Penn Engineering
University of Pennsylvania

FA26

Special Seminar on
Multi-Modal AI

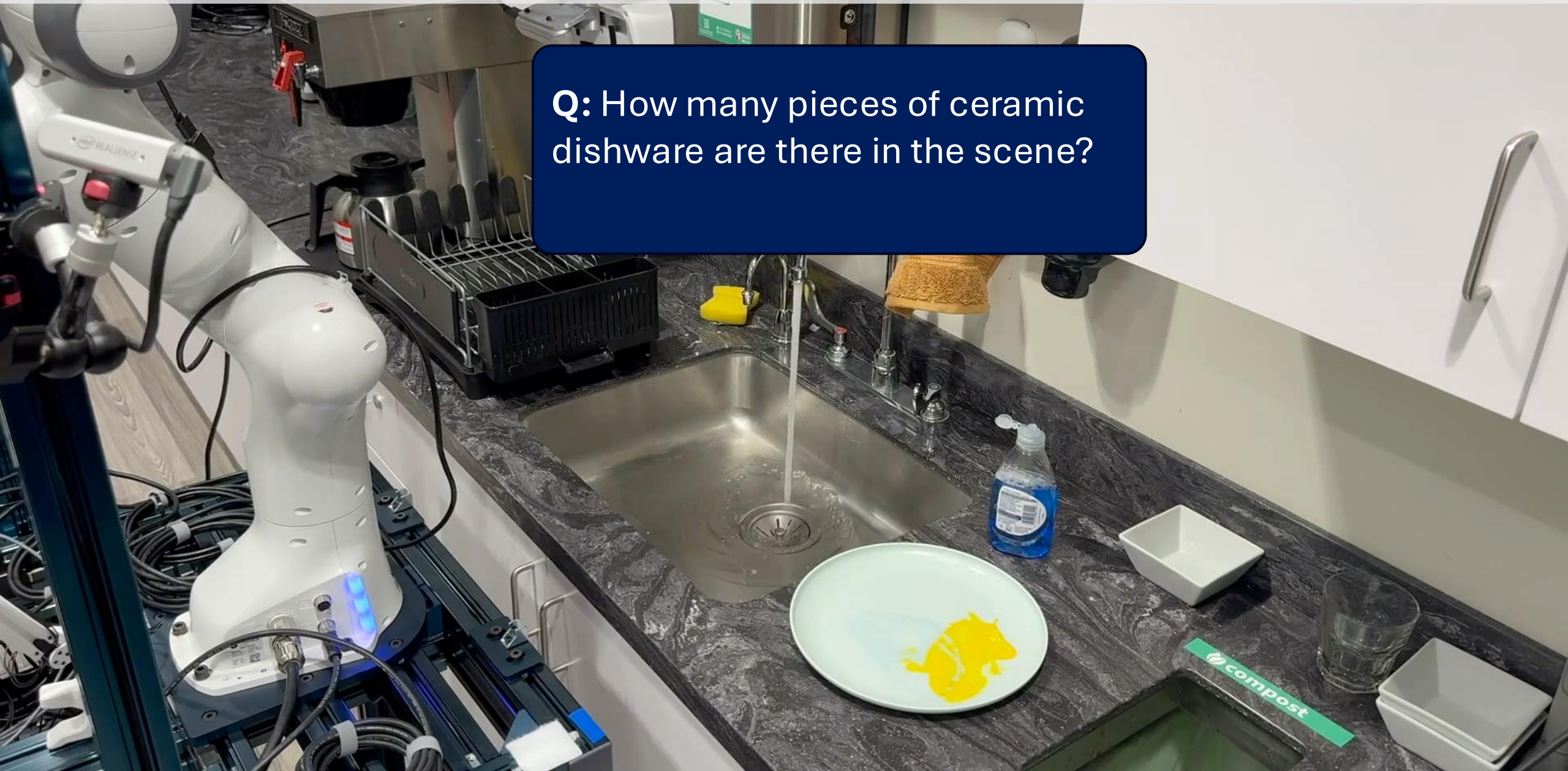
Jiayuan Mao

Department of CIS, GRASP

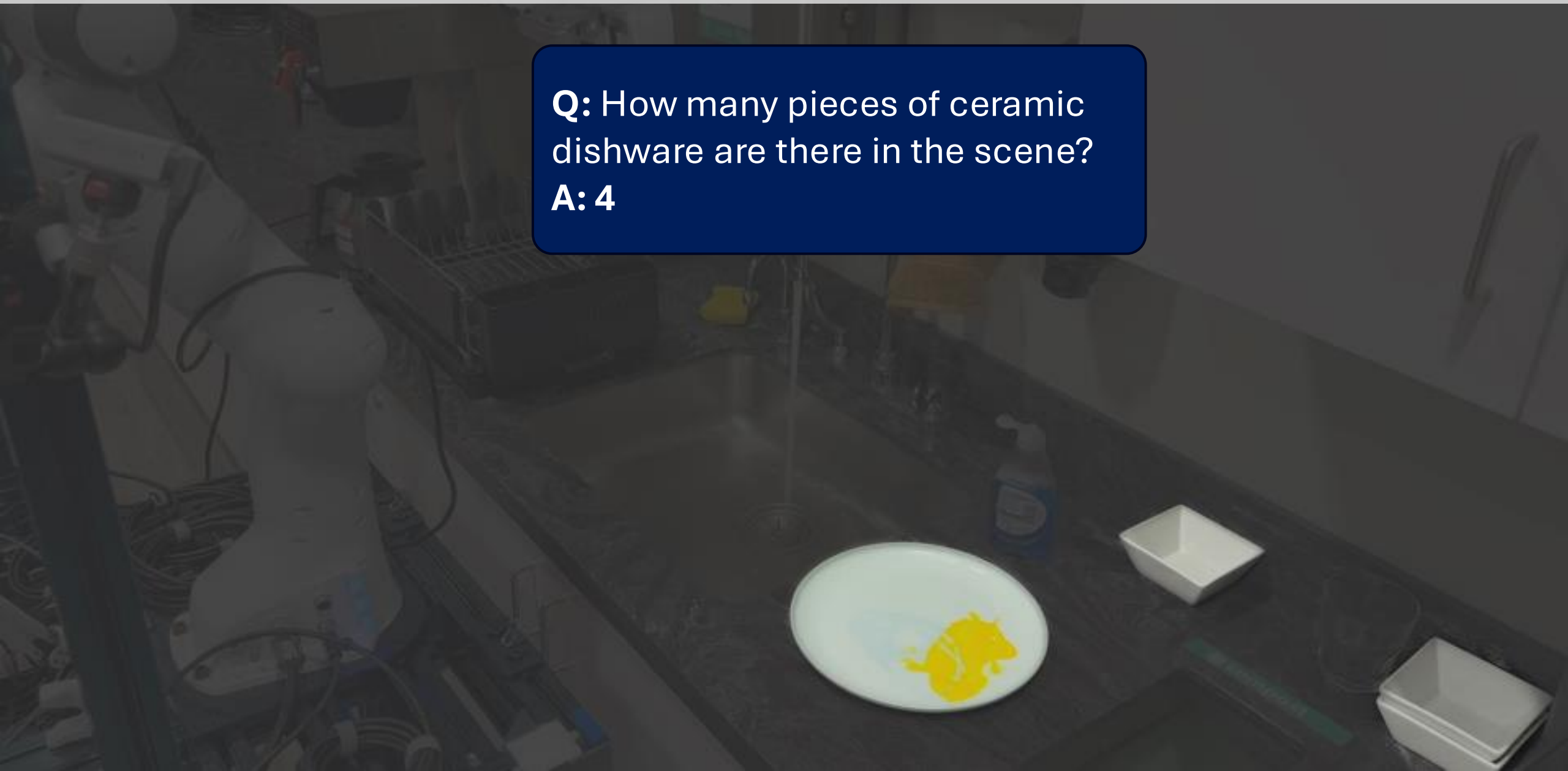
https://cis7000-fall26.thelabone.org/

Generalist Agents Requires Multi-Modal Understanding

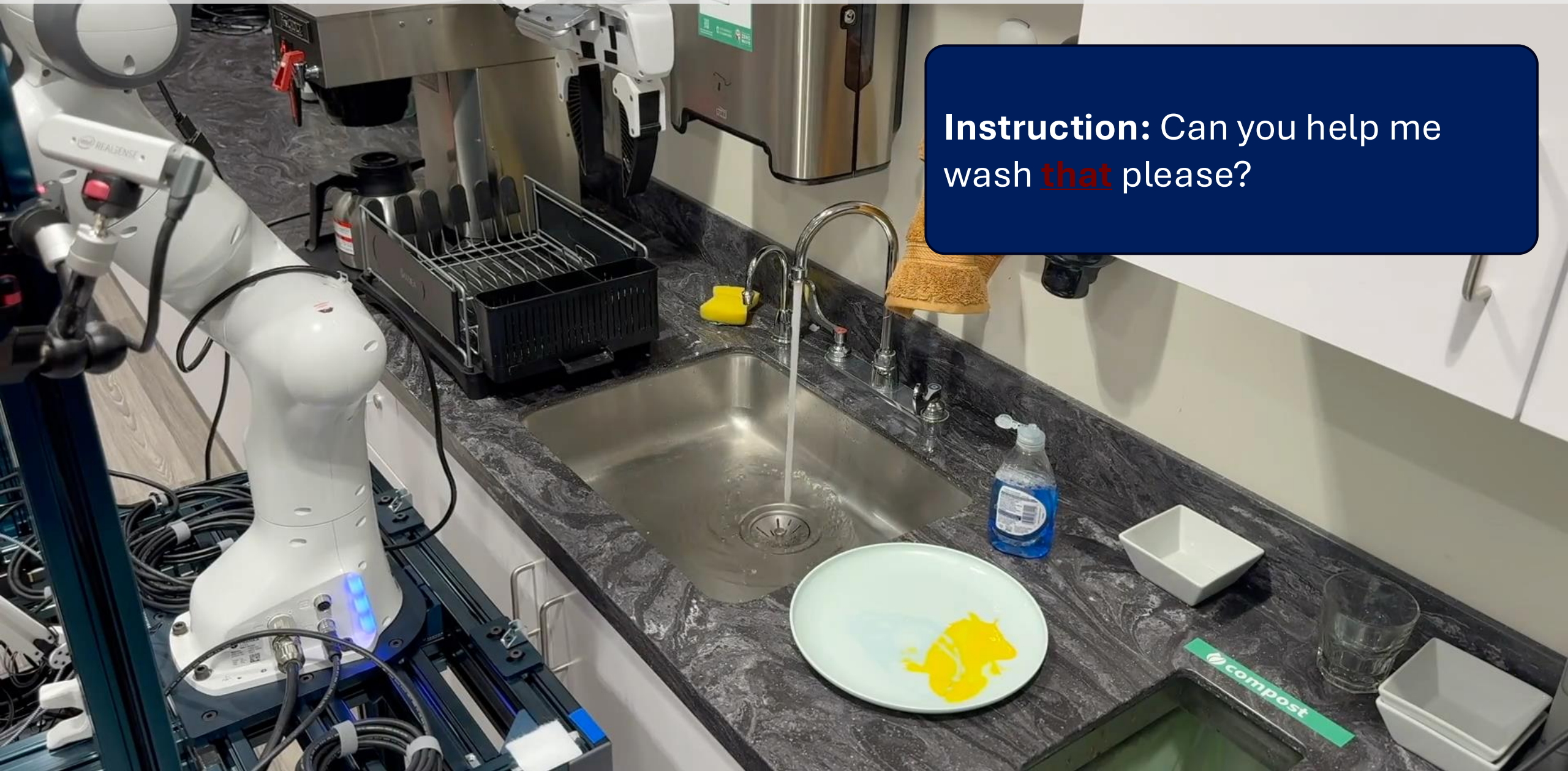
Generalist Agents Requires Multi-Modal Understanding



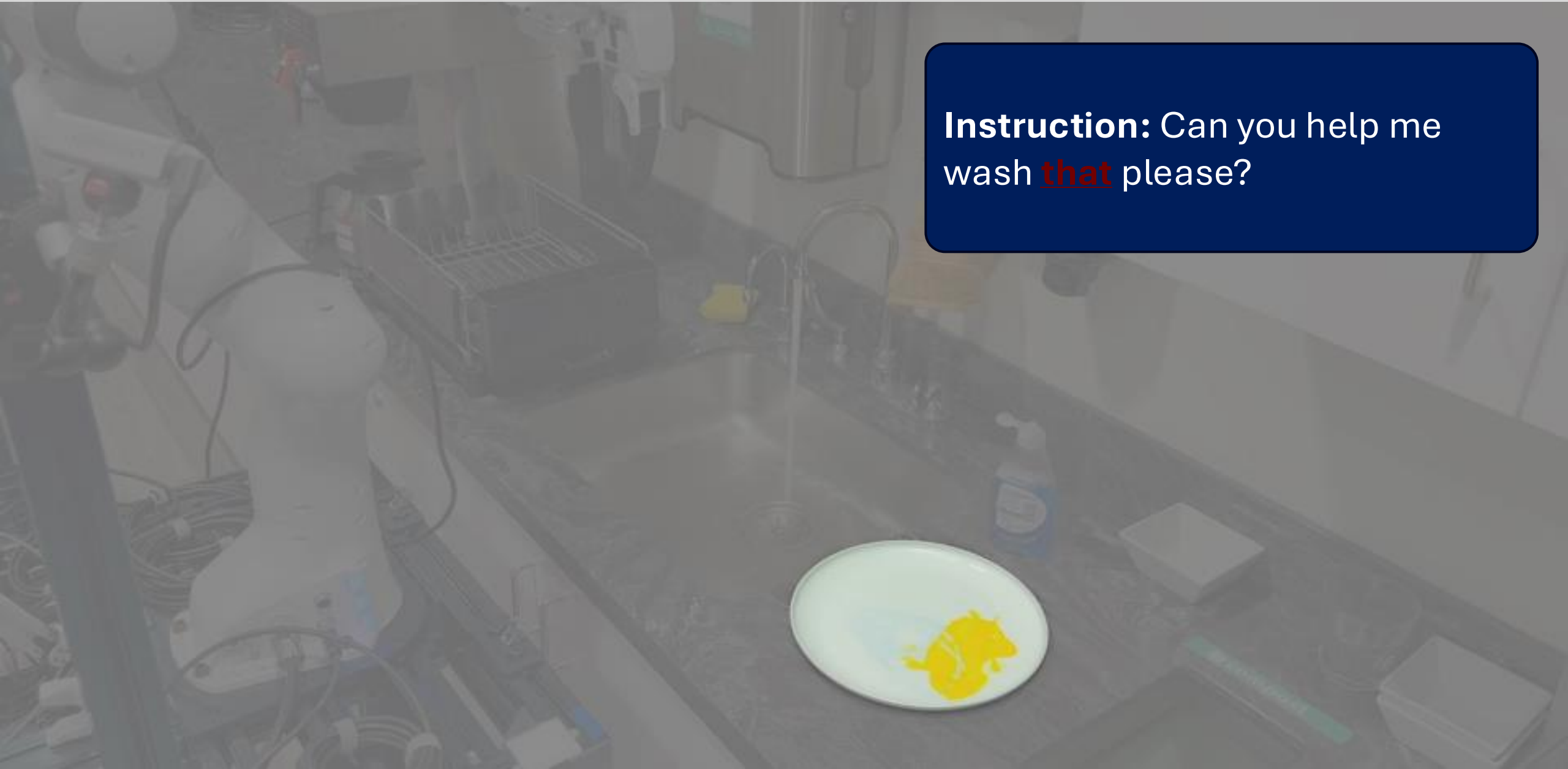
Generalist Agents Requires Multi-Modal Understanding



Generalist Agents Requires Multi-Modal Understanding



Generalist Agents Requires Multi-Modal Understanding



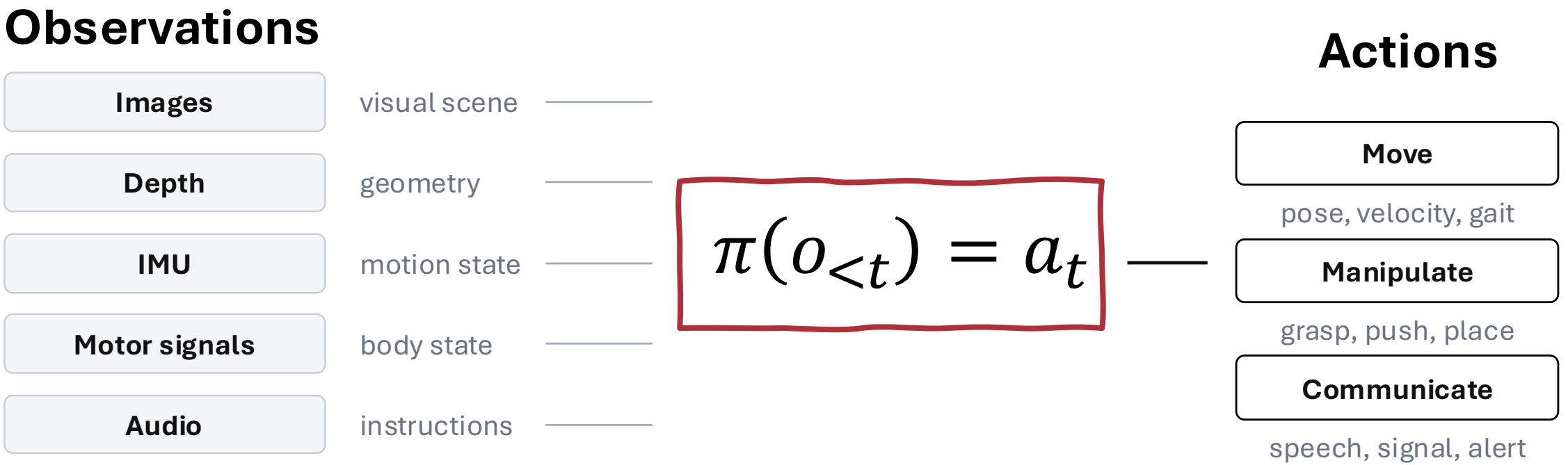
Generalist Agents Requires Multi-Modal Understanding

Goal of This Course

Use Machine Learning Techniques to Build Such Systems

Abstract View of An Agent

$$\pi: \text{observations} \rightarrow \text{action}$$



Notation: t – time. $o_{<t}$ - all observations before t. a_t - action at t.

Goal of This Course

$$\pi: \text{observations} \rightarrow \text{action}$$

- How to formulate a multi-modal AI (learning) problem
- What are the principles for solving these problems
  - with machine learning

Outline for Today

Part 1: Overview of the Course

Part 2: Formulation: Representation and Computation of Multi-Modal AI

Part 3: Recap of Machine Learning and Deep Learning Basics

Part 4: Self-Introductions and Q&A

Outline for Today

Part 1: Overview of the Course

Part 2: Formulation: Representation and Computation of Multi-Modal AI

Part 3: Recap of Machine Learning and Deep Learning Basics

Part 4: Self-Introductions and Q&A

Overview: Course Format

- Enrollment Cap: 20
- Working on extending this to 45

Note: Master's theses, capstone projects are allowed
The project may be a (subset of) a paper for publication, but it should be related to this course

Overview: Course Format

- **Lecture:**
  - MW 3:30 – 5:00 PM. Tonwe 337.
  - Each lecture will be ~1 hours of talk + ~30mins of QA/discussion
  - Come with your own questions and answers
- **Final Project:**
  - Max 2 students per team (individual projects welcome).
  - Week 3: Initial Proposal Due
  - Week 6: Final Proposal Due
  - Week 9: Mid-Project Check-In (Through Office Hours)
  - Week 14: Final Project Presentation

Note: Master's theses, capstone projects are allowed
The project may be a (subset of) a paper for publication, but it should be related to this course

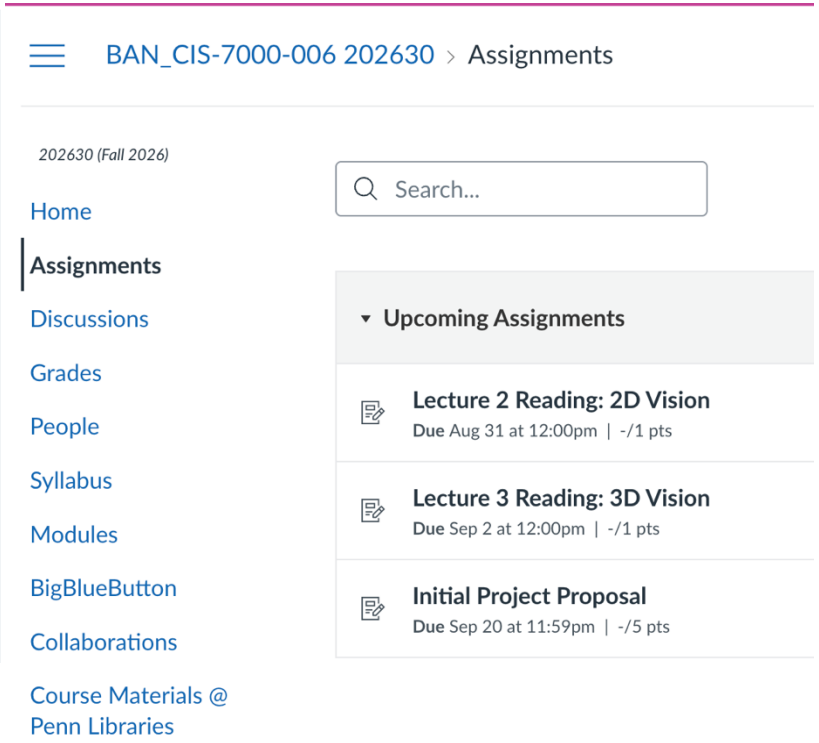
Overview: Lecture Format

- **Part 1: Talk**
- The lecturer or guest lecturer will present a topic, such as 2D vision
- Each lecture will have around four assigned papers
- Please read these papers before class and post your questions online
- If you have a paper that you want to present, contact the lecturer before class

- **Part 2: Discussion**
- Discuss how the topic relates to your research interests (e.g., medical imaging)
- What shared principles and gaps do you see (e.g., GCaMP is event-driven)

Readings and Discussion Questions

Lec#	Date	Topic	Reading
1	**Wed Aug 26**	Course overview and multimodal AI formulation	None
2	**Mon Aug 31**	2D Vision	Fully Convolutional Networks for Semantic Segmentation (2015) Mask R-CNN (2017) Unsupervised Learning of Depth and Ego-Motion from Video (2017) An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale (2021) Optional: Mask DINO: Towards A Unified Transformer-based Framework for Object Detection and Segmentation (2023)
3	**Wed Sep 2**	3D Vision	PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation (2017) VGGT: Visual Geometry Grounded Transformer (2025) 3D Dynamic Scene Graphs: Actionable Spatial Perception with Places, Objects, and Humans (2020) Where2Act: From Pixels to Actions for Articulated 3D Objects (2021) Optional: SayPlan: Grounding Large Language Models using 3D Scene Graphs for Scalable Robot Task Planning (2023)



Overview: Prerequisites

- **Prerequisites:**
  - Basic machine learning and deep learning concepts:
    - Classification, Regression, Gaussian Distributions, CNN, Transformer
    - Optional: PGMs, FCNs, LSTMs, Diffusion Models, Policy Gradient
  - Python experience; ability to run and debug ML experiments
  - Comfort reading deep learning (e.g., new model architecture) and machine learning papers (e.g., new probabilistic models)

Note: Talk to the instructor if you are unsure about preparation

Overview: Final Project

- **Topic**
- Any topic that is related to multi-modal AI
- The topic may be connected to your own research field, but it does not have to be
- If you have questions about the scope or topic, feel free to chat with me

- **Check-ins**
- The initial and final project submissions should both be submitted through Canvas
- I will provide feedback. Optionally, we can also schedule office hours

- **Final Submission**
- A presentation in Week 14 and a written report of at least 4 pages

Overview: Assessment

Component	Weight	Description
Participation	25%	Active engagement in lectures, discussions, and peer feedback
Reading Responses	20%	Timely and thoughtful responses to assigned readings, in particular, discussion question posts
Project proposal	10%	Problem formulation, related work, proposed methodology, and evaluation
Final project	35%	Research execution, experiments, analysis, and final report
Final presentation	10%	Clear communication of the project motivation, method, results, and limitations

Overview: Topics

- Understand the major representation and computation paradigms for different representations, individually
- Explore basic ideas about how modalities are aligned through contrastive learning, cross-attention, generative objectives, and multi-task learning

Week	Theme	Topics
1	Foundations	Overview of Multimodal AI; representation, alignment, and fusion
2	2D/3D Vision and Perception	Visual representations; 3D and spatial perception
3	Vision and Text	Language foundations; vision-language learning
4	Text-to-Image	Generative vision; controllable and structured generation
5	Audio	Speech/audio representation; audio-language and generation
6	Video	Video understanding; video generation and prediction

https://cis7000-fall26.thelabone.org/

Overview: Topics

- Derive shared principles for processing data of different modalities, and learn to build "unified" models with tokenization

Week	Theme	Topics
7	Tokenization	Tokenization across modalities; discrete vs. continuous representations
8	Multimodal Architectures	Connecting modalities to LLMs; fusion architectures; unified multimodal models
9	Large Multimodal Models	Scaling, multimodal data, pretraining, post-training, and evaluation

https://cis7000-fall26.thelabone.org/

Overview: Topics

- Generalize existing principles and approaches for many more domains
- Understand advanced topics in grounding, causal discovery, and world modeling in multimodal systems
- Identify broader research opportunities in privacy, safety, alignment, HCI, science, and beyond

Week	Theme	Topics
10	Structured & Scientific Domains	Charts, tables, and scientific modalities
11	Tactile and Action	Tactile/sensor learning; action representations; vision-language-action models
12	Structure & Reasoning	Grounding, planning, causal discovery, and world models
13	Advanced Topics	Agents, HCI, privacy, safety, and alignment

https://cis7000-fall26.thelabone.org/

Outline for Today

Part 1: Overview of the Course

Part 2: Formulation: Representation and Computation of Multi-Modal AI

Part 3: Recap of Machine Learning and Deep Learning Basics

Part 4: Self-Introductions and Q&A

Let’s Formalize Everything

- Input-Output specifications:
- Input modalities: images, text, audio, video, depth, proprioception, actions, etc.
- Output modalities: labels, captions, controls, retrieved items, generated media

- Training signal:
- Supervised learning, unsupervised learning, reinforcement learning

- Evaluation:
- What counts as success for the intended behavior

Let’s Formalize Everything

Component	Notation	Question to Ask
Observations	$x = x^1, \ldots, x^M$	Which modalities are observed, missing, noisy, or asynchronous?
Representation	$z = f_\theta(x)$	What information should be preserved, compressed, or made invariant?
Computation	$y = g_\phi(z)$	What decision, prediction, generation, or action is produced?
Learning Objective	$\text{minimize } L(y, y^*)$	Where does supervision or feedback come from?
Evaluation	$metric(\text{task}, \text{behavior})$	Does success require accuracy, grounding, robustness, or usefulness?

Visual Modality

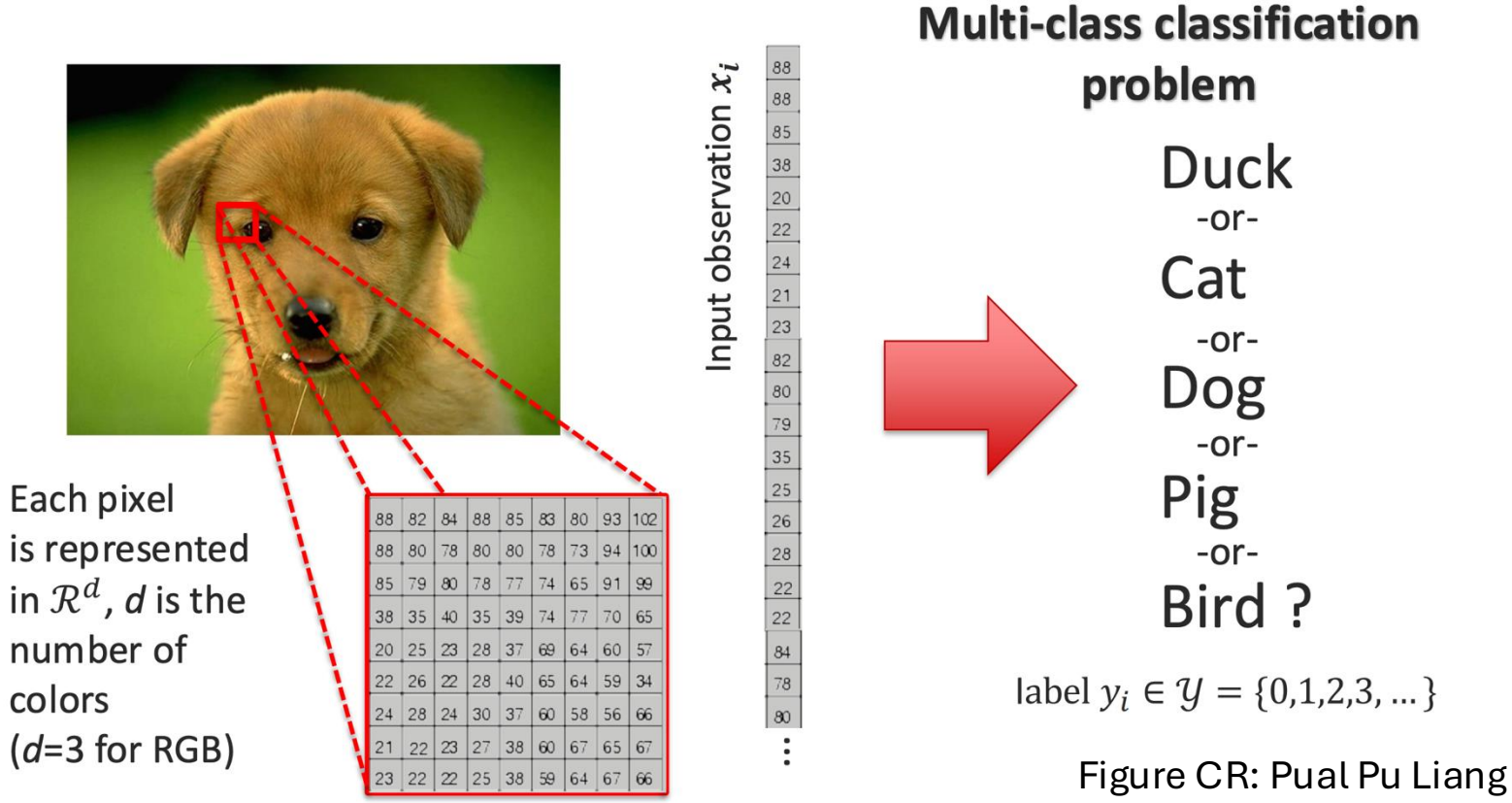


Figure CR: Pual Pu Liang

Component	Realization
Observations	x^{V}
Representation	An array z of shape [256, 256, 3]
Computation	$y = g_\phi(z)$ of shape [1000]
Objective	$\mathrm{minimize\ Cross-Entropy}(y, y^*)$
Evaluation	$\mathrm{Accuracy}(\mathrm{argmax}\, y, y^*)$

Text Modality

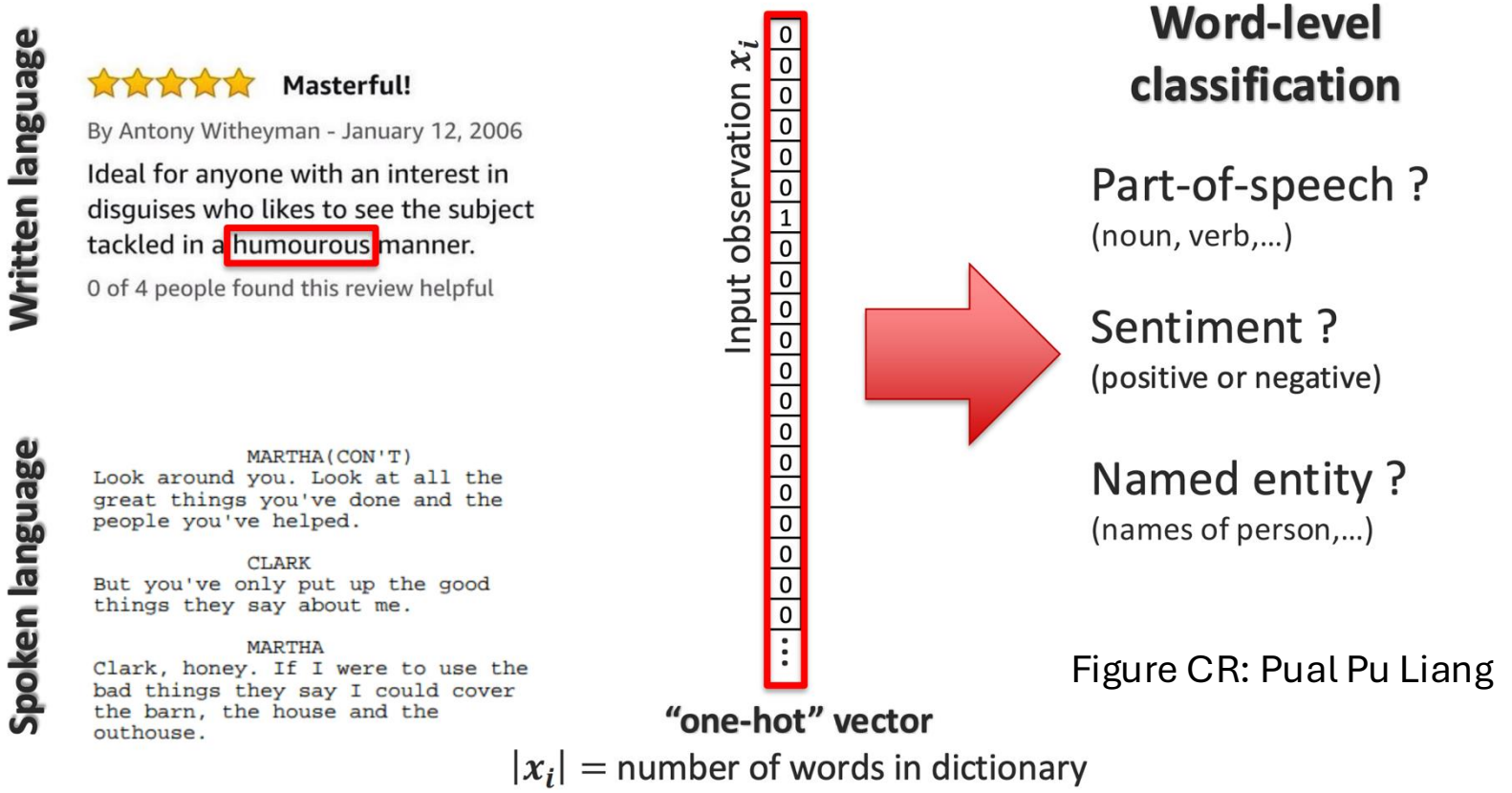
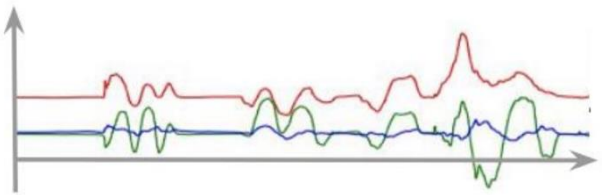


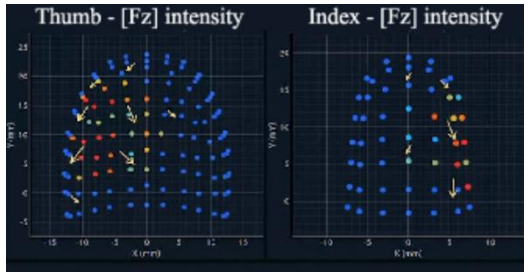
Figure CR: Pual Pu Liang

Component	Realization
Observations	x^{T}
Representation	An array z of shape [seq_len, vocab_size]
Computation	$y = g_\phi(z)$ of shape [2] (e.g., Sentimant)
Objective	$\text{minimize Cross} - \text{Entropy}(y, y^*)$
Evaluation	$\text{Accuracy}(\text{argmax}\, y, y^*)$

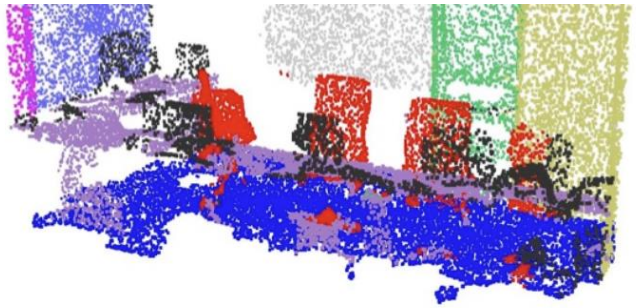
There Are Many Other Modalities



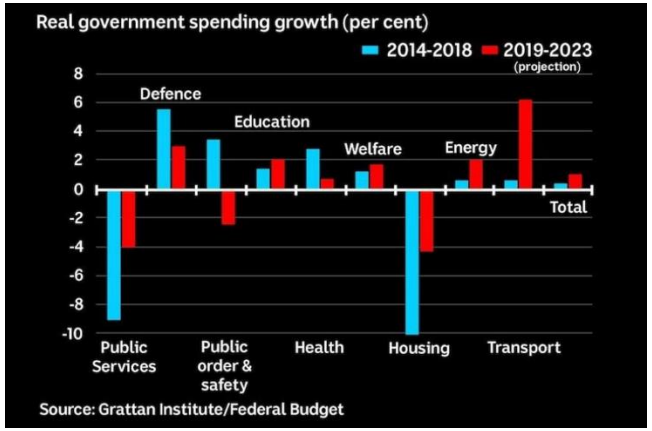
Force-Torque sensors



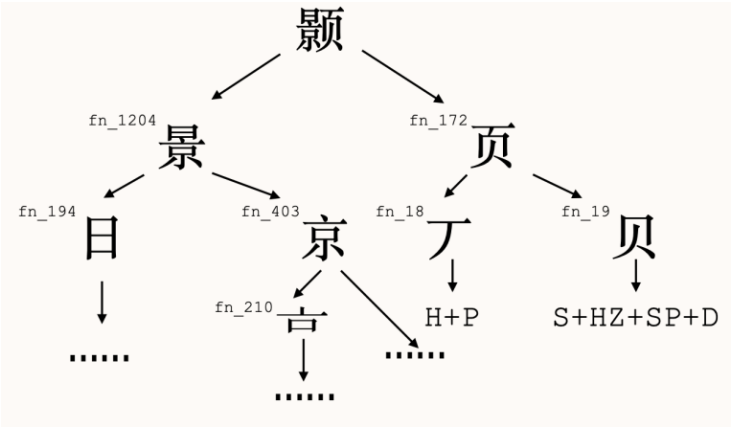
Tactile sensors



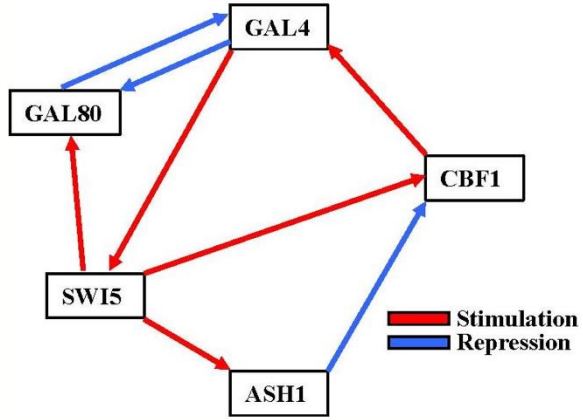
Point Clouds



Charts and Figures



Logographic Writing

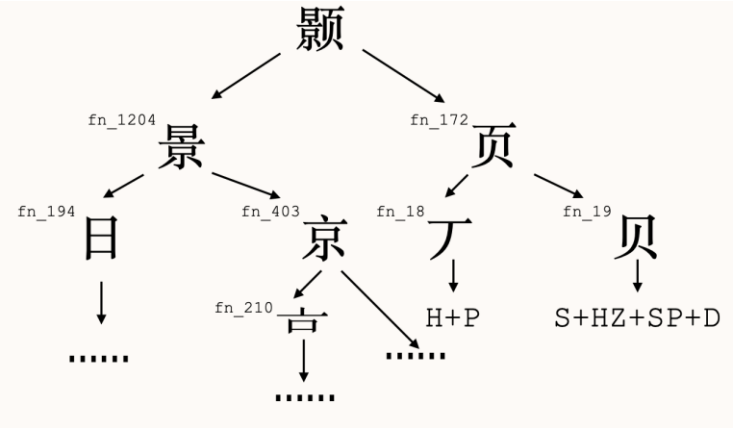


Gene Networks

Figure CR: https://arxiv.org/abs/2504.05506, https://arxiv.org/abs/2607.14487, https://arxiv.org/abs/1612.00593, https://pubmed.ncbi.nlm.nih.gov/19327819/

What Makes a “New” Modality?

顥



- For example, can we just treat each Chinese character as an image?
- Yes, we can.

- A modality is characterized by our representation and computation
- And different choices come with trade-offs

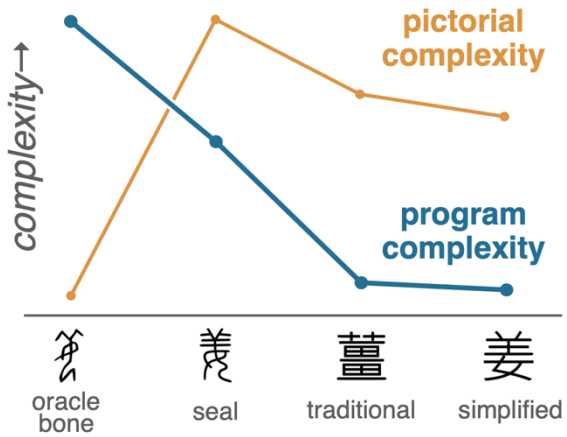
Representation Choices Trade Off Detail and Abstraction

Representation	What It Preserves	Typical Computation
Pixels / Waveforms	Local signal detail	Convolution, filtering
Tokens	Discrete unit	Transformer attention and sequence modeling
Embeddings	Semantic similarity	Retrieval, classification, alignment
Structured States	Entities and relations	Arithmetic computation, logic reasoning

Representation Choices Trade Off Detail and Abstraction

顥

Representation	What It Preserves	Advantage and Limitations
Pixels / Waveforms	Local signal detail	Can measure visual similarities with other characters No semantic similarity
Tokens	Discrete unit	Does not preserve any visual or semantic similarity Can apply sequence modeling tools for language structures
Embeddings	Semantic similarity	Can compute semantic similarities, such as $\langle emb(顥), emb(天) \rangle$
Structured States	Entities and relations	Can analyze radical structures and structured complexity



Outline for Today

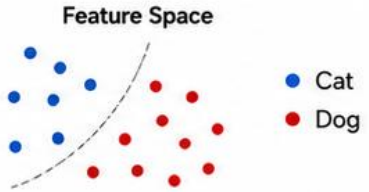
Part 1: Overview of the Course

Part 2: Formulation: Representation and Computation of Multi-Modal AI

Part 3: Recap of Machine Learning and Deep Learning Basics

Part 4: Self-Introductions and Q&A

Learning Objectives Shape the Representation

Objective	Signal	Example Illustration	Advantages
Predictive	Labels	 Model (Classifier) Cat 0.95 Dog 0.05 Ground-truth label: Cat 	• Task-specific features • Clear decision boundaries

Learning Objectives Shape the Representation

Objective	Signal	Example Illustration	Advantages
Predictive	Labels	 Model (Classifier) → Cat 0.95, Dog 0.05. Ground-truth label: Cat. Feature Space: Cat, Dog	• Task-specific features • Clear decision boundaries
Contrastive	Paired modalities (image ↔ T)	 Image Encoder, Text Encoder. Text candidates: A dog catching a frisbee ✓, A red sports car ✗, A bowl of ramen ✗. Shared Embedding Space: Image (dog+frisbee), "A dog catching a frisbee", Pull together, Push apart, "A red sports car", "A bowl of ramen"	• Shared embedding spaces for retrieval and matching • Better generalization across modalities

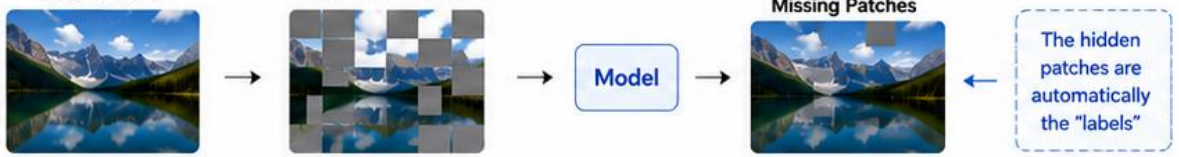
Learning Objectives Shape the Representation

Objective	Signal	Example Illustration	Advantages
Predictive	Labels	 → Model (Classifier) → Cat 0.95, Dog 0.05 Ground-truth label: Cat Feature Space: Cat, Dog	• Task-specific features • Clear decision boundaries
Contrastive	Paired modalities	 Image Encoder, Text Encoder Text candidates: A dog catching a frisbee ✓, A red sports car ✗, A bowl of ramen ✗ Shared Embedding Space: Image (dog+frisbee), "A dog catching a frisbee", Pull together, Push apart, "A red sports car", "A bowl of ramen"	• Shared embedding spaces for retrieval and matching • Better generalization across modalities
Generative	Reconstruct or predict data	"A red robot holding a blue cup" (text prompt) → Generative Model (Diffusion) →  Data Distribution Learned Model learns $p(data \mid text)$	• Rich probabilistic modeling • Generates entire data (e.g., images, text)

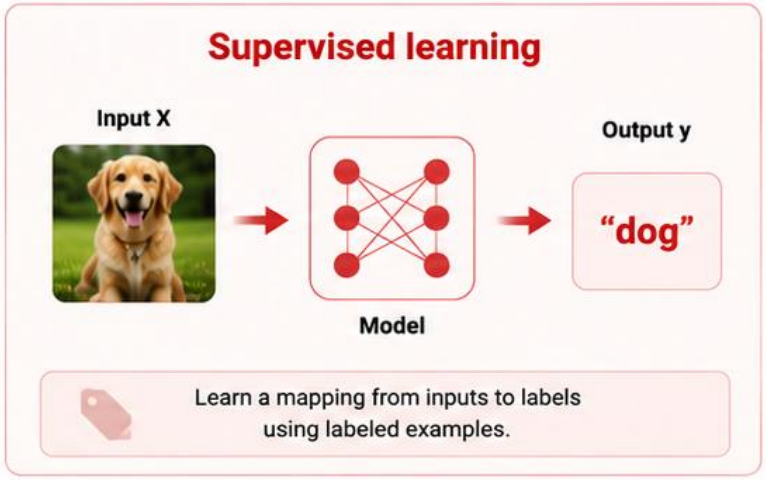
Learning Objectives Shape the Representation

Objective	Signal	Example Illustration	Advantages
Predictive	Labels	 Model (Classifier) → Cat 0.95, Dog 0.05. Ground-truth label: Cat. Feature Space: Cat, Dog	• Task-specific features • Clear decision boundaries
Contrastive	Paired modalities	 Image Encoder, Text Encoder. Text candidates: A dog catching a frisbee ✓, A red sports car ✗, A bowl of ramen ✗. Shared Embedding Space: Image (dog+frisbee), "A dog catching a frisbee", Pull together, Push apart, "A red sports car", "A bowl of ramen"	• Shared embedding spaces for retrieval and matching • Better generalization across modalities
Generative	Reconstruct or predict data	"A red robot holding a blue cup" (text prompt) → Generative Model (Diffusion) →  Data Distribution Learned: Model learns $p(data \mid text)$	• Rich probabilistic modeling • Generates entire data (e.g., images, text)
Reinforcement Learning	Reward of actions	Observe (State) → Choose Action → Environment (Next State) ⋯ Agent interacts with environment. Reward: Success +1, Failure: -1	• Does not need step-by-step labels • Learns from interaction and feedback

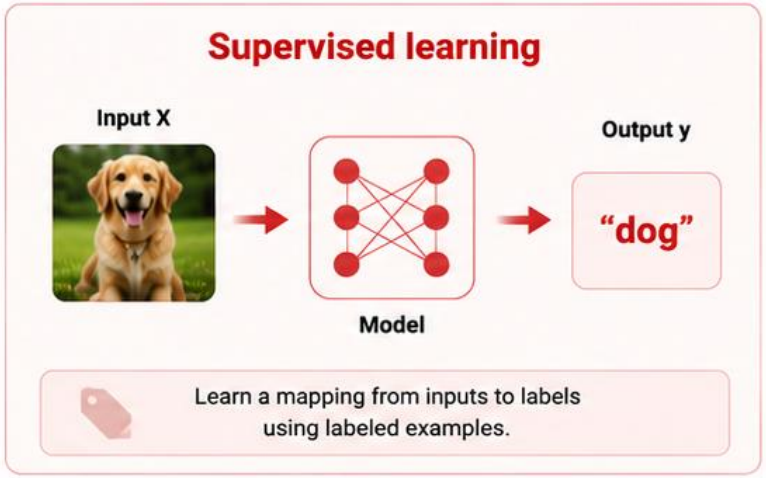
Learning Objectives Shape the Representation

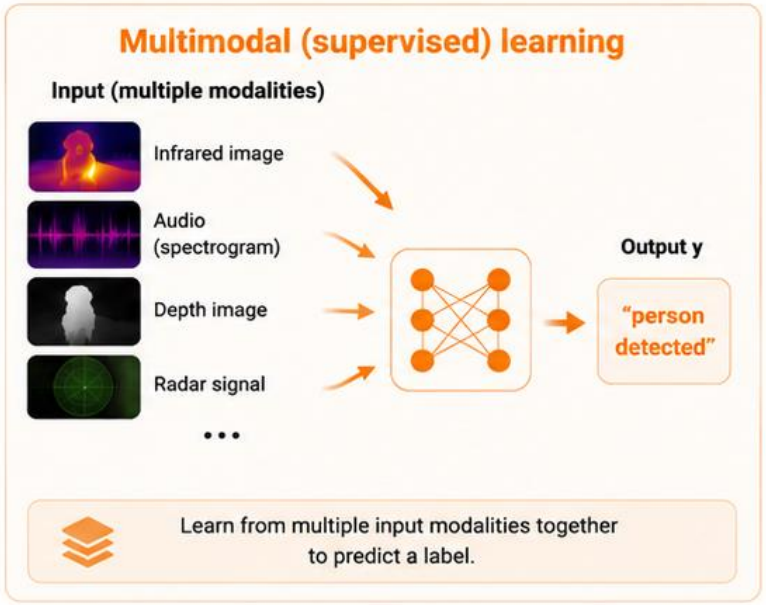
Objective	Signal	Example Illustration	Advantages
Predictive	Labels	 → Model (Classifier) → Cat 0.95, Dog 0.05 Ground-truth label: Cat Feature Space: Cat, Dog	• Task-specific features • Clear decision boundaries
Contrastive	Paired modalities	 Image Encoder, Text Encoder Text candidates: A dog catching a frisbee ✓, A red sports car ✗, A bowl of ramen ✗ Shared Embedding Space: Image (dog+frisbee), "A dog catching a frisbee", Pull together, Push apart, "A red sports car", "A bowl of ramen"	• Shared embedding spaces for retrieval and matching • Better generalization across modalities
Generative	Reconstruct or predict data	"A red robot holding a blue cup" (text prompt) → Generative Model (Diffusion) →  Data Distribution Learned Model learns $p(data \mid text)$	• Rich probabilistic modeling • Generates entire data (e.g., images, text)
Reinforcement Learning	Reward of actions	Observe (State) → Choose Action → … → Environment (Next State)  Agent interacts with environment Reward: Success +1, Failure: -1	• Does not need step-by-step labels • Learns from interaction and feedback
Self-supervised	Raw data structure	Original Image → Mask 40% of Patches → Model → Reconstruct Missing Patches  The hidden patches are automatically the "labels"	• Scalable pretraining without dense labels • Learns useful representations from raw data

More Learning Problem Formulations

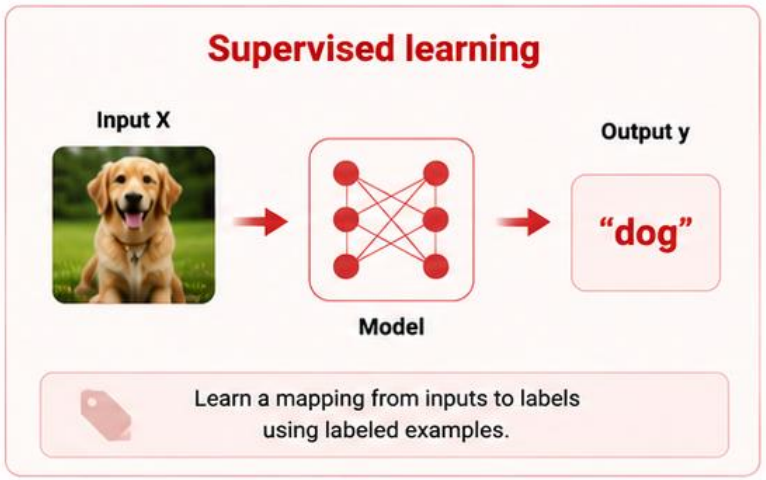


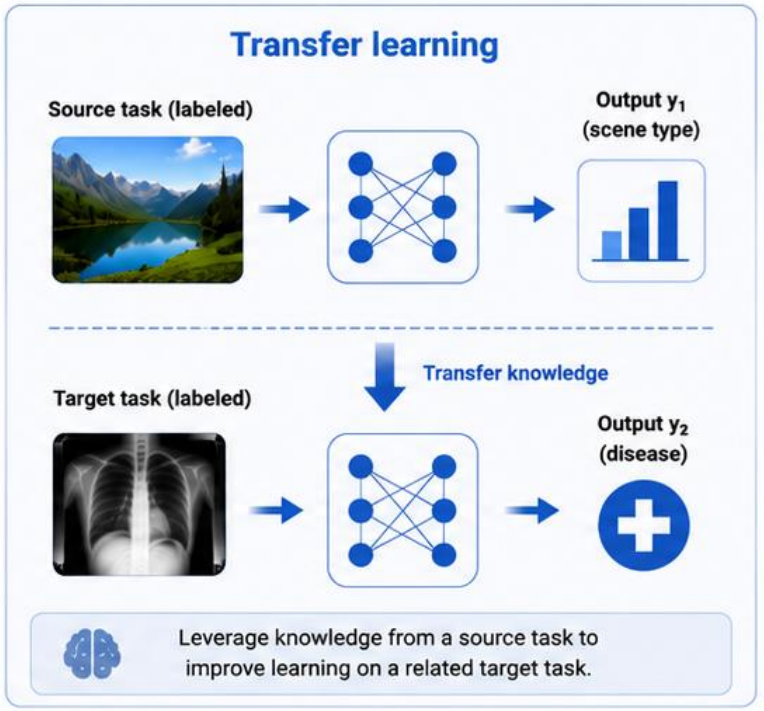
More Learning Problem Formulations

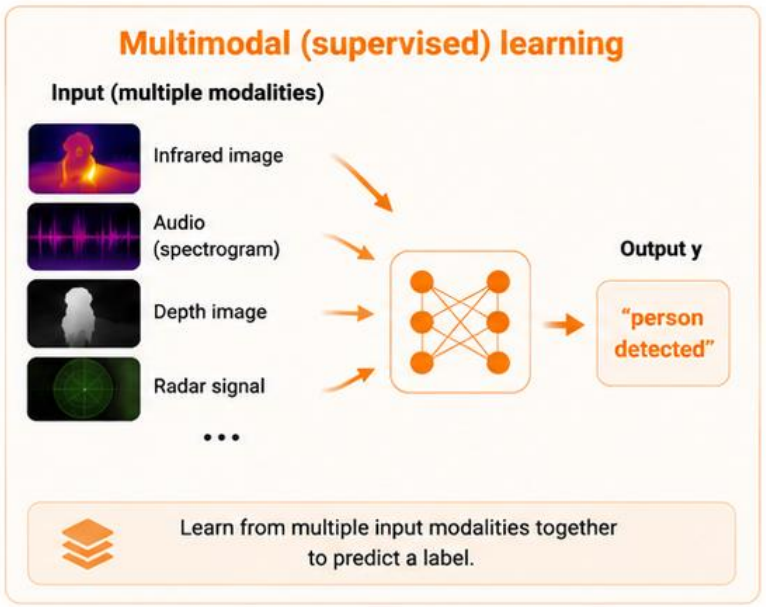




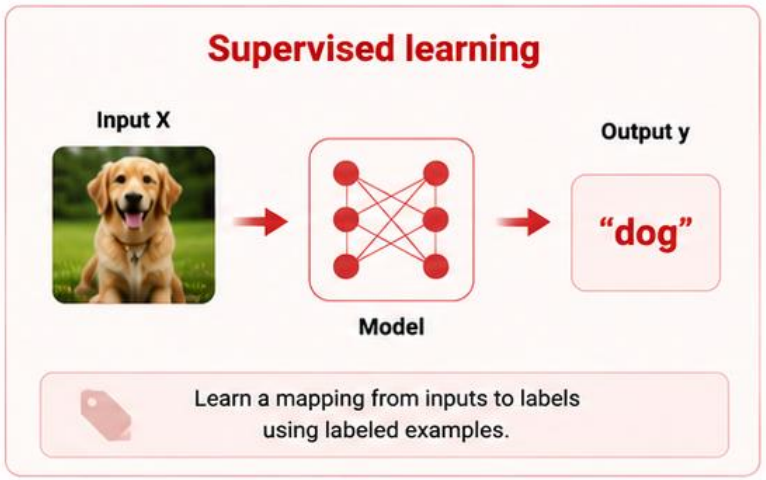
More Learning Problem Formulations

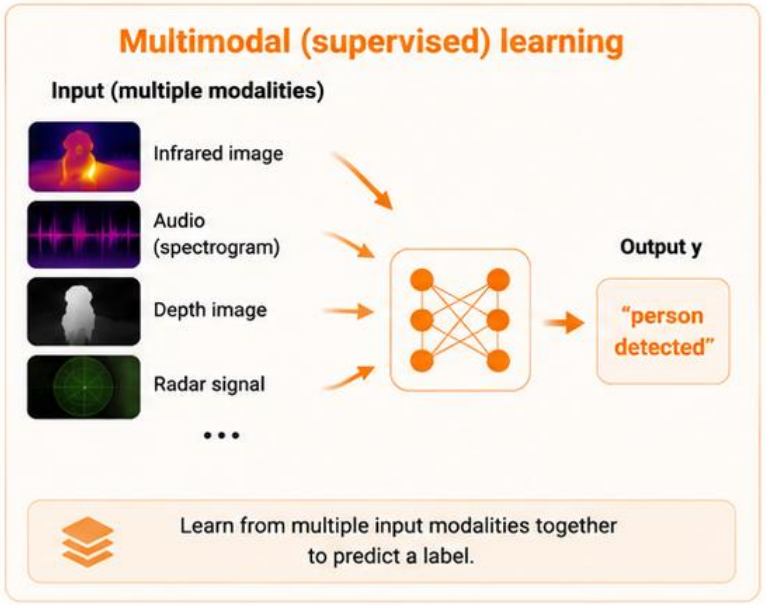


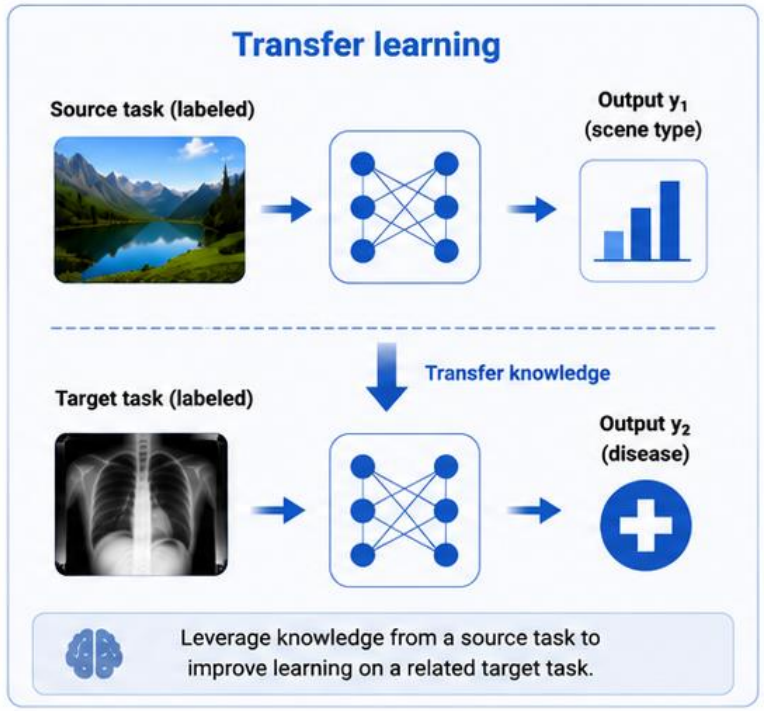


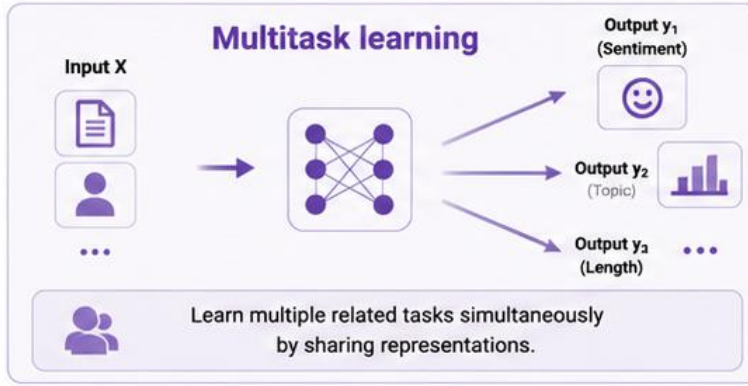


More Learning Problem Formulations

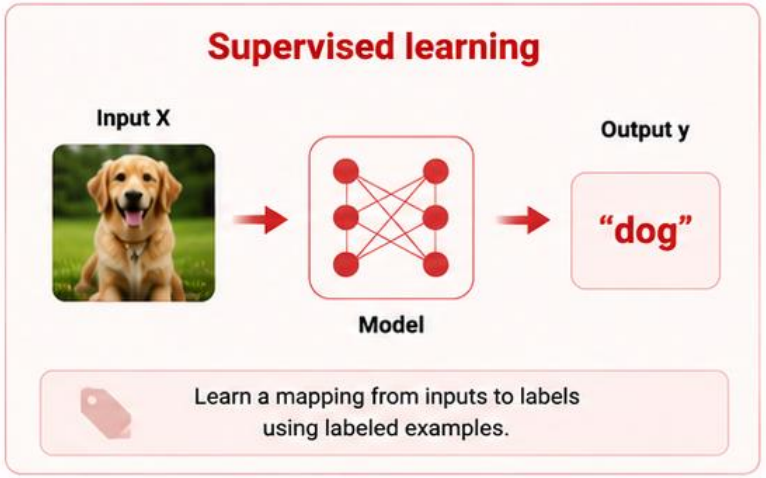


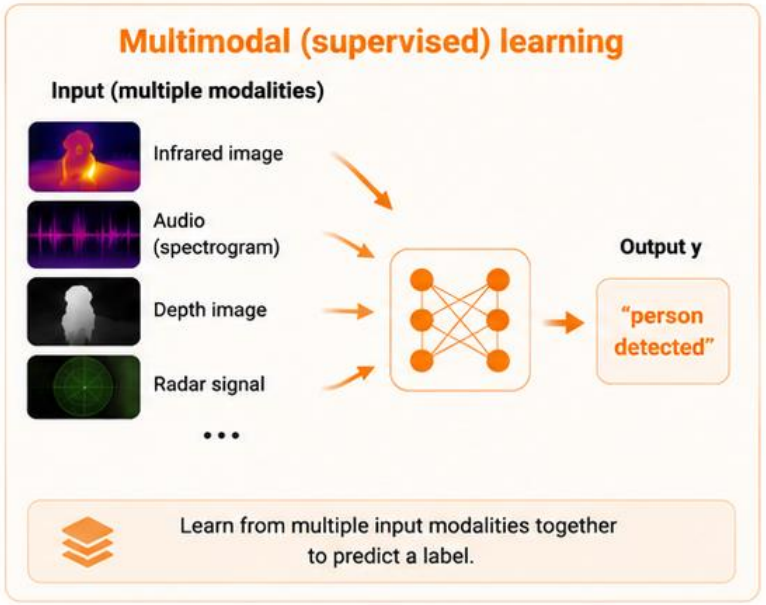


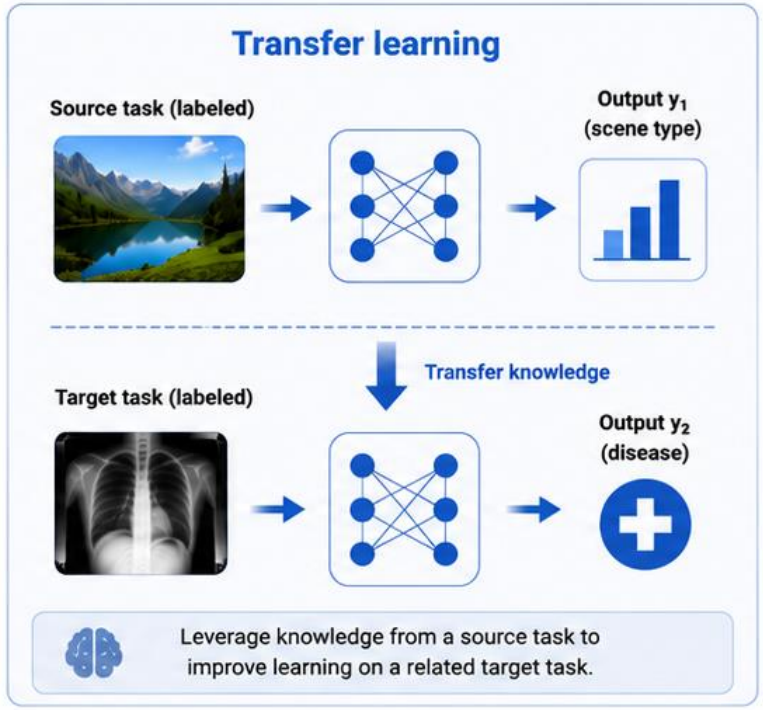


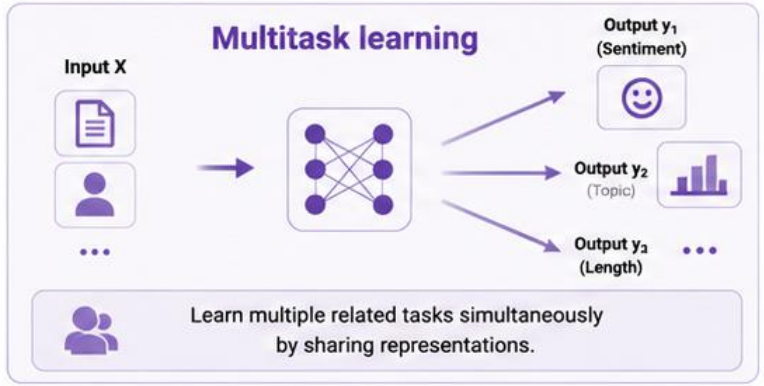


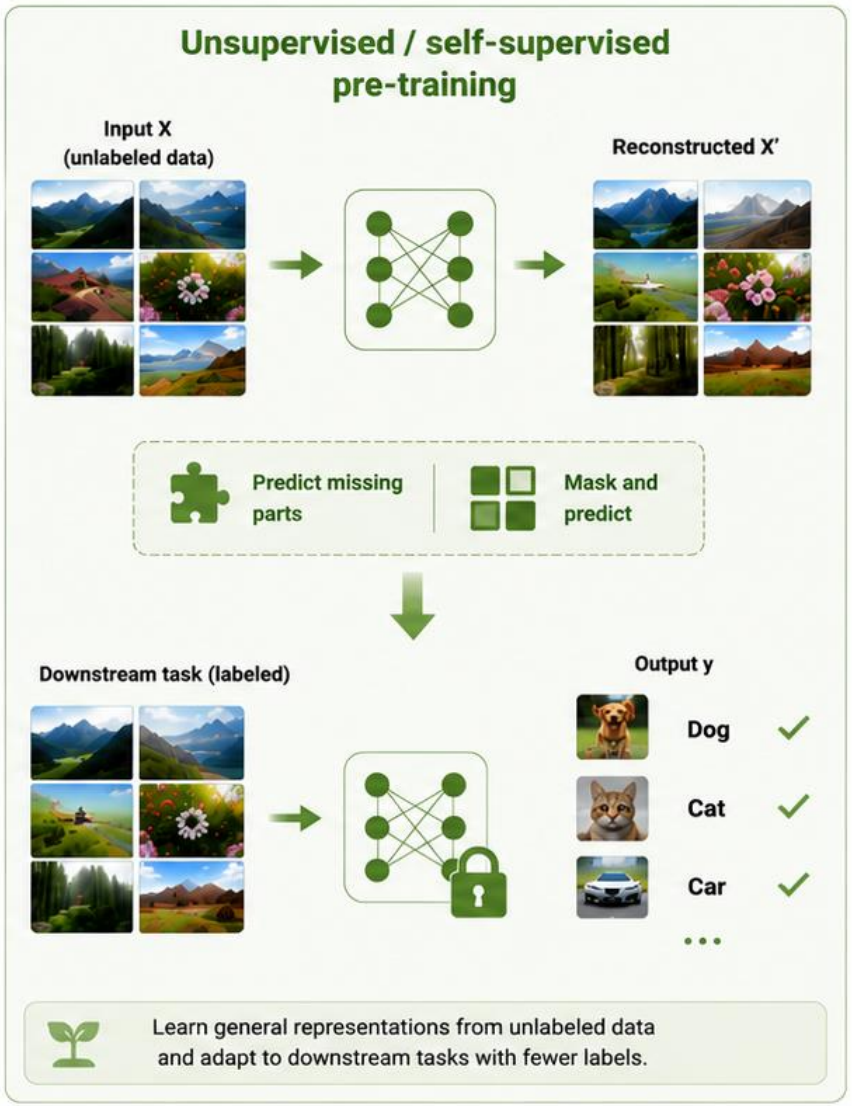
More Learning Problem Formulations











Evaluation is Part of the Formulation

Evaluation Lens	Example Metric	What It Tests
Prediction	Accuracy, F1	Can the model infer the right answer?
Alignment	Retrieval recall, matching score	Do modalities refer to the same concept?
Generation	Human preference, fidelity, diversity	Can the model create useful multimodal outputs?
Robustness	Missing modality, shift, noise tests	Does the system fail gracefully?
Calibration	Calibration error	Do the system's predicted probabilities align with real-world observations?

Recurring Questions in this Class

- How to formulate a multi-modal AI problem
  - What's the inputs, outputs, objectives, and evaluation metrics?
- What are the principles for solving these problems
  - What's the principles, advantages, and disadvantages of the representation and the computation models?

Outline for Today

Part 1: Overview of the Course

Part 2: Formulation: Representation and Computation of Multi-Modal AI

Part 3: Recap of Machine Learning and Deep Learning Basics

Part 4: Self-Introductions and Q&A

Self-Introductions!

- Your name, department, research interest
- Make a name tag and bring to future classes
  - Fold a sheet of paper into thirds
  - Write your **name in large letters** on the front

3. **Get your coloured card and fold it about a centimeter at the bottom.**

4. **Fold the whole piece of card into 3 parts.**

5. **The last quarter, put in front of the cm fold and then write your name on the front.** Stand it up and there you have it.