

Ph.D. Dissertation Defense

Elevating Large-Scale Pre-Trained Models into Foundation Models for Mobile Sensing

Hyungjun Yoon, Ph.D. Candidate, KAIST Electrical Engineering

Committee Members

Prof. Sung-Ju Lee (KAIST)

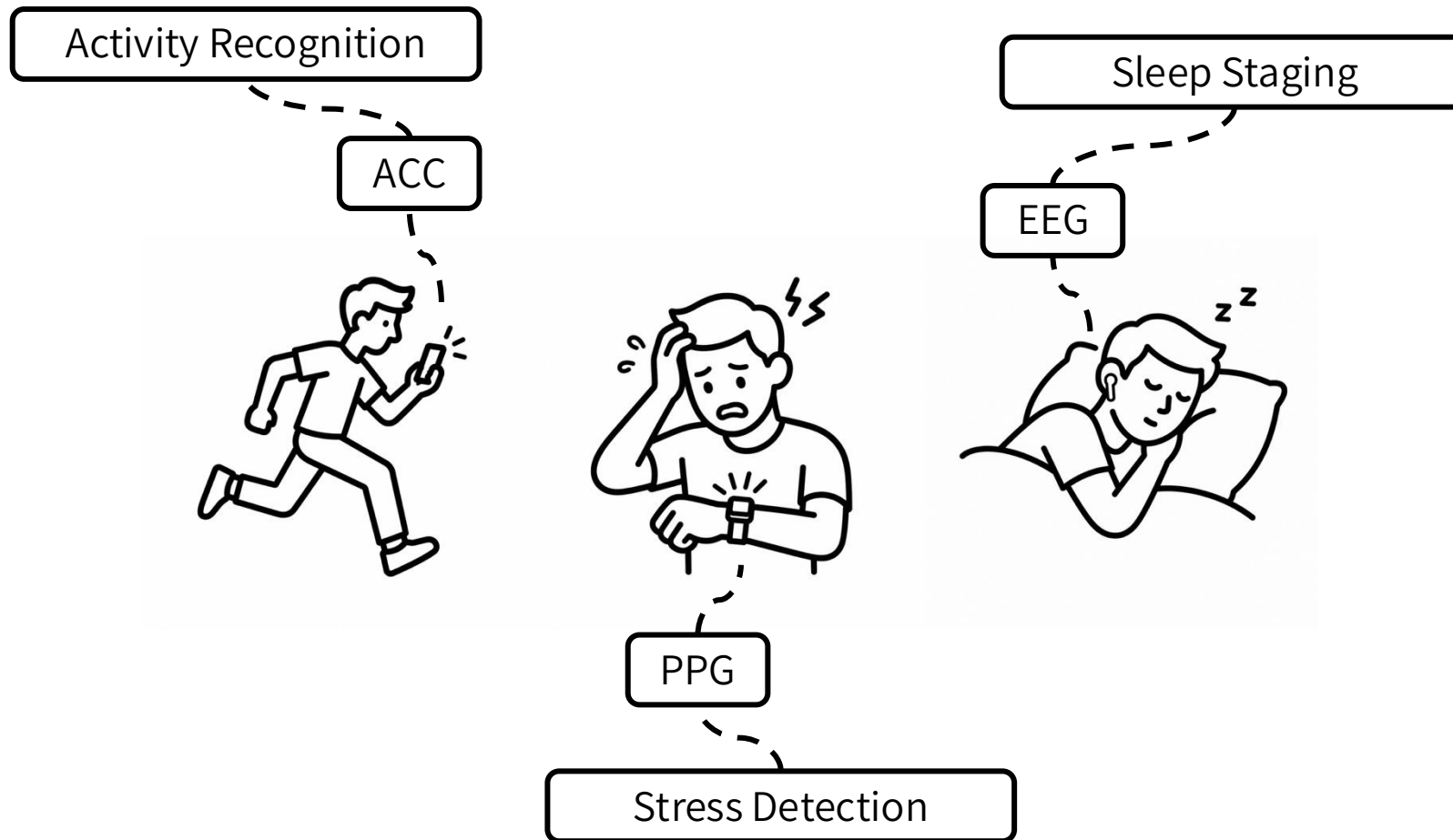
Prof. Edward Choi (KAIST)

Prof. Kimin Lee (KAIST)

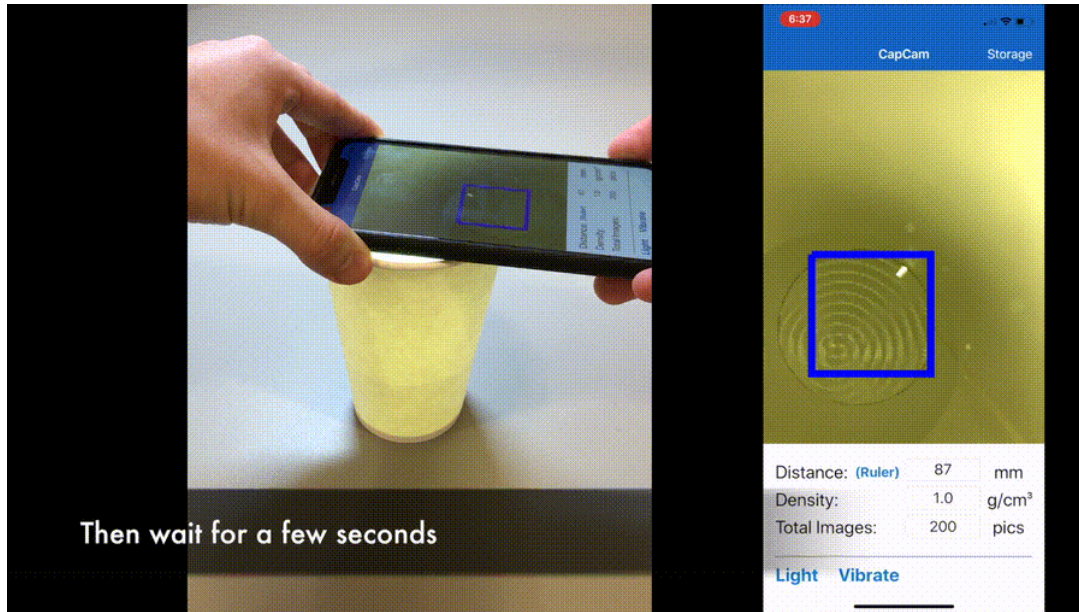
Prof. Mo Li (HKUST)

Prof. Lili Qiu (The University of Texas at Austin)

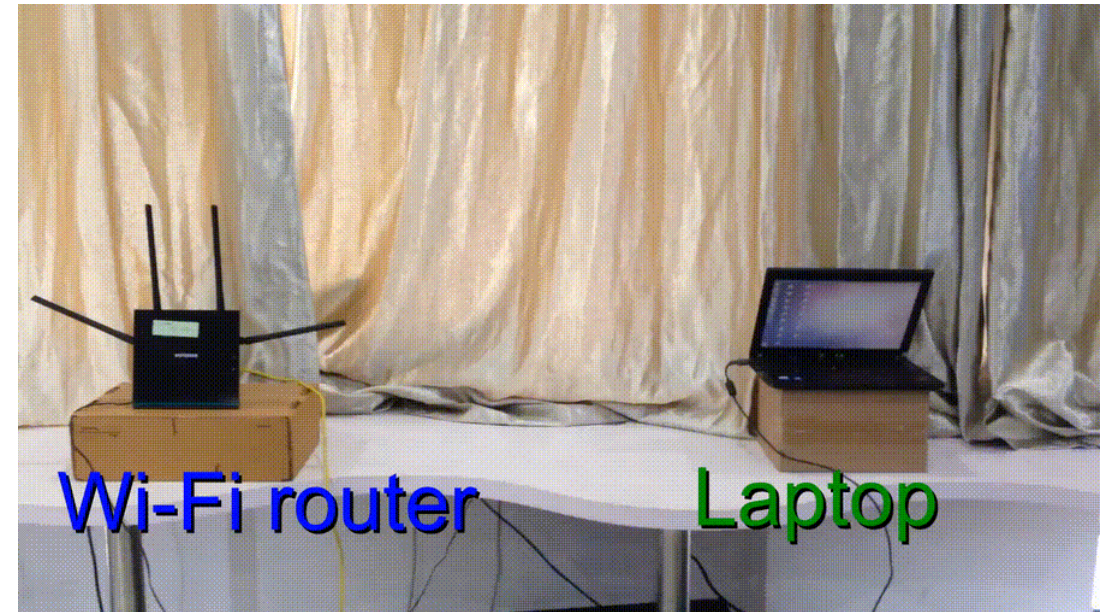
Background: Mobile Sensing Applications



Background: Mobile Sensing Applications

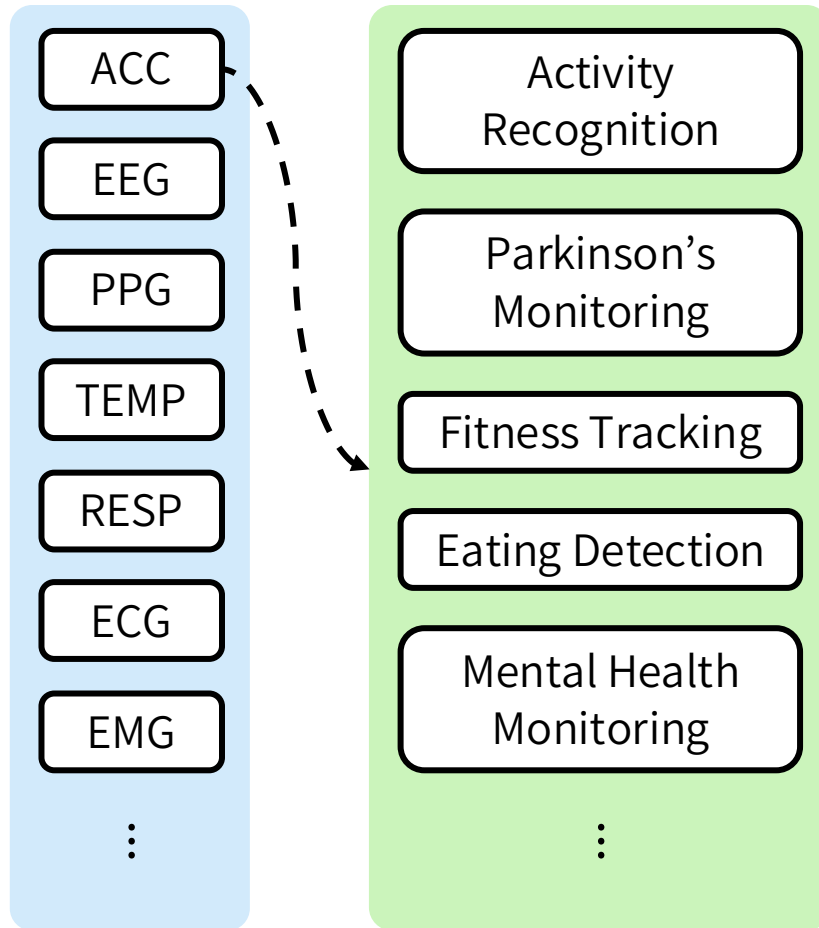


Liquid Testing with Your Smartphone
Yue et al., MobiSys '19

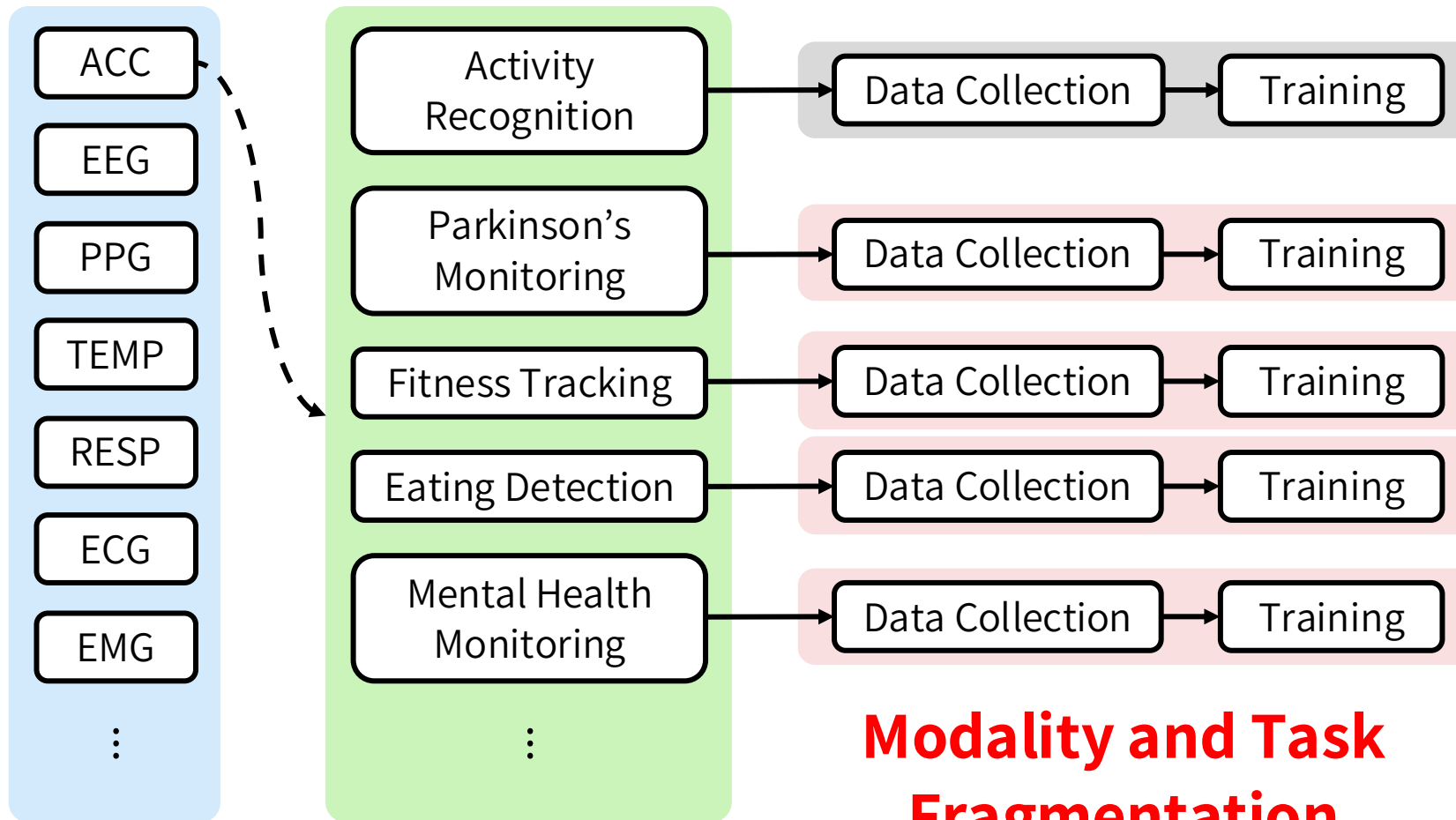


*Understanding and Modeling of WiFi Signal
Based Human Activity Recognition*
Wang et al., MobiCom '15

Background: Deep Learning + Mobile Sensing

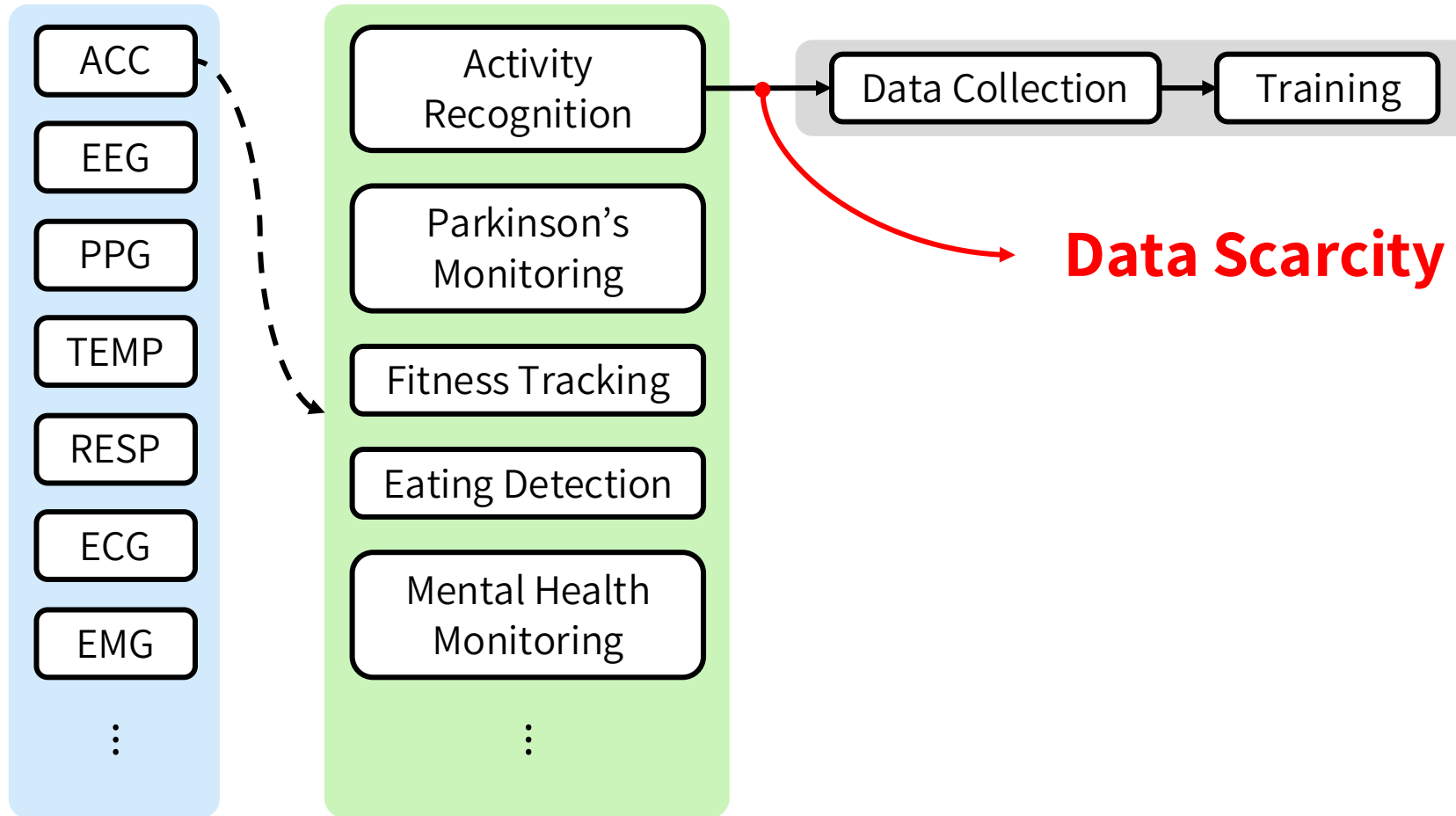


Scalability Challenges: Narrow, Task-Specific Pipelines

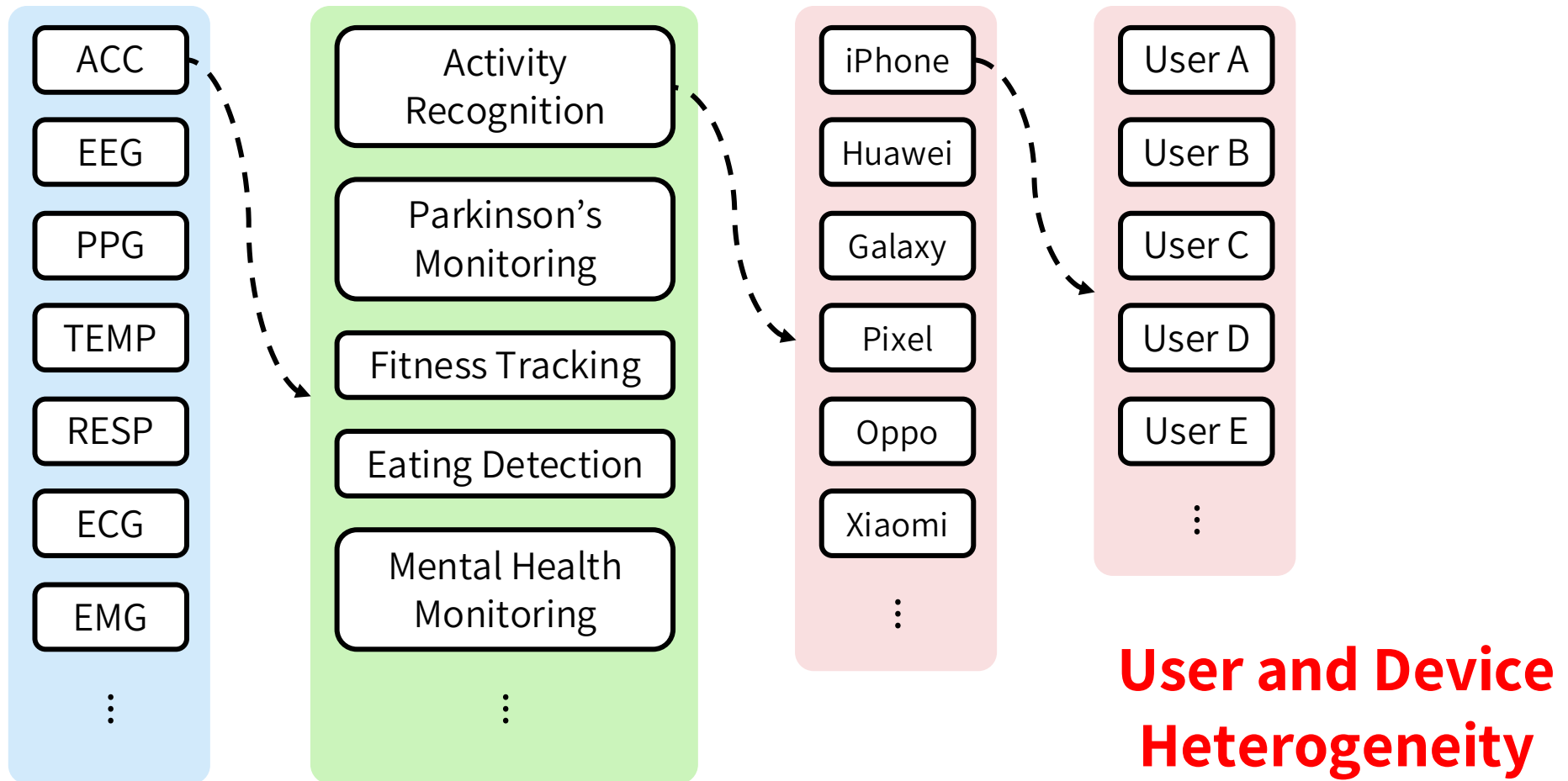


**Modality and Task
Fragmentation**

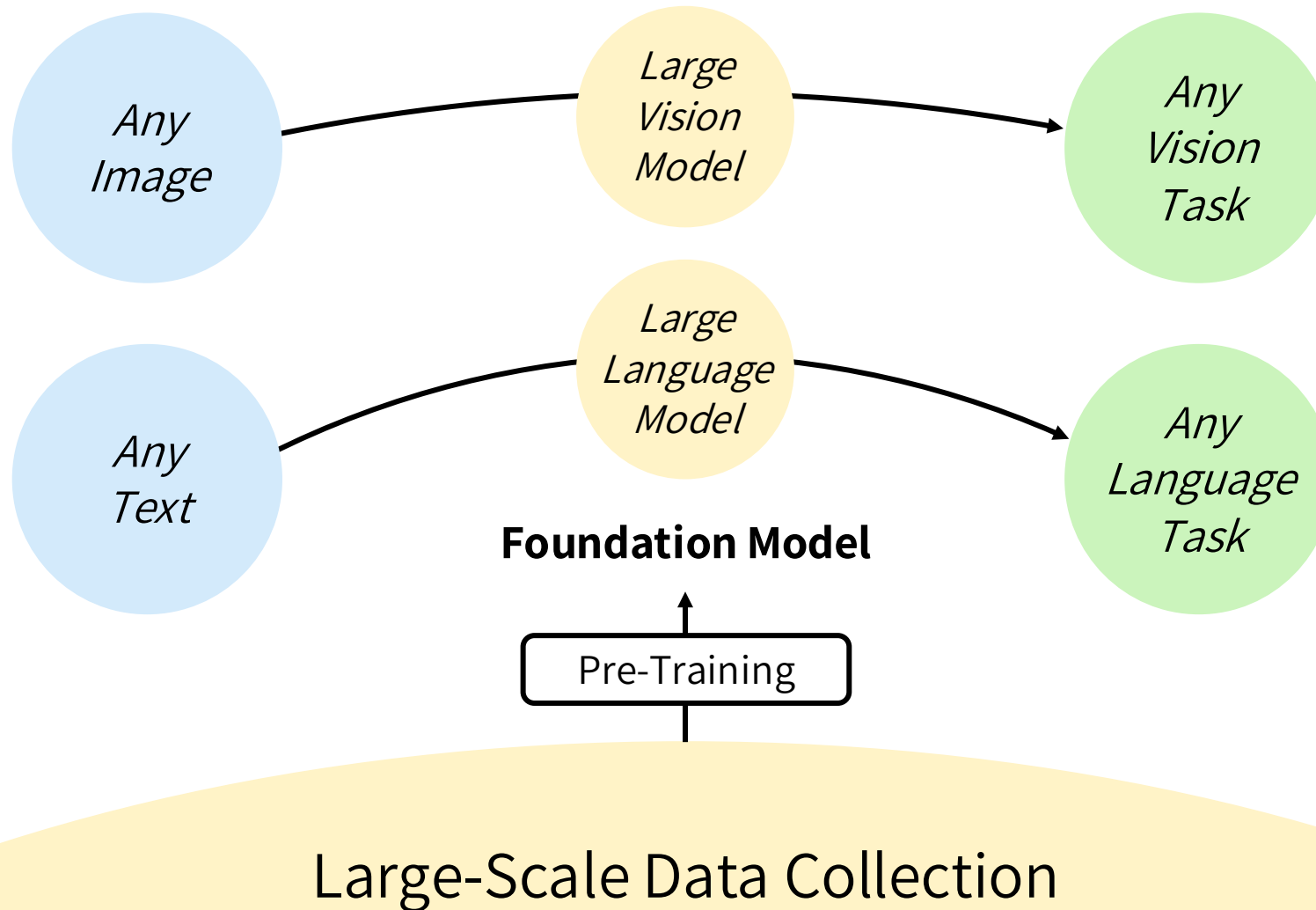
Scalability Challenges: Narrow, Task-Specific Pipelines



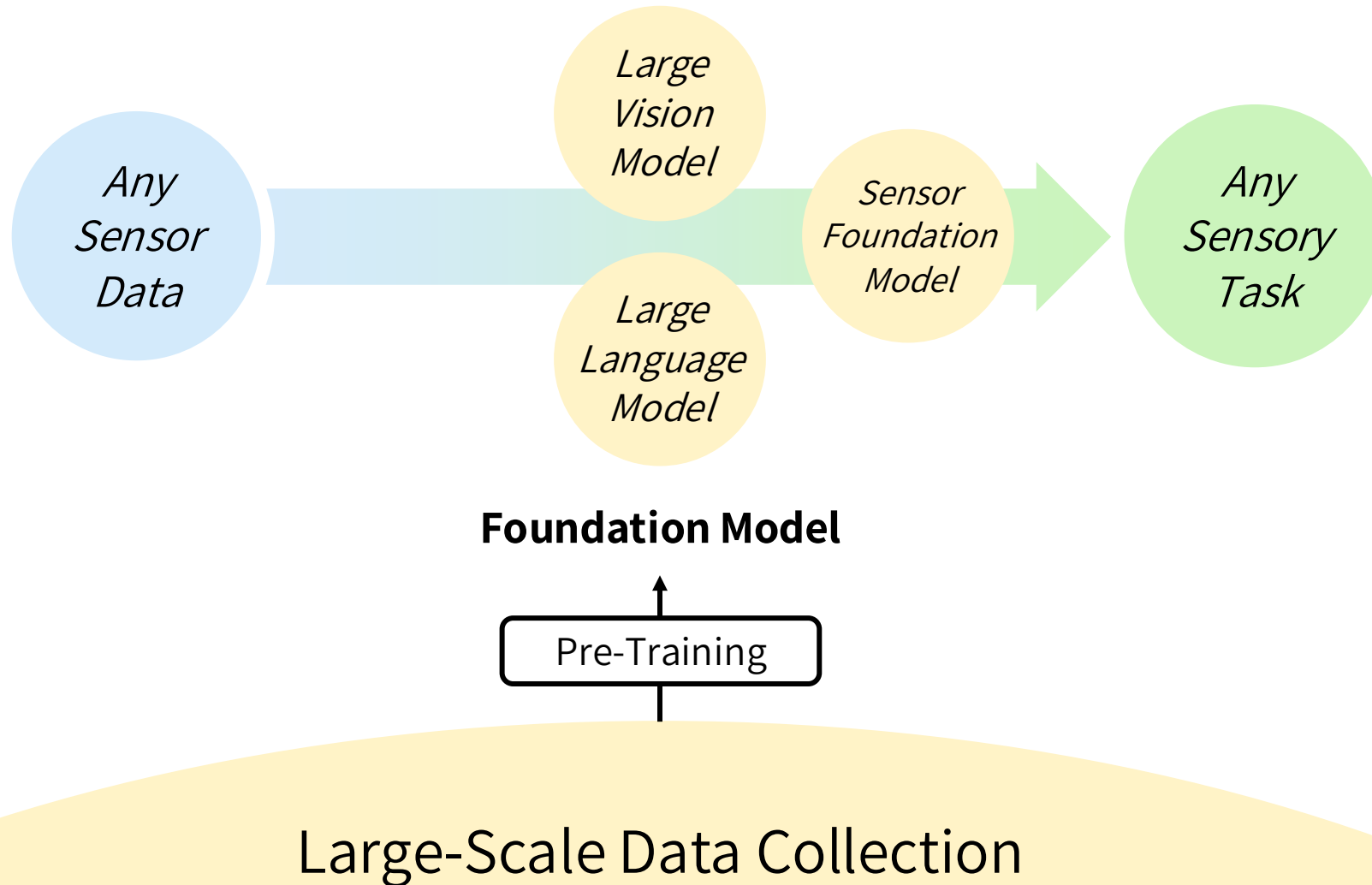
Scalability Challenges: Narrow, Task-Specific Pipelines



Opportunity: Large-Scale Pre-Trained Models



Opportunity: Large-Scale Pre-Trained Models



Dissertation Statement

This dissertation demonstrates how **large-scale pre-trained models** can serve as a foundation for **mobile sensing**, moving the field beyond narrow task-specific pipelines toward systems that **generalize across tasks, users, and modalities**.

Dissertation Outline

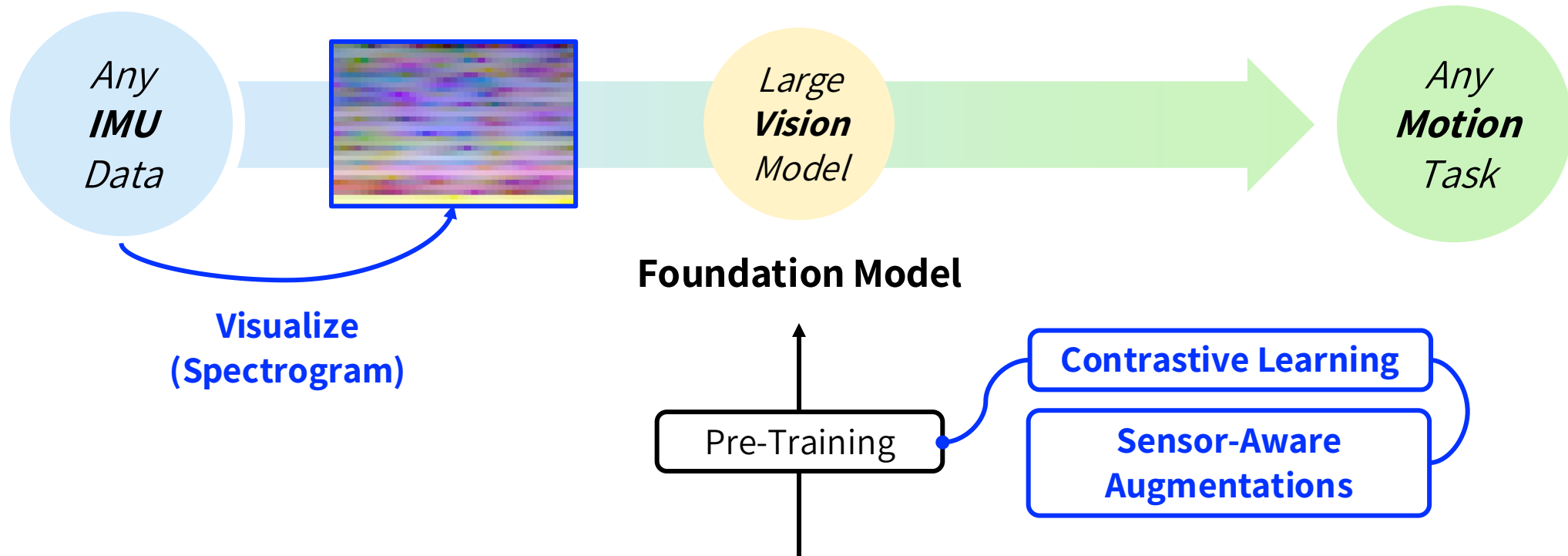
- **[Chapter 1] From Vision to Motion** (IEEE TMC '25)
Translating Large-Scale Vision Knowledge to Data-Scarce IMU Applications
- **[Chapter 2] SelfReplay** (SenSys '25)
Adapting Pre-Trained Sensory Models to Heterogeneous Users
- **[Chapter 3] By My Eyes** (EMNLP '24)
Grounding Multimodal Large Language Models with Sensor Data
- **[Chapter 4] ConSensus** (ACL '26 Findings)
Multi-Agent Collaboration for Multimodal Sensing Tasks
- **[Chapter 5] Beyond Hearing** (ICLR '26)
Learning a Sensor Foundation Model for Earphone-Based Physiological Signals

Dissertation Outline

- **[Chapter 1] From Vision to Motion** (IEEE TMC '25)
Translating Large-Scale Vision Knowledge to Data-Scarce IMU Applications
- **[Chapter 2] SelfReplay** (SenSys '25)
Adapting Pre-Trained Sensory Models to Heterogeneous Users
- **[Chapter 3] By My Eyes** (EMNLP '24)
Grounding Multimodal Large Language Models with Sensor Data
- **[Chapter 4] ConSensus** (ACL '26 Findings)
Multi-Agent Collaboration for Multimodal Sensing Tasks
- **[Chapter 5] Beyond Hearing** (ICLR '26)
Learning a Sensor Foundation Model for Earphone-Based Physiological Signals

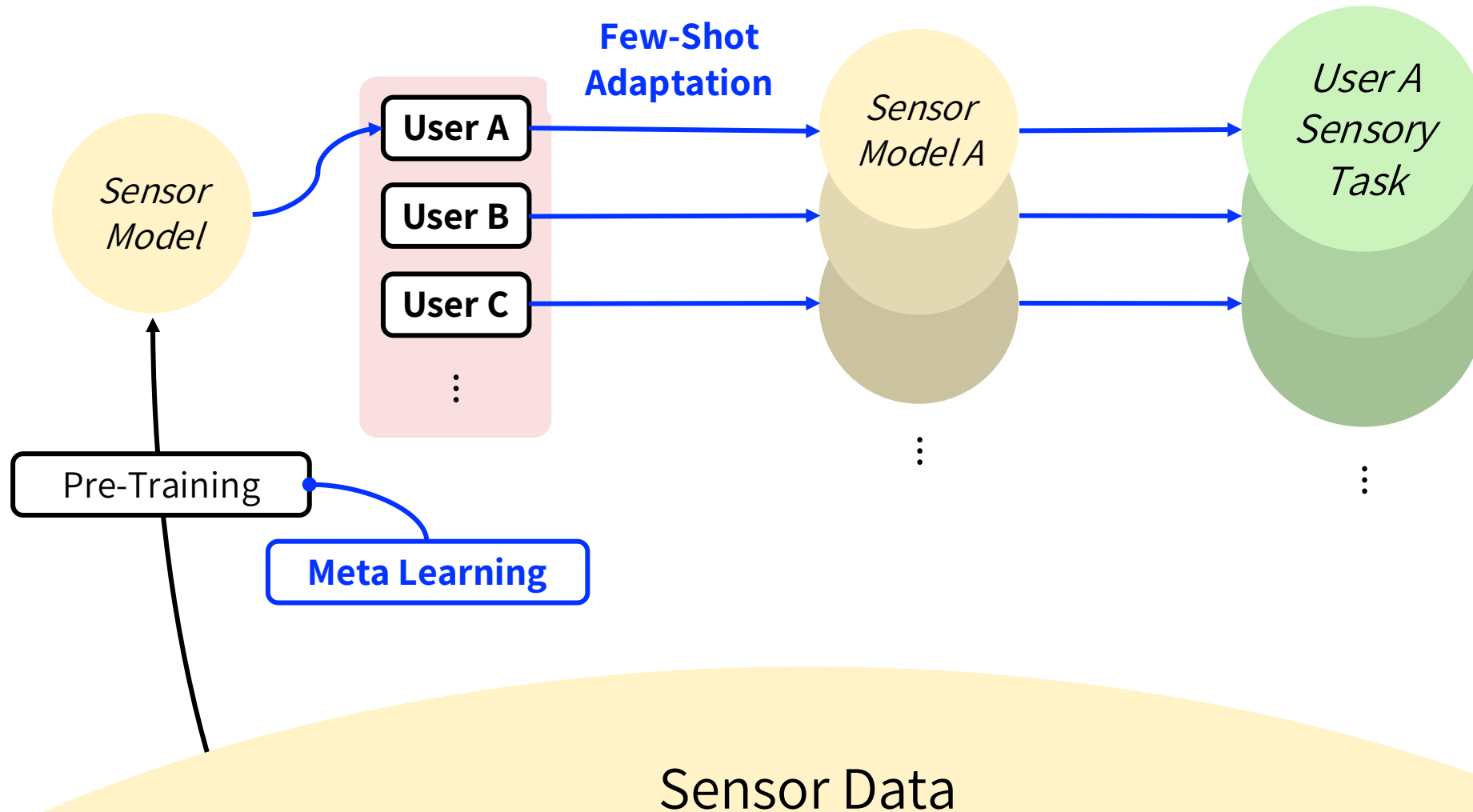
Chapter 1: From Vision to Motion

Translating Large-Scale Vision Knowledge to Data-Scarce IMU Applications



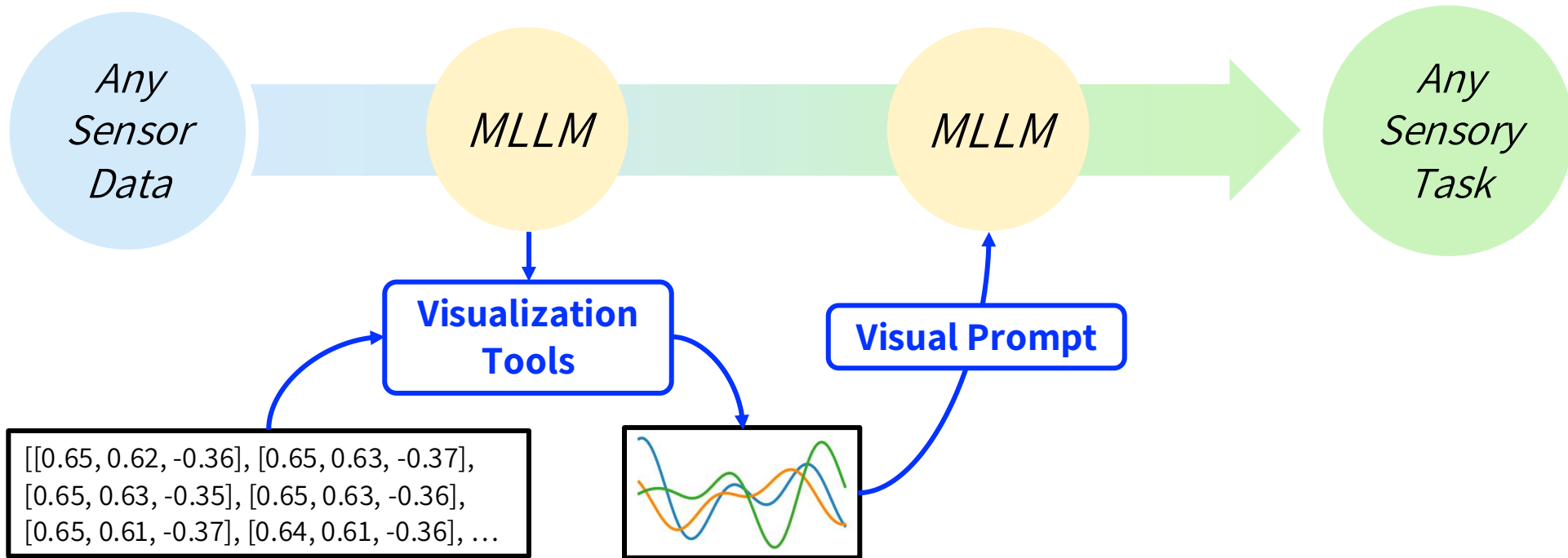
Chapter 2: SelfReplay

Adapting Pre-Trained Sensory Models to Heterogeneous Users



Chapter 3: By My Eyes

Grounding Multimodal Large Language Models with Sensor Data



World Knowledge

Chapter 3: By My Eyes

Grounding Multimodal Large Language Models with Sensor Data

By My Eyes:
Grounding Multimodal Large Language Models
with Sensor Data via Visual Prompting

Smartwatch Demo

Updates from Proposal Feedback

→ **Real-World Applications**

- Multimodal sensing scenarios (Chapter 4)
- Real-world data collection (Chapter 5)

→ **Sensor Foundation Model Challenges**

- Hardware–data–model co-design (Chapter 5)

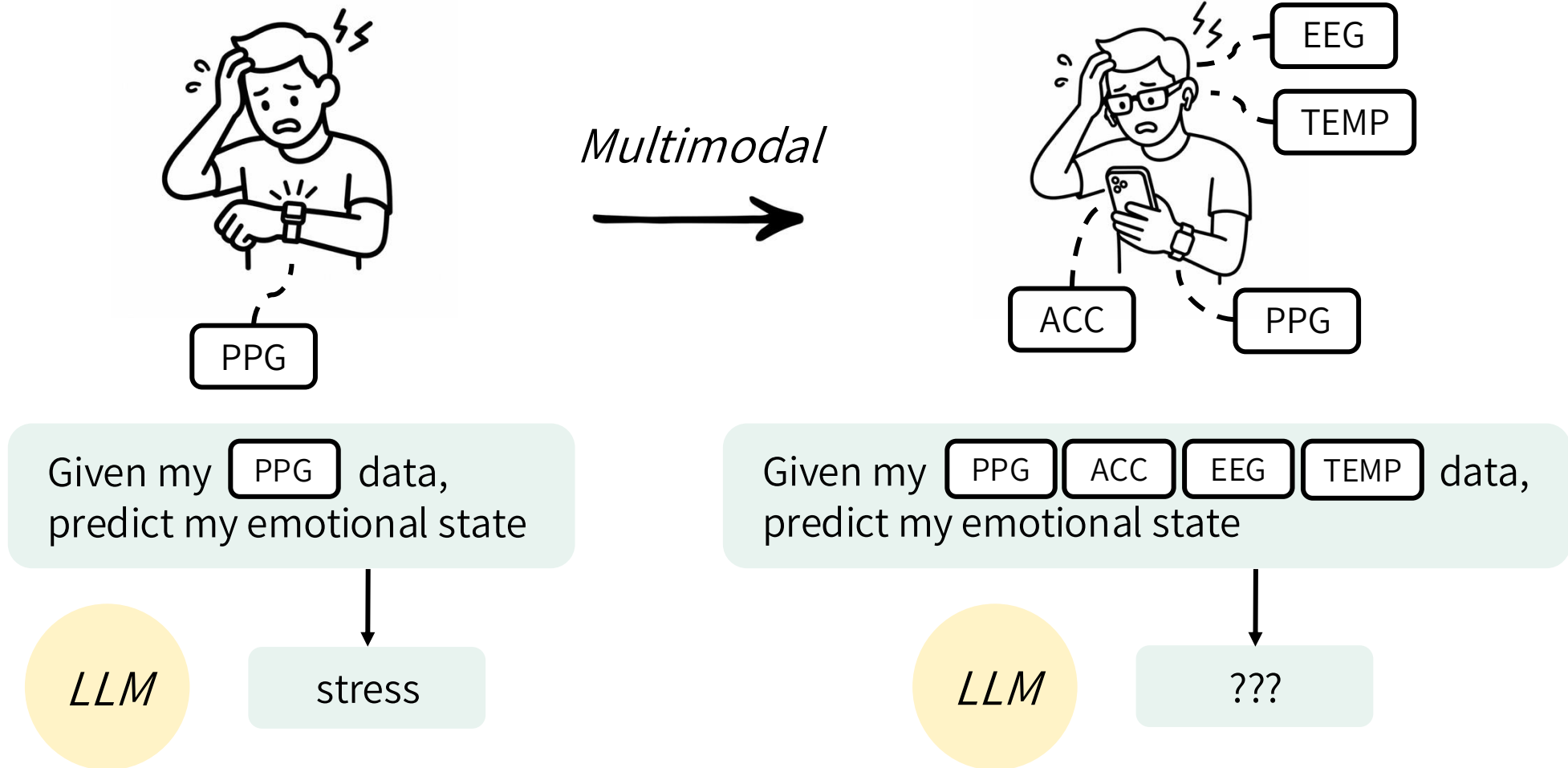
→ **Future Directions**

- Personalization of health LLMs with physiological data
- Extending physical AI toward Sense-Vision-Language-Action Models

Dissertation Outline

- [Chapter 1] **From Vision to Motion** (IEEE TMC '25)
Translating Large-Scale Vision Knowledge to Data-Scarce IMU Applications
- [Chapter 2] **SelfReplay** (SenSys '25)
Adapting Pre-Trained Sensory Models to Heterogeneous Users
- [Chapter 3] **By My Eyes** (EMNLP '24)
Grounding Multimodal Large Language Models with Sensor Data
- [Chapter 4] **ConSensus** (ACL '26 Findings)
Multi-Agent Collaboration for Multimodal Sensing Tasks
- [Chapter 5] **Beyond Hearing** (ICLR '26)
Learning a Sensor Foundation Model for Earphone-Based Physiological Signals

Grounding LLMs with **Multimodal** Sensor Data



Observation 1) Single-Agent Inference

Ground truth: amusement

User: Given the following sensor features, predict the emotional state from [baseline, stress, amusement].

Accelerometer: {ACC features}

ECG: {ECG features}

EMG: {EMG features}

EDA: {EDA features}

Temperature: {TEMP features}

Respiration: {RESP features} ...

Assistant: The *heart-rate* is elevated (≈ 75 bpm) compared to baseline. *EDA dynamic range* (0.12) less than amusement. *Respiratory stretch* is markedly increased (≈ 18 s), indicating stress-hyperventilation. Overall, the pattern aligns best with **stress**.

A single agent often produces an **incomplete interpretation** in multimodal sensing tasks.

To ensure completeness, we decompose interpretation into **multiple agents**.

Observation 2) Modality Agents + Judge Agent

Ground truth: amusement

(after observing {ECG features})

ECG agent: Features show a very low high-frequency (HF) HRV component ($3.02e-03$) and a very high low-frequency to HF ratio (15.22), indicating **stress**.

ACC agent: ... is **amusement**.

EMG agent: ... is **amusement**.

EDA agent: ... is **amusement**.

TEMP agent: ... is **amusement**.

...

Judge agent: The most important modality is ECG. With its very low HF HRV component, the answer is **stress**.

An LLM-based judge is **prone to knowledge bias**, even when robust evidence exists in other modalities.

Majority voting can mitigate knowledge bias by leveraging statistical aggregation.

Observation 3) Data with 30% Missingness

Ground truth: amusement (missing)

ACC agent: ... is
baseline.

ECG agent: ... is
stress.

EMG agent: ... is
stress.

EDA agent: ... is
amusement.

TEMP agent: ... is
amusement.

RESP agent: ... is
stress.

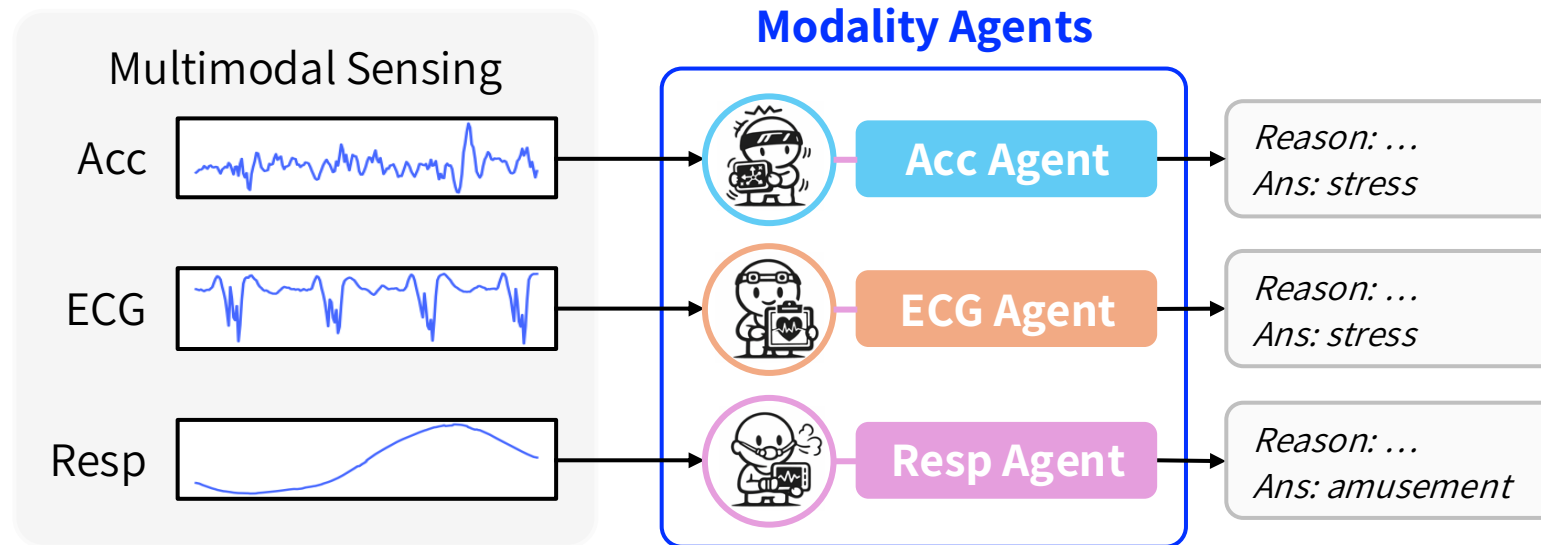
→ Majority Voting: **stress**

Judge agent: *The ECG and respiration features appear unreliable. In contrast, the EDA modality provides robust evidence, and the temperature features show consistent indicators. Based on these signals, the predicted emotional state is **amusement**.*

Majority voting is prone to **sensor failure**, often producing unreliable votes.

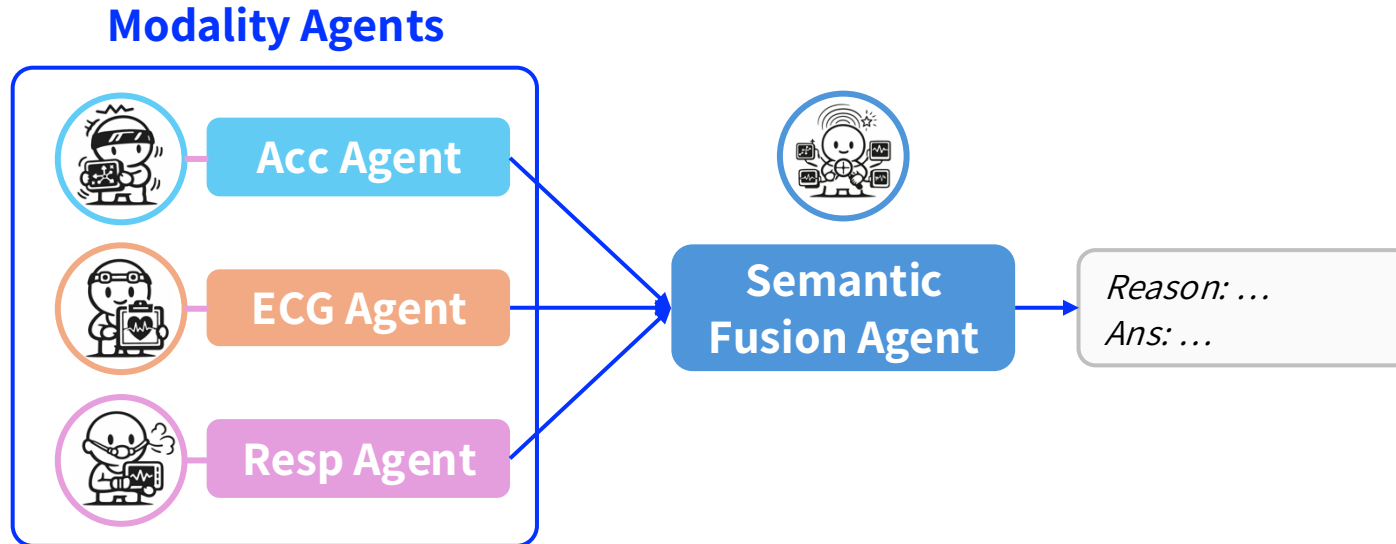
Observations 1-3 jointly expose a fundamental tension between **semantic-level** and **statistical aggregation**.

ConSensus: Multi-Agent Collaboration for Multimodal Sensing



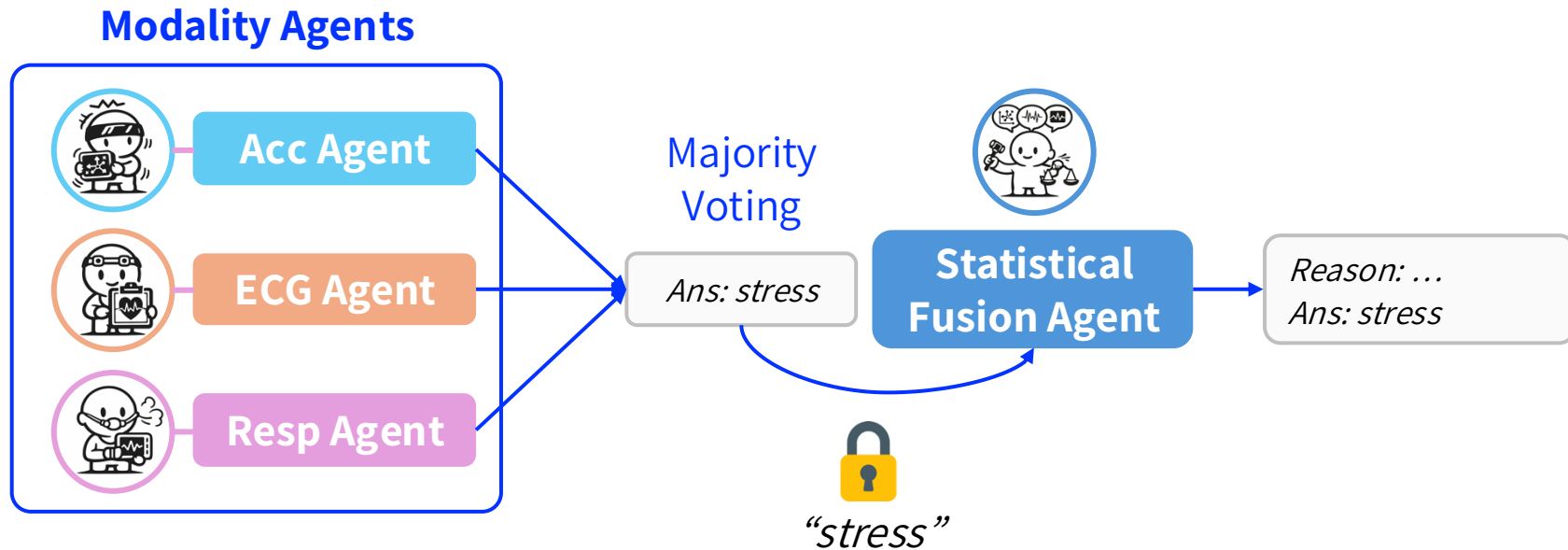
We define **Modality Agents**,
where each agent is assigned to a single modality
and produces sensor interpretation for the task

ConSensus: Multi-Agent Collaboration for Multimodal Sensing



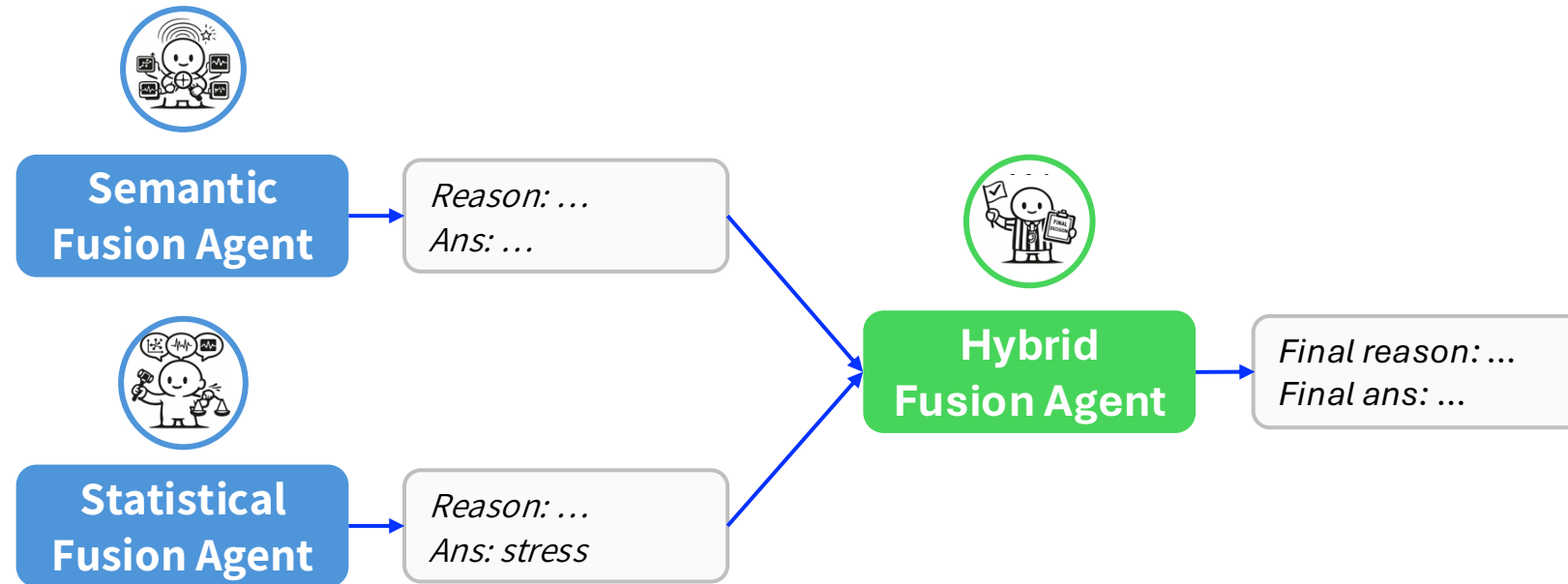
A **Semantic Fusion Agent** performs a knowledge-ground aggregation that synthesizes semantic evidence across modality agents

ConSensus: Multi-Agent Collaboration for Multimodal Sensing



A **Statistical Fusion Agent** performs a consensus-grounded aggregation by anchoring its answer to a majority-voted answer

ConSensus: Multi-Agent Collaboration for Multimodal Sensing



A **Hybrid Fusion Agent** arbitrates the outputs from semantic and statistical fusion agents as a coordinator role.

Experiments: Datasets

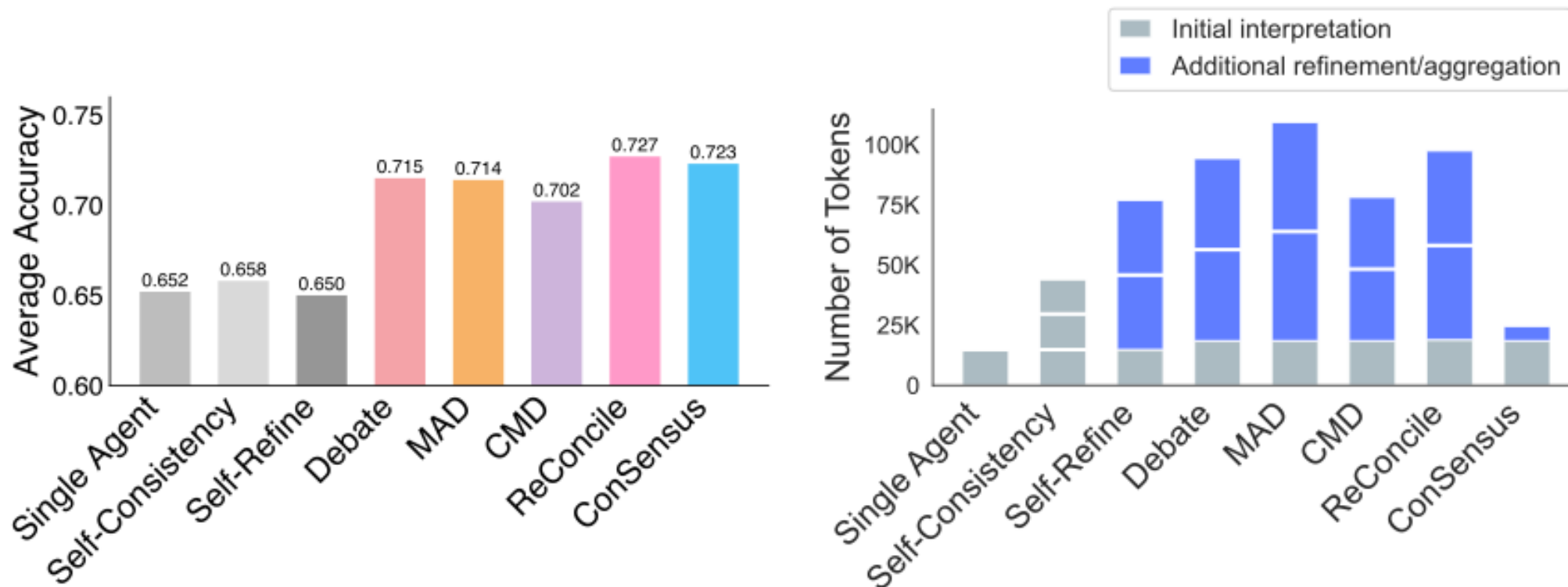
Dataset	Task	#Classes	Sensor modalities
WESAD	Affective state recognition	3	ACC ECG EMG EDA TEMP RESP (Chest) ACC PPG EDA TEMP (Wrist)
SleepEDF	Sleep stage classification	5	EEG (Fpz-Cz) EEG (Pz-Oz) EOG EMG RESP
ActionSense	Kitchen activity recognition	5	ACC ANG EMG (Left arm), ACC ANG EMG (Right arm)
MMFit	Gym exercise recognition	10	ACC GYR × (Ear Left wrist Right wrist Waist) MAG (Waist) HR × (Left wrist Right wrist)
PAMAP2	Daily activity recognition	12	ACC GYR × (Hand Chest Ankle)

5 different multimodal sensing benchmarks
12 distinct sensing modalities

Experiments: Setup

- **Prompt:** Extracted Features + 1-Shot Examples + Chain-of-Thought (CoT)
- **Model:** gpt-oss-20B
- **Baselines:**
 - *Single-Agent Baselines*
 - CoT (NeurIPS '22)
 - Self-Consistency (ICLR '23)
 - Self-Refine (NeurIPS '23)
 - *Multi-Agent Debate Baselines*
 - Debate (ICML '24): Majority voting after discussion rounds
 - MAD (EMNLP '24): Judge makes the final decision
 - CMD (ACL '24): Group-based voting
 - ReConcile (ACL '24): Confidence-based voting

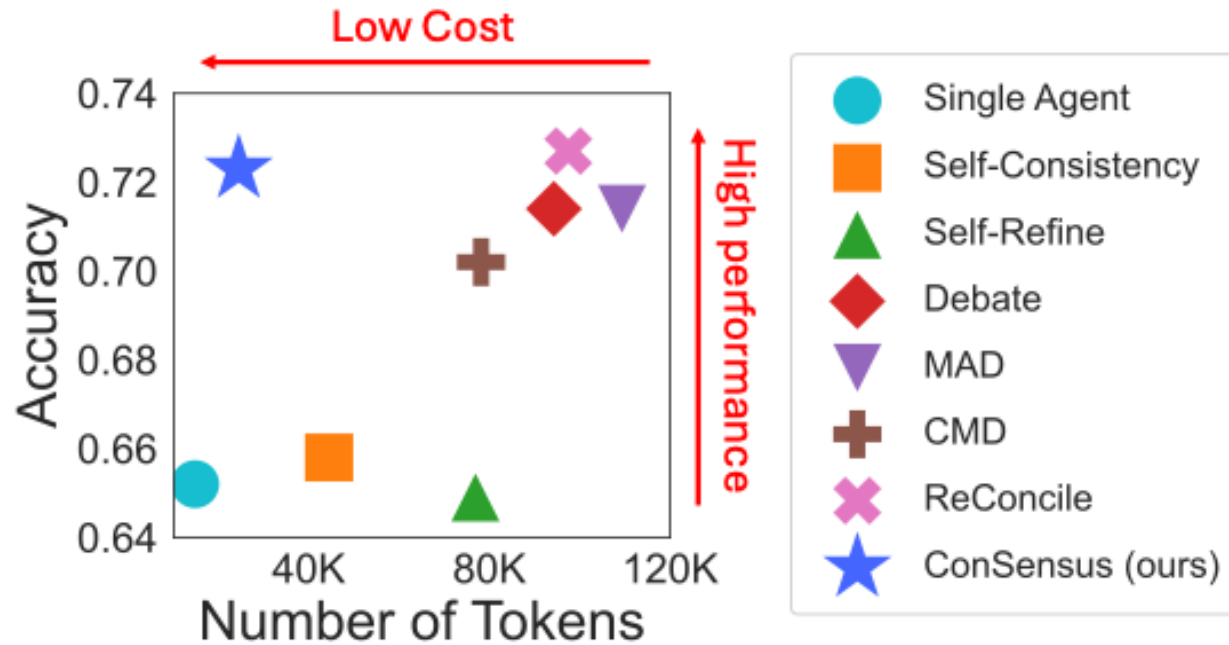
Experiments: Main Results



+7.1% accuracy gain over the single-agent baseline

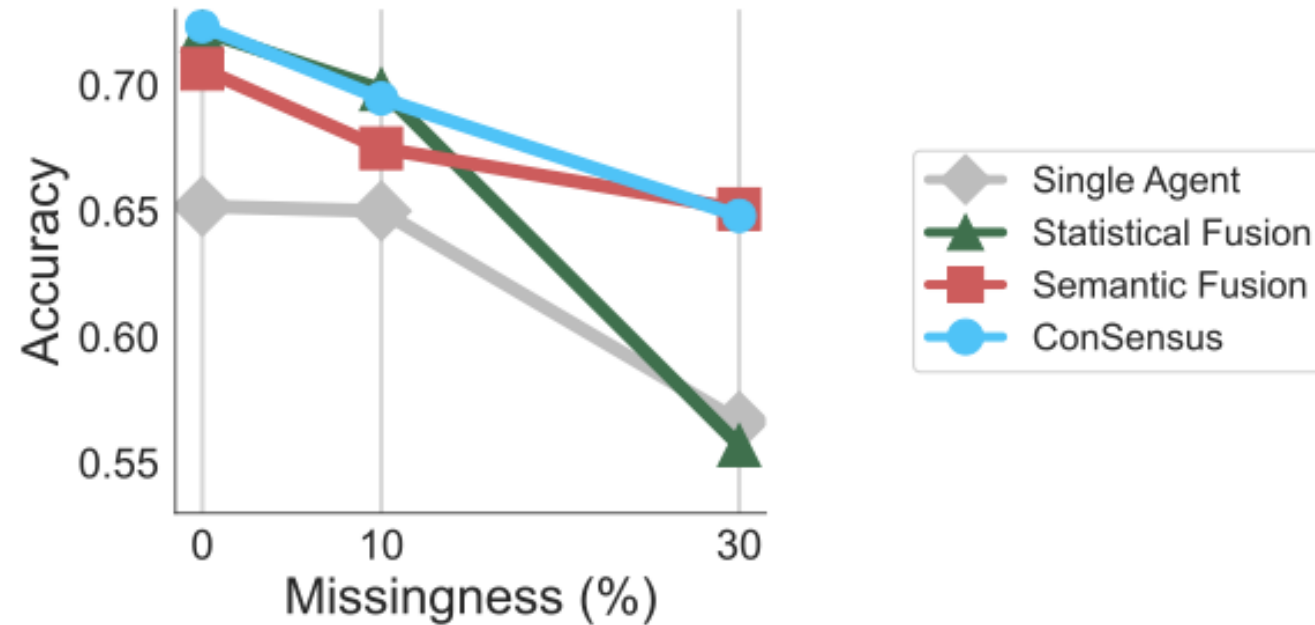
12.7x lower token cost than multi-agent debate methods

Experiments: Main Results



+7.1% accuracy gain over the single-agent baseline
12.7x lower token cost than multi-agent debate methods

Experiments: Robustness to Sensor Failure



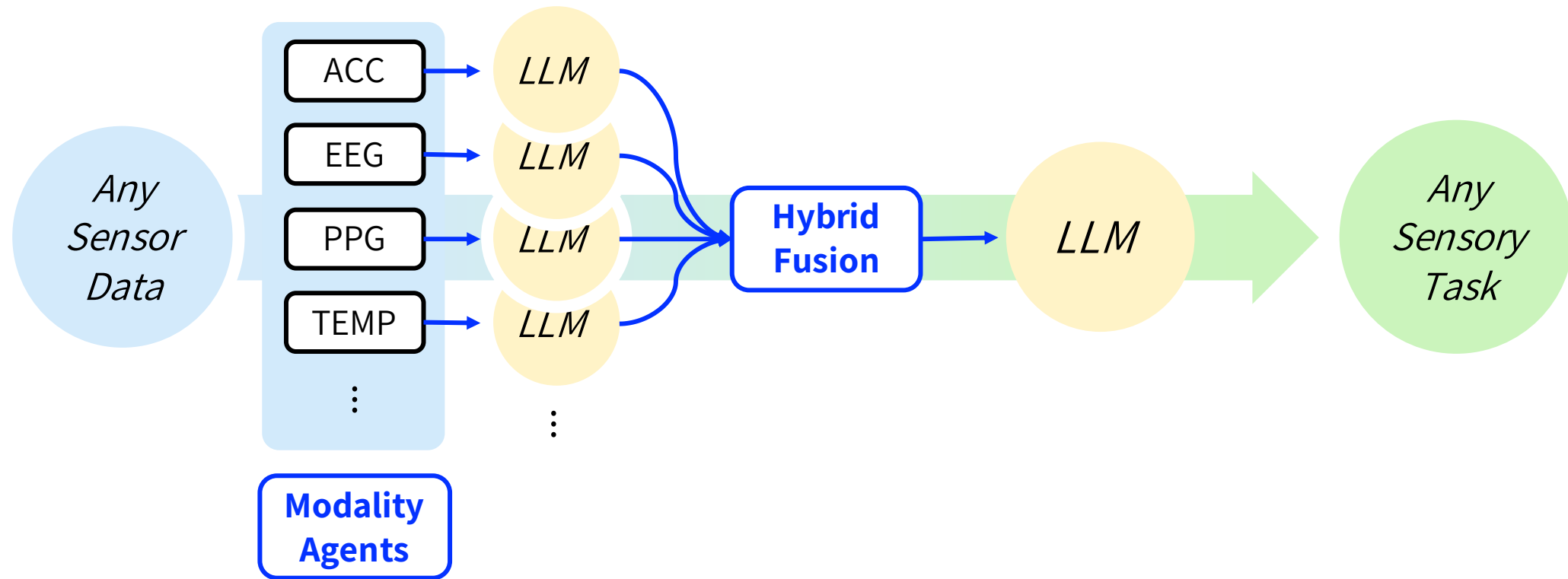
On clean data, Statistical > Semantic

On missing data, Semantic > Statistical

ConSensus achieves the best overall performance

Chapter 4: ConSensus

Multi-Agent Collaboration for Multimodal Sensing Tasks

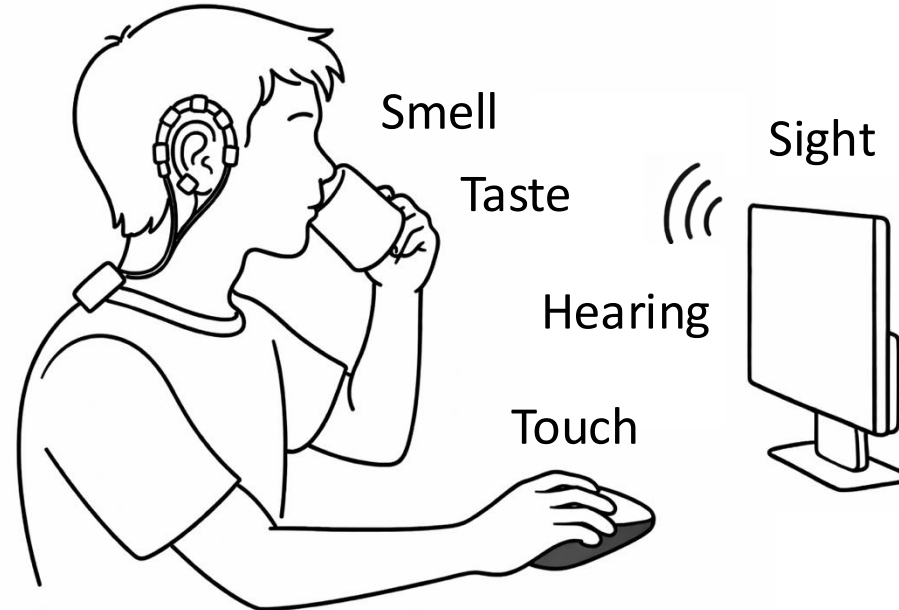


TL;DR Modality agents + hybrid semantic/statistical fusion enable robust and efficient multimodal sensor reasoning.

Dissertation Outline

- [Chapter 1] **From Vision to Motion** (IEEE TMC '25)
Translating Large-Scale Vision Knowledge to Data-Scarce IMU Applications
- [Chapter 2] **SelfReplay** (SenSys '25)
Adapting Pre-Trained Sensory Models to Heterogeneous Users
- [Chapter 3] **By My Eyes** (EMNLP '24)
Grounding Multimodal Large Language Models with Sensor Data
- [Chapter 4] **ConSensus** (ACL '26 Findings)
Multi-Agent Collaboration for Multimodal Sensing Tasks
- [Chapter 5] **Beyond Hearing** (ICLR '26)
Learning a Sensor Foundation Model for Earphone-Based Physiological Signals

Toward **General-Purpose** Sensor Foundation Models



**Goal: General-purpose physiological sensing
across 5 fundamental human senses**

Background: ExG Signals

EMG
from muscles



ECG
from the heart



EEG
from the brain



EOG
from the eyes



**Goal: General-purpose ExG sensing
across 5 fundamental human senses**

Challenges in Building ExG Foundation Models

Hardware

- Bulky lab equipment
- Unsuitable for everyday use

→ **Earphone-based sensing device design**

Data

- Lab-controlled datasets
- Limited real-world diversity

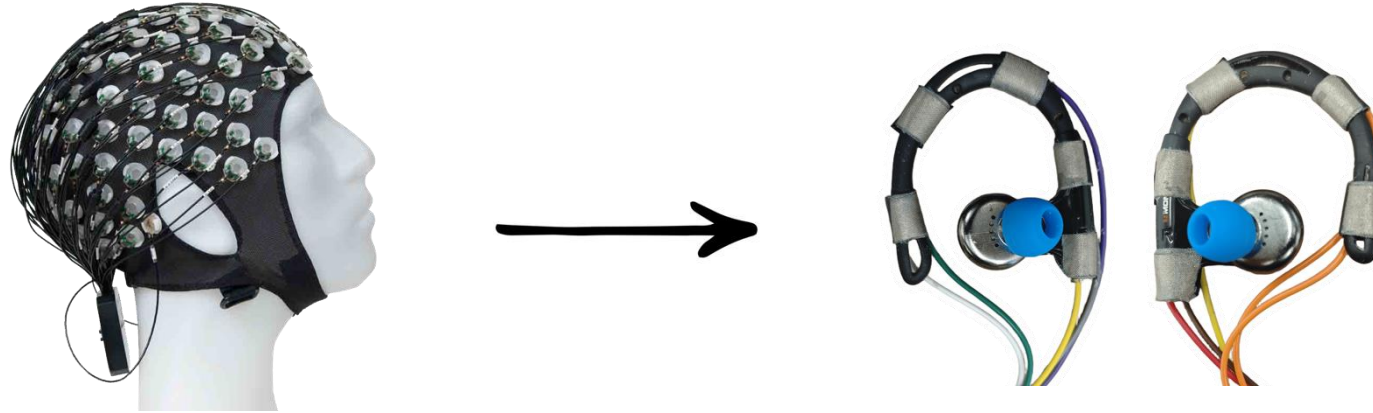
→ **Free-living dataset collection**

Model

- Task-specific model designs
- Hand-crafted frequency filters

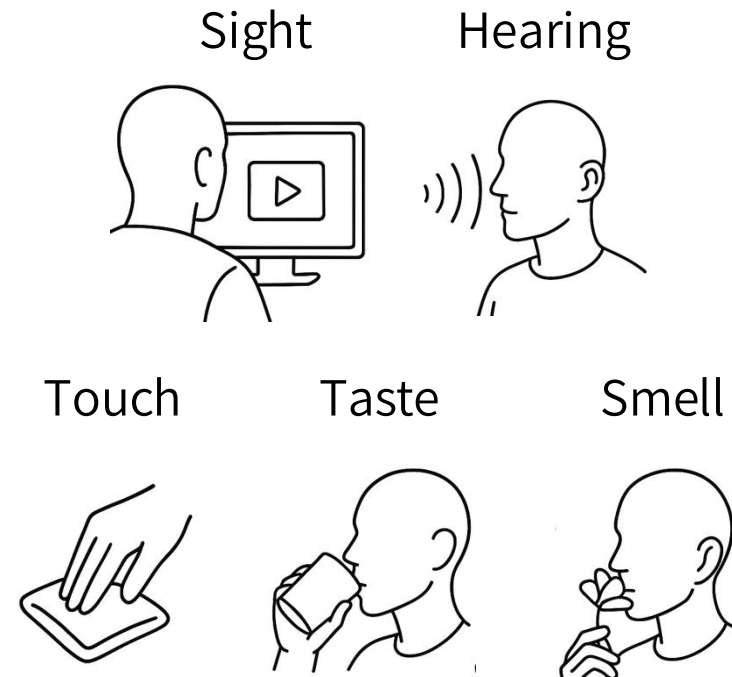
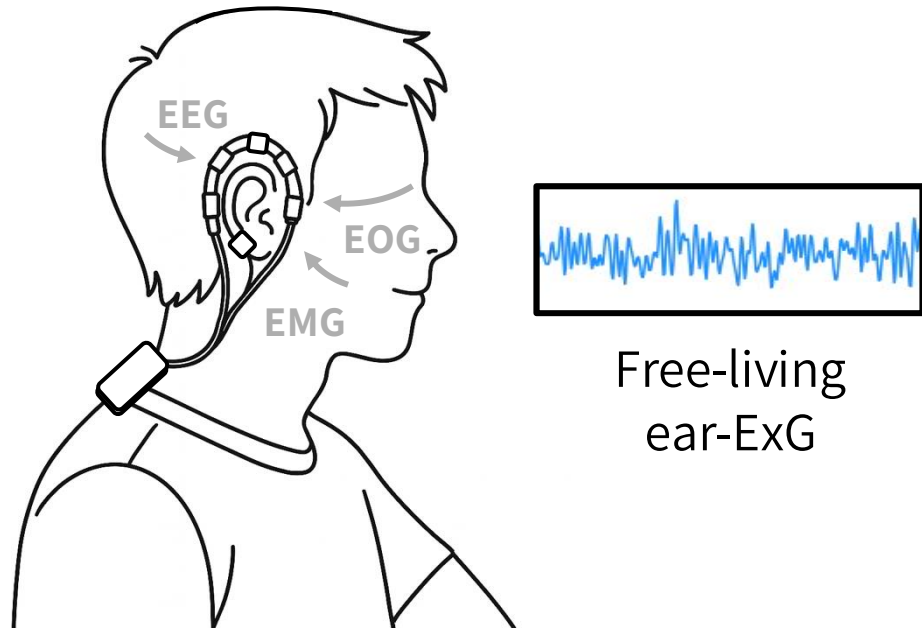
→ **Physiology-informed tokenization**

NeuroBuds: Earphone-Based ExG Sensing



- Designed **NeuroBuds**, an **earphone-shaped** ExG sensing device
- NeuroBuds is comfortable for **daily use** and **long-term** data collection

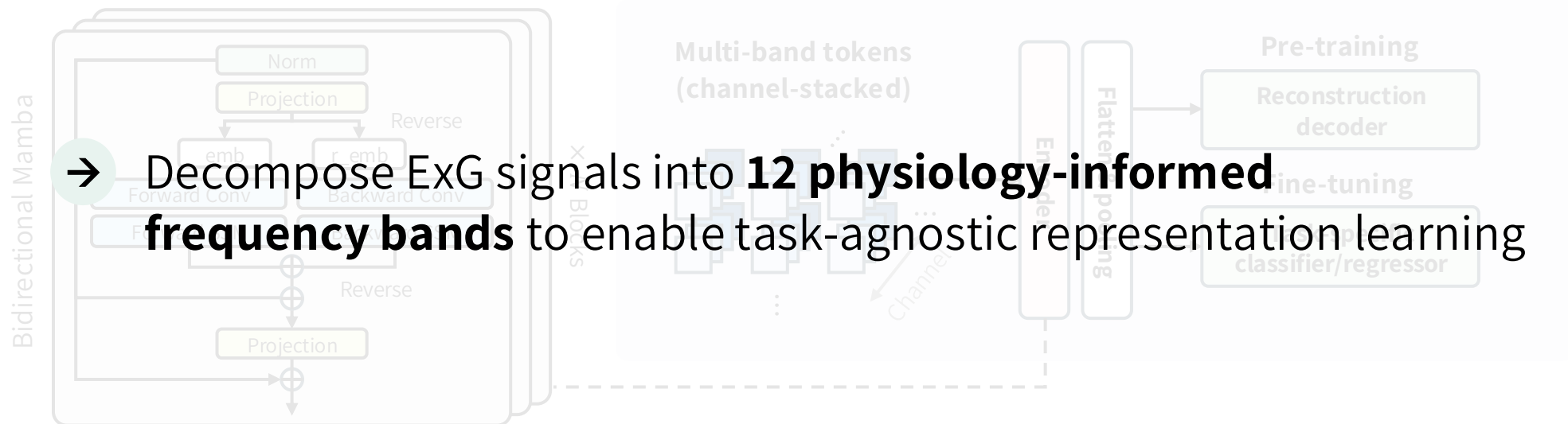
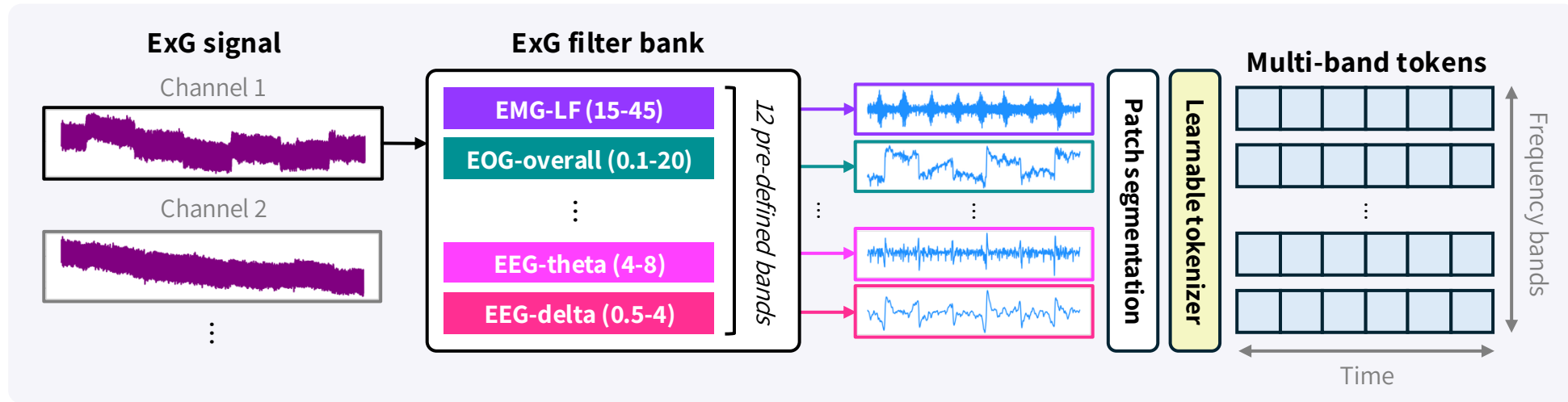
DailySense: Free-Living ExG Data



→ Collected **50 hours of free-living ExG data** from 22 participants

→ **20 hours of labeled recordings** across five human senses

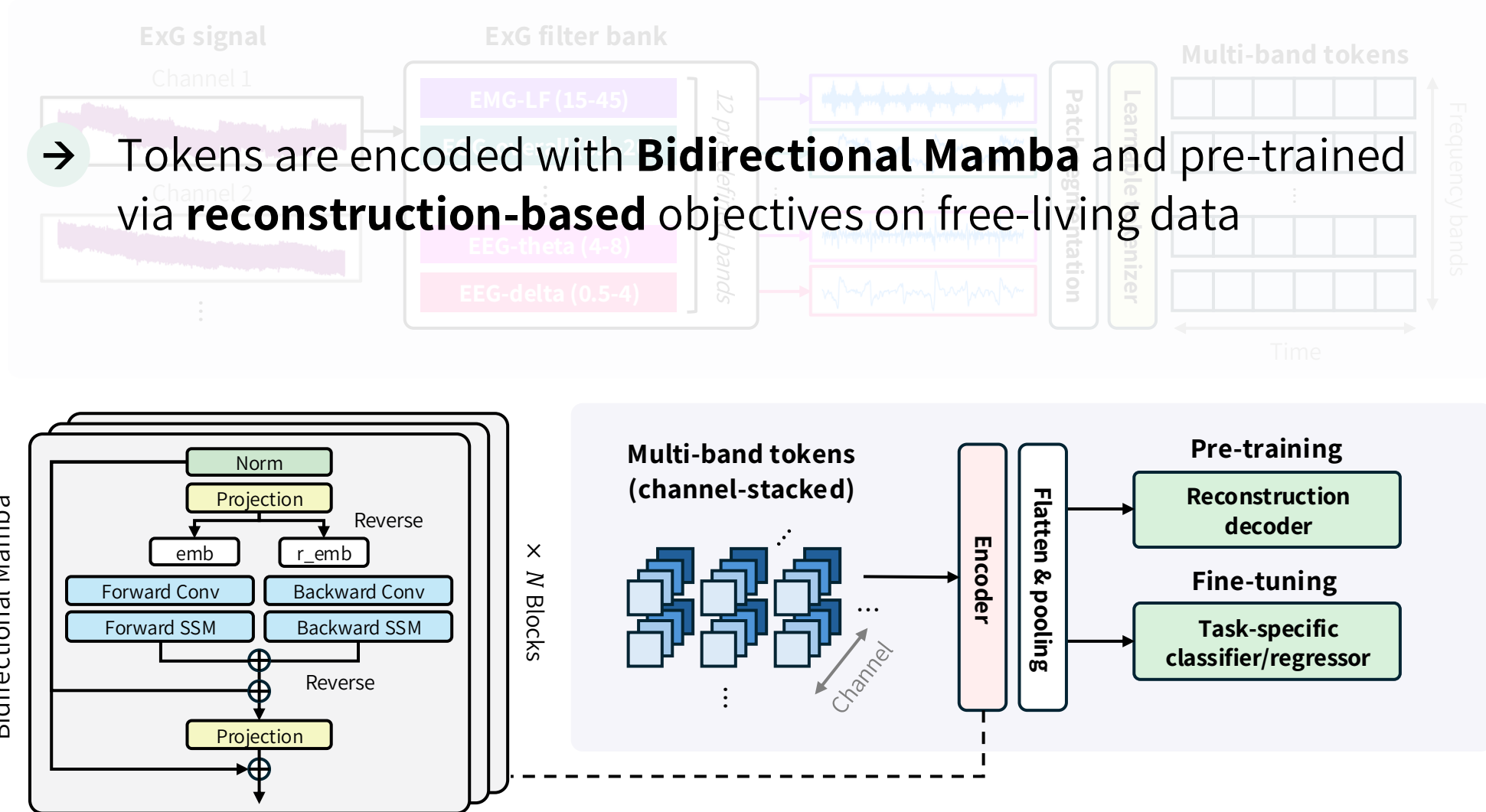
PiMT: Physiology-informed Multi-band Tokenization



→ Decompose ExG signals into **12 physiology-informed frequency bands** to enable task-agnostic representation learning

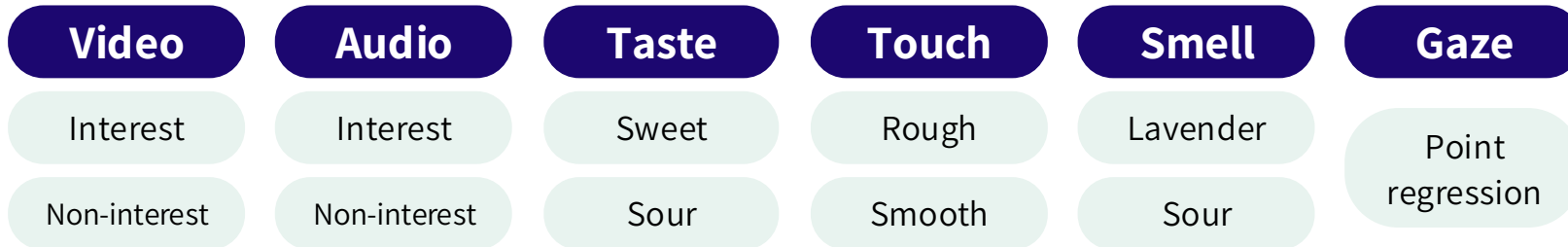
PiMT: Physiology-informed Multi-band Tokenization

→ Tokens are encoded with **Bidirectional Mamba** and pre-trained via **reconstruction-based** objectives on free-living data



Experiments: Setup

→ **6 task-specific** data across **5 fundamental human senses**



→ Conducted experiments on **4 public datasets**

- DREAMER: Emotion recognition
- SEED: Emotion recognition
- Sleep-EDF: Sleep staging
- BCI Competition IV 2b: Motor imagery

→ Compared PiMT against **7 baselines**

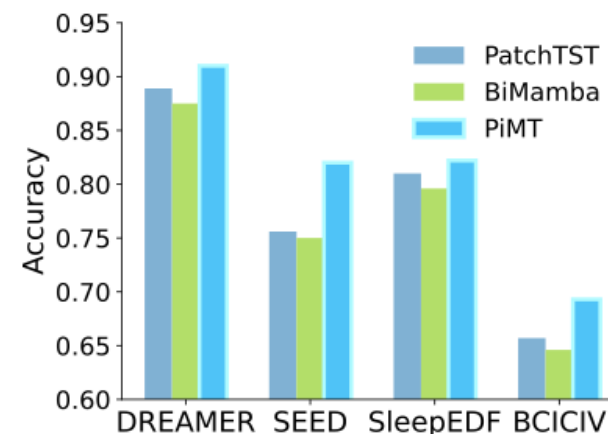
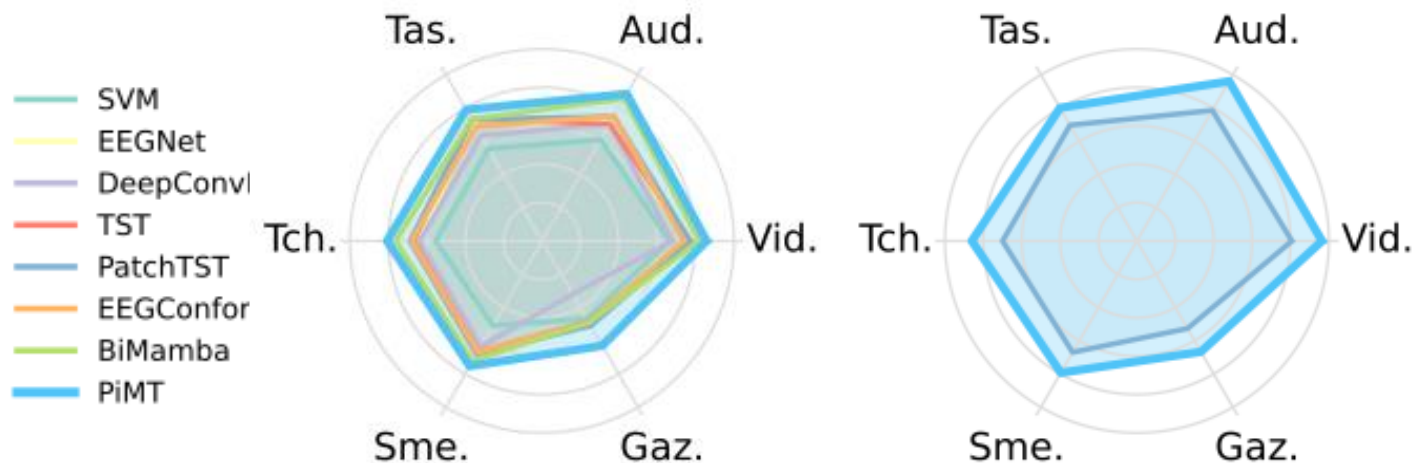
- SVM, EEGNet, DeepConvNet, TST, PatchTST, EEGConformer, BiMamba

Experiments: Main Results

w/o pre-training

with pre-training

Public datasets

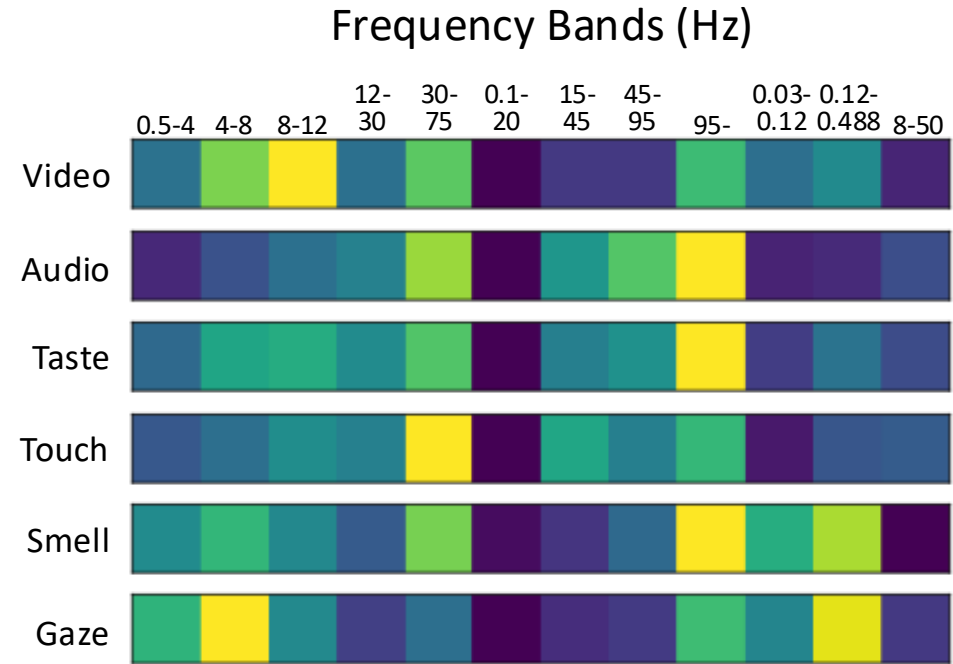
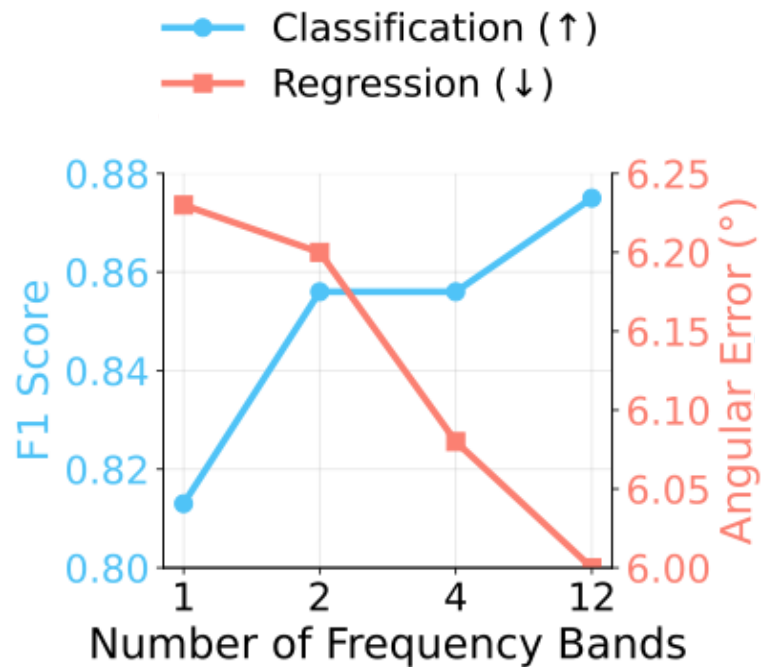


PiMT (no pre-train): **81.9%** (+4% over baselines)

Best on public datasets

PiMT + pre-training: **87.6%** accuracy (+14.4%), **6.00°** gaze tracking error

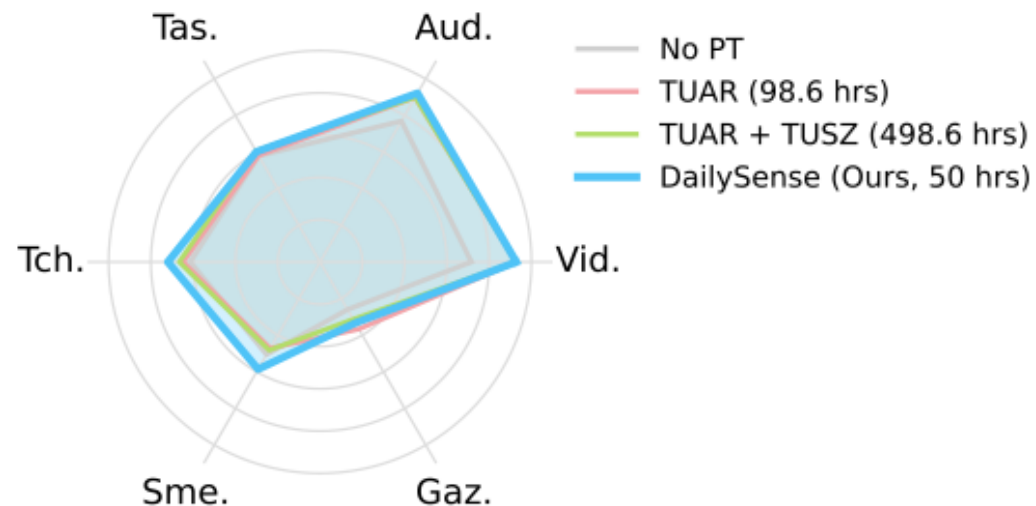
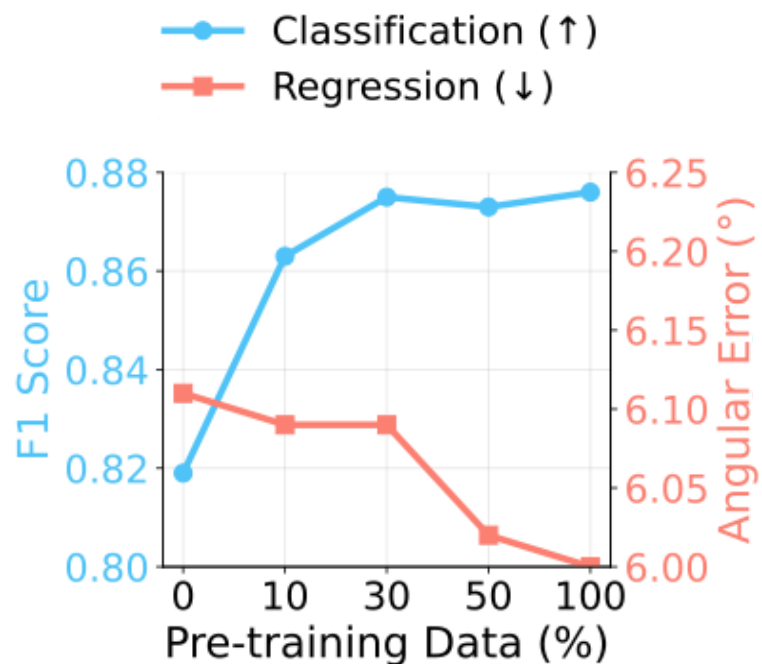
Experiments: Effect of Frequency Bands



12 physiology-informed bands outperform fewer bands

PiMT dynamically captures task-relevant frequency bands

Experiments: Effect of Pre-Training Data

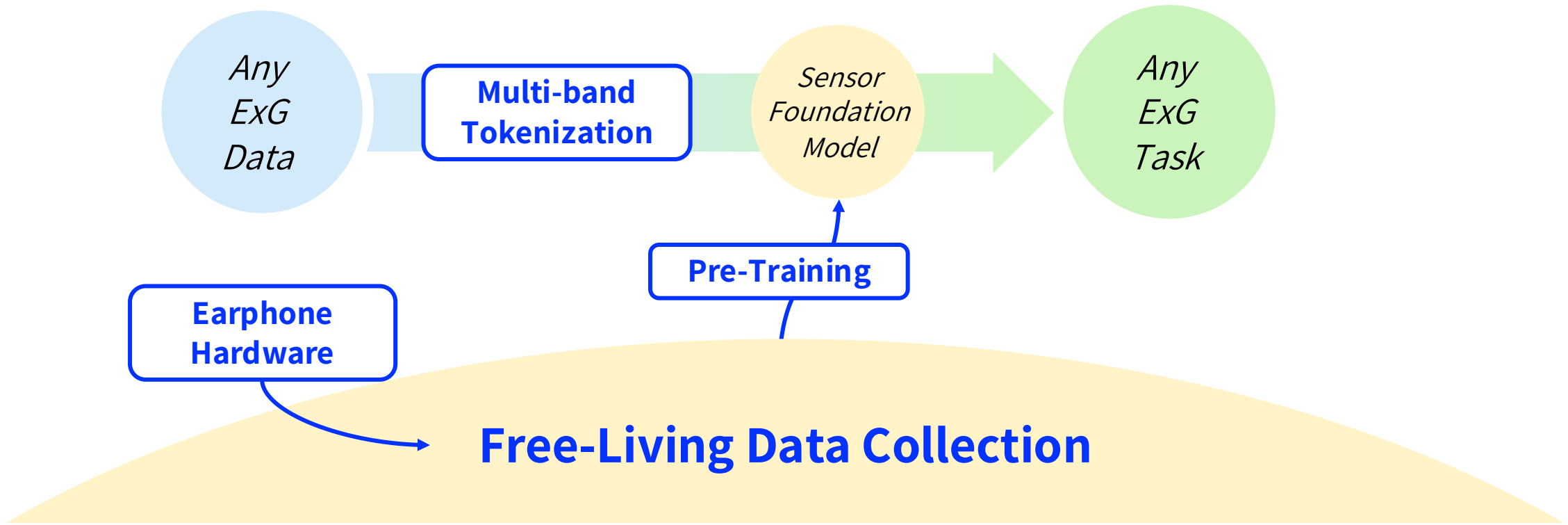


More pre-training →
Better performance (scaling law)

Pre-training on free-living data outperforms
lab-controlled pre-training data

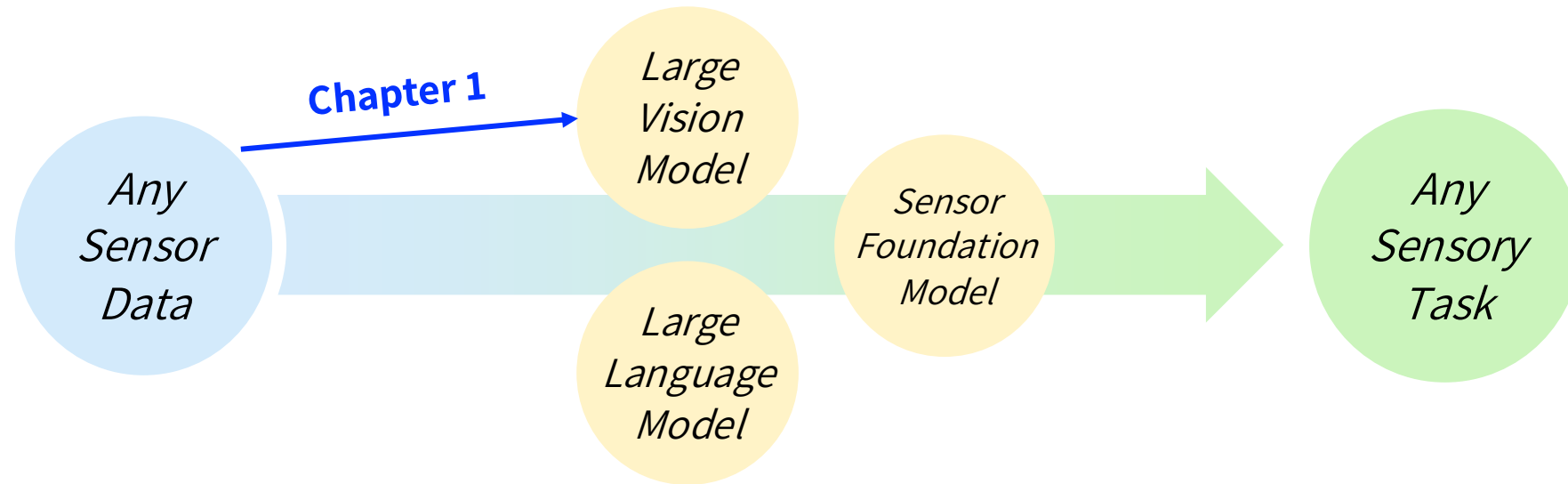
Chapter 5: Beyond Hearing

Learning a Sensor Foundation Model for Earphone-Based Physiological Signals



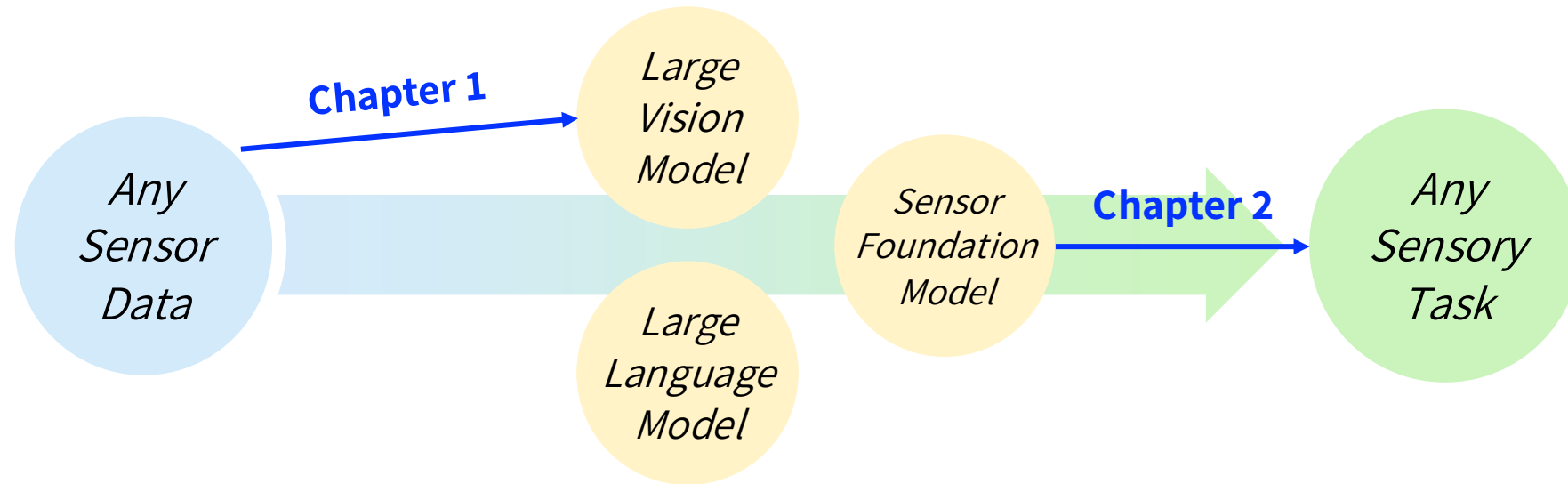
TL;DR Earphone ExG + Free-Living Data + Physiology-Informed Tokenization
→ Task-Agnostic ExG Model

Contribution: Elevating Large-Scale Pre-Trained Models into Foundation Models for Mobile Sensing



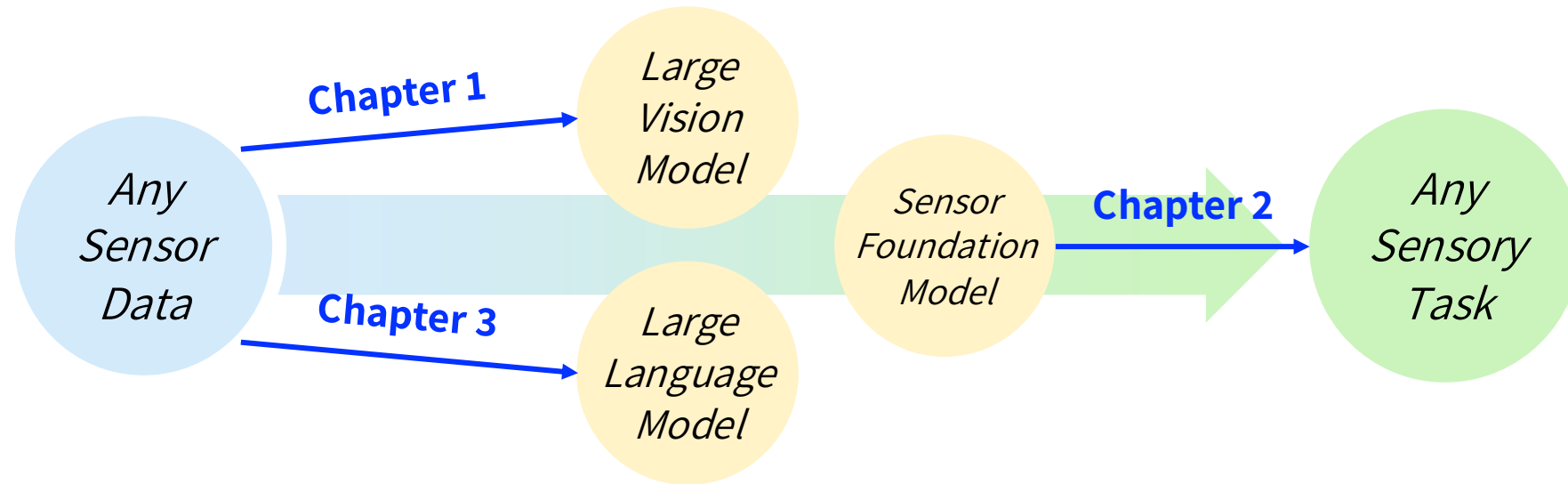
This dissertation contributes a method for bridging the gap between data-rich vision domains and data-scarce sensor domains

Contribution: Elevating Large-Scale Pre-Trained Models into Foundation Models for Mobile Sensing



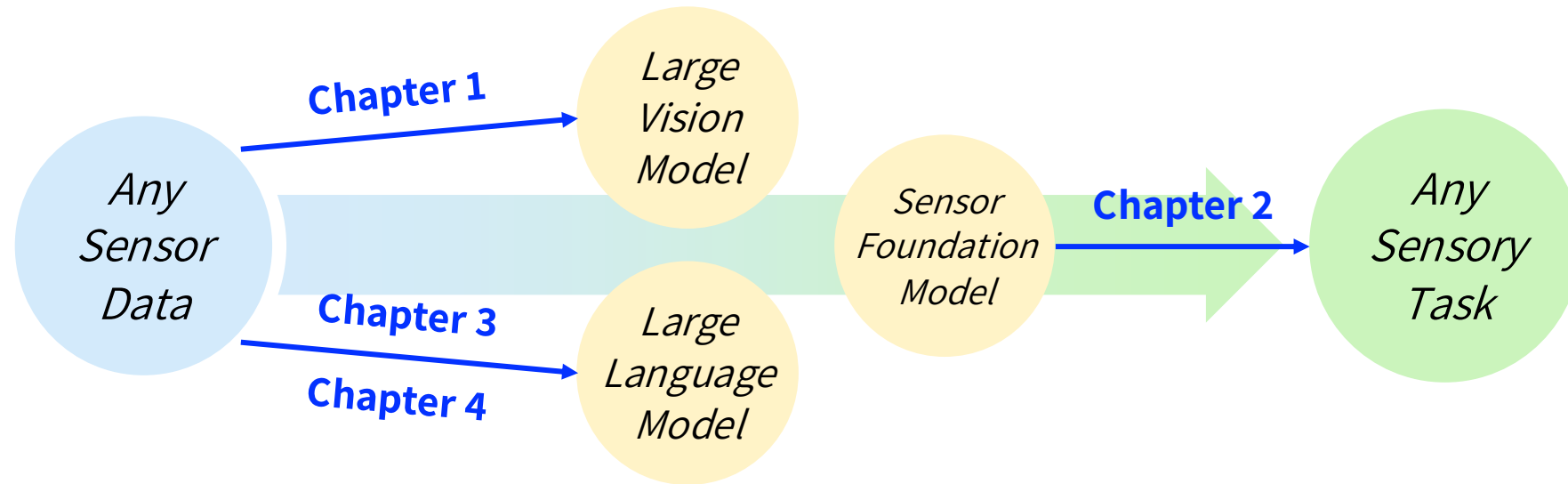
This dissertation contributes an adaptation framework for deploying pre-trained models across heterogeneous users and devices

Contribution: Elevating Large-Scale Pre-Trained Models into Foundation Models for Mobile Sensing



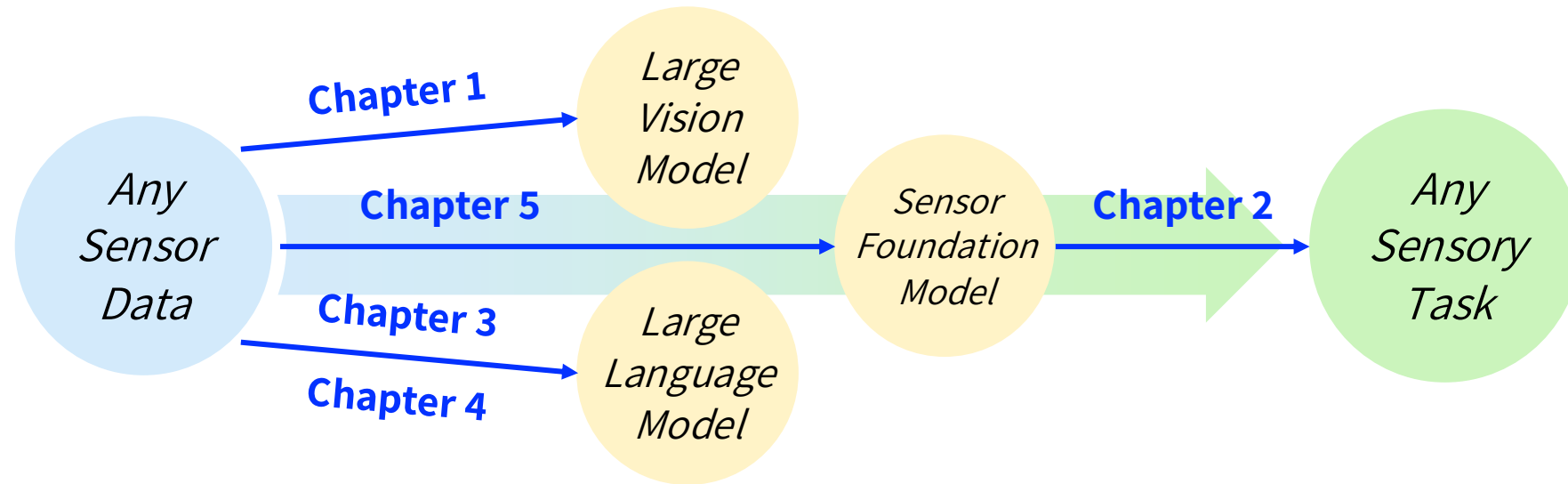
This dissertation contributes an approach for grounding multimodal large language models with sensor data

Contribution: Elevating Large-Scale Pre-Trained Models into Foundation Models for Mobile Sensing



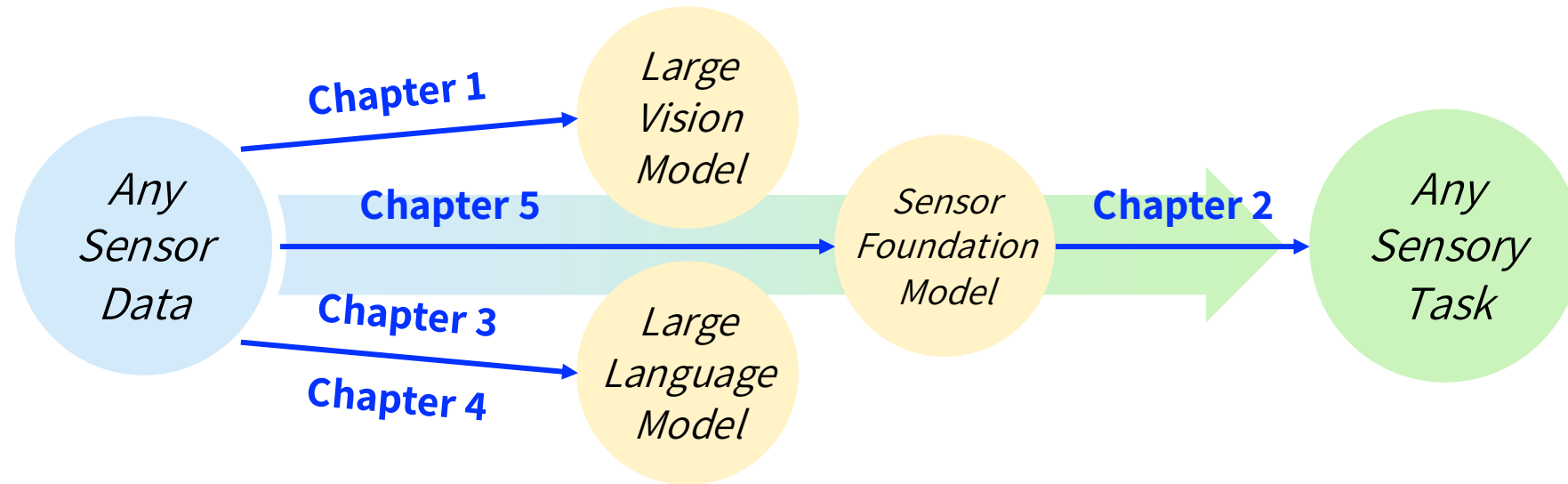
This dissertation contributes a multi-agent framework for multimodal sensing

Contribution: Elevating Large-Scale Pre-Trained Models into Foundation Models for Mobile Sensing



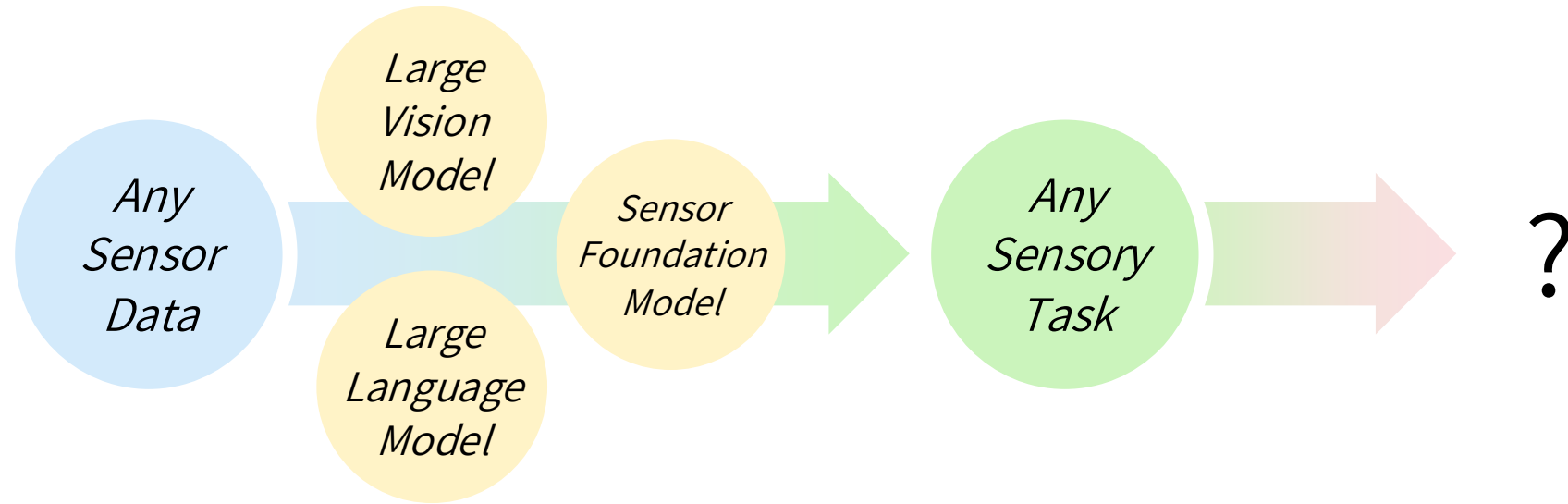
This dissertation presents a sensor foundation model for earphone-based physiological sensing

Contribution: Elevating Large-Scale Pre-Trained Models into Foundation Models for Mobile Sensing

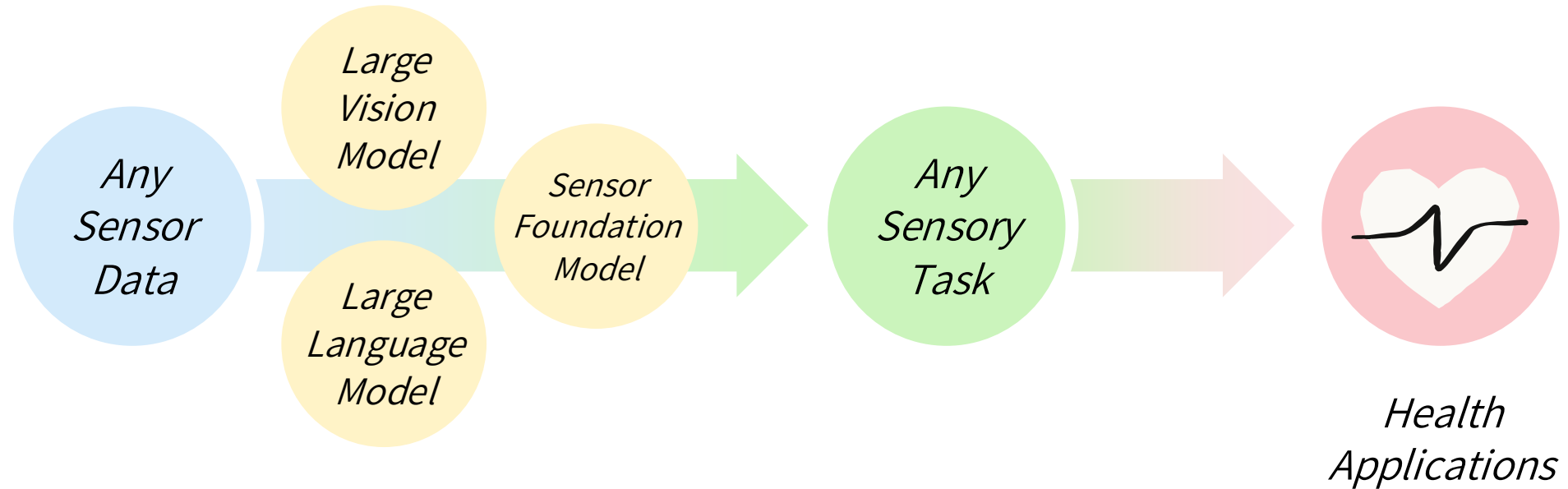


Foundation models extend beyond vision and language into mobile sensing, the very modality through which computing meets the physical world

Vision: Toward Sensor-Grounded Intelligence



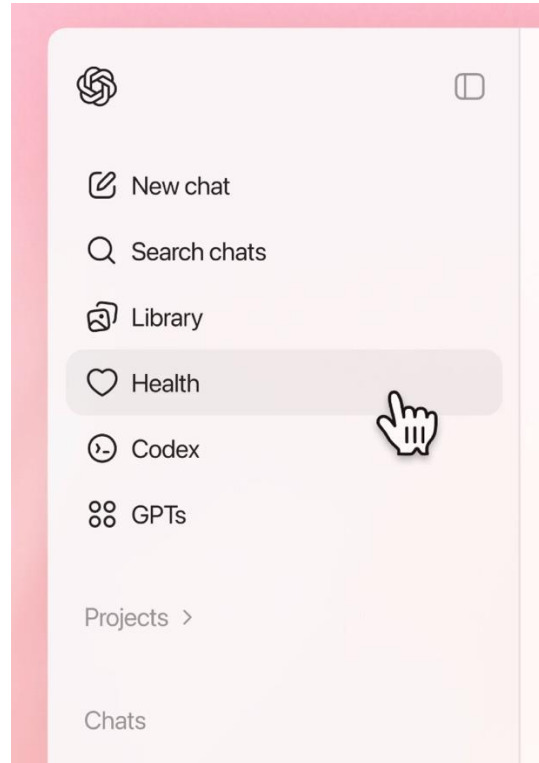
Future Direction: Toward Sensor-Grounded Intelligence



Future Direction: Toward Sensor-Grounded Intelligence



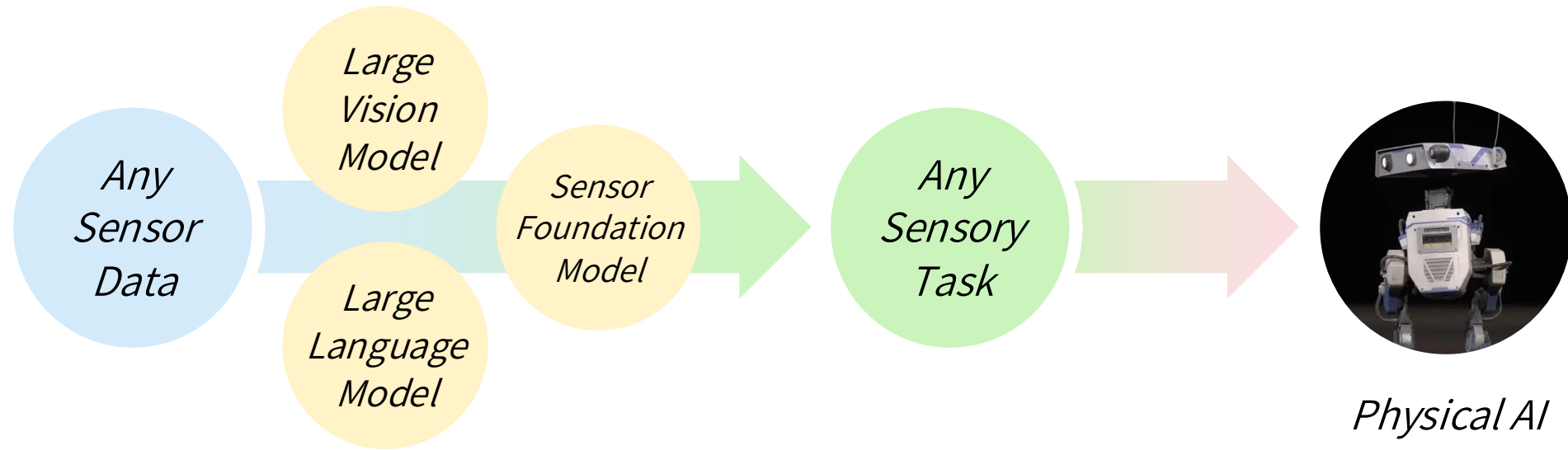
Google Health app
2026



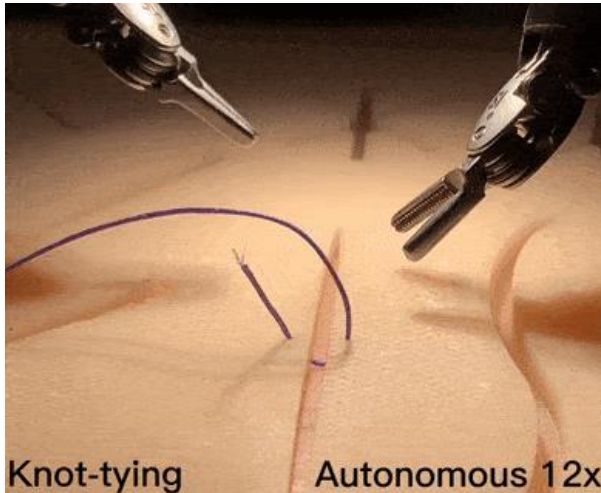
ChatGPT Health
2026

Sensor grounding and personalization of **Health LLMs**

Future Direction: Toward Sensor-Grounded Intelligence



Future Direction: Toward Sensor-Grounded Intelligence



Surgical Robot Transformer
Kim et al., CoRL '24

+



Knocker
Gong et al., UbiComp '19

Robots that knock, push, and sense the physical world
→ From Vision-Language-Action models to
Vision-Sense-Language-Action models

Publications

- **[IEEE TMC '25]** From Vision to Motion: Translating Large-Scale Knowledge for Data-Scarce IMU Applications
Hyungjun Yoon, Hyeongheon Cha, Canh Hoang Nguyen, Taesik Gong, and Sung-Ju Lee
- **[SenSys '25]** SelfReplay: Adapting Self-Supervised Sensory Models via Adaptive Meta-Task Replay
Hyungjun Yoon, Jaehyun Kwak, Biniyam Aschalew Tolera, Gaole Dai, Mo Li, Taesik Gong, Kimin Lee, and Sung-Ju Lee
- **[EMNLP '24]** By My Eyes: Grounding Multimodal Large Language Models with Sensor Data via Visual Prompting
Hyungjun Yoon, Biniyam Aschalew Tolera, Taesik Gong, Kimin Lee, and Sung-Ju Lee
- **[ACL '26 Findings]** ConSensus: Multi-Agent Collaboration for Multimodal Sensing
Hyungjun Yoon, Mohammad Malekzadeh, Sung-Ju Lee, Fahim Kawsar, and Lorena Qendro
- **[ICLR '26]** Beyond Hearing: Learning Task-Agnostic ExG Representations from Earphones via Physiology-Informed Tokenization
Hyungjun Yoon*, Seungjoo Lee*, Yu Wu*, Xiaomeng Chen*, Taiting Lu, Freddy Yifei Liu, TaECKYUNG Lee, Hyeongheon Cha, Haochen Zhao, Gaoteng Zhao, Sung-Ju Lee, Dongyao Chen, Cecilia Mascolo, and Lili Qiu