruction
query
Y
Real-World Use Cases
Multimedia
Affect
Sciences
Healthcare
HCI

Research Projects on AI Reasoning

Motivation: Robust, reliable, interpretable reasoning in (multimodal) LLMs.

Challenges:

- Fine-grained and compositional reasoning
- Neuro-symbolic reasoning
- Emergent reasoning in foundation models



Potential models and dataset to start with

- LLM reasoning survey: https://arxiv.org/abs/2501.09686
- Can LLMs actually reason and plan? https://arxiv.org/abs/2403.04121
- Emotion-o1: https://arxiv.org/abs/2505.22548
- Social genome: https://arxiv.org/abs/2502.15109

Media Description – MS COCO

- Microsoft Common Objects in COntext (MS COCO)
- 120000 images
- Each image is accompanied with five free form sentences describing it (at least 8 words)
- Sentences collected using crowdsourcing (Mechanical Turk)
- Also contains object detections, boundaries and keypoints

The man at bat readies to swing at the pitch while the umpire looks on.

A large bus sitting next to a very tall building.

Visual Question Answering – VQA

- Task - Given an image and a question, answer the question (http://www.visualqa.org/)

What color are her eyes?
What is the mustache made of?

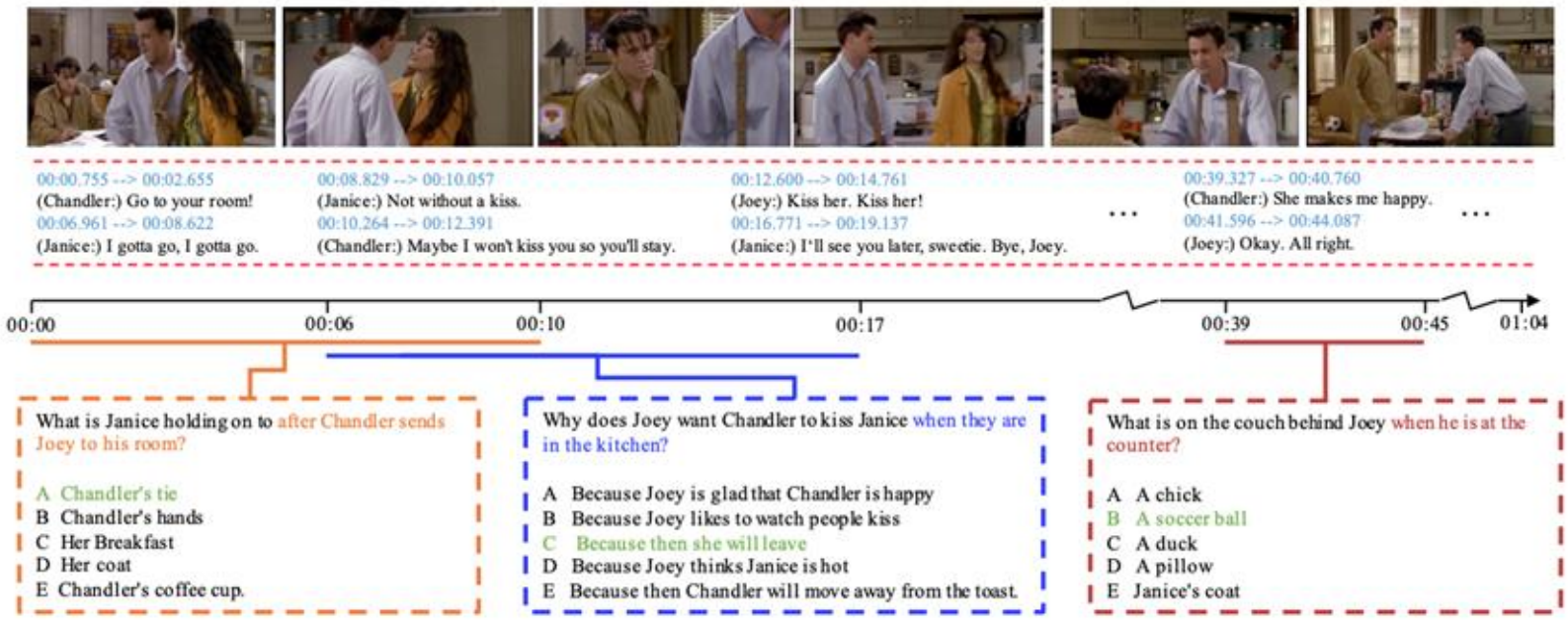
How many slices of pizza are there?
Is this a vegetarian pizza?

Is this person expecting company?
What is just under the tree?

Does it appear to be rainy?
Does this person have 20/20 vision?

Multimodal QA

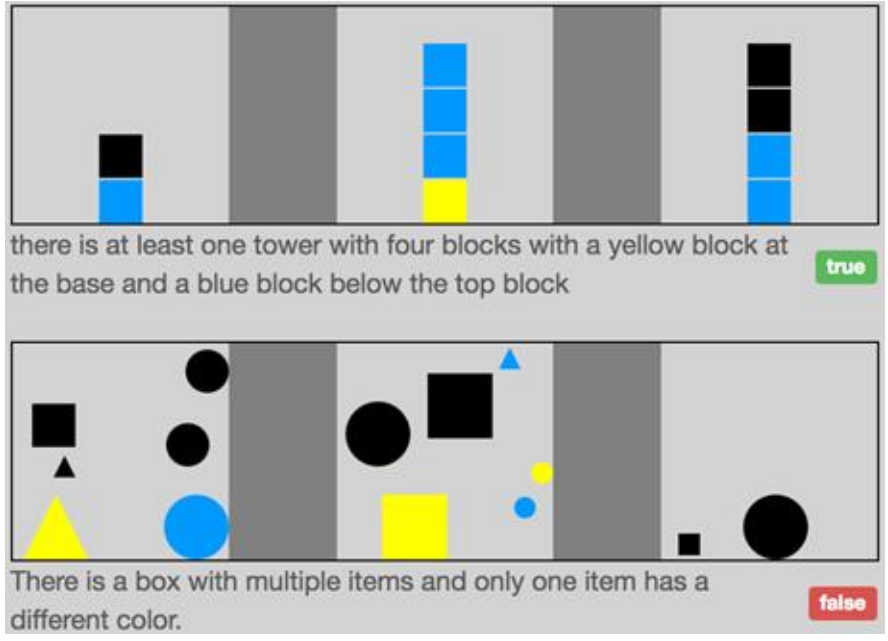
- TVQA
  - Video QA dataset based on 6 popular TV shows
  - 152.5K QA pairs from 21.8K clips
  - Compositional questions



multisensory intelligence

Multimodal QA – Visual Reasoning

- Cornell NLVR
  - 92,244 pairs of natural language statements grounded in synthetic images
  - Determine whether a sentence is true or false about an image



Multimodal QA – Visual Reasoning

- Cornell NLVR2
  - Same as NLVR but with >100k real images

Winoground

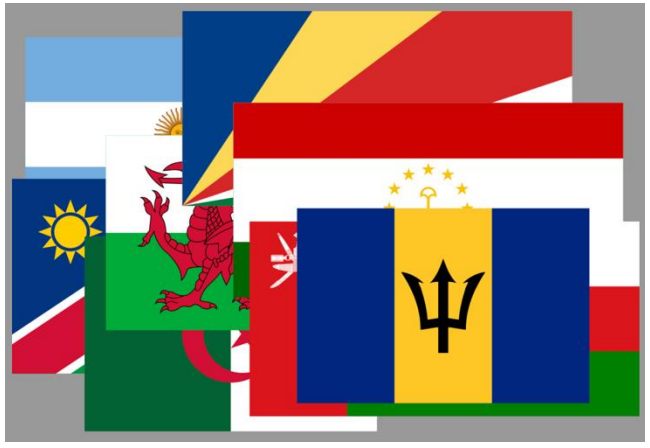
- Github
- Same words, different order, different images. Intended to test the compositionality of vision-language models

(a) some plants surrounding a lightbulb

(b) a lightbulb surrounding some plants

PuzzleWorld

"Top-Down Processing"



GPT o1 output

- From top-down, the flags are Barbados, Oman, Tajikistan, Seychelles, Wales, Algeria, Namibia, Argentina ✓
- Extracting first letter gives B, O, T, S, W, A, N, A ✓
- the puzzle hints at the name of the country **BOTSWANA** ✓

3/3

#1: Recognize the images as country flags	#2: Identify the country names of the flags	#3: Arrange the country names' first letters in order
Visual Reasoning	***External Knowledge***	***Text Reasoning***

[Li et al., PuzzleWorld: A Benchmark for Multimodal, Open-ended Reasoning in PuzzleHunts. ICLR 2026]

PuzzleWorld

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

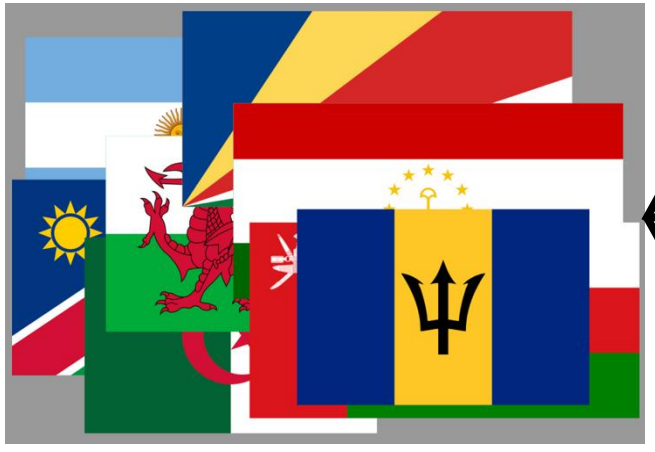
Puzzles require:

Advanced multimodal understanding

Diverse reasoning strategies

Extensive exploration

"Top-Down Processing"



- Input Modality: text, visual, structured
- Reasoning Skills: logic, knowledge, spatial, ...
- Solution Steps: #1 → #2 → #3 ...

~500 puzzles in total, annotated for human reasoning steps

[Li et al., PuzzleWorld: A Benchmark for Multimodal, Open-ended Reasoning in PuzzleHunts. ICLR 2026]

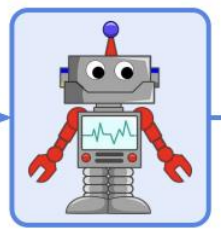
multisensory intelligence

ScienceQA

Question: Which type of force from the baby's hand opens the cabinet door?

Options: (A) pull (B) push

Context: A baby wants to know what is inside of a cabinet. Her hand applies a force to the door, and the door opens.



Answer: The answer is A.

BECAUSE:

Lecture: A force is a push or a pull that one object applies to a second object. The direction of a push is away from the object that is pushing. The direction of a pull is toward the object that is pulling.

Explanation: The baby's hand applies a force to the cabinet door. This force causes the door to open. The direction of this force is toward the baby's hand. This force is a pull.

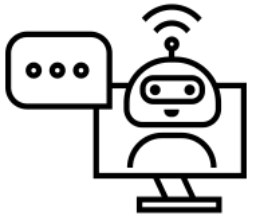
[Lu et al., Learn to Explain: Multimodal Reasoning via Thought Chains for Science Question Answering. NeurIPS 2022]

Research Projects on Interactive Agents

Motivation: Grounding AI models in the web, computer, or other virtual worlds to help humans with digital tasks.

Challenges:

- Instructions and language grounded in web images, tools, APIs
- Search over environment, plan long-term actions and effects
- Asking for human clarification, human-in-the-loop
- Potential risks of interactive agents



Potential models and dataset to start with

- WebArena: https://arxiv.org/pdf/2307.13854.pdf
- AgentBench: https://arxiv.org/pdf/2308.03688.pdf
- ToolFormer: https://arxiv.org/abs/2302.04761
- SeeAct: https://osu-nlp-group.github.io/SeeAct/
- Socially intelligent AI agents: https://arxiv.org/abs/2404.11023

multisensory intelligence

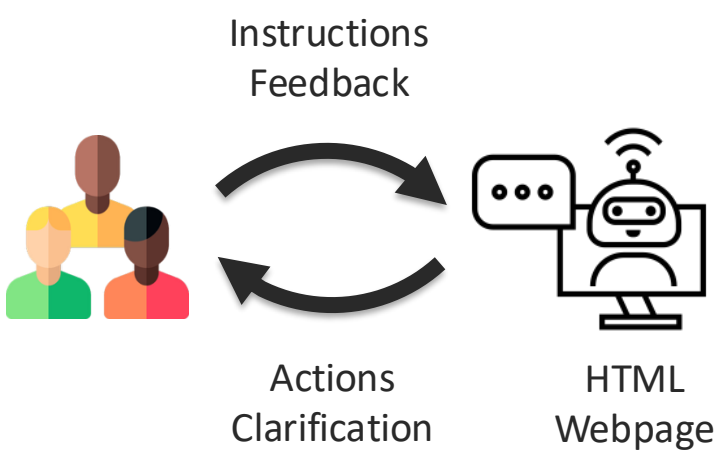
WebQA

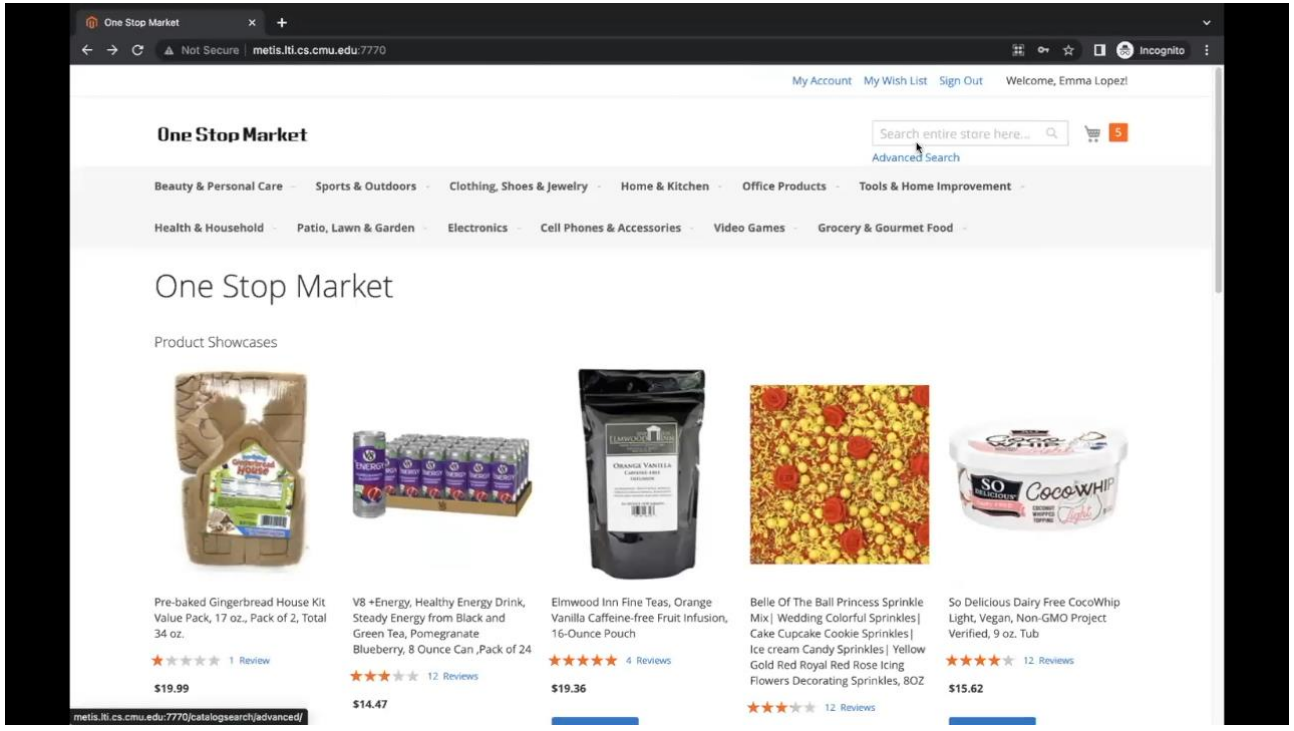
- https://webqna.github.io/
- Given a question Q, and a list of sources S = {s1, s2, ...}, a system must a) identify the sources from which to derive the answer, and b) generate an answer as a complete sentence.

multisensory intelligence

WebArena Environment

Example task: Purchase a set of earphones with at least 4.5 stars in rating and ship it to me.

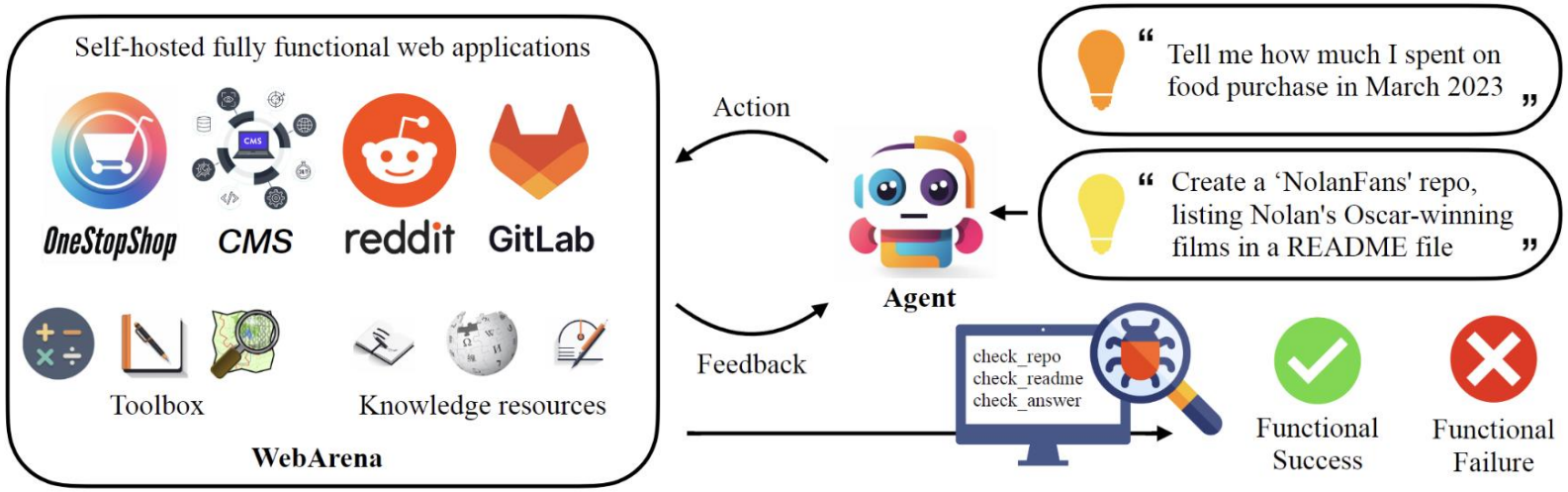




[Zhou et al., WebArena: A Realistic Web Environment for Building Autonomous Agents. ICLR 2024]

multisensory intelligence

WebArena Environment

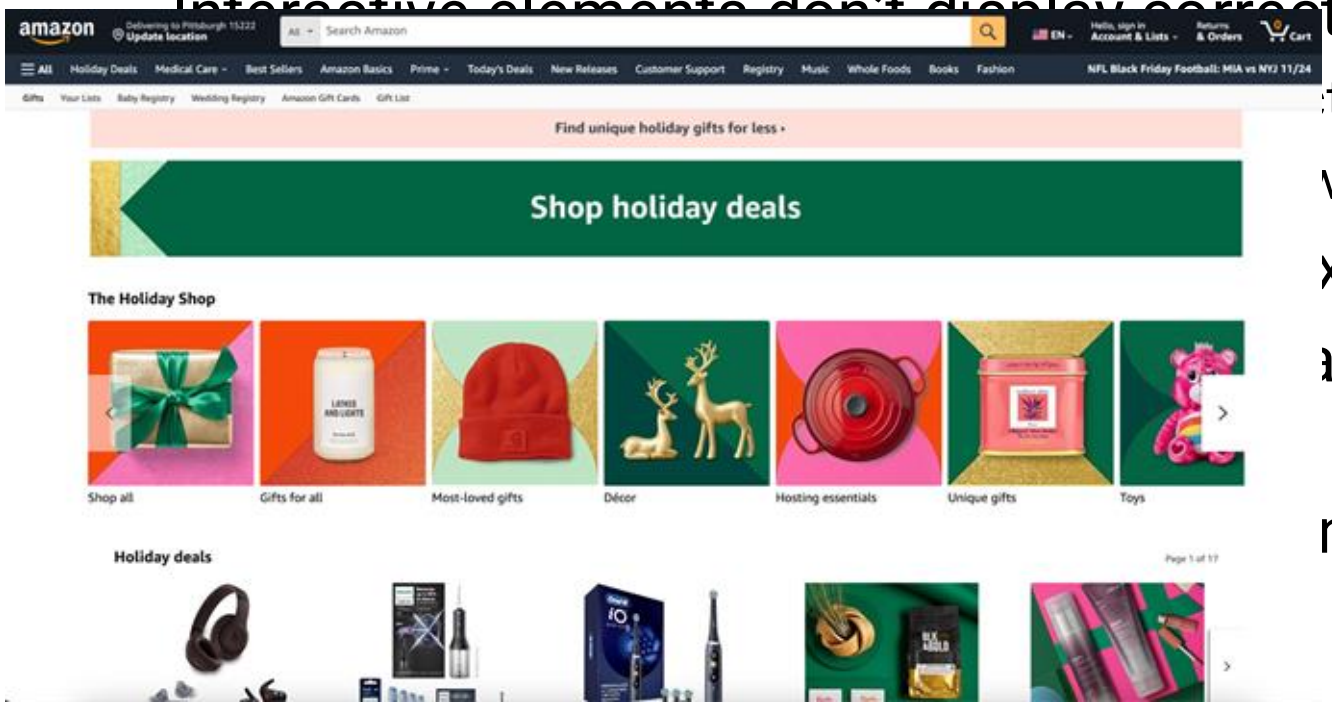


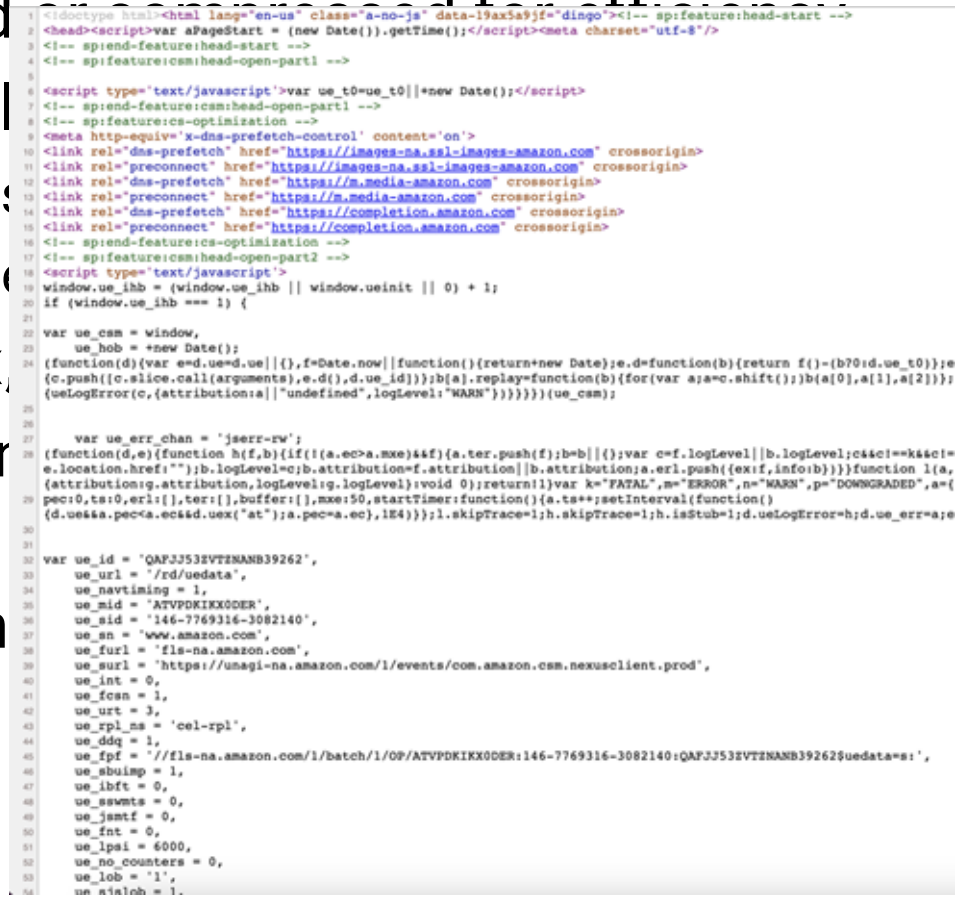
- Websites from four popular categories (shopping, CMS, Reddit, GitLab)
  - Self-hosted open source re-implementations
  - Data from real websites (Amazon, Reddit, GitHub)
- Tasks easy for humans (78% success) but difficult for LLM agents (14%)
- **But:** Tasks are designed to use just text and HTML source code

[Zhou et al., WebArena: A Realistic Web Environment for Building Autonomous Agents. ICLR 2024]

HTML is Insufficient

- Messy HTML, JavaScript: usually minified or compressed for efficiency
- Interactive elements don't display correctly

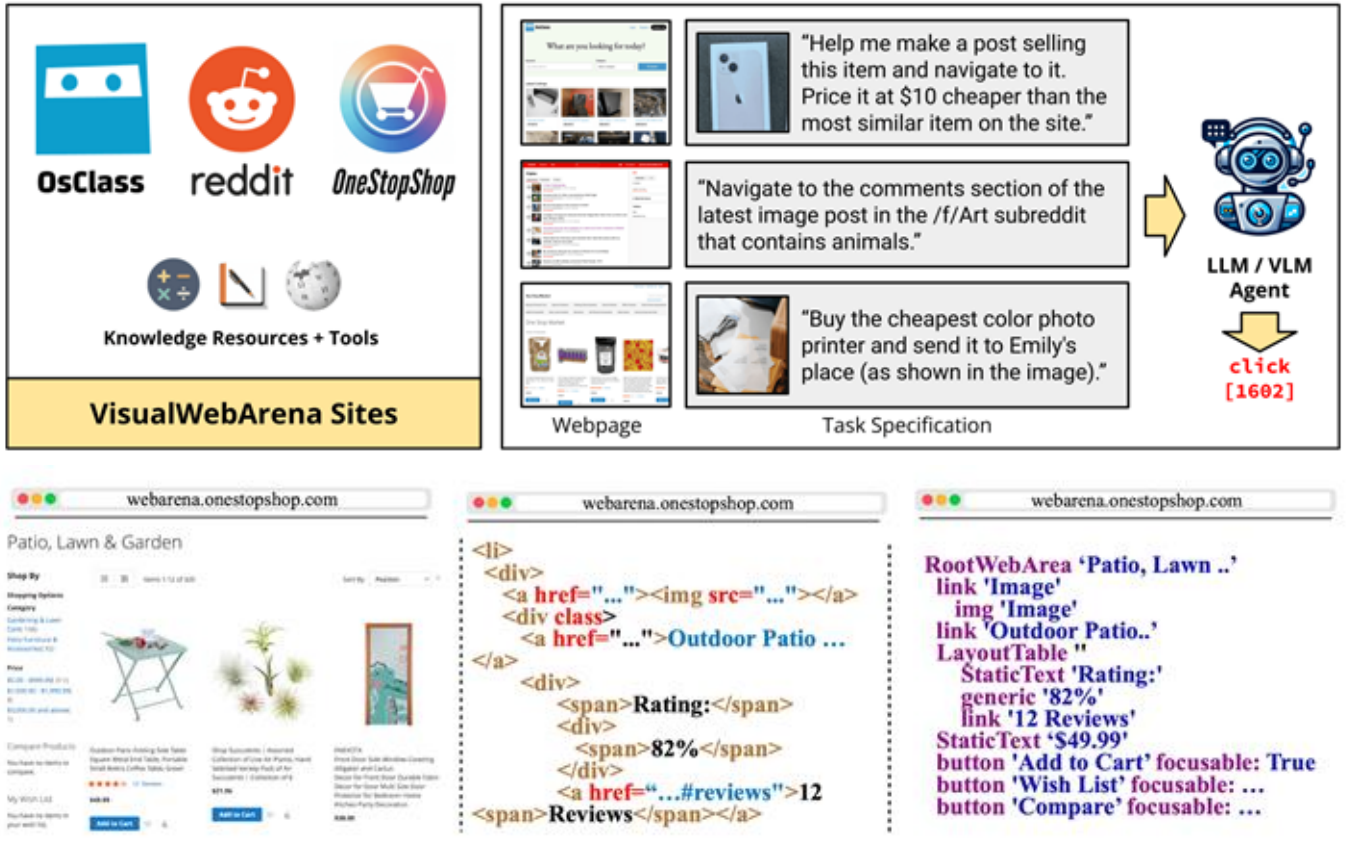




multisensory intelligence

VisualWebArena: A Visually Grounded Benchmark

- Benchmark and track the progress of **multimodal agents**

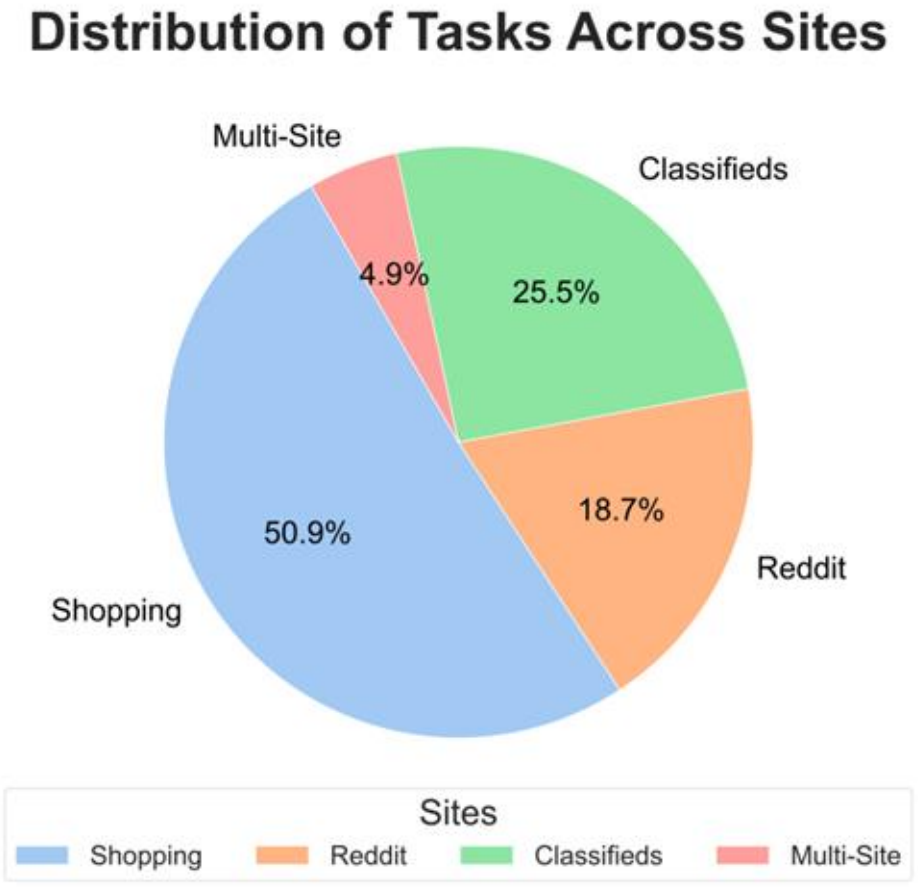


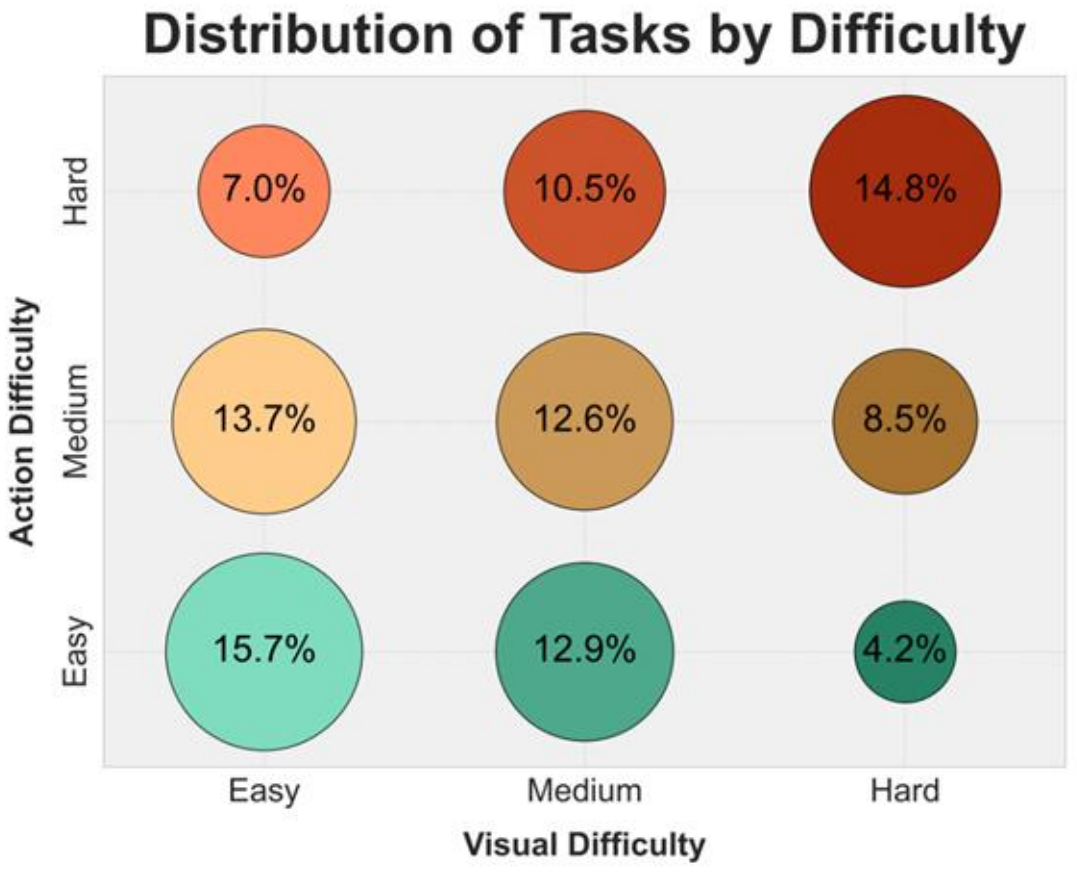
Action Type a	Description
`click [elem]`	Click on element `elem`.
`hover [elem]`	Hover on element `elem`.
`type [elem] [text]`	Type `text` on element `elem`.
`press [key_comb]`	Press a key combination.
`new_tab`	Open a new tab.
`tab_focus [index]`	Focus on the i-th tab.
`tab_close`	Close current tab.
`goto [url]`	Open `url`.
`go_back`	Click the back button.
`go_forward`	Click the forward button.
`scroll [up|down]`	Scroll up or down the page.
`stop [answer]`	End the task with an optional output.

[Koh et al., VisualWebArena: Evaluating Multimodal Agents on Realistic Visual Web Tasks. ACL 2024]

multisensory intelligence

VisualWebArena: A Visually Grounded Benchmark



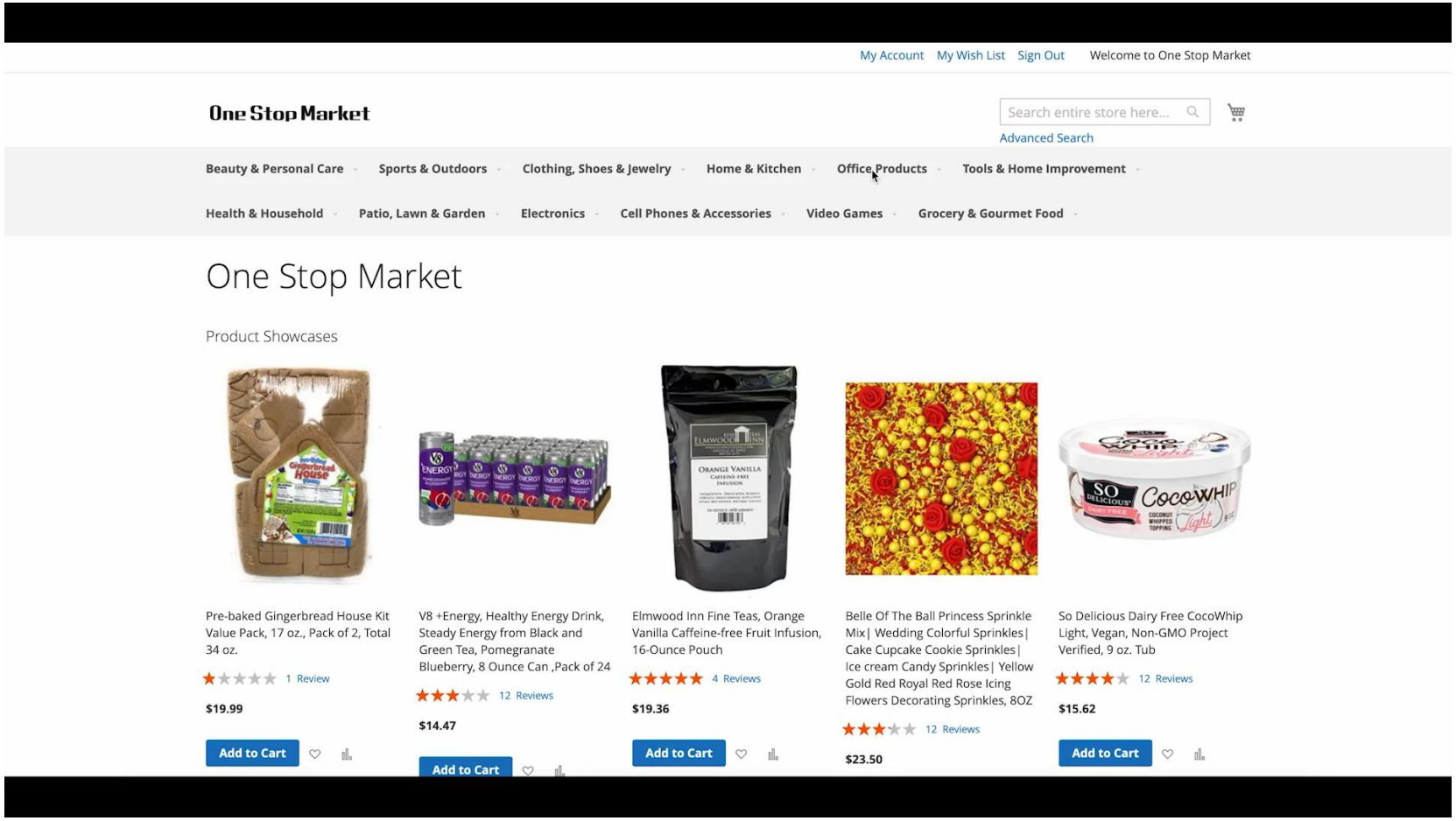


[Koh et al., VisualWebArena: Evaluating Multimodal Agents on Realistic Visual Web Tasks. ACL 2024]

multisensory intelligence

VisualWebArena Shopping Example

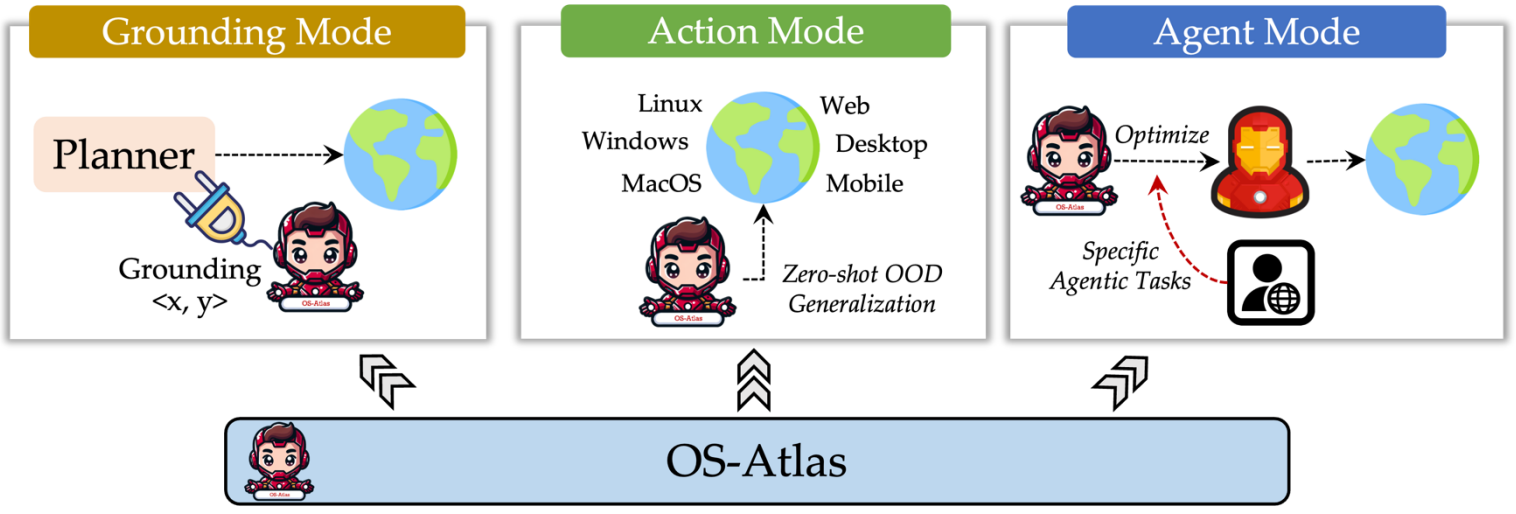
Task: Buy the cheapest color photo printer and send it to Emily's place (as shown in the image).

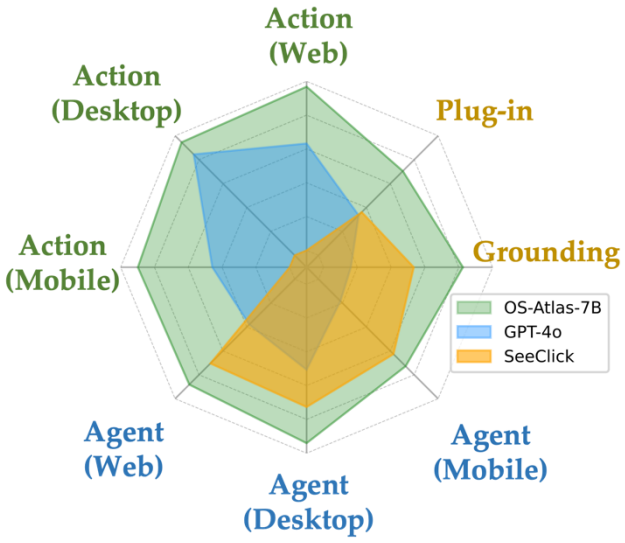


[Koh et al., VisualWebArena: Evaluating Multimodal Agents on Realistic Visual Web Tasks. ACL 2024]

OS-Atlas

Web + desktop + mobile GUI grounding



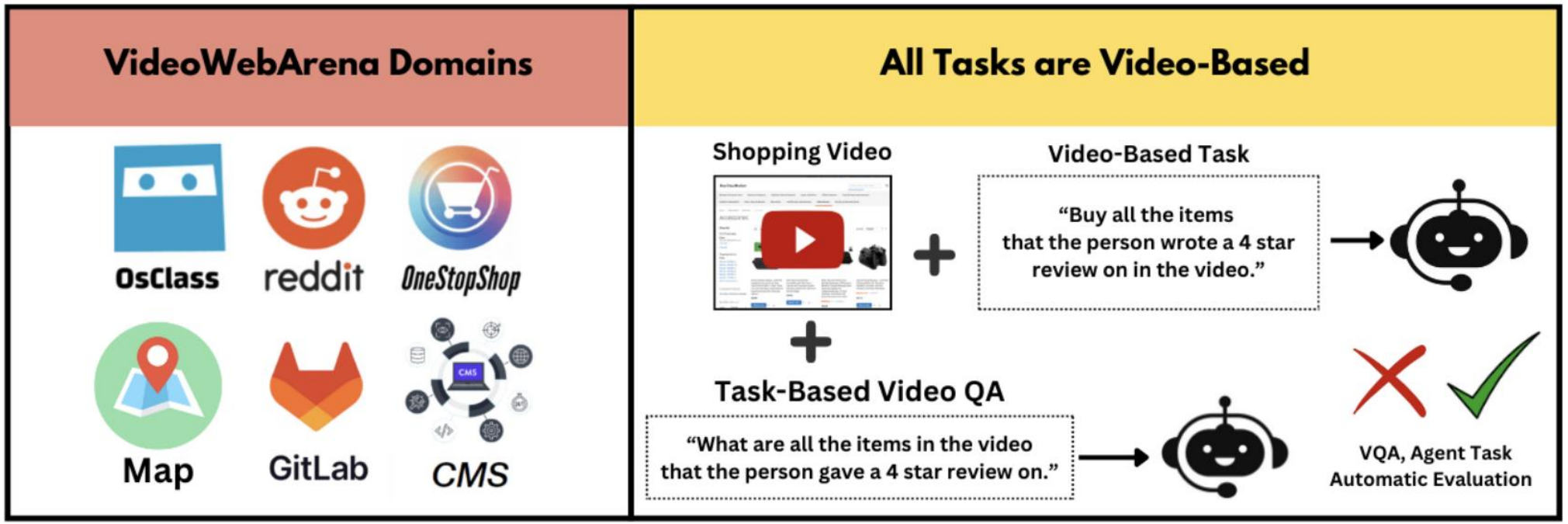


[Wu et al., OS-Atlas: A Foundation Action Model for Generalist GUI Agents. ICLR 2025]

multisensory intelligence

VideoWebArena

Extension to video understanding web tasks

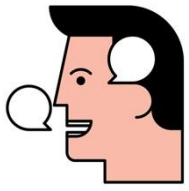


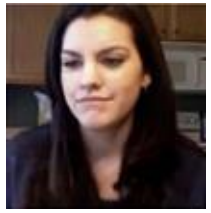
[Jang et al., VideoWebArena: Evaluating Multimodal Agents on Video Understanding Web Tasks. ICLR 2025]

multisensory intelligence

Research Projects on Socially Intelligent AI

Motivation: Building AI that can understand and interact with humans in social situations.





Challenges:

- Social interaction, reasoning, and commonsense.
- Building social relationships over months and years.
- Theory-of-Mind and multi-party social interactions.

Potential models and dataset to start with

- Multimodal WereWolf: https://github.com/SALT-NLP/PersuationGames
- Ego4D: https://arxiv.org/abs/2110.07058
- MMToM-QA: https://openreview.net/pdf?id=jbLM1yvxaL
- 11866 Artificial Social Intelligence: https://cmu-multicomp-lab.github.io/asi-course/spring2023/

Affect Recognition

- Three AVEC challenge datasets 2011/2012, 2013/2014, 2015, 2016, 2017, 2018
- Audio-Visual emotion recognition
- Labeled for dimensional emotion (per frame)
- 2011/2012 has transcripts
- 2013/2014/2016 also includes depression labels per subject
- 2013/2014 reading specific text in a subset of videos
- 2015/2016 includes physiological data
- 2017/2018 includes depression/bipolar

AVEC 2011/2012

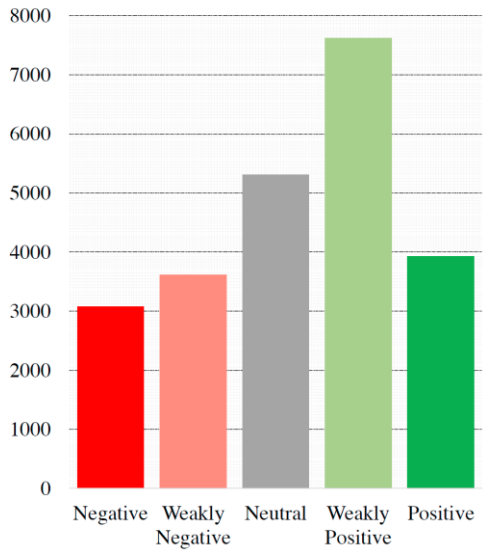
AVEC 2013/2014

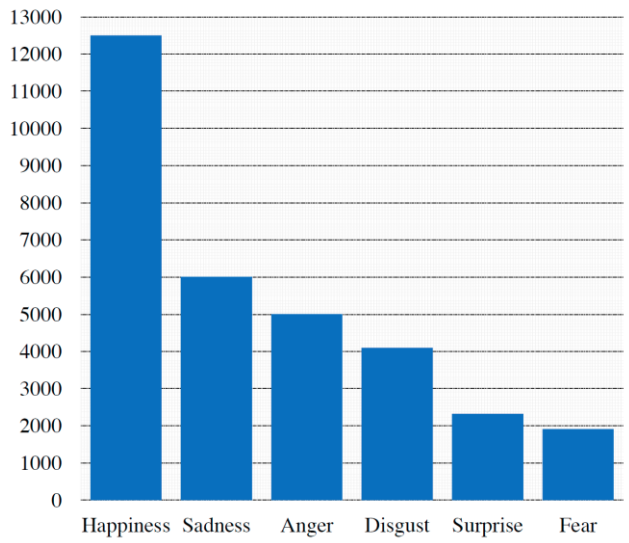
AVEC 2015/2016

multisensory intelligence

Multimodal Sentiment Analysis

- Multimodal sentiment and emotion recognition
- CMU-MOSEI : 23,453 annotated video segments from 1,000 distinct speakers and 250 topics

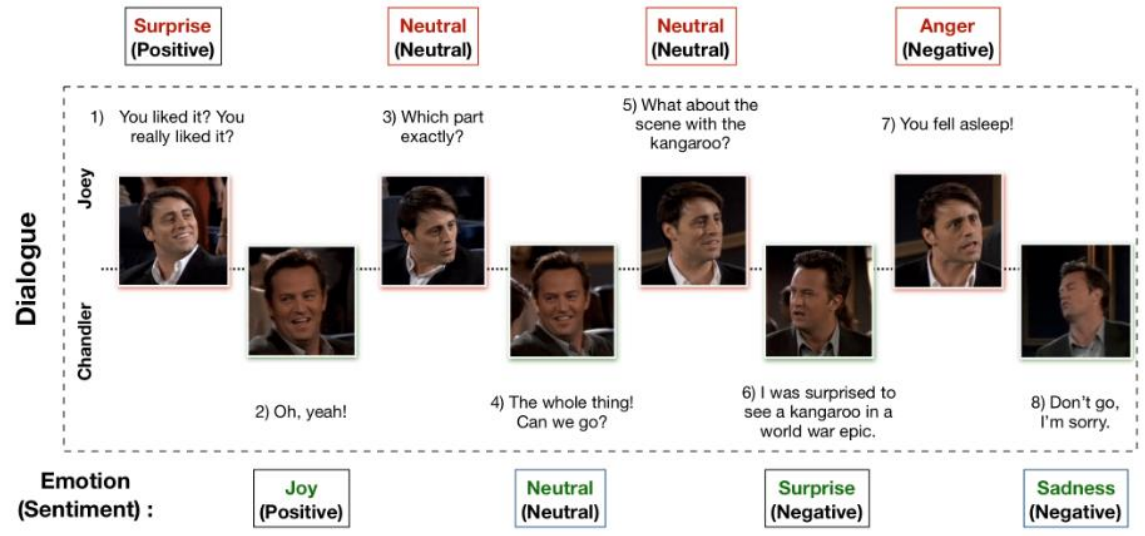




multisensory intelligence

Multi-Party Emotion Recognition

- MELD: Multi-party dataset for emotion recognition in conversations



Utterance: *"Become a drama critic!"*
Emotion: *Joy* **Sentiment:** *Positive*

Text	Audio	Visual
Ambiguous	Joyous tone	Smiling Face

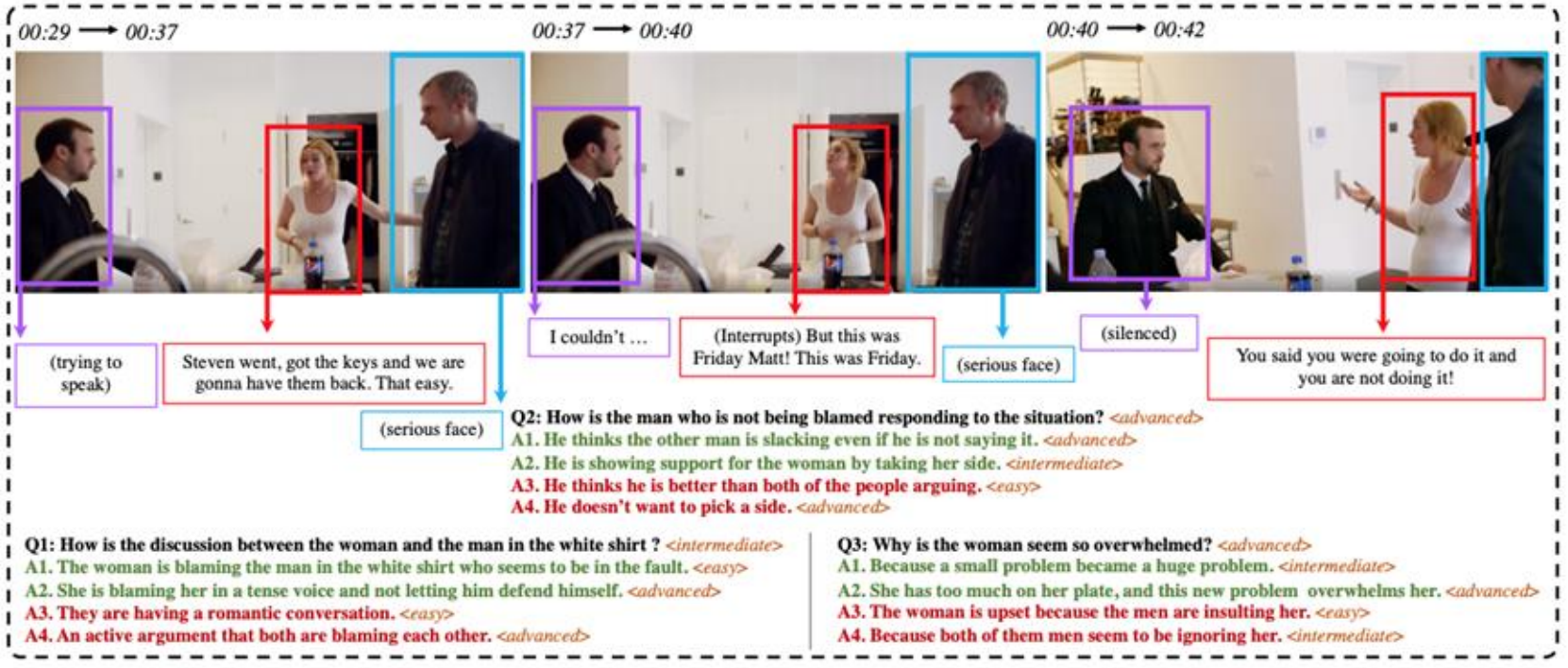
Utterance: *"Great, now he is waving back"*
Emotion: *Disgust* **Sentiment:** *Negative*

Text	Audio	Visual
Positive/Joy	Flat tone	Frown

multisensory intelligence

Social Interaction Q&A

- Social-IQ: 1.2k videos, 7.5k questions, 50k answers
- Questions and answers centered around social behaviors



MimeQA: Social Reasoning without Language

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

Grounding the Imaginary | **Object Recognition**

Recognizing imaginary objects and activities portrayed with abstract gestures and body movements.

Q: What is the person in the black shirt holding?

A: The person is holding a heavy object, likely a stone.

[Li et al., MimeQA: Towards Socially-Intelligent Nonverbal Foundation Models. NeurIPS 2025]

multisensory intelligence

MimeQA: Social Reasoning without Language

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

Scene-Level

Temporal Reasoning

Connecting events, reading emotions, and decoding intentions across a short video scene.

Q: What caused the person reading a book to trip?

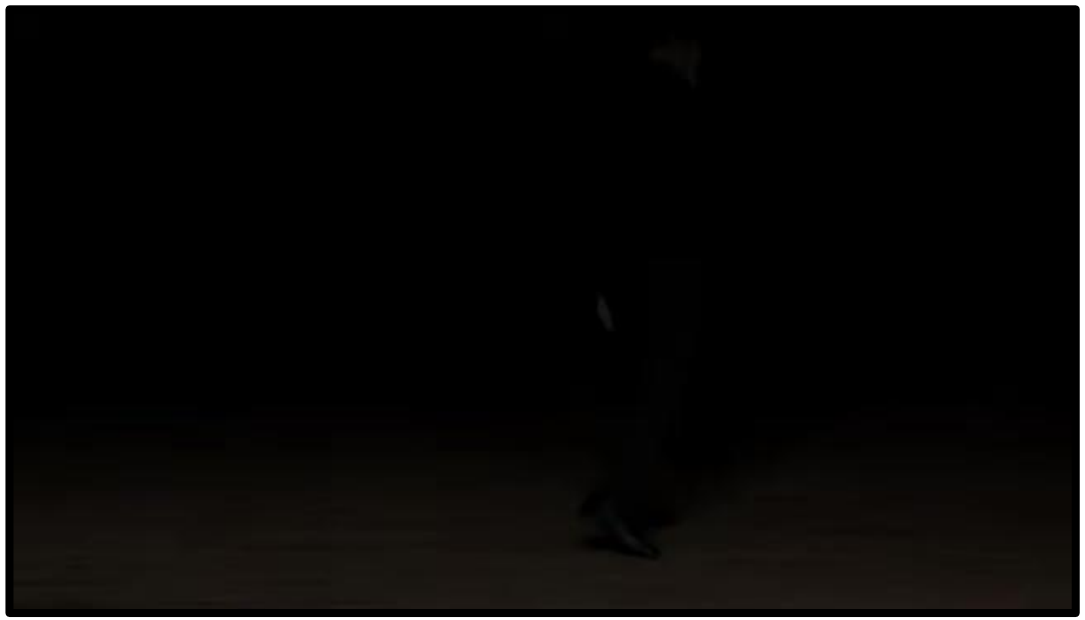
A: The person tripped over a heavy object left by someone earlier.

[Li et al., MimeQA: Towards Socially-Intelligent Nonverbal Foundation Models. NeurIPS 2025]

multisensory intelligence

MimeQA: Social Reasoning without Language

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



Global-Level | **Social Judgment**

Synthesizing and reasoning social information across multiple scenes to form higher-order interpretations

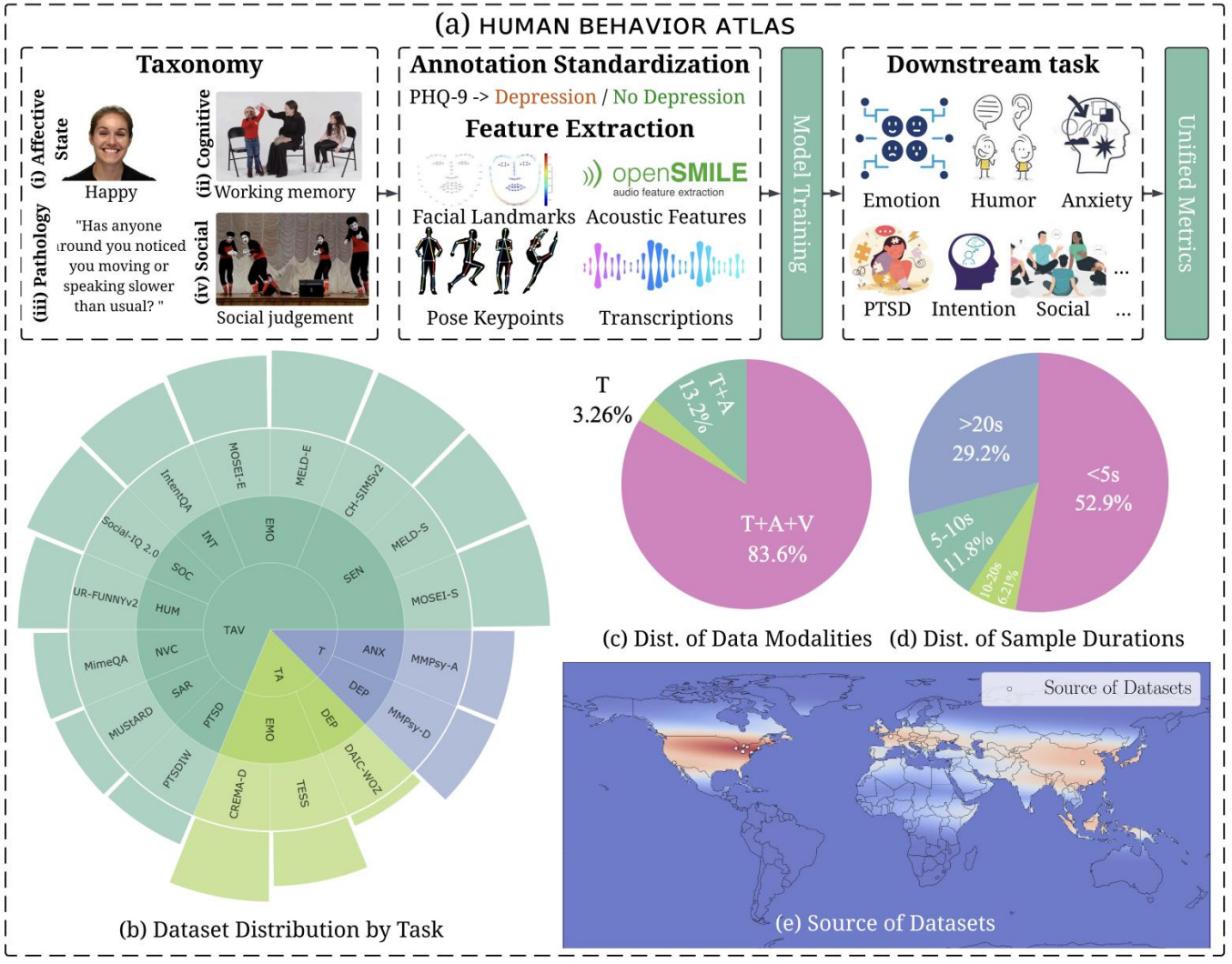
Q: How do the person's actions demonstrate his personality?

A: He is a mischievous person, as he pranks people walking by.

[Li et al., MimeQA: Towards Socially-Intelligent Nonverbal Foundation Models. NeurIPS 2025]

multisensory intelligence

Human Behavior Atlas

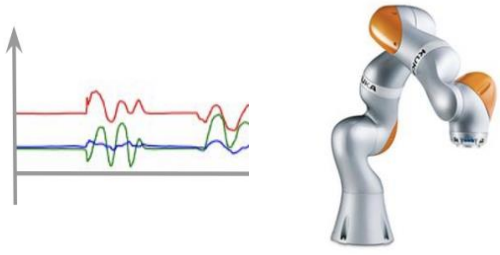


[Ong et al., Human Behavior Atlas: Towards Socially-Intelligent Nonverbal Foundation Models. ICLR 2026]

multisensory intelligence

Research Projects on Embodied and Tangible AI

Motivation: Building tangible and embodied AI systems that help humans in physical tasks.



Challenges:

- Perception, reasoning, and interaction
- Connecting sensing and actuation
- Efficient models that can run on hardware
- Understanding influence of actions on the world (world model)

Potential models and dataset to start with

- Virtual Home: http://virtual-home.org/paper/virtualhome.pdf
- Habitat 3.0 https://ai.meta.com/static-resource/habitat3
- RoboThor: https://ai2thor.allenai.org/robothor
- LangSuite-E: https://github.com/bigai-nlco/langsuite
- Language models and world models: https://arxiv.org/pdf/2305.10626.pdf

multisensory intelligence

Navigating in a Virtual House

Visually-grounded natural language navigation in real buildings

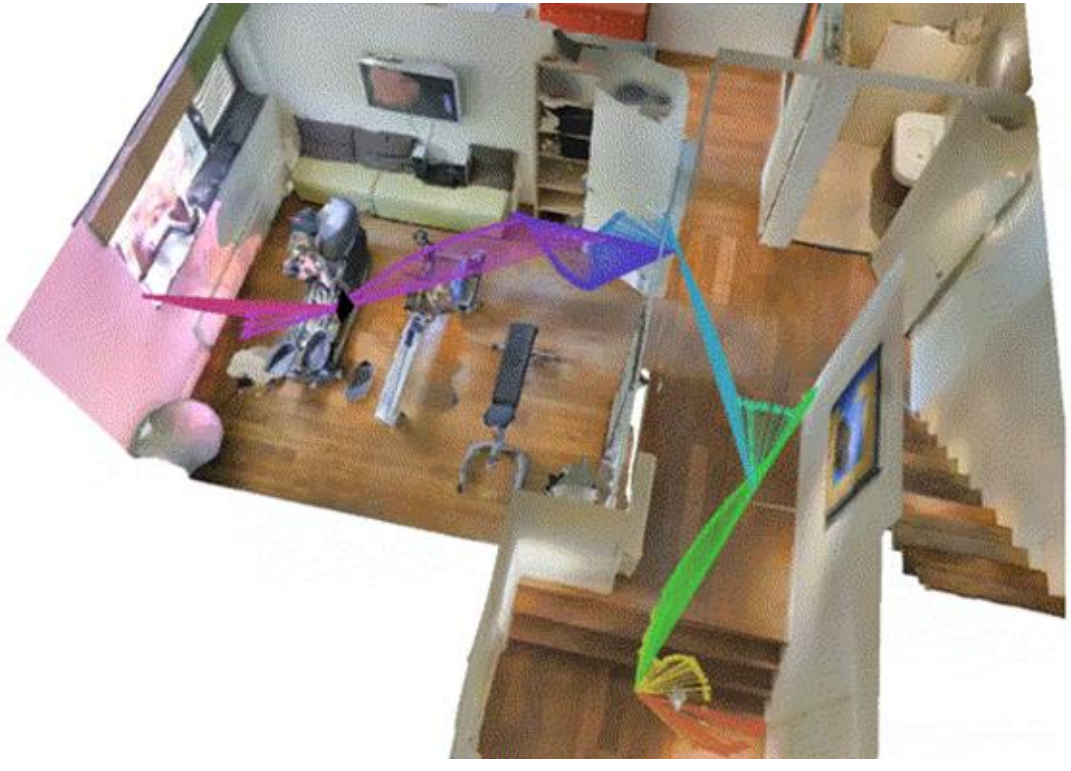
- Room-2-Room: 21,567 open vocabulary, crowd-sourced navigation instructions

Instruction: Head upstairs and walk past the piano through an archway directly in front. Turn right when the hallway ends at pictures and table. Wait by the moose antlers hanging on the wall.

multisensory intelligence

Room-Across-Room

- Github
- Similar to Room-to-Room (D1) except larger, multilingual, with longer paths



Now you are standing in-front of a closed door, turn to your left, you can see two wooden steps, climb the steps and walk forward by crossing a wall paint which is to your right side, you can see open door enter into it. This is a gym room, move forward, walk till the end of the room, you can see a grey colored ball to the corner of the room, stand there, that's your end point.

EPIC-Kitchens

- Dataset
- Large-scale dataset in first-person (egocentric) vision; multi-faceted, audio-visual, non-scripted recordings in native environments - i.e. the wearers' homes

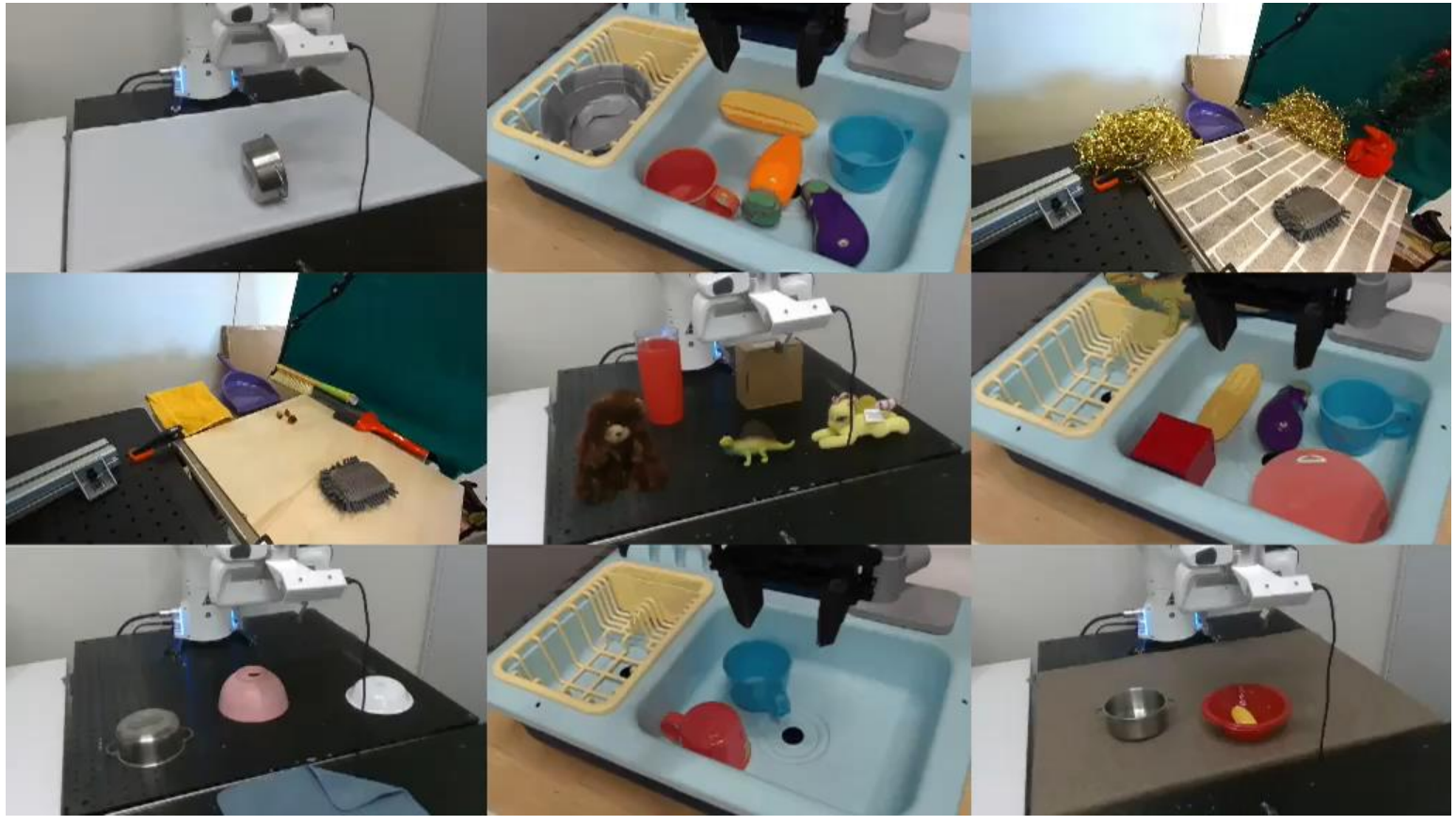
multisensory intelligence

Ego4D

multisensory intelligence

Embodied Agents

Generate precise robotics control directly via trained vision language models.

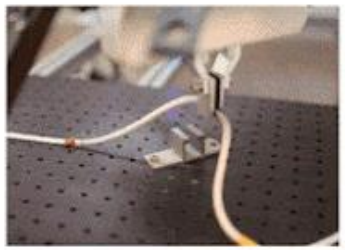


[Kim et al., OpenVLA: An Open-Source Vision-Language-Action Model. 2024]

multisensory intelligence

Google RT-X

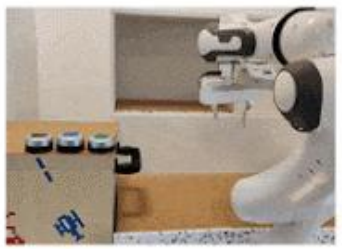
CLVR, USC



RAIL, UC Berkeley

CILVR, NYU

AUTOLab, UC Berkeley



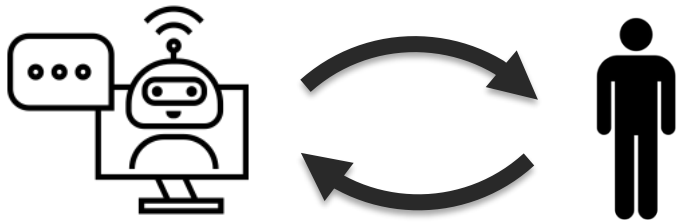
AiS, University of Freiburg

https://deepmind.google/discover/blog/scaling-up-learning-across-many-different-robot-types/

multisensory intelligence

Research Projects on Human-AI Interaction

Motivation: What is the right medium for human-AI interaction? How can we really trust AI? How do we enable collaboration and synergy?



Challenges:

- Modeling and conveying model uncertainty – text input uncertainty, visual uncertainty, multimodal uncertainty? cross-modal interaction uncertainty?
- Asking for human clarification, human-in-the-loop, types of human feedback and ways to learn from human feedback through all modalities.
- New mediums to interact with AI. New tasks beyond imitating humans, leading to collaboration.

Potential models and dataset to start with

- MMHal-Bench: https://arxiv.org/pdf/2309.14525.pdf aligning multimodal LLMs
- HACL: https://arxiv.org/pdf/2312.06968.pdf hallucination + LLM

multisensory intelligence

Research Projects on Ethics and Safety

Motivation: Large AI models are can emit unsafe text content, generate or retrieve biased images.

Challenges:

- Taxonomizing types of biases: text, vision, audio, generation, etc.
- Tracing biases to pretraining data, seeing how bias can be amplified during training, fine-tuning.
- New ways of mitigating biases and aligning to human preferences.

Potential models and dataset to start with

- Many works on fairness in LLMs -> how to extend to multimodal?
- Mitigating bias in text generation, image-captioning, image generation

How to do Literature Review and Read a Paper

1. Google scholar
2. Papers with code, Github, Huggingface
3. Recent conference proceedings
4. Blog posts
5. Survey papers, tutorials, courses

Testing Research Ideas

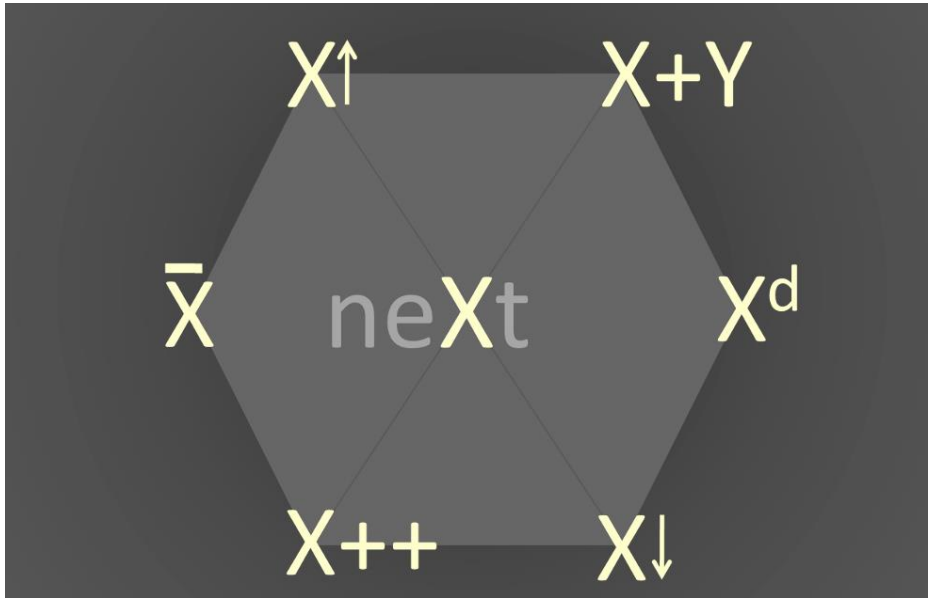
1. Gather and process dataset, visualize data, gather labels, do data splits.
2. Implement the simplest pipeline and get it working.

-> Pipeline = data loading + basic model + eval function + loss/visualization/deployment

3. Change one component of the model at a time, repeat x10 (top-down or bottom-up).
4. Find what works best, and exploit.
5. Scale up experiments, repeat across multiple datasets.
6. Careful ablation studies.
7. Qualitative comparisons and visualizations.
8. Repeat until successful.

More resources

https://github.com/pliang279/awesome-phd-advice
https://github.com/jbhuang0604/awesome-tips
https://www.cs197.seas.harvard.edu/
https://medium.com/spotprobe/the-hexagon-of-ideas-02e5b770d75e



multisensory intelligence

Available Tools

- Use available tools in your research groups
  - Or pair up with someone that has access to them
- Find some GPUs!
- We will be getting AWS credit for some extra computational power
- Google Cloud Platform credit as well

Some Advice About Multimodal Datasets

- Text, speech, audio, video: Space will become an issue working with image/video data. Some datasets are in 100s of GB (compressed)
- Memory for processing it will become an issue as well, won't be able to store it all in memory
- Time to extract features and train algorithms will also become an issue
- Plan accordingly! Sometimes tricky to experiment on a laptop (might need to do it on a subset of data)

Assignments for This Coming Week

Project preference form

- *To help with team matching*
- *Google Form link will be sent out*

Homework 1: Multimodal data processing and visualization.
- Find, process, visualize, and label your unique multimodal dataset of interest.

Next Tuesday: tutorial on **data processing and basic ML coding**