COMP 6611C: Advanced Topics in Embedded AI Systems

Xiaomin Ouyang

Assistant Professor

Department of Computer Science and Engineering

Hong Kong University of Science and Technology

香港科技大學
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Outline

- **Course logistics**
- Embedded AI overview and examples

Course Arrangement

- **Instructor: Xiaomin Ouyang**
  - Email: xmouyang@cse.ust.hk (with subject [COMP6611C]))
  - Lecture: Tuesday and Thursday, 4:30pm - 5:50pm; Room 2304, Lift 17/18
  - Office Hour: Friday 4:30pm - 5:30pm, Room 3562, Lift 27/28
- **Teaching Assistant:**
  - Mr. Runxi Huang (runxi.huang@connect.ust.hk)
  - Mr. Wenjie Du (wduaj@cse.ust.hk)
- **Course Website:**
  - https://xmouyang.github.io/teaching/comp6611c-2025-spring/ : for updating slides and schedule
  - https://canvas.ust.hk/courses/: for updating slides/schedule and submitting reports/reviewers/slides

Prerequisites & Enrollment

- This is COMP-6XXX course, which means it is a graduate level course
  - UG student can enroll upon approval
- All enrolled students must have basic understanding of machine learning, and coding experience with python and pytorch.
- Familiarity, experience or interest with the fundamentals of embedded or mobile systems are preferred.

Course Objectives

- Understand challenges in embedded AI systems
- Gain hands-on experience implementing state-of-the-art algorithms
- Develop critical thinking for solving embedded AI problems

What we will cover

- This course will enable students to have an in-depth understanding of **embedded AI algorithms and their implementation in real systems and applications**.

  The major topics include:

  - 1) basics on machine learning;
  - 2) data and system challenges in embedded AI;
  - 3) AI techniques and their implementation on cutting-edge platforms;
  - 4) real-world applications, such as smart health and smart buildings.

Course Overview

- The course structure will primarily consist of:
  - **Lectures**
    - The instructor will introduce background, challenges, techniques on the above topics
  - **Student paper presentations, reviews and discussions**
    - Students will read and discuss the latest publications in the areas of embedded AI, Internet of Things, mobile systems, and ubiquitous computing
  - **A course project**
    - Students will work on an individual or team **project** (**1-3 students per team**) to build an end-to-end embedded AI system.

Tentative Syllabus

- Lectures
- Paper Presentations
- Project Presentations

Please follow the update in both Canvas and website.

Date	Topics	Note
Feb 4 (Tuesday)	Course Introduction and Overview	
Feb 6 (Thursday)	Machine Learning Basics	
Feb 11 (Tuesday)	Challenges in Embedded AI Systems	
Feb 13 (Thursday)	Challenges in Embedded AI Systems	
Feb 18 (Tuesday)	Unsupervised Learning	
Feb 20 (Thursday)	Unsupervised Learning	Paper Presentation
Feb 25 (Tuesday)	**Project Proposal Presentation**	8min pre+ 2min QA
Feb 27 (Thursday)		Cancelled (Makeup on Mar 8/9)
Mar 4 (Tuesday)	Multimodal Sensing and Learning	
Mar 6 (Thursday)	Multimodal Sensing and Learning	Paper Presentation
Mar 11 (Tuesday)	Federated Learning	
Mar 13 (Thursday)	Federated Learning	Paper Presentation
Mar 18 (Tuesday)	Efficient Deep Learning on the Edge	
Mar 20 (Thursday)	Efficient Deep Learning on the Edge	Paper Presentation
Mar 25 (Tuesday)	**Midterm Project Presentation**	15min pre+ 5min QA
Mar 27 (Thursday)	**Midterm Project Presentation**	15min pre+ 5min QA
April 1 (Tuesday)		No class (Midterm Break)
April 3 (Thursday)		No class (Midterm Break)
April 8 (Tuesday)	LLMs and Foundation Models on the Edge	
April 10 (Thursday)	LLMs and Foundation Models on the Edge	Paper Presentation
April 15 (Tuesday)	LLMs and Foundation Models on the Edge	Paper Presentation
April 17 (Thursday)	Physics-strengthened AI for Sensing Systems	
April 22 (Tuesday)	Physics-strengthened AI for Sensing Systems	Paper Presentation
April 24 (Thursday)	Applications	
April 29 (Tuesday)	Applications	Paper Presentation
May 1 (Thursday)	No class (Public Holiday)	
May 3 (Saturday)	**Final Project Presentation (160 mins)**	15min pre+ 5min QA
May 6 (Tuesday)		Cancelled (Makeup on May 5)
May 8 (Thursday)	**Report Deadline**	

Course Assessment

Attendance and Discussion	**20%**
(Attendance@10 + Questions@10)	
Paper Presentation	**10%**
(1 paper)	
Paper Reviews	**10%**
(5 reviews * 2)	
Course Project	**60%**
(proposal presentation@10+midterm presentation@10+final presentation@20+final report@20)	
Total	**100%**

Project

- Students will work as groups for course projects **(1-3 students, 8-10 groups).**
- Each group will propose their own project topics.

- **Project Proposal presentation @10**
- **Midterm project presentation @10**
- **Final project presentation @20**
- **Final project report @20**

Project Hardware Platform

- A small computer cluster (contact me for accounts)

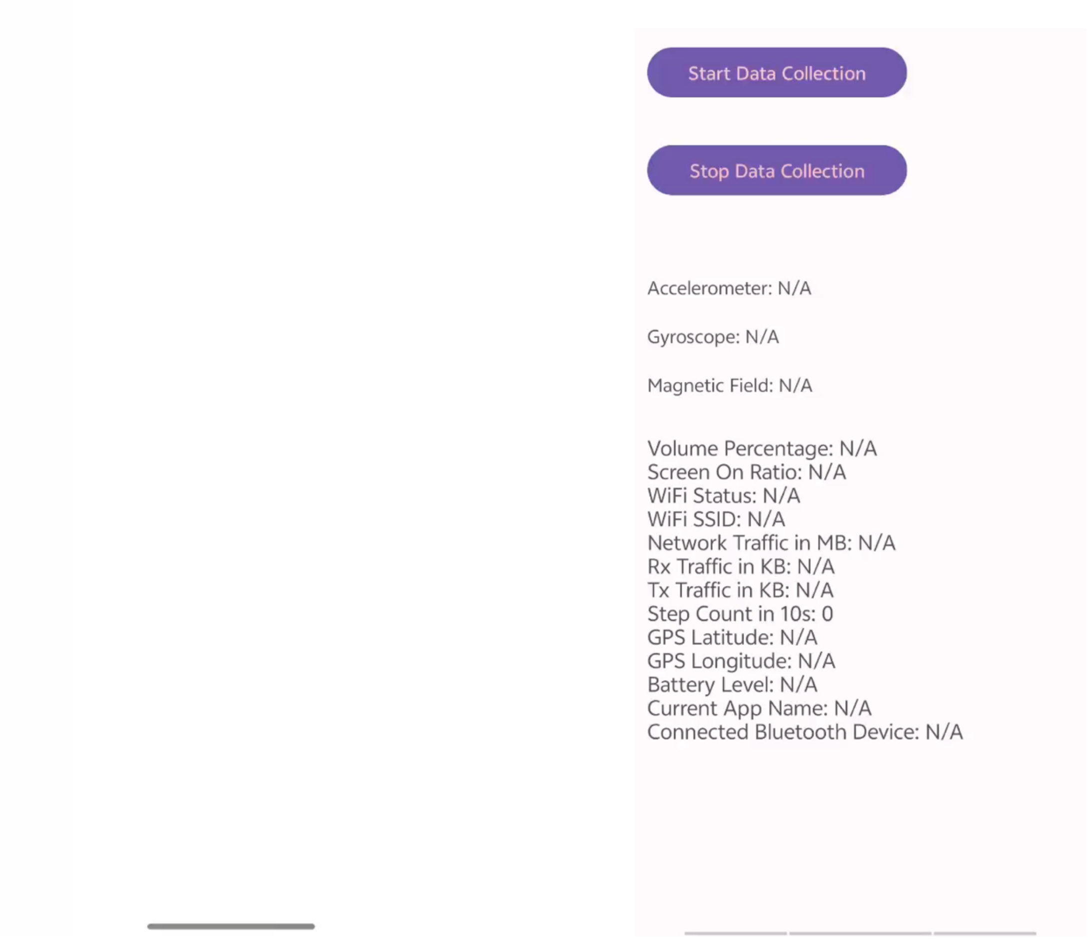
- Edge computer: Raspberry Pi, Nvidia Xavier Nx (limited number available, only on needed basis)

- Your own PC / laptop / Smartphone / Smartwatch

- Sensors: RGBD camera, radar, microphone (limited number available, only on needed basis)

Course Datasets and Challenges

- ➢ MobiBox APP and Dataset



- ➢ Data Collection
  - IMU (Accelerometer / Gyroscope / Magnetic), GPS, Screen & App Usage, Battery Status, Bluetooth connection, Network Traffic, Step Count, Wi-Fi Connections
- ➢ Daily activity summary
- ➢ Bump-up Suggestion

Course Datasets and Challenges

- MobiBox APP and Dataset
  - Step 1 As participants: install the APP for at least one month
  - Step 2 Before project proposal: release some sample data, propose project ideas based on data
    - Missing data, noisy labels, multimodal alignment, real-time inference
  - Step 3 Midterm Project Presentation: preliminary design and evaluations on the data
  - Step 4: Final project: design and evaluations on the full data

Project Ideas

- **Based on the course datasets, or other public datasets, or the system you propose**
  - **Problem-driven projects – Solve a specific challenging problem**
    - Example: Improve inference efficiency of LLMs running on the edge
  - **Sensor-driven projects – Enhance the sensing quality of a specific sensor**
    - Example: Leverage AI to improve the quality of UWB signals under mobility
  - **Application-driven projects – Build a system for specific application**
    - Example: Embedded AI system for breath/occupant/environment monitoring
  - **Measurement-driven projects – Experimental evaluation of a system/network**
    - Example: Performance of ML algorithms for different hardware and tasks

Project Proposal Presentation

- Tentatively schedule: **Feb 25**, **10 mins** for each group (8min pre+ 2min QA)
- May present **more than one topic;** the instructor will work with the group to finalize the choice.
- **The presentation should include**
  - Motivation (why do you want to work on this topic? Why it is important?)
  - Literature study (what has been done on this topic? Why can’t the existing solutions work?)
  - Technical approach (what are the main challenges of this project? how will you tackle them? What’s the novelty of your proposed approach?)
  - Project timelines (what are milestones of this project and when do you plan to accomplish each of them
- Upload slides to canvas (“Project Proposal Presentation”) after presentations

Midterm & Final Project Presentation

- Tentatively schedule:
  - Midterm pre on **Mar 25&27**, Final pre on **May 5** (makeup for May 6&8)
  - **20 mins** for each group (15min pre+ 5min Q&A)
- Each group will give a presentation on the project and demonstrate the system built for the project.
- Evaluation criteria – **Novelty, system challenges, functionality, experimental evaluation**
- Upload slides to canvas (“Midterm or Final Project Presentation”) after presentations

Project Report

- Due on **May 8, 11:59pm.** A written report should be submitted to canvas ("Project Report") for each project.
- Suggested sections of the report:
  - introduction, related work, motivation and application scenarios, design, implementation, and evaluation.
- The report should be **no less than 8 pages** (A4, 11 point font, single or double column, single spacing) excluding references. It must include **a URL to your source code**.
- Group project: **states detailed contribution of each member in the report.**

Paper Presentations

- Each student will pick one paper to present at class from a provided paper list.
    - MobiCom / MobiSys / SenSys / UbiComp / IPSN / IoTDI
    - NeurIPS / ICML / ICLR / CVPR / ACL
- Each presentation takes 30 minutes plus 10 minutes for Q&A.
- The presentation should not only cover the in-depth discussion of the paper, but also all necessary background and related work for the class to fully understand the technical approach described in the paper.
- The evaluation criteria of presentation include **clarity, organization, technical content, and question answering**.

Paper Presentation Criteria

- **Clarity: _____ / 2**
  - Was the speaker clear and logical or confused and disorganized? Was the speed OK? or, too fast, too slow?
  - Are there too many texts on each slide? Did the presenter try to cram too much information on each slide?
- **Organization: _____ / 2**
  - Was the presentation clearly organized and well planned? Were there necessary transitions between slides?
  - Did the presentation finish on time?
- **Technical content: _____ / 4**
  - Did the presentation describe enough background and related work for audience to understand the problem(s) to be solved?
  - Did the presentation contain enough technical details of the paper?
  - Did the presentation give necessary examples to explain difficult technical issues?
  - Did the presentation clearly explain the experimental settings and results?
- **Question answering: _____ / 2**
  - Did the presenter(s) clearly answer the questions?
  - Did the presenter(s) refer to useful resources for the questions that are challenging or beyond the scope of the paper?

Audiences – Paper Reviews

- **Students will write a review for 5 papers to be presented during class** (5 reviews * 2)
  - No less than half A4 page, 11 font, single spacing.
  - Submit to canvas ("Paper Reviews") **before the paper is presented**.
  - Guidance for the reviews
    - Summarizing the motivation, research problems, and contributions of the paper
    - Pros and Cons of the paper: motivation, design, evaluations
    - Open research problems, holes/limitations of the paper
    - Other considerations:
      - Are the research problems significant enough?
      - Whether the assumptions/models used in the paper are reasonable and realistic.
      - Are there any problems in experimentation settings?
      - Are the claims supported by experimental results?
      - Any ways to improve solutions described in the paper

Audiences – In-Class Q&A

- **Students will raise at least 5 questions during the class** (5 questions * 2)
  - Paper Presentations
  - Lectures
  - Project presentations

Tentative Syllabus

- ➢ Lectures
- ➢ Paper Presentations
- ➢ Project Presentations

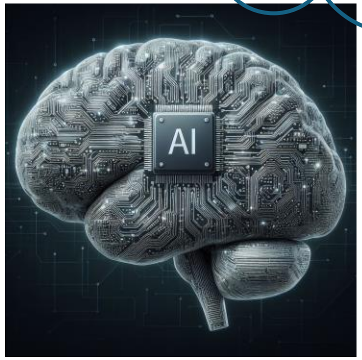
Please follow the update in both Canvas and website.

Date	**Topics**	**Note**
Feb 4 (Tuesday)	Course Introduction and Overview	
Feb 6 (Thursday)	Machine Learning Basics	
Feb 11 (Tuesday)	Challenges in Embedded AI Systems	
Feb 13 (Thursday)	Challenges in Embedded AI Systems	
Feb 18 (Tuesday)	Unsupervised Learning	
Feb 20 (Thursday)	Unsupervised Learning	Paper Presentation
Feb 25 (Tuesday)	**Project Proposal Presentation**	8min pre+ 2min QA
Feb 27 (Thursday)		Cancelled (Makeup on Mar 8/9)
Mar 4 (Tuesday)	Multimodal Sensing and Learning	
Mar 6 (Thursday)	Multimodal Sensing and Learning	Paper Presentation
Mar 11 (Tuesday)	Federated Learning	
Mar 13 (Thursday)	Federated Learning	Paper Presentation
Mar 18 (Tuesday)	Efficient Deep Learning on the Edge	
Mar 20 (Thursday)	Efficient Deep Learning on the Edge	Paper Presentation
Mar 25 (Tuesday)	**Midterm Project Presentation**	15min pre+ 5min QA
Mar 27 (Thursday)	**Midterm Project Presentation**	15min pre+ 5min QA
April 1 (Tuesday)		No class (Midterm Break)
April 3 (Thursday)		No class (Midterm Break)
April 8 (Tuesday)	LLMs and Foundation Models on the Edge	
April 10 (Thursday)	LLMs and Foundation Models on the Edge	Paper Presentation
April 15 (Tuesday)	LLMs and Foundation Models on the Edge	Paper Presentation
April 17 (Thursday)	Physics-strengthened AI for Sensing Systems	
April 22 (Tuesday)	Physics-strengthened AI for Sensing Systems	Paper Presentation
April 24 (Thursday)	Applications	
April 29 (Tuesday)	Applications	Paper Presentation
May 1 (Thursday)	No class (Public Holiday)	
May 3 (Saturday)	**Final Project Presentation (160 mins)**	15min pre+ 5min QA
May 6 (Tuesday)		Cancelled (Makeup on May 5)
May 8 (Thursday)	**Report Deadline**	

Outline

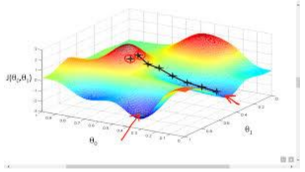
- Course logistics
- **Embedded AI overview and examples**

Today's AI



computer vision, natural language processing, generative AI, …

Cloud

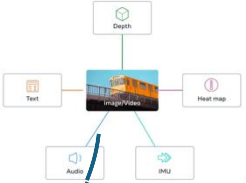


Learning Algorithm

Data

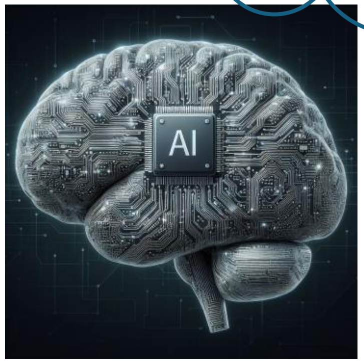
Model

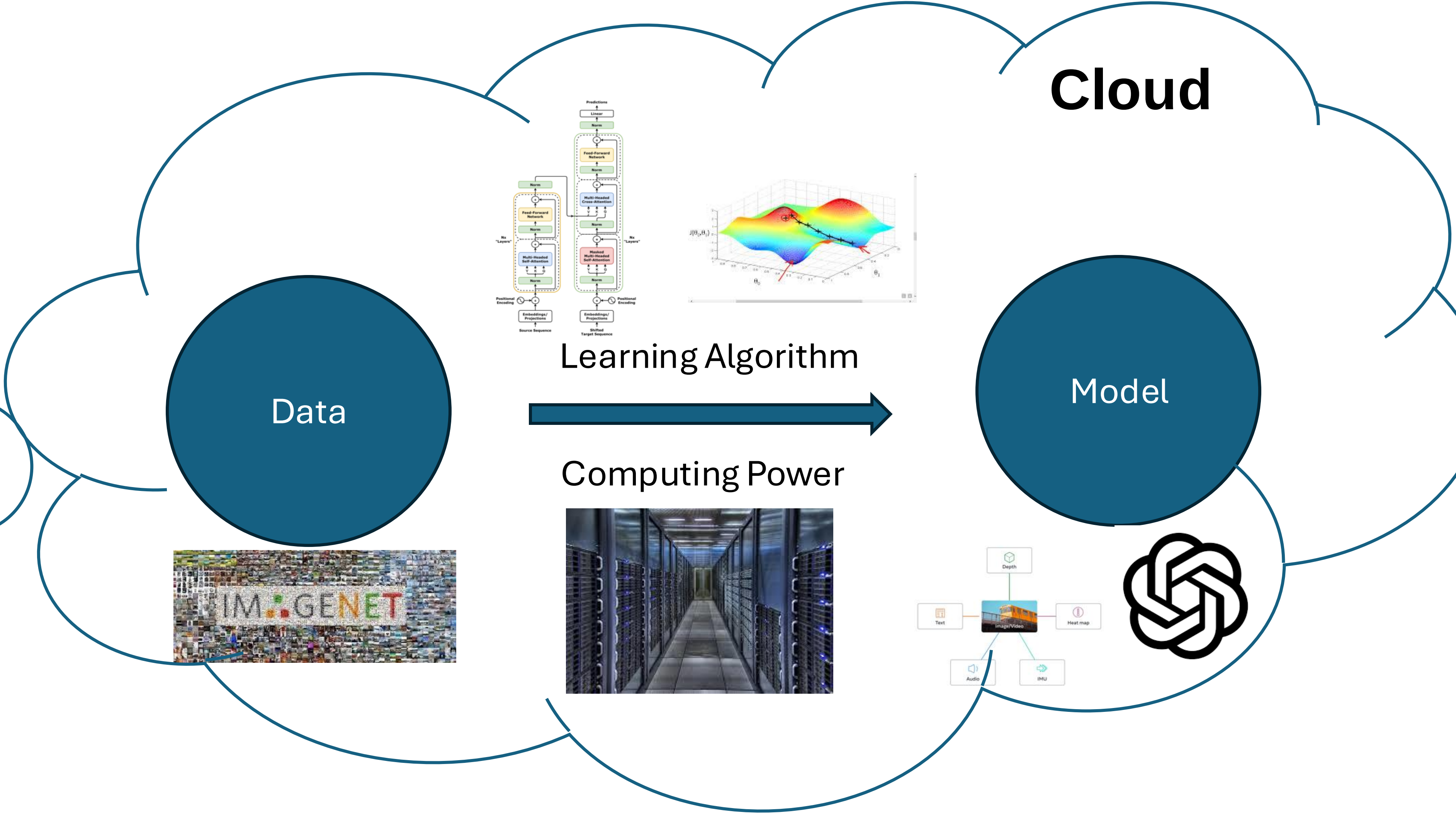
Computing Power



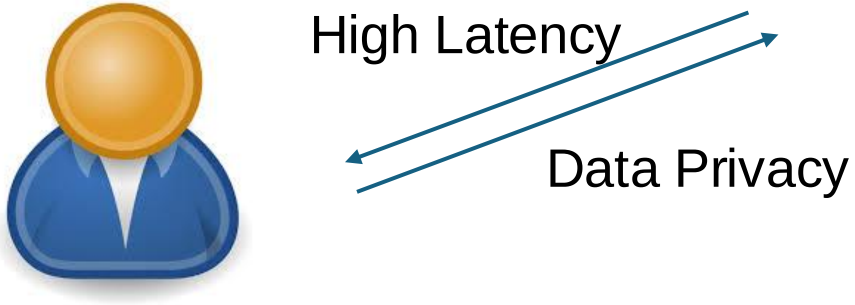
- ➢ Large Data, Large Computing Resources, Large Models
- ➢ Vision and Language
- ➢ Cloud-based architecture

Today's AI





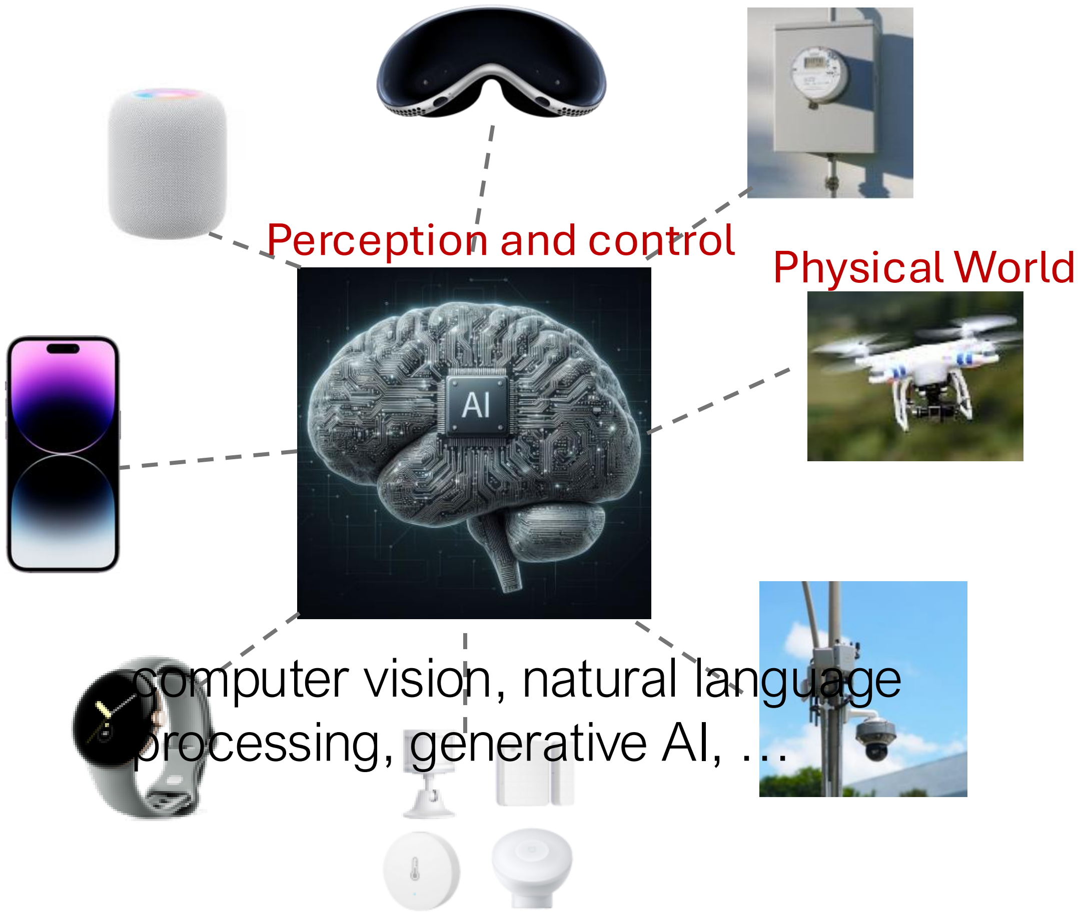
computer vision, natural language processing, generative AI, …



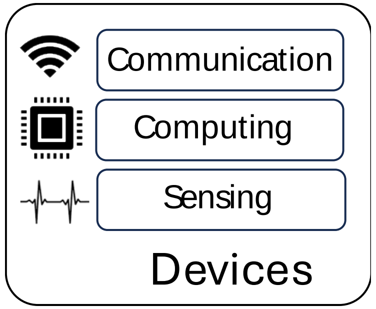
What is Embedded AI

➢ AI on network edge, physical devices & "Things"

- Vehicles, mobiles, wearables, robots, sensors



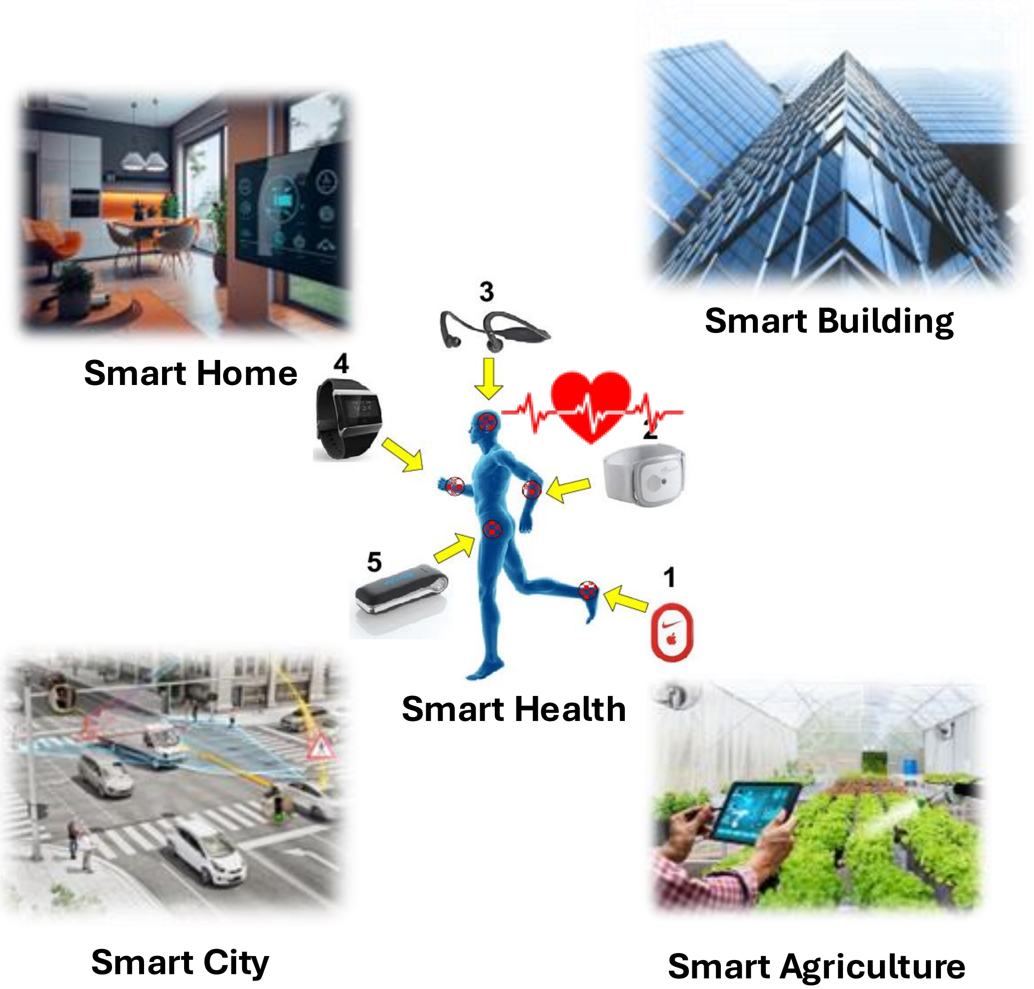
Environment with **Ambient Intelligence**



- In-situ sensing
- On-device computing
- Networked computing

Applications with Embedded AI systems

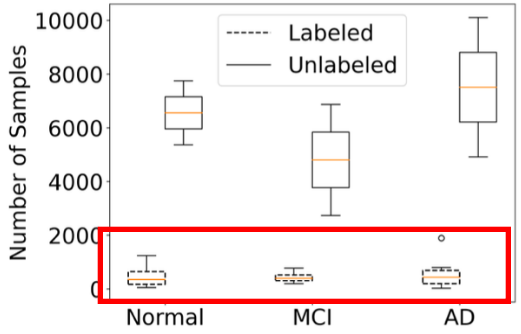
- Data-Intensive, Real-Time, Mission-Critical, Privacy-Sensitive



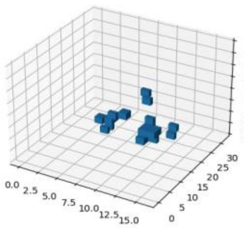
Major Challenges of Embedded AI Systems

➢ Data Challenges: How to harness **distributed and imperfect data**?

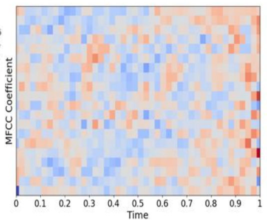
- Limited labeled data
- Fusing heterogeneous modalities
- Data Privacy



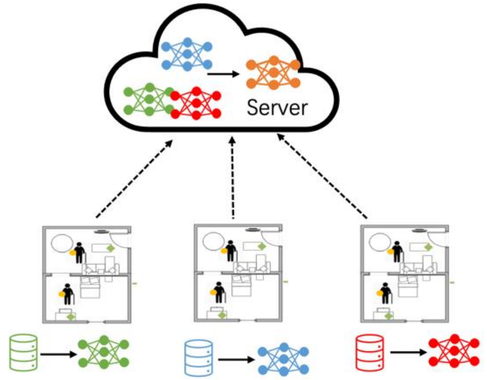
Depth Image

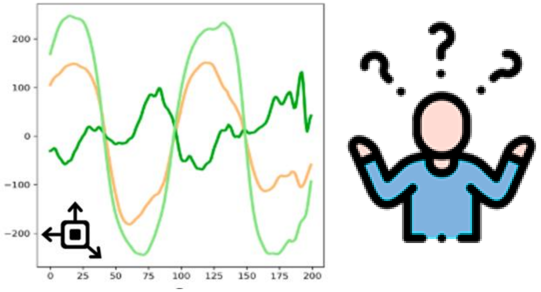


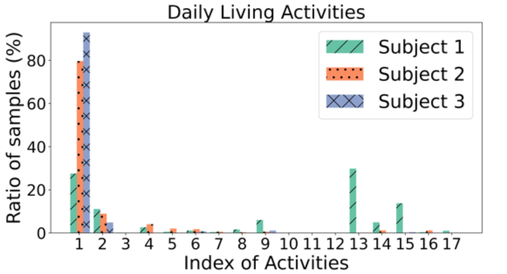
Radar Data



MFCC of Audio







Unsupervised Learning

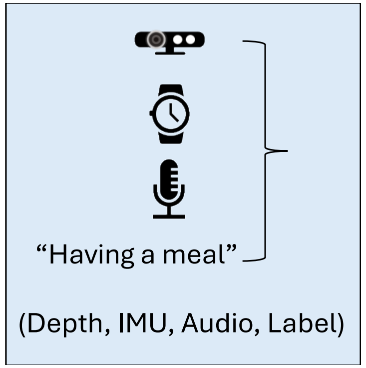
Multimodal Sensing and Learning
LLMs and Foundation Models on the Edge

Federated Learning

Multimodal Learning in Embedded AI

➢ Conventional Approaches

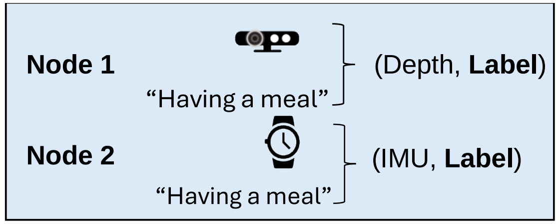
- Complete and synchronized data



$D=(x_1, x_2, x_3, y)$

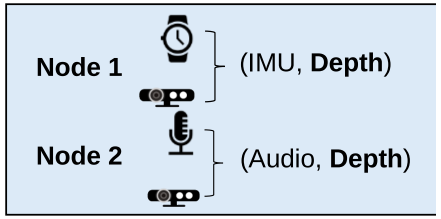
➢ Data Collected by Distributed Devices

- Incomplete and heterogeneous
  - Modality missing:



$D1=(x_1, \mathbf{y})$,
$D2=(x_2, \mathbf{y})$

  - Label missing:



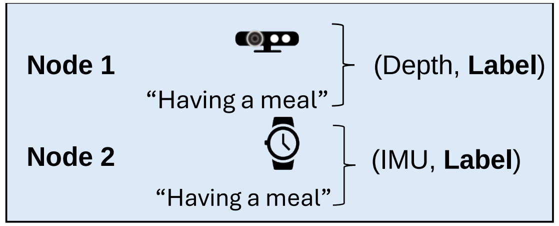
$D1=(x_1, \mathbf{x_2})$,
$D2=(\mathbf{x_2}, x_3)$

- Distribution shift:
  - Collected by distributed nodes at different times and locations

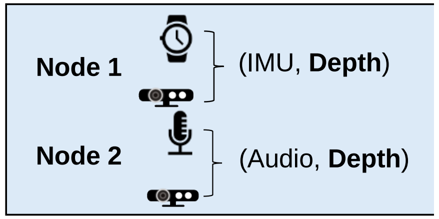
Multimodal Learning with Distributed and Incomplete Data

➢ Key Question:

- Can we learn joint multimodal embeddings with **distributed and incomplete data in IoT**?



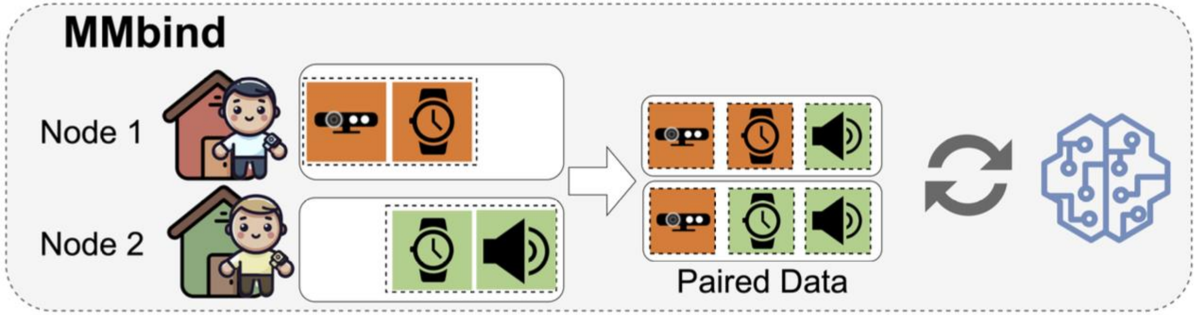
$D1=(x_1, \mathbf{y})$,
$D2=(x_2, \mathbf{y})$



$D1=(x_1, \mathbf{x_2})$,
$D2=(\mathbf{x_2}, x_3)$

➢ Key Idea:

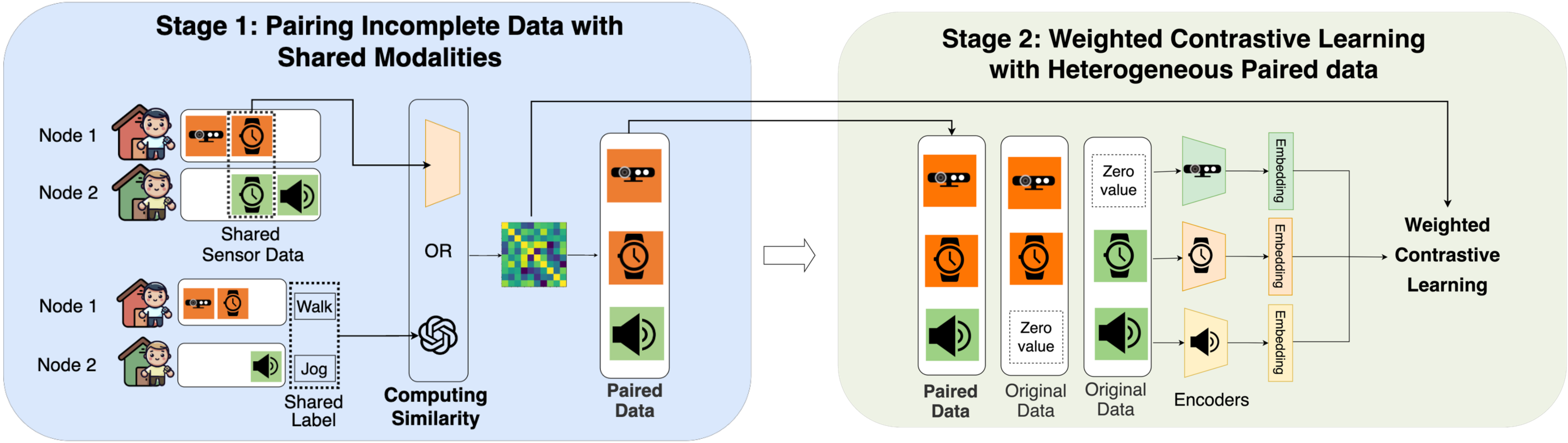
- Bind data from disparate sources and incomplete modalities with **the shared modality**
  - Shared modality: **sensor data or labels**



MMBind: System Overview

➤ **Construct Pseudo-Paired Data**

➤ **Learning with Heterogeneous Paired Data**



Imperfect paring

- Data of different modalities observing similar events can be effectively used for multimodal training.

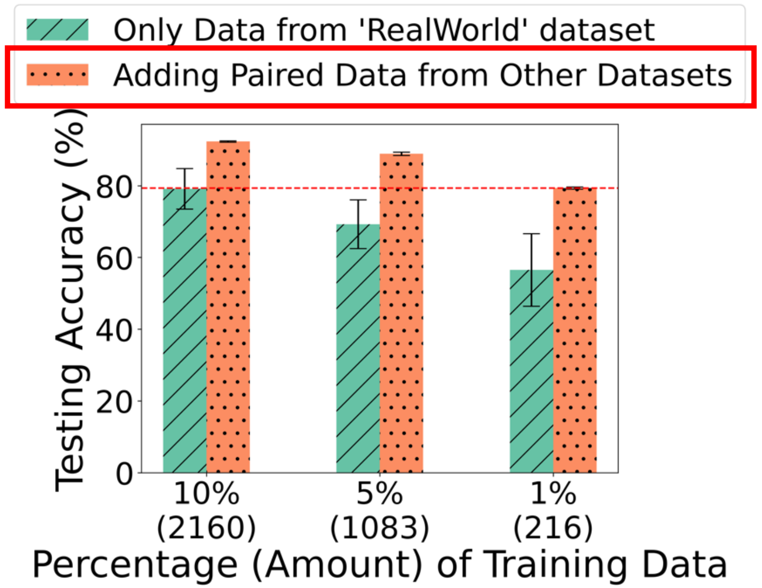
MMBind: Unleashing the Potential of Distributed and Heterogeneous Data for Multimodal Learning in IoT. (Accepted by SenSys ’25)

Evaluation

➢ Intra-Dataset Binding

Datasets	UTD-MHAD		MM-FI		PAMAP2	
	Acc	F1	Acc	F1	Acc	F1
Lower Bound	40.41	0.380	65.74	0.654	64.51	0.609
Unimodal	69.04	0.646	53.91	0.532	59.44	0.528
MIM	62.23	0.590	68.31	0.676	63.38	0.567
MPM	69.74	0.666	70.71	0.701	64.15	0.592
CMG	61.69	0.592	72.17	0.722	61.62	0.577
DCM	59.25	0.563	68.26	0.678	64.43	0.597
MMBind	**78.86**	**0.763**	**77.72**	**0.775**	**69.08**	**0.654**
Upper Bound	78.68	0.768	72.45	0.720	68.87	0.636

➢ Cross-Dataset Binding

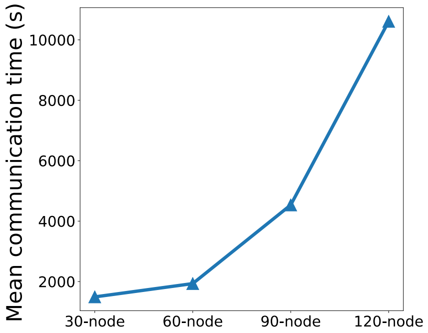


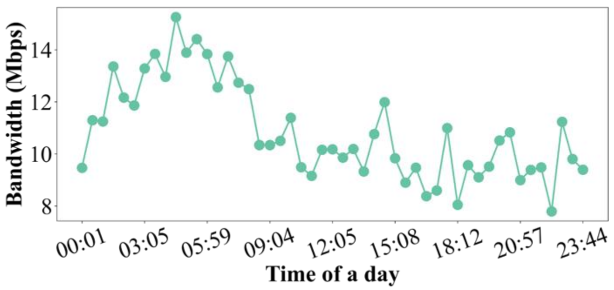
- Adding pseudo-paired data samples significantly boosts model performance.
  - Generate a **foundational dataset** for IoT applications

Major Challenges of Embedded AI Systems

➢ System Challenges: How to make the system more **scalable, resource-efficient and robust to real-world dynamics**?

- Scalability
- Limited and dynamic resources
- Heterogeneous platforms



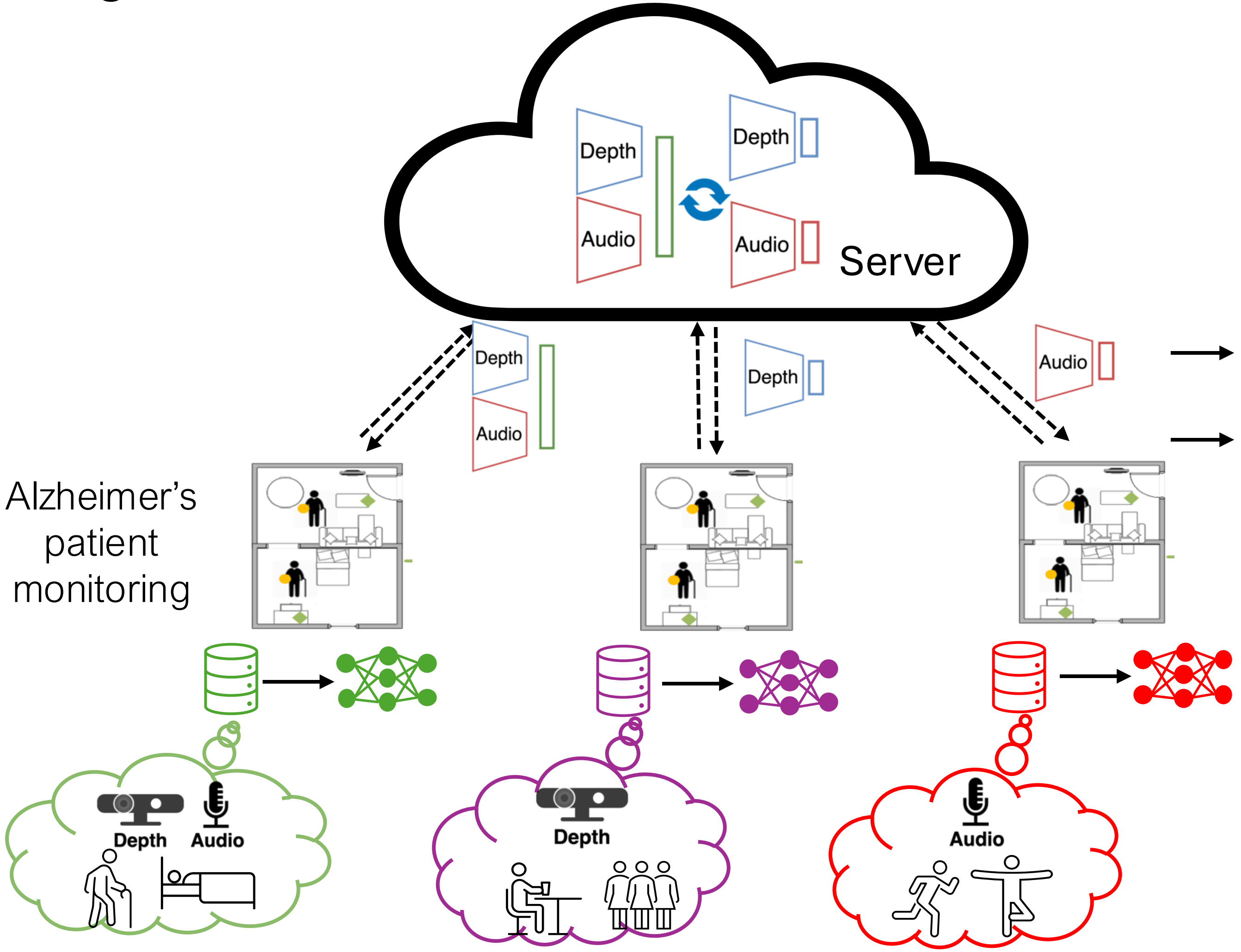


Federated Learning

- Efficient Deep Learning on the Edge
- LLMs and Foundation Models on the Edge
  - On-device inference
  - On-device training
  - Offloading

Heterogeneous Multi-Modal Federated Learning

- Challenges

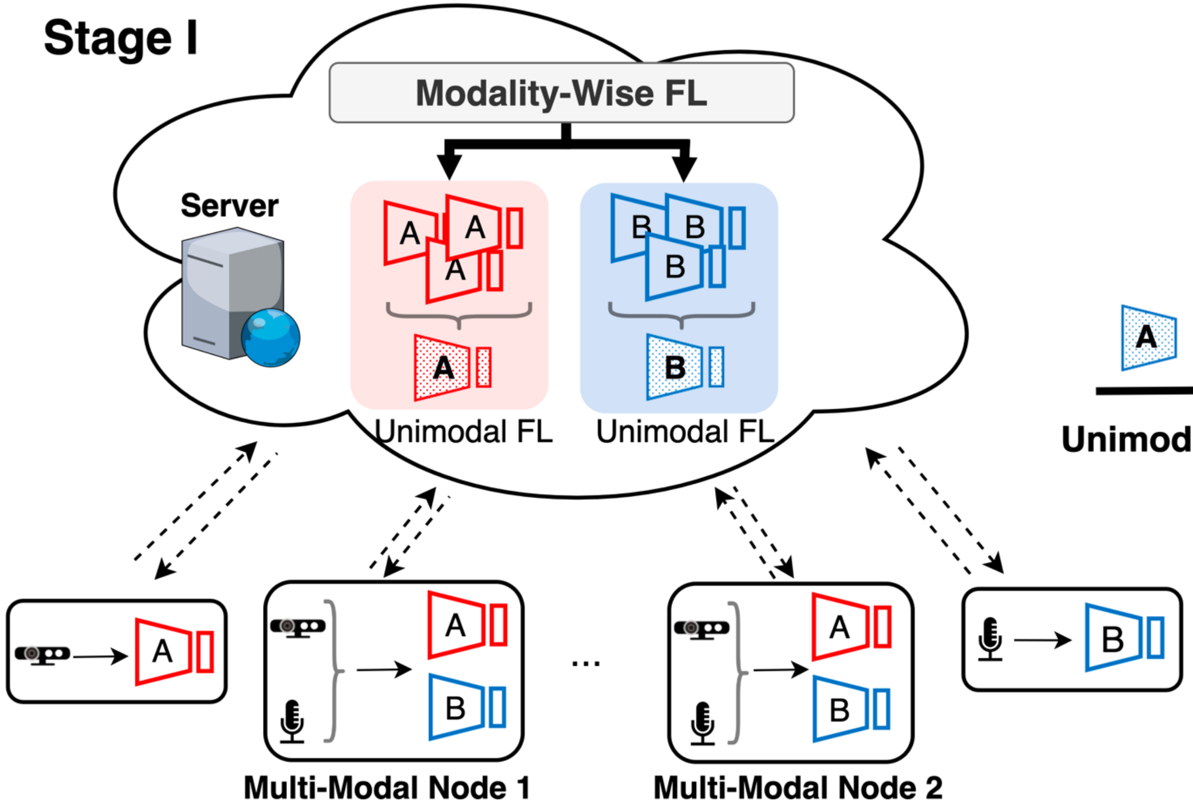


- Large model divergence
- Significant training latency

- Modality heterogeneity
- Distribution heterogeneity

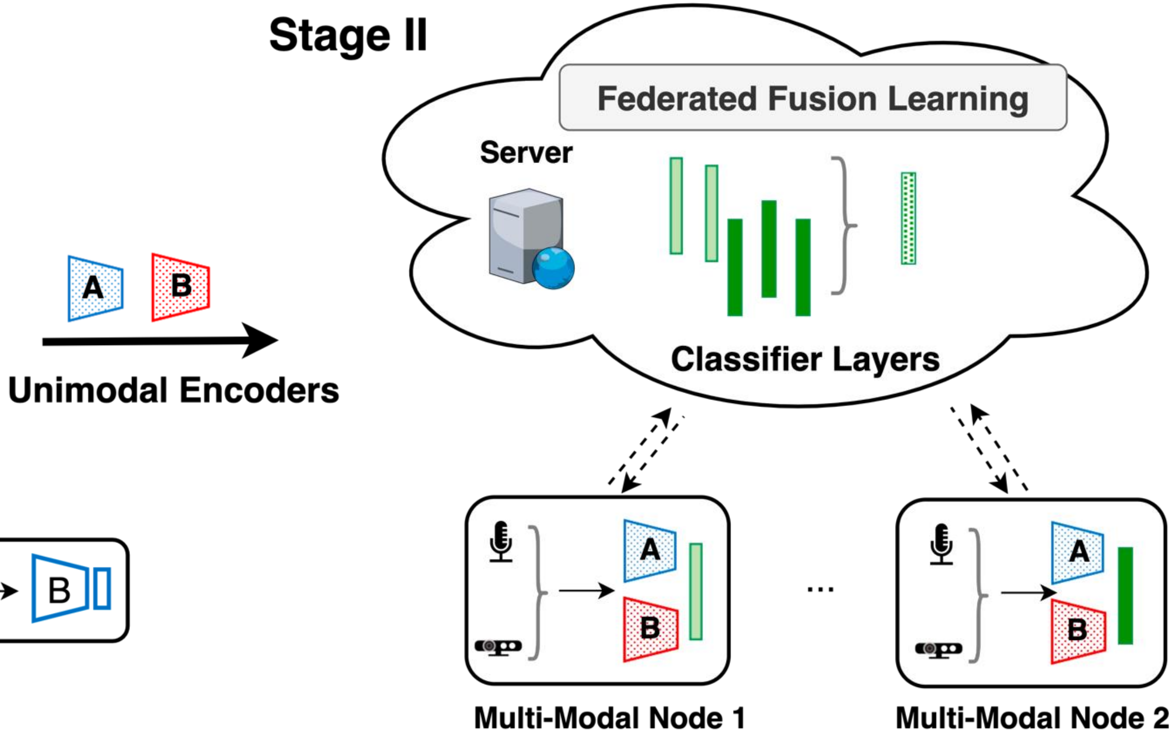
A Two-Stage Framework for Multi-Modal FL

➢ Modality-Wise Federated Learning



- Collaboratively train **unimodal encoders**

➢ Federated Fusion Learning

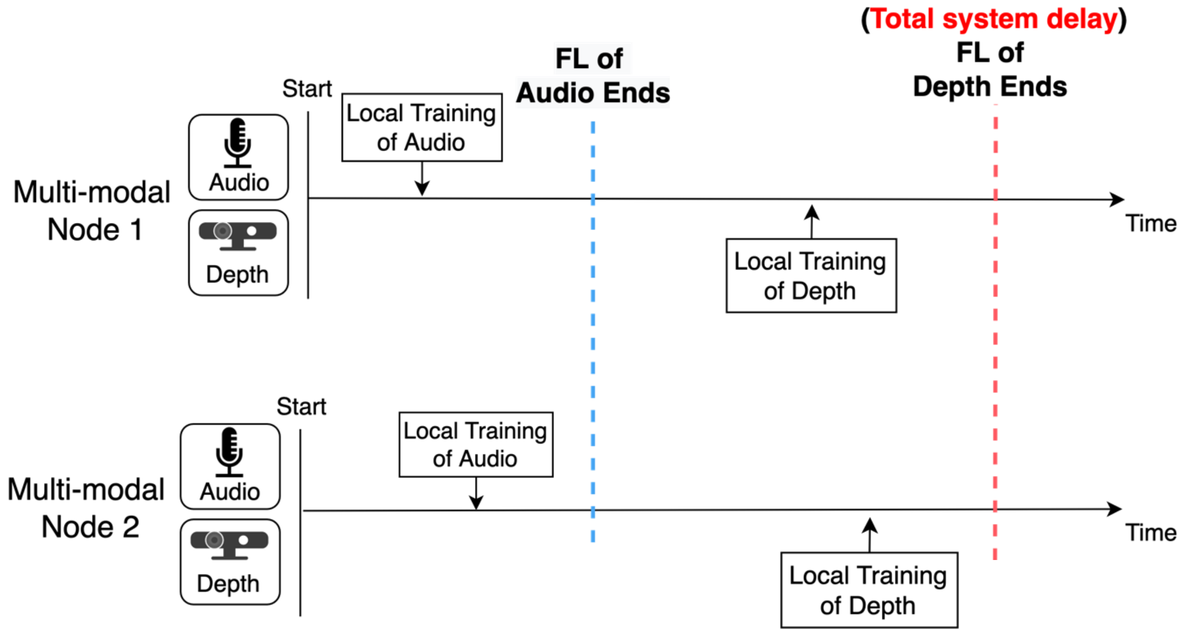


- Collaboratively train the **multi-modal classifier**

Harmony: Heterogeneous Multi-Modal Federated Learning through Disentangled Model Training. (MobiSys '23)

Modality-Wise Federated Learning

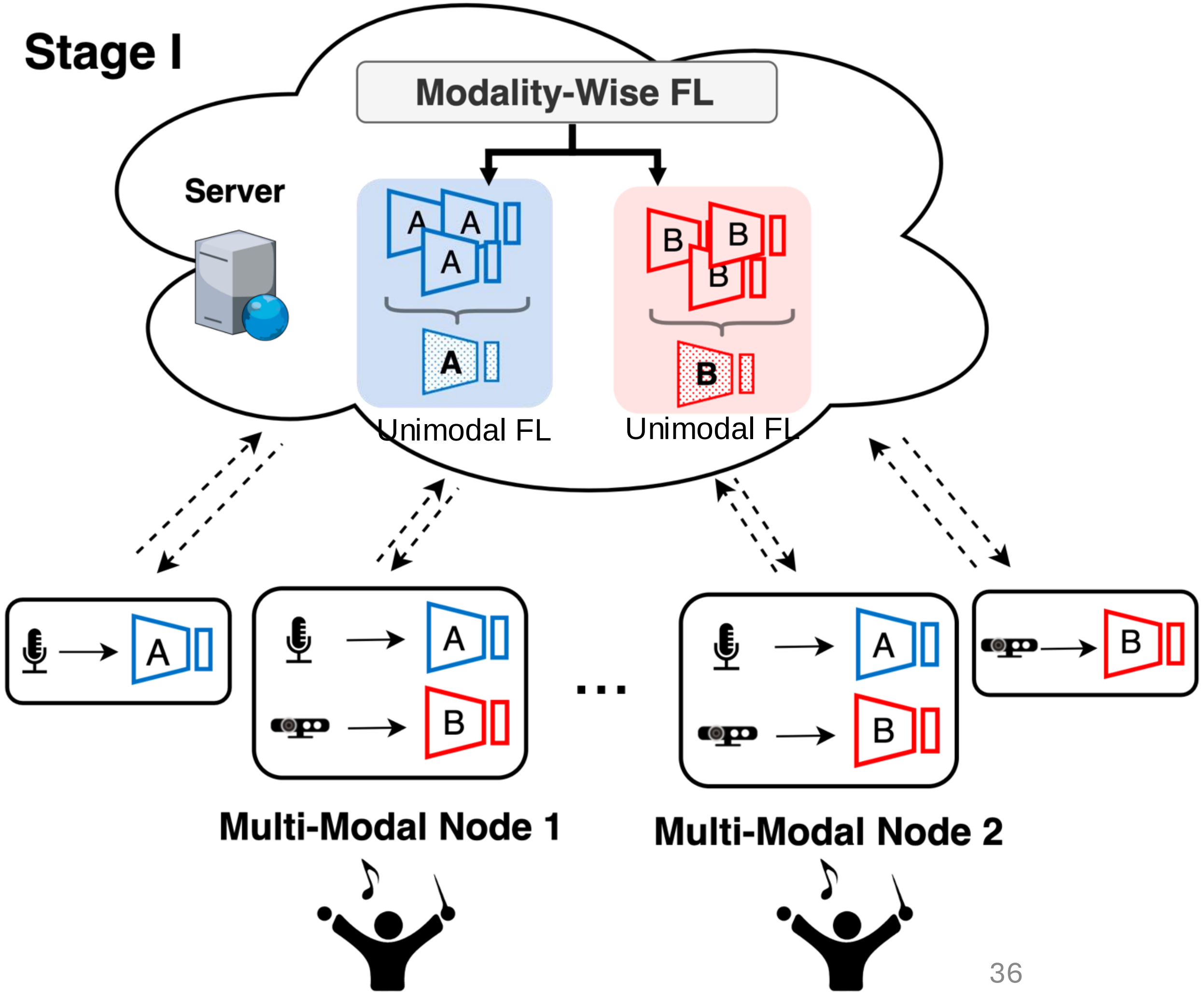
➢ Imbalanced Training Delays



➢ Balance-Aware Resource Allocation

Less resources

More resources



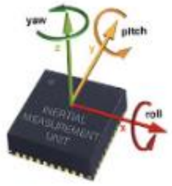
Major Challenges of Embedded AI Systems

➢ Challenges related to specific sensor modalities

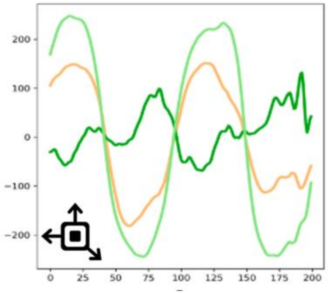
- Vision Sensors

Depth

- Motion Sensors

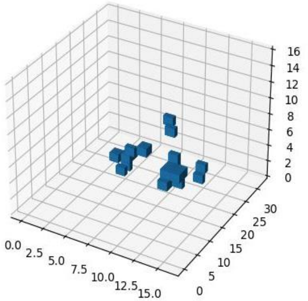


IMU



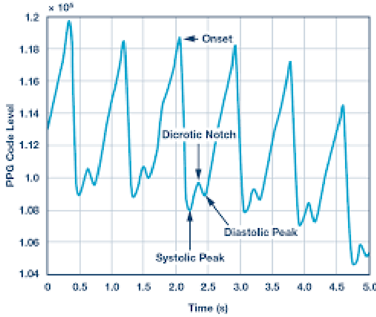
- Radio frequency (RF)

mmWave Radar



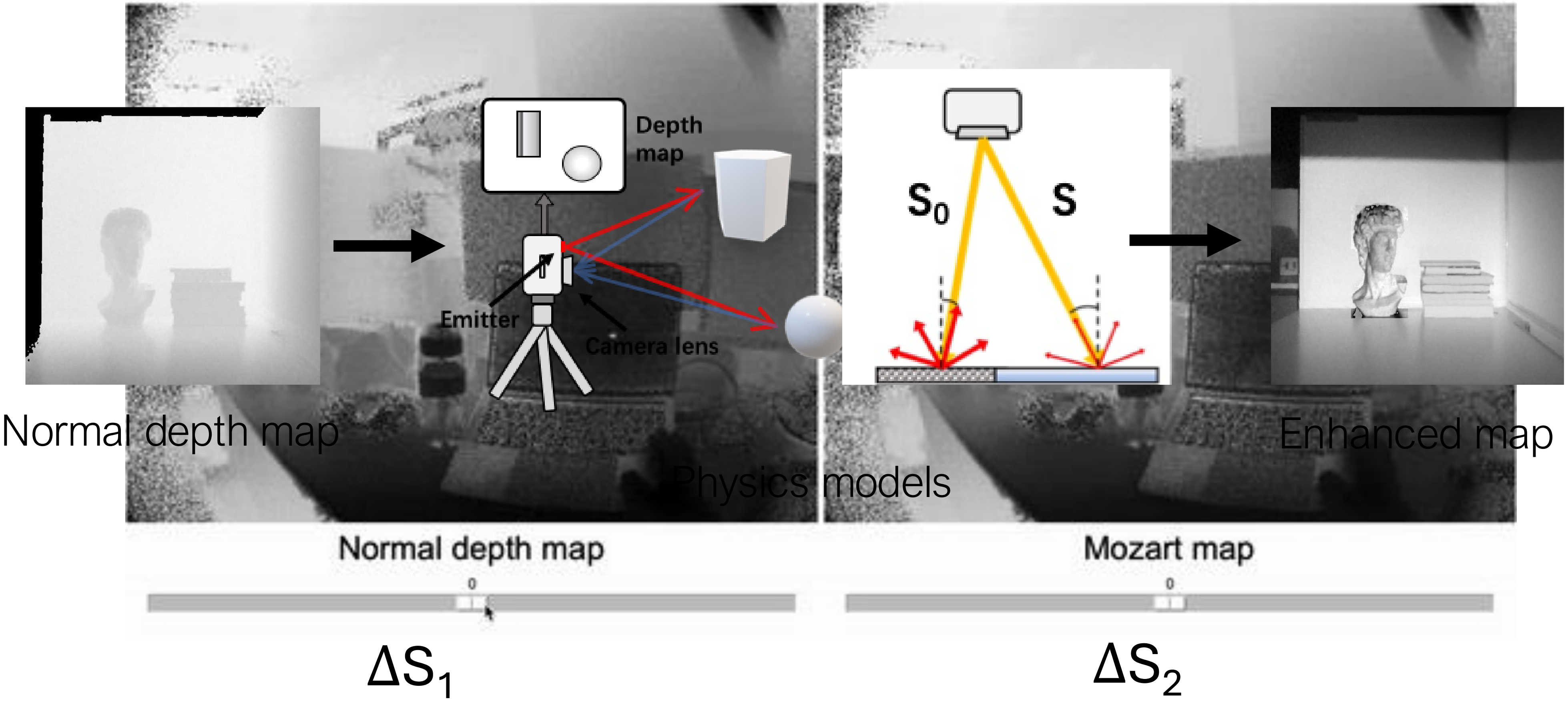
- Biological Sensors

PPG



Physics-Strengthened AI for Robust Sensing

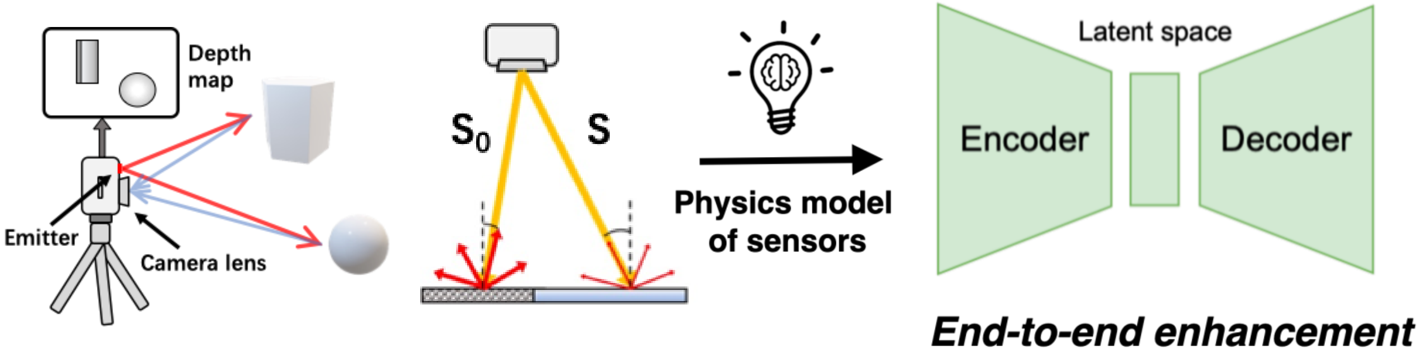
- Enhancing ToF Depth Sensing with Lambertian Reflection Model

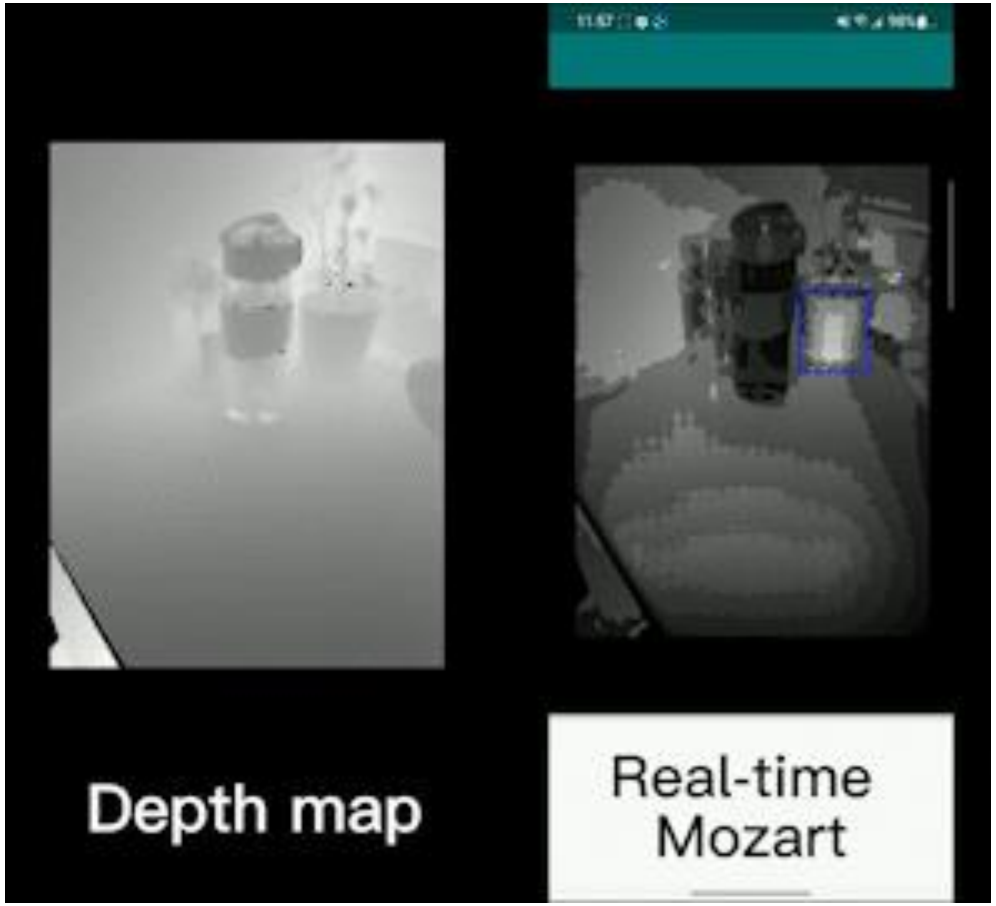


Mozart: A Mobile ToF System for Sensing in the Dark through Phase Manipulation (**MobiSys '23 Best Paper**).

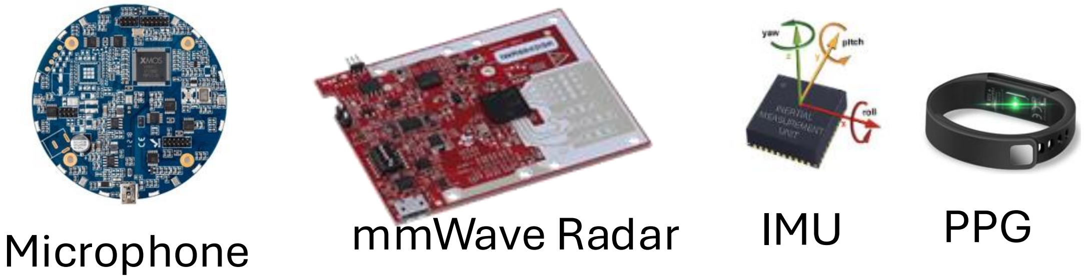
Physics-Strengthened AI for Robust Sensing

- Integrate First-principle Model with ML





- Enhancing Mobile Sensing



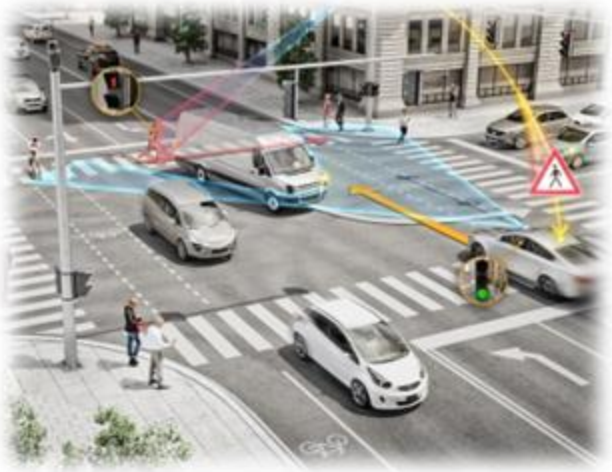
Mozart: A Mobile ToF System for Sensing in the Dark through Phase Manipulation (**MobiSys '23 Best Paper**).

Major Challenges of Embedded AI Systems

➢ Challenges related to specific applications

- Smart Health

- Autonomous Driving

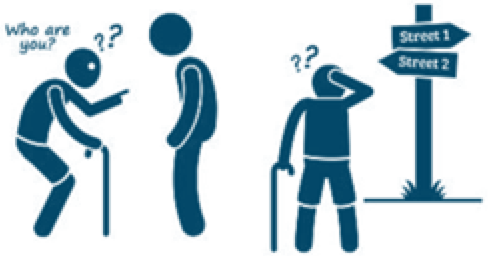


- Localization

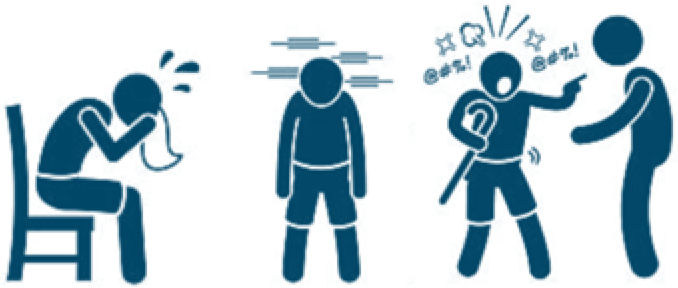
Digital Biomarkers for Early AD Diagnosis

- Leverage AI and sensor devices to capture **physiological, behavioral and lifestyle** symptoms of AD in natural living environments.

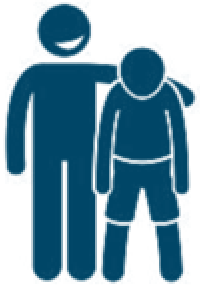
Activities of Daily Living



Cognition



Behavioral and Psychological Symptoms of Dementia (BPSD)



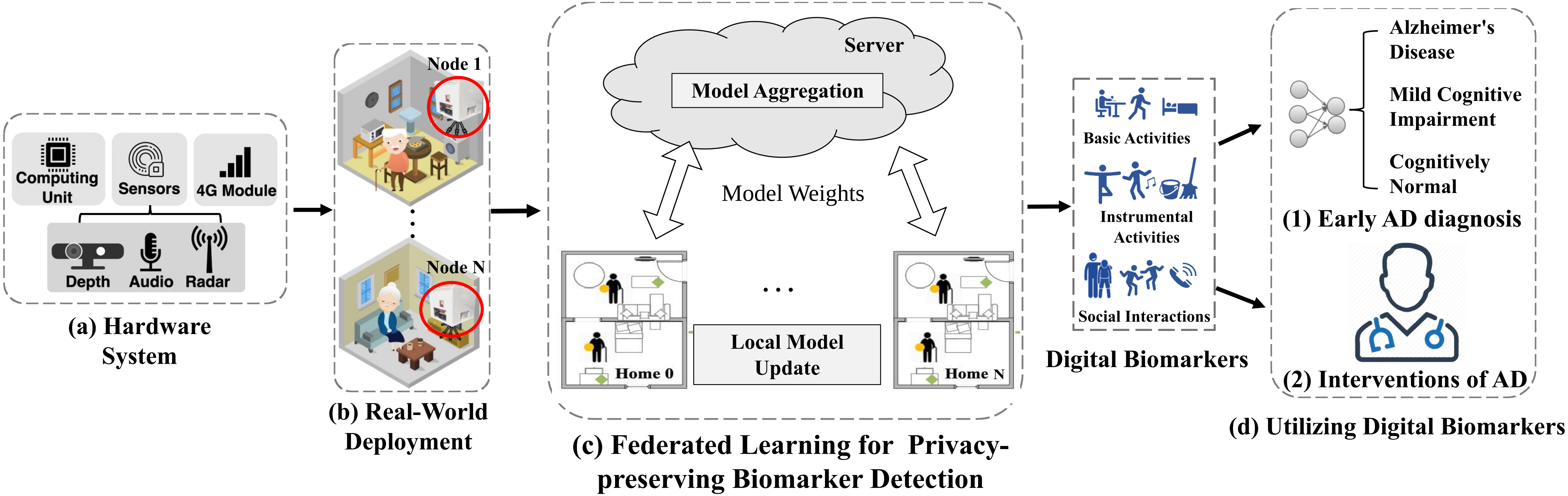
Social Interaction

- Multi-dimensional
- Complex and dynamic

⇒ Need multiple sensor modalities

ADMarker: System Overview

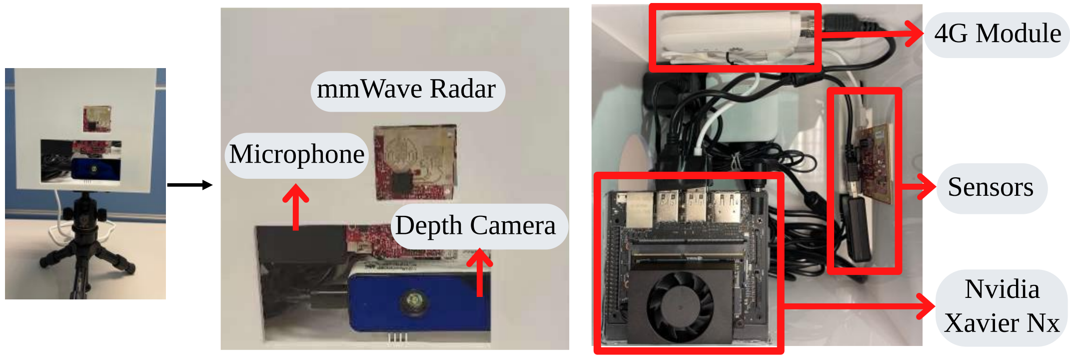
➢ An end-to-end system that integrates **multi-modal sensors** and **new machine learning** systems for detecting **multi-dimensional AD digital biomarkers** in home environments.



ADMarker: A Multi-Modal FL System for Monitoring Digital Biomarkers of Alzheimer's Disease. (MobiCom '24)

Hardware and Deployment

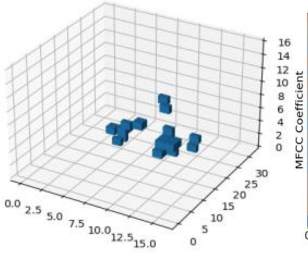
➢ Multi-modal hardware systems



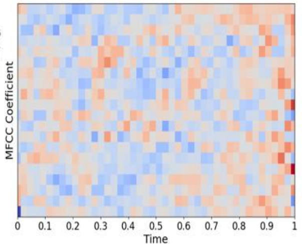
- Sensing + Computing + Communication

➢ Examples of recorded data

Depth Image



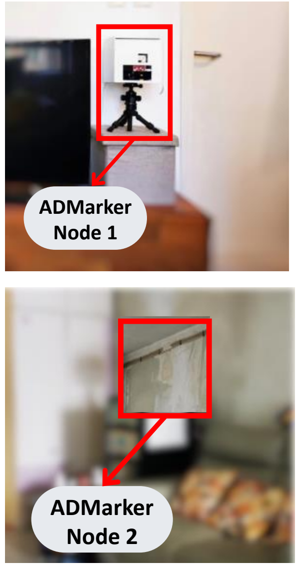
Radar Data

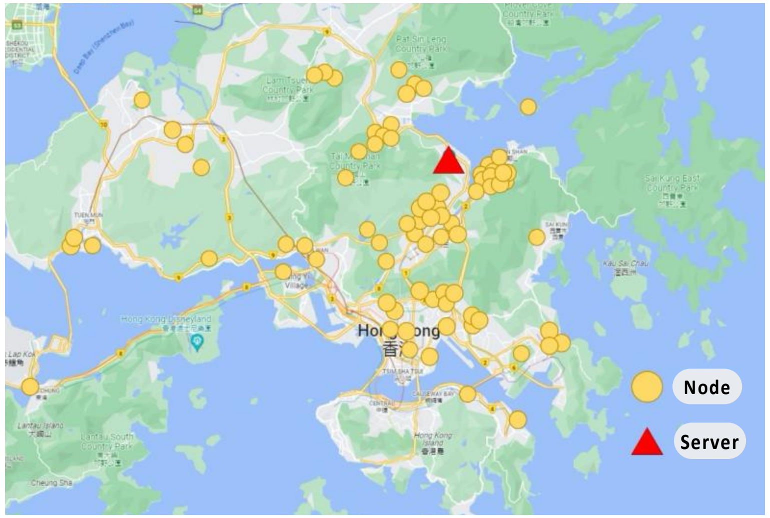


MFCC of Audio

➢ Clinical deployment

- Participants (N=91): 31 AD, 30 mild cognitive impairment, 30 cognitively normal
  - 61-93 (average 76.1) years old
  - 43 females and 48 males
- Deployment: four weeks





Research Problems

- **Problem-driven projects – Solve a specific challenging problem**
  - Example: Improve inference efficiency of LLMs running on the edge
- **Sensor-driven projects – Enhance the sensing quality of a specific sensor**
  - Example: Leverage AI techniques to improve the quality of UWB signals under mobility
- **Application-driven projects – Build a system for specific application**
  - Example: Embedded AI system for breath/occupant/environment monitoring
- **Measurement-driven projects – Experimental evaluation of a system/network**
  - Example: Performance of ML algorithms for different hardware and tasks

Related Techniques

- **Unsupervised Learning, Multimodal Learning, Federated Learning**
- **Task scheduling, Model compression/ quantization /finetuning**
- **Physics-strengthened AI**

Break

- **Next lecture: Machine Leaning Basics**
- **A shared spreadsheet to be released on canvas, please remember to select papers for presentation**
- **Course APP and dataset to be released next week**
- **Any questions?**