Representation Learning from Multimodal Sensor Data

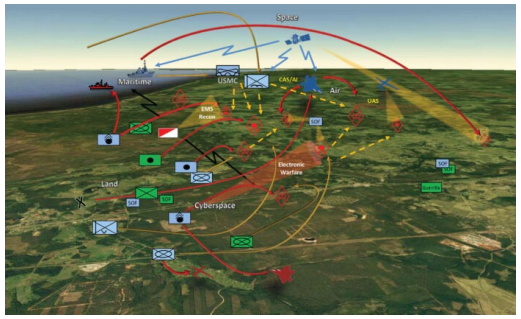
Tomoyoshi Kimura (tkimura4), Yuheng Pan (yuhengp2), Hongjue Zhao (hongjue2)

CS 537 AIoT G4 - Spring 2026

Agenda

1. **Overview and Background**
2. Self-Supervised Representation Learning for IoT Signals
3. Multimodal Representation with Incomplete Sensor Signals
4. Multimodal Sensing Applications and Age of LLM
5. Conclusion + Q & A

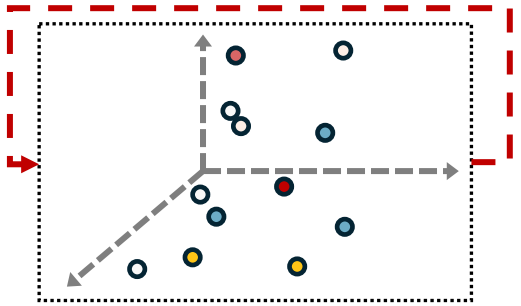
Representation Learning for IoT



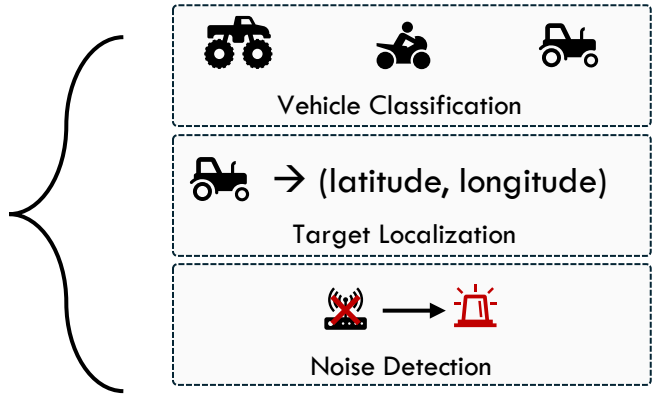
Unlabeled Data

Encoder

Self-supervision



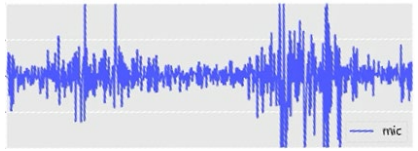
Representation/Latent Space



Downstream Tasks

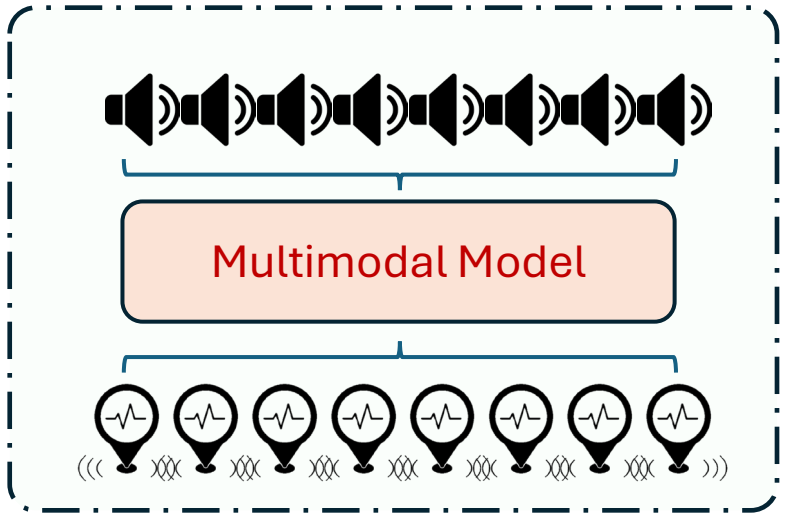
- Representation Learning or Self-Supervised Learning
  - Learn an **encoder** mapping: raw sensing signals → generalized representation
  - Semantically structured latent space correspond to physical events and activities
- Multimodal Representation Learning for IoT
  - Labeled data is **scarce**, expensive, and task-specific
  - IoT Signals are multimodal capture heterogeneous properties of the event

Geophone signal

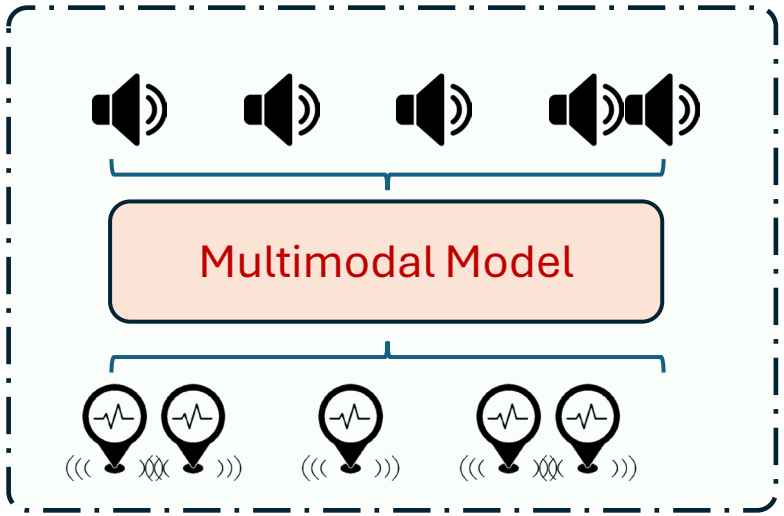


Microphone signal

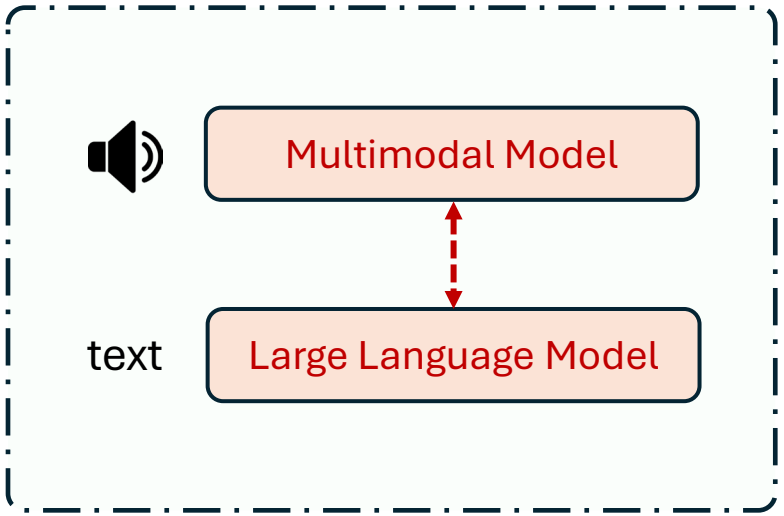
Presentation Overview



Multimodal Representation Learning



Multimodal SSL with Incomplete Multimodal Pairs

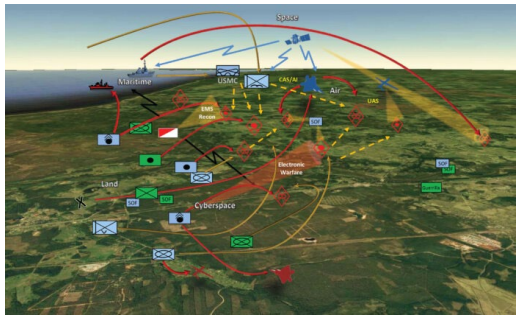


Multimodal SSL with LLM and Benchmarking

Agenda

1. Overview and Background
2. **Self-Supervised Representation Learning for IoT Signals**
3. Multimodal Representation with Incomplete Sensor Signals
4. Multimodal Sensing Applications and Age of LLM
5. Conclusion + Q & A

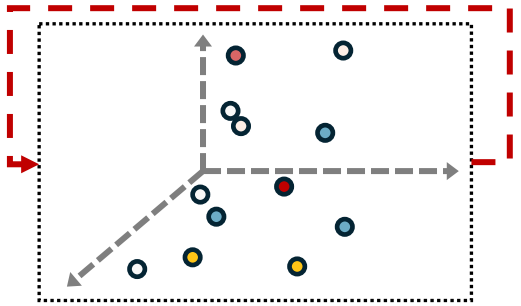
Self-Supervised Representation Learning



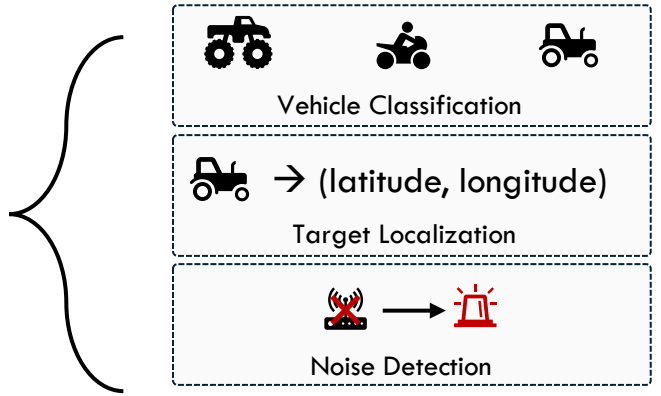
Unlabeled Data

Encoder

Self-supervision



Representation/Latent Space



Downstream Tasks

Loss Function

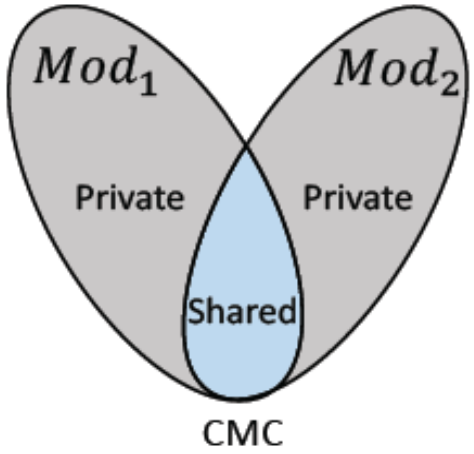
$$\mathcal{L} = -\log \frac{\exp(\langle z_i^{(1)}, z_i^{(2)} \rangle / \tau)}{\sum_k \exp(\langle z_i^{(1)}, z_k^{(2)} \rangle / \tau)}$$

$$h = E(x)$$

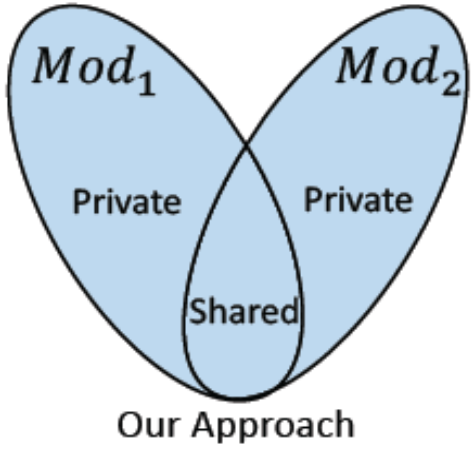
$$z = P(h) = P(E(x))$$

FOCAL: Motivation

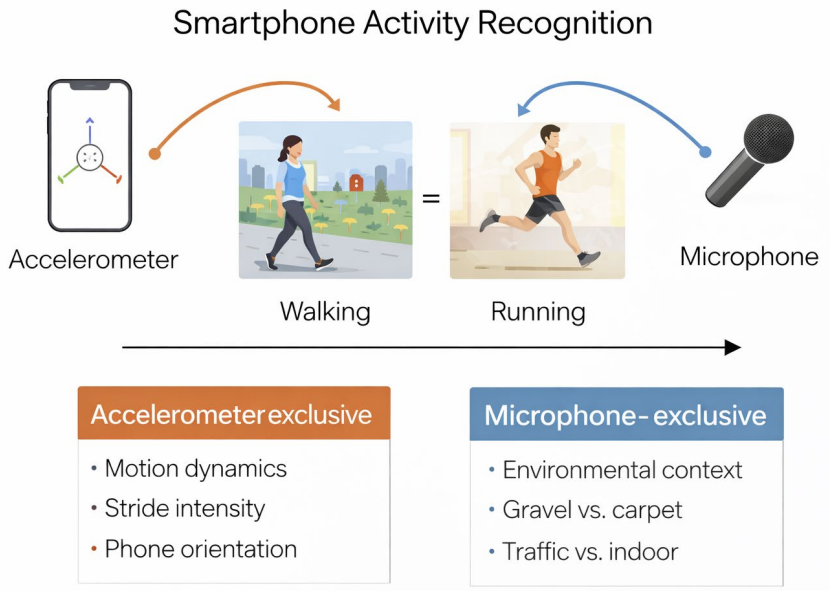
- Problem1: Ignoring exclusive modality features



CMC

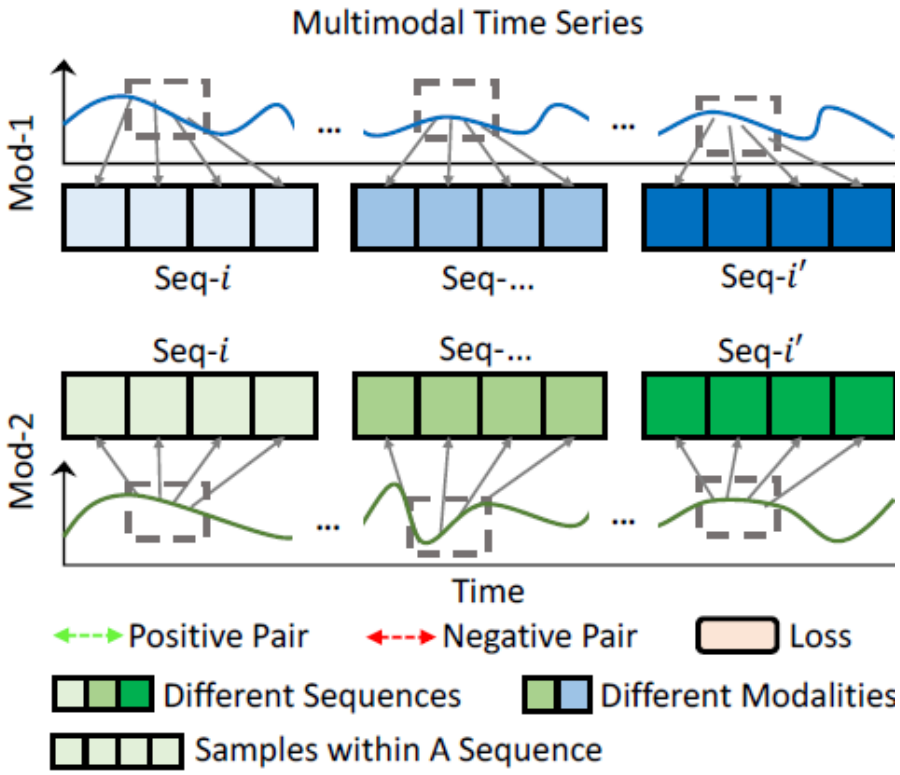


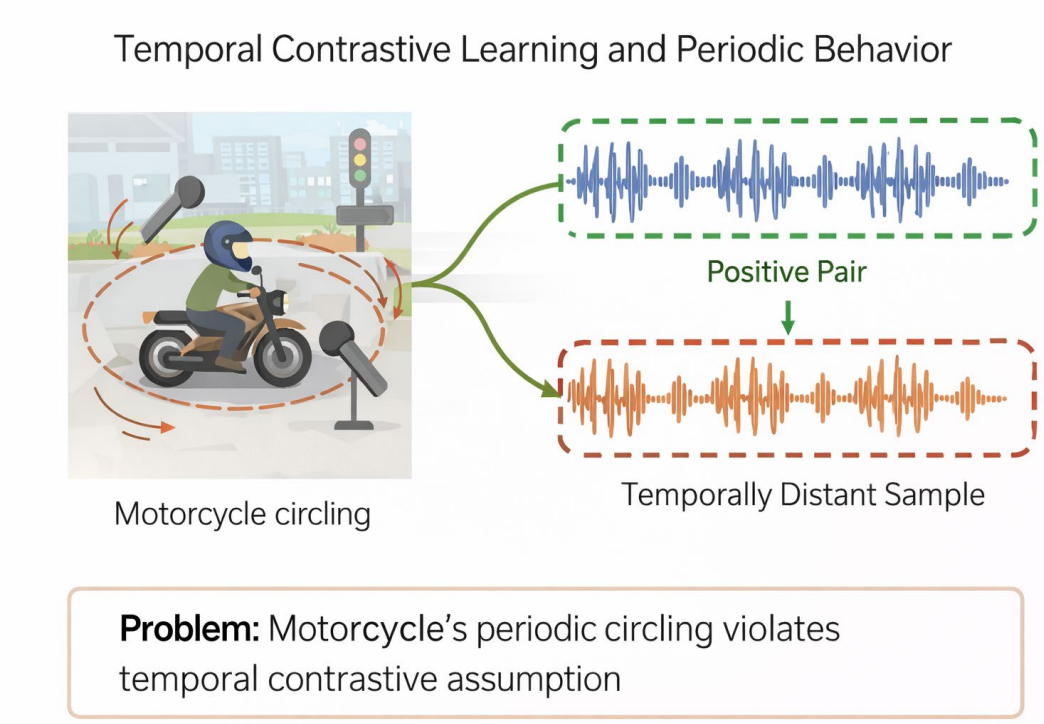
Our Approach



FOCAL: Motivation

- Problem2: Insufficient handling of temporal constraints





FOCAL: Overview

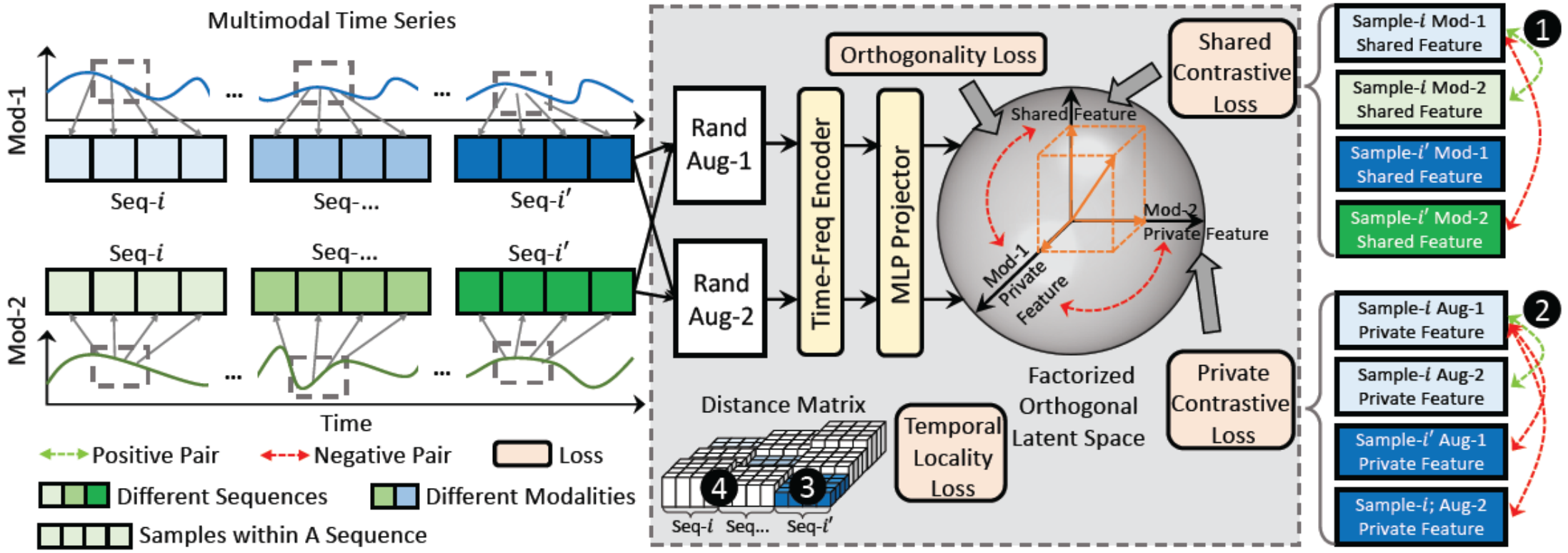
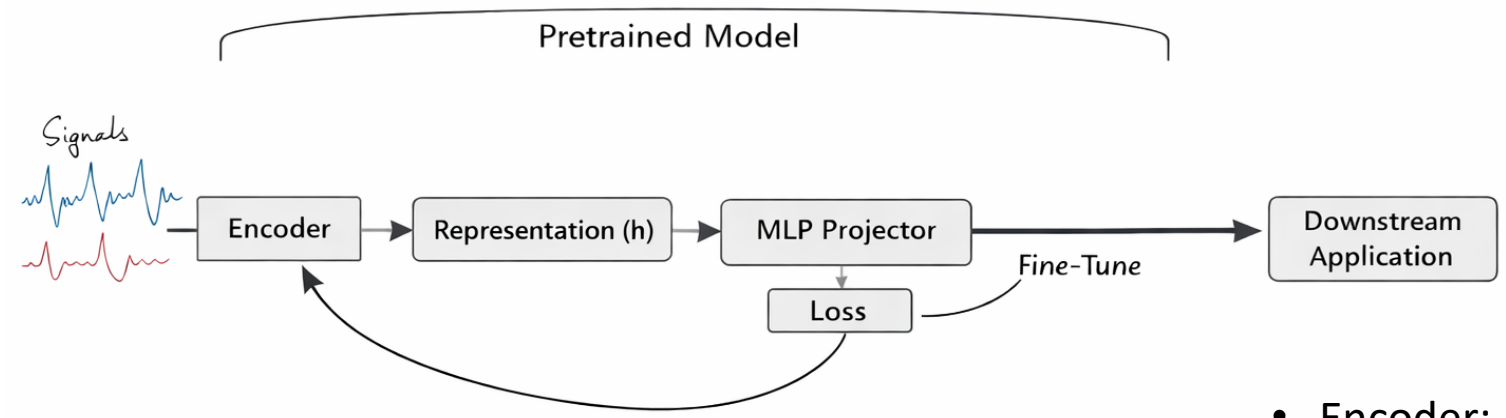


Figure 1: Overview of the FOCAL framework. Best viewed in color.

FOCAL: Experimental Setup



$$\mathcal{L}_{\text{shared}} = -\sum_{i}\sum_{M_j \in \mathcal{M}}\sum_{M_{j'} \in \mathcal{M}, j' \neq j} \log \frac{\exp\left(\left\langle h_{ij}^{\text{shared}}, h_{ij'}^{\text{shared}} \right\rangle / \tau\right)}{\sum_{i' \in \mathcal{B}} \exp\left(\left\langle h_{ij}^{\text{shared}}, h_{i'j'}^{\text{shared}} \right\rangle / \tau\right)}$$

$$\mathcal{L}_{\text{private}} = -\sum_{i}\sum_{M_j \in \mathcal{M}} \log \frac{\exp\left(\left\langle h_{ij}^{\text{private}}, \hat{h}_{ij}^{\text{private}} \right\rangle / \tau\right)}{\sum_{i' \in \mathcal{B}, i' \neq i} \exp\left(\left\langle h_{ij}^{\text{private}}, h_{i'j}^{\text{private}} \right\rangle / \tau\right) + \sum_{i' \in \mathcal{B}} \exp\left(\left\langle h_{ij}^{\text{private}}, \hat{h}_{i'j}^{\text{private}} \right\rangle / \tau\right)}$$

$$\mathcal{L}_{\text{orthogonal}} = \sum_{i}\sum_{M_j \in \mathcal{M}} \left\langle h_{ij}^{\text{shared}}, h_{ij}^{\text{private}} \right\rangle + \sum_{i}\sum_{M_j \in \mathcal{M}}\sum_{M_{j'} \in \mathcal{M}, j' \neq j} \left\langle h_{ij}^{\text{private}}, h_{ij'}^{\text{private}} \right\rangle$$

$$\mathcal{L}_{\text{temporal}} = \sum_{s}\sum_{s' \neq s} \max(\bar{D}_{ss} - \bar{D}_{ss'} + \text{margin}, 0)$$

$$\mathcal{L} = \mathcal{L}_{\text{shared}} + \lambda_p \cdot \mathcal{L}_{\text{private}} + \lambda_o \cdot \mathcal{L}_{\text{orthogonal}} + \lambda_t \cdot \mathcal{L}_{\text{temporal}}$$

- Encoder: DeepSense and SWIN-Transformer
- Datasets:
  - MOD/ACIDS: vehicle acoustic and seismic signals for vehicle classification
  - RealWorld-HAR/PAMAP2: accelerometer, gyroscope, magnetometer, and light signals for physical activity classification

FOCAL: Experiment Results

Table 1: Statistical Summaries of Evaluated Datasets.

Dataset	Classes	Modalities (Freq)	Sample Length	Interval (Overlap)	#Samples	#Labels
MOD	7	acoustic (8000Hz), seismic (100Hz)	2 sec	0.2 sec (0%)	39,609	7,335
ACIDS	9	acoustic, seismic (both 1025Hz)	1 sec	0.25 sec (50%)	27,597	27,597
RealWorld-HAR	8	acc, gyro, mag, lig (all 50Hz)	5 sec	1 sec (50%)	12,887	12,887
PAMAP2	18	acc, gyr, mag (all 100Hz)	2 sec	0.4 sec (50%)	9,611	9,611

Table 2: Finetune Results with Linear Classifier

Dataset		MOD		ACIDS		RealWorld-HAR		PAMAP2	
Encoder	Framework	Acc	F1	Acc	F1	Acc	F1	Acc	F1
DeepSense	Supervised	0.9404	0.9399	**0.9566**	0.8407	0.9348	**0.9388**	**0.8849**	**0.8761**
	SimCLR	0.8855	0.8855	0.7438	0.6101	0.7138	0.6841	0.6802	0.6583
	MoCo	0.8808	0.8812	0.7717	0.6205	0.7859	0.7708	0.7559	0.7387
	CMC	0.9196	0.9186	0.8443	0.7244	0.7975	0.8116	0.7906	0.7706
	MAE	0.5981	0.5993	0.6644	0.5618	0.7565	0.7515	0.7114	0.6158
	Cosmo	0.8989	0.8998	0.8511	0.6929	0.8956	0.8888	0.8356	0.8135
	Cocoa	0.8774	0.8764	0.6644	0.5359	0.8465	0.8488	0.7603	0.7187
	MTSS	0.4153	0.3582	0.4352	0.2441	0.2989	0.1405	0.3541	0.1795
	TS2Vec	0.7669	0.7648	0.5224	0.3587	0.6595	0.5984	0.5729	0.4715
	GMC	0.9257	0.9267	0.9096	0.7929	0.8869	0.8948	0.8119	0.7860
	TNC	0.9518	0.9528	0.8237	0.6936	0.8892	0.8971	0.8387	0.8143
	TS-TCC	0.8707	0.8735	0.7667	0.6164	0.8073	0.8010	0.7776	0.7250
	FOCAL	**0.9732**	**0.9729**	0.9516	**0.8580**	**0.9382**	0.9290	0.8588	0.8463
SW-T	Supervised	0.8948	0.8931	0.9137	0.7770	0.9313	0.9278	**0.8612**	0.8384
	SimCLR	0.9250	0.9247	0.9128	0.8144	0.7046	0.7220	0.7705	0.7424
	MoCo	0.9390	0.9384	0.9174	0.8100	0.7813	0.8024	0.7717	0.7313
	CMC	0.9129	0.9105	0.8128	0.6857	0.8840	0.8955	0.8080	0.7901
	MAE	0.7803	0.7772	0.8516	0.7023	0.8829	0.8813	0.7910	0.7606
	Cosmo	0.3429	0.3378	0.7110	0.6086	0.8604	0.8169	0.7741	0.7366
	Cocoa	0.7040	0.7038	0.7096	0.5794	0.8892	0.8861	0.7689	0.7317
	MTSS	0.4206	0.4163	0.3429	0.2250	0.5136	0.4370	0.2847	0.1714
	TS2Vec	0.7254	0.7174	0.7183	0.5748	0.6151	0.5955	0.6195	0.5426
	GMC	0.8640	0.8611	0.9402	0.7766	0.9319	0.9379	0.8312	0.8083
	TNC	0.8533	0.8539	0.8352	0.7372	0.8817	0.8784	0.8013	0.7506
	TS-TCC	0.8734	0.8735	0.9041	0.7547	0.8731	0.8454	0.7997	0.7260
	FOCAL	**0.9805**	**0.9800**	**0.9489**	**0.8262**	**0.9451**	**0.9503**	0.8580	**0.8401**

Table 4: Benefits of Temporal Constraints to SOTA baselines on ACIDS

Metrics	SimCLR		MoCo		CMC		Cocoa		GMC	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
wTemp	**0.7461**	**0.6938**	**0.7836**	**0.6618**	**0.8690**	0.7090	**0.8543**	**0.7665**	**0.9347**	**0.8109**
Vanilla	0.7438	0.6101	0.7717	0.6205	0.8443	**0.7244**	0.6644	0.5359	0.9096	0.7929

Table 5: Benefits of Temporal Constraints to SOTA baselines on PAMAP2

Metrics	SimCLR		MoCo		CMC		Cocoa		GMC	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
wTemp	**0.7129**	**0.6884**	**0.7800**	**0.7602**	0.7804	0.7583	**0.8442**	**0.8146**	**0.8253**	**0.8114**
Vanilla	0.6802	0.6583	0.7559	0.7387	**0.7906**	**0.7706**	0.7603	0.7187	0.8119	0.7860

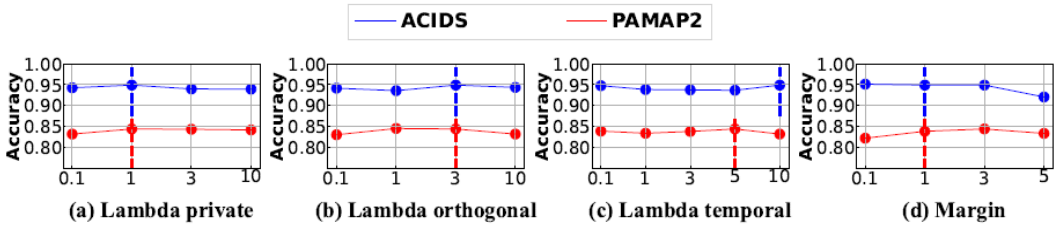


Figure 9: Loss weights sensitivity test.

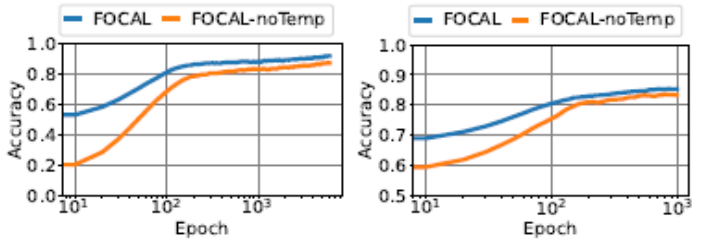


Figure 8: Convergence curves.

FOCAL: Key Takeaway

- What is this paper trying to solve?
- Exclusive part of each modality are not used.
- Temporal structural constraints are too simplistic.

- How does this paper solve these problem?
- Explicitly disentangling shared and private modality information and enforcing structured constraints on temporal and cross-modal relationships.

SemiCMT: Motivation

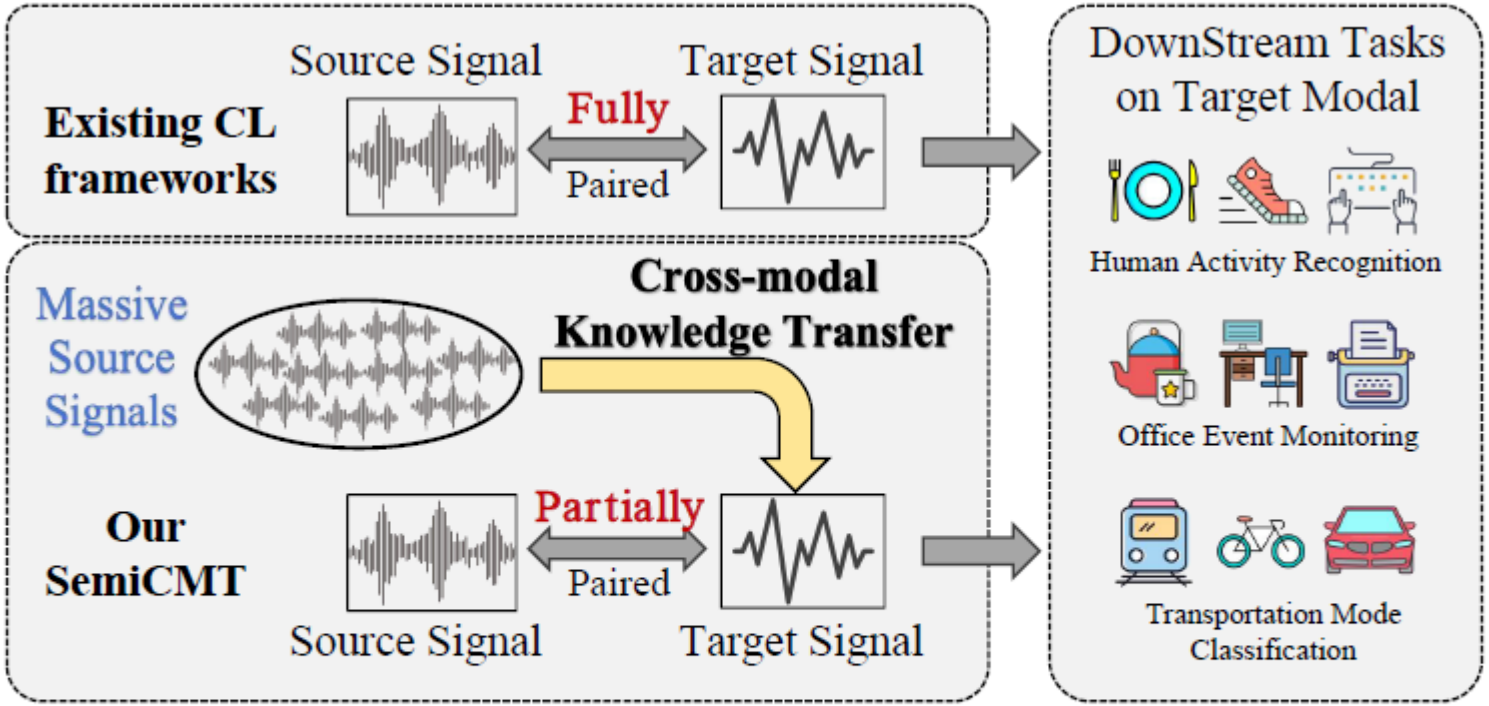


Fig. 2. The difference between existing multi-modal contrastive learning and the proposed SemiCMT.

SemiCMT: Challenges

- How to handle highly imbalanced data distributions between the two modalities
- How to integrate the cross-modal knowledge-transfer objective with the self-supervised learning paradigm such that the transferred knowledge to the target modality can still be easily calibrated to downstream tasks
- How to deal with the information gap between the source and target modalities, since the information among different sensory modalities is only partially shared but not fully overlapped

SemiCMT: Insights

- Insight 1: Don’t try to force the two modalities to become identical
  - There will always be **information that exists in one sensor but not the other**.
- Insight 2: With few paired samples, standard contrastive alignment is an inefficient way to transfer knowledge
  - We have **very few matched pairs**, so there are **not enough positive examples** to learn a strong cross-modal mapping.
- Insight 3: Train the source encoder with transfer in mind, not just to be good at itself
  - Because the source encoder might become strong by using **source-only cues** that don’t exist in the target modality.

SemiCMT: Overview

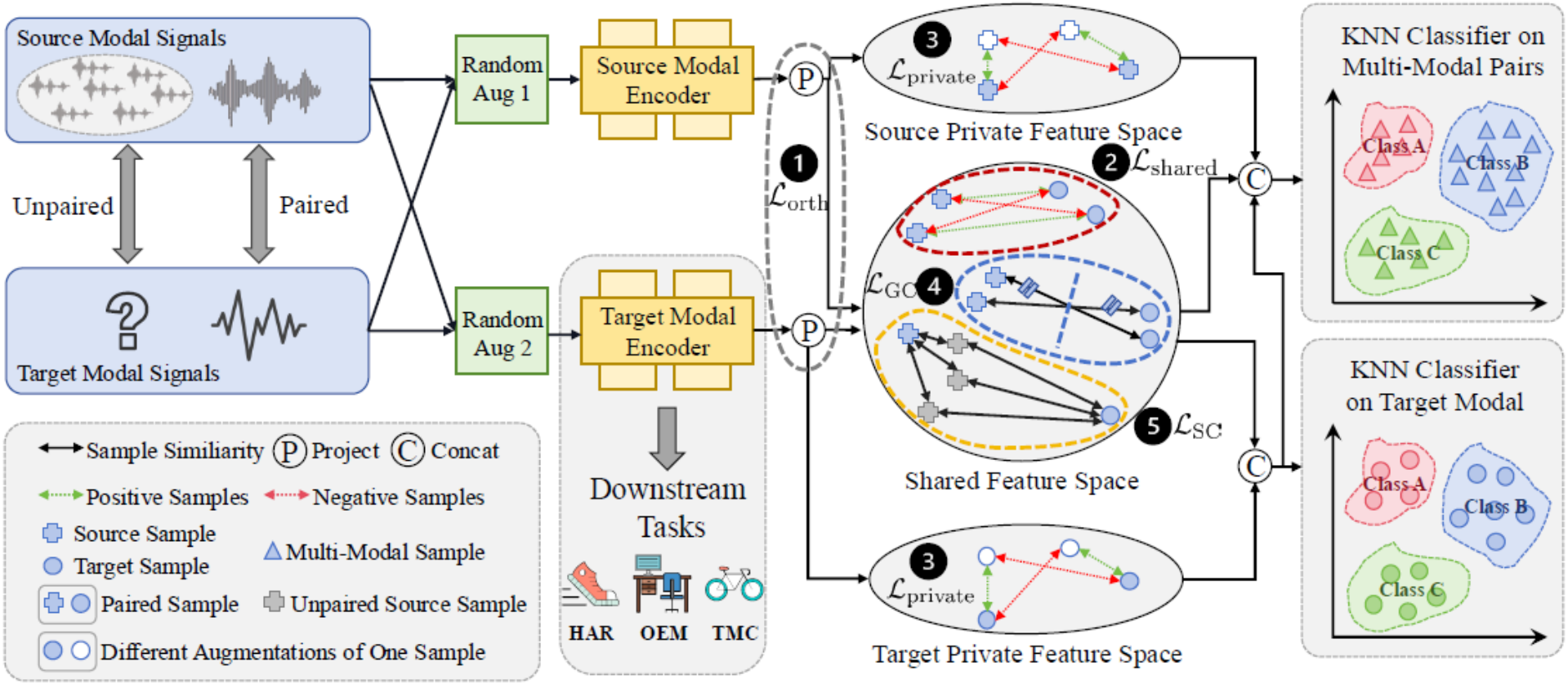


Fig. 7. Overview of the proposed SemiCMT framework for self-supervised cross-modal knowledge transfer, where a large set of unpaired source-modal data and a small set of multi-modal pairs are available.

SemiCMT: Overview

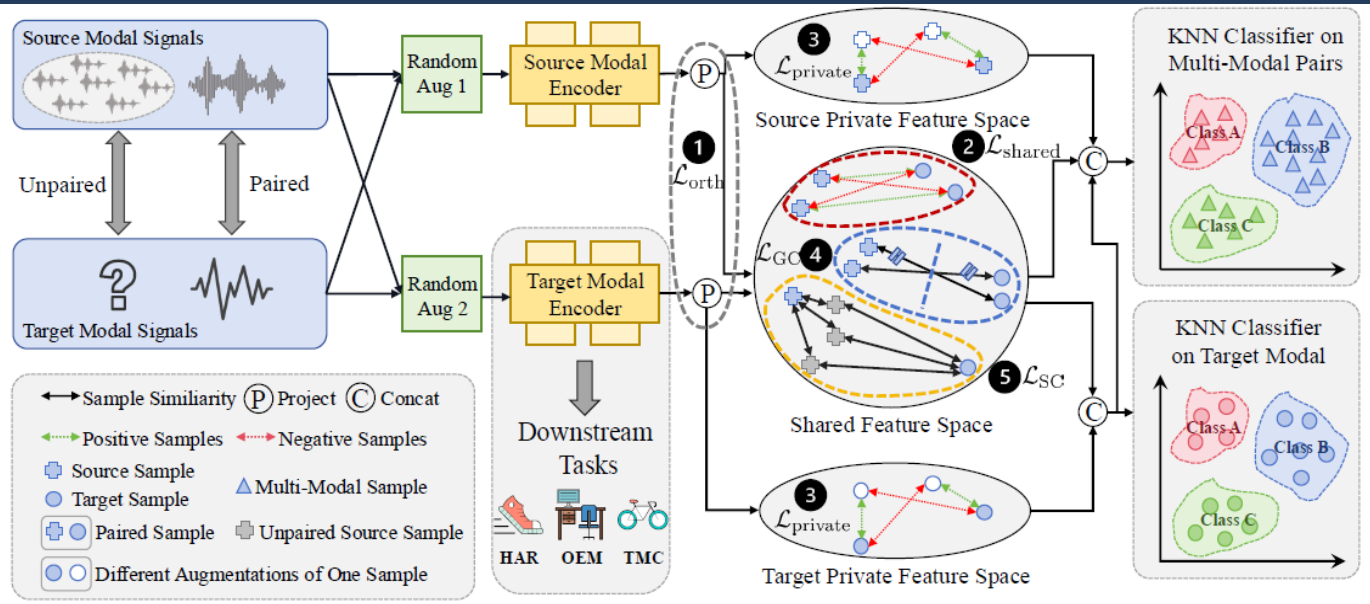


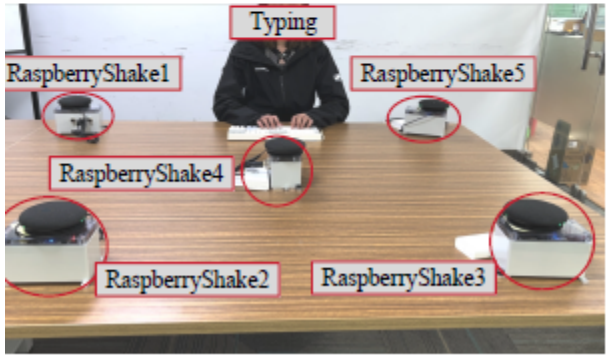
Fig. 7. Overview of the proposed SemiCMT framework for self-supervised cross-modal knowledge transfer, where a large set of unpaired source-modal data and a small set of multi-modal pairs are available.

1. **Orthogonality Constraint**: We impose an orthogonality constraint that requires both the shared and private features within each modality, as well as the shared features across two modalities, to be orthogonal to one another.
2. **Modality Consistency in Shared Space**: Within the shared feature space, we aim to align the features of the source and target modalities for each paired sample more closely than the modality features from mismatched samples.
3. **Modality Consistency in Private Space**: We encourage the modality features under two random augmentations to be more similar to each other than the modality features from two different samples.
4. **Cross-Modal Geometric Consistency**: We ensure that the source similarities and the target similarities between the same two samples are aligned as closely as possible.
5. **Cross-Modal Semantical Consistency**: We align the source-to-anchor similarities with the target-to-anchor similarities by treating all the unpaired source samples as anchors within a single batch.

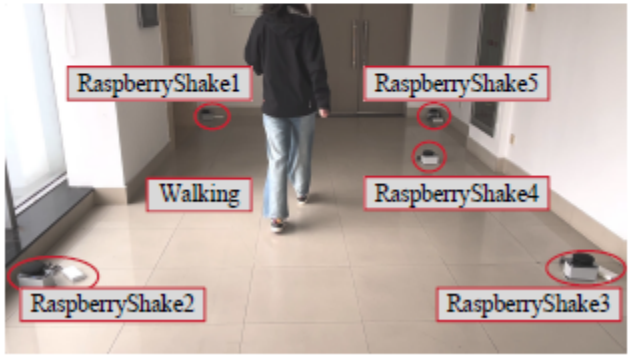
SemiCMT: Experimental Setup

Table 4. Summaries of five human-related datasets in evaluation.

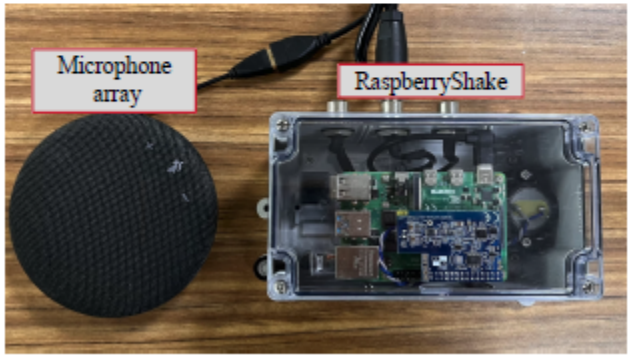
Dataset	Classes	Modalities (Freq)	Sample Length	Overlap	Samples
SHL	8	acc (100Hz), gyro (100Hz), mag (100Hz)	2 seconds	50%	59,044
InOffice	11	acoustic (48KHz), seismic (100Hz)	2 seconds	0	18,535
AudioIMU	23	acoustic (22kHz), gyro (50Hz), mag (50Hz)	2 seconds	50%	14,347
RealWorld-HAR	8	acc (50Hz), gyro (50Hz), mag (50Hz), lig(50Hz)	5 seconds	50%	12,877
PAMAP2	18	acc (100Hz), gyro (100Hz), mag (100Hz)	2 seconds	50%	9,611



a) Data collection scenario on the table



(b) Data collection scenario on the floor



(c) The sensing device constitution

SemiCMT: Results

Table 5. Finetuning performance with KNN and linear classifiers on **multi-modal samples**. The best and the second best performances are denoted in **bold** and underline.

Classifier	Method	InOffice		AudioIMU		SHL		PAMAP2		RealWorld-HAR	
		Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
Supervised [26]		0.9003	0.8285	0.6716	0.6656	0.8046	0.8132	0.7349	0.7039	0.6615	0.4892
Linear	MAE [14]	0.8796	0.8051	0.6330	0.6328	0.6553	0.6381	0.6888	0.6727	0.7613	0.7691
	UnpairedMR [15]	0.5043	0.4931	0.4219	0.3855	0.5151	0.4020	0.5511	0.5392	0.6250	0.6218
	SimCLR [3]	0.5417	0.4270	0.6134	0.6084	0.7030	0.6884	0.7351	0.7138	0.7745	0.7709
	CMC [45]	0.7678	0.6913	0.5273	0.5184	0.6611	0.6714	0.7665	0.7524	0.8107	0.8116
	CMCD [22]	0.7815	0.7189	0.4469	0.4361	0.6851	0.6741	0.7650	0.7526	0.7836	0.7865
	CCD [54]	0.1722	0.0430	0.0778	0.0140	0.2955	0.1932	0.4376	0.3171	0.5465	0.3864
	WTB [48]	0.1312	0.0511	0.5649	0.5653	0.6359	0.6068	0.7245	0.7038	0.7006	0.6759
	ColloSSL [16]	0.8672	0.8102	0.6659	0.6589	0.7977	0.8020	0.7272	0.7141	0.8907	0.9003
	COCOA [5]	0.8866	0.8241	0.6412	0.6386	0.6129	0.5479	0.6996	0.6663	0.7430	0.5791
	FOCAL [25]	0.8724	0.8096	0.6611	0.6579	0.7288	0.7273	0.7549	0.7393	0.7882	0.7783
	CroSSL [4]	0.7843	0.7167	0.4591	0.4404	0.4981	0.4299	0.3077	0.2006	0.1799	0.0509
	SemiCMT (ours)	0.8848	0.8428	0.7076	0.7039	0.7926	0.7906	0.7503	0.7493	0.8674	0.8769
KNN	MAE [14]	0.7472	0.7231	0.5371	0.5387	0.6456	0.6218	0.5336	0.5060	0.6149	0.5950
	UnpairedMR [15]	0.4710	0.4122	0.3824	0.3349	0.4356	0.4212	0.4590	0.4933	0.5528	0.5311
	SimCLR [3]	0.4836	0.3923	0.6024	0.5927	0.7636	0.7535	0.6538	0.6159	0.7484	0.7344
	CMC [45]	0.7983	0.7182	0.5440	0.5390	0.7732	0.7741	0.6380	0.6048	0.6613	0.6323
	CMCD [22]	0.7970	0.7225	0.6602	0.6488	0.7804	0.7834	0.6414	0.6049	0.6596	0.6652
	CCD [54]	0.4487	0.3859	0.1642	0.1619	0.3054	0.2577	0.4248	0.3464	0.4576	0.4282
	WTB [48]	0.5035	0.4276	0.5929	0.5873	0.6188	0.5750	0.6590	0.6438	0.5929	0.5423
	ColloSSL [16]	0.8158	0.7334	0.6466	0.6388	0.7151	0.7412	0.6126	0.5888	0.7117	0.6993
	COCOA [5]	0.7694	0.6656	0.6333	0.6263	0.5844	0.5424	0.5773	0.5111	0.5944	0.5627
	FOCAL [25]	0.7986	0.7169	0.8371	0.8357	0.8604	0.8663	0.7783	0.7638	0.8428	0.8471
	CroSSL [4]	0.7664	0.7002	0.2951	0.2603	0.5165	0.4545	0.2802	0.2215	0.3266	0.2812
	SemiCMT (ours)	**0.9321**	**0.9152**	**0.8850**	**0.8776**	**0.9017**	**0.9007**	**0.7801**	**0.7695**	**0.9003**	**0.9070**

SemiCMT: Key Takeaway

- What is this paper trying to solve?
- Too many unpaired modality data

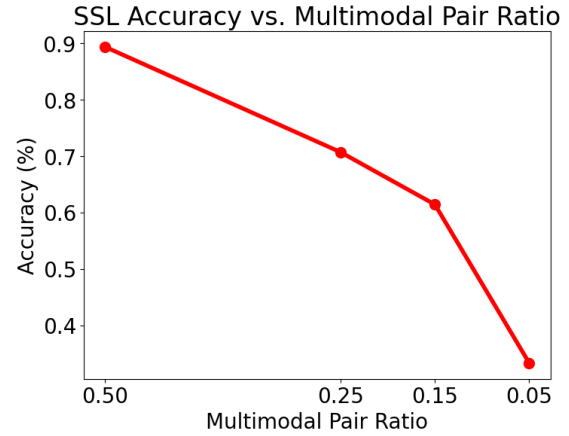
- How does this paper solve these problem?
- Using unpaired source samples as anchors to distill their similarity structure (soft semantic relations) into the target embedding space, while only using the limited paired samples to bridge modalities via shared-feature alignment and geometric consistency.

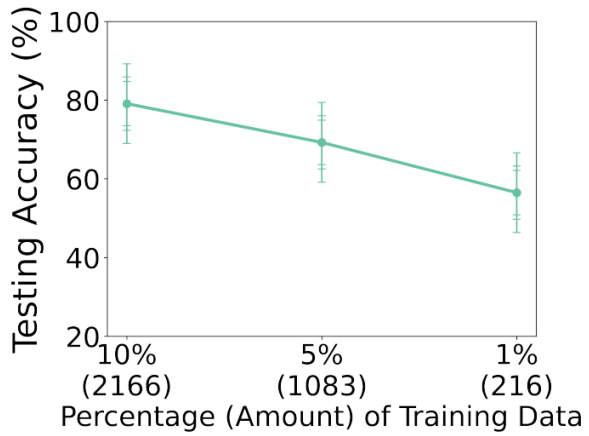
Agenda

1. Overview and Background
2. Self-Supervised Representation Learning for IoT Signals
3. **Multimodal Representation with Incomplete Sensor Signals**
4. Multimodal Sensing Applications and Age of LLM
5. Conclusion + Q & A

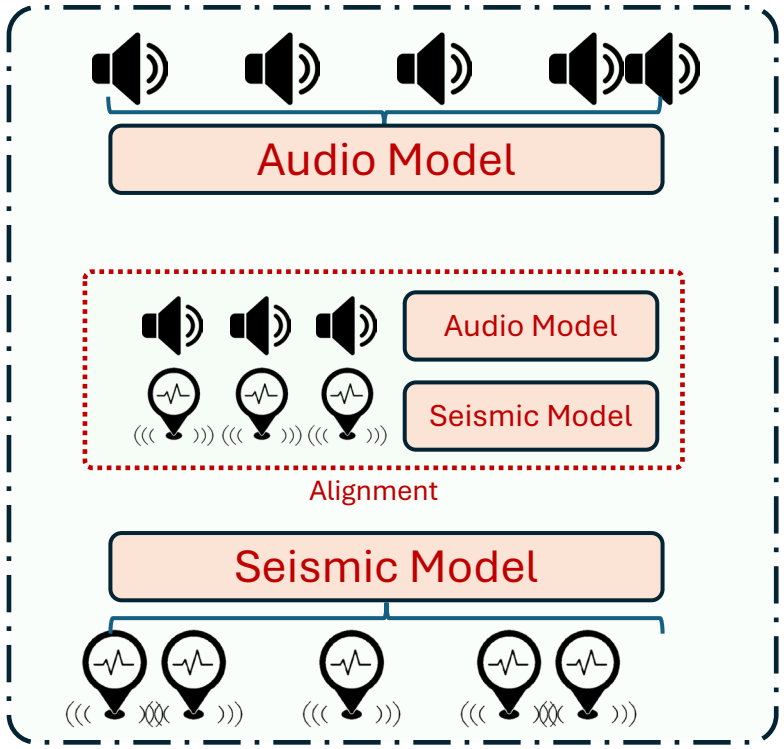
Challenges in Multimodal SSL for IoT

- Why are Multimodal Pairs Incomplete in IoT Sensing?
  - **Distributed & Heterogeneous Deployment**
    - Different locations, times, and perspectives
      → heterogeneous and partially overlapping subsets
  - **Asynchrony & Heterogeneity of Sensory Modalities**
    - Different sampling rates, clocks, and windowing schemes
      → Weak, noisy, or unusable cross-modal correspondences
  - **Sensor Failures**
    - Powerless, communication, environmental interference
      → Temporally inconsistent modality availability
- Most multimodal learning frameworks implicitly assume large amounts of time-aligned, co-located, full-modality data.

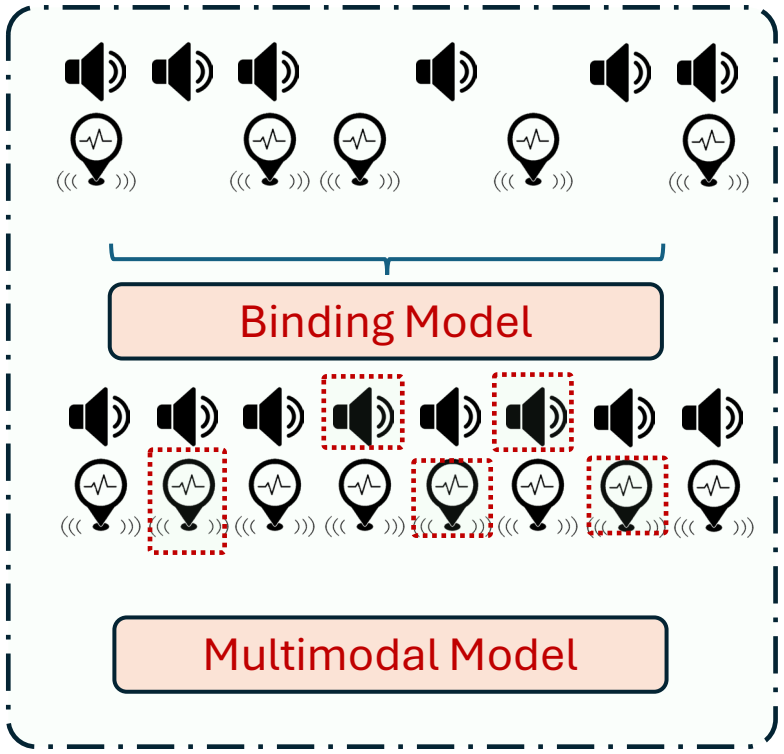




Binding Between Multimodal Data



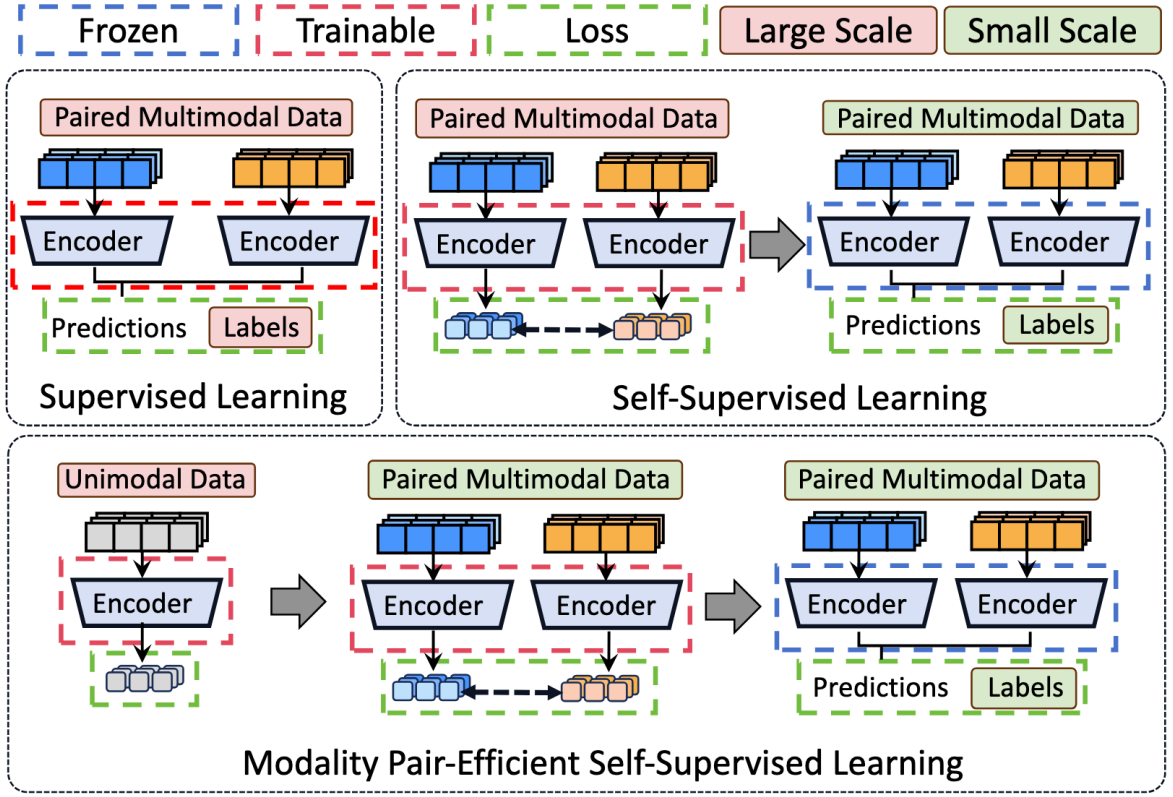
Model Binding



Data Binding

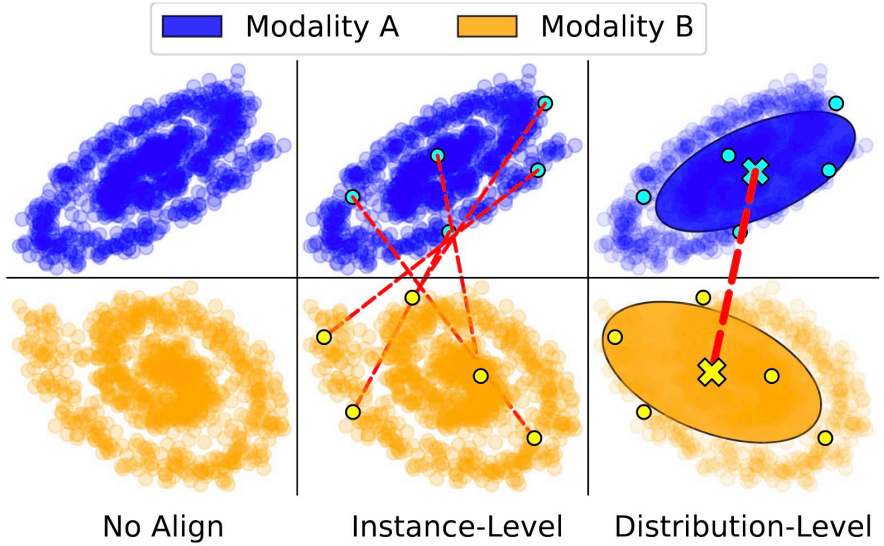
Pair-Efficient Self-Supervised Learning

- Modality Pair as Supervision
  - Supervised Learning
    - Manual labels for task mapping
    - Heavy dependence on annotations
  - Multimodal Self-Supervised Learning
    - Use data properties (pairing) as supervision
    - Minimal labels for finetuning
- Pair-Efficient Self-Supervised Learning
  - Unimodal Pretraining (large scale)
  - Cross-modal Alignment (small scale pairing)



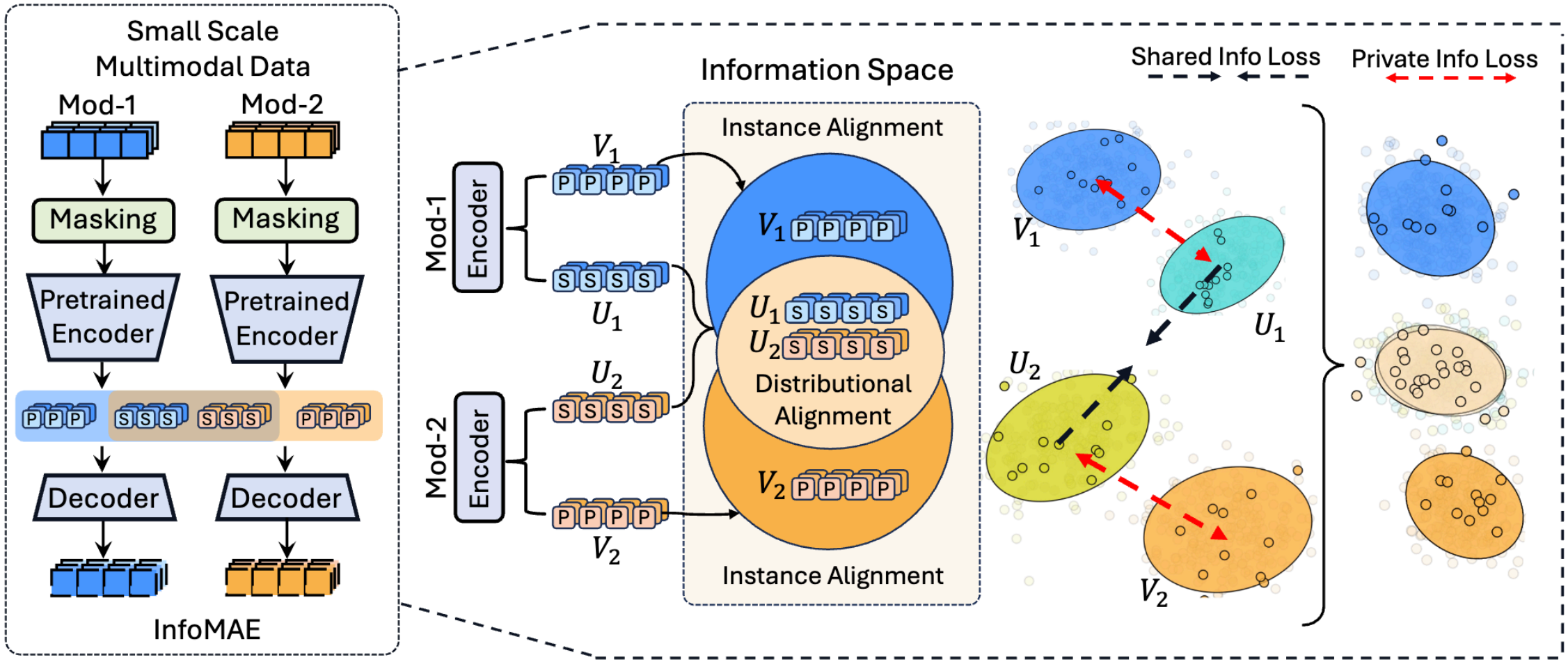
Efficient Cross-modal Alignment

- Instance-Level Alignment
  - Contrastive Learning
  - Sample-to-sample similarity
- Distribution-Level Alignment
  - Distribution-to-Distribution Similarity
  - Information-theory based Formulation



Shared Representation	Private Representation
Sufficient Common Variable (W) with minimal entropy min(H(W))	For each modality, find private variable V with minimal entropy
W contains information shared with all modalities	V contains remaining information unique to each modality
$I(\boldsymbol{X};\boldsymbol{Y}\|\boldsymbol{W}^{(x)}) + \beta I(\boldsymbol{X};\boldsymbol{W}^{(x)}) + \alpha d(\boldsymbol{W}^{(x)} - \boldsymbol{W}^{(y)})$	$I(\boldsymbol{X};\boldsymbol{U}) + I(\boldsymbol{Y};\boldsymbol{V})$

InfoMAE Overview



Key Takeaway

- Information-theoretic formulation
- Leverage large amounts of unimodal data
- High multimodal data efficiency
- Flexibility as a standard multimodal learning framework with abundant multimodal pairs

Experimental Setup

- Encoder: SWIN-Transformer
- Real-world Multimodal Applications
  - Moving Object Detection (MOD)
    - Separate Independent Domains (M, G, T)
    - Seismic and Acoustic data
    - Vehicle classification (accuracy, F1-score)
  - Human Activity Recognition (HAR)
    - RealWorld-HAR & PAMAP2
    - Accelerometer, Gyroscope, and Magnetometer
    - Activity classification (accuracy, F1-score)

Application	Modalities	Domains
Moving Object Detection (MOD)	seismic acoustic	domain M domain G domain T
Human Activity Recognition (HAR)	accelerometer gyroscope magnetometer	PAMAP2 RealWorld-HAR

InfoMAE: Pair-Efficient Multimodal Alignment

InfoMAE: Multi-modal Representation Learning **using mostly unimodal data**

Seismic & Acoustic pretrained independently on three domains --- G, M, T
A small portion of Domain M data (5%) is used for alignment

	Framework	Aligned Domains		T_{Sei} || M_{Aco}		G_{Sei} || T_{Aco}		T_{Sei} || T_{Aco}		G_{Sei} || M_{Aco}		T_{Sei} || G_{Aco}	
		Joint Pretrain	Modal Alignment	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
Unimodal: 100% Multimodal: 0%	Unimodal Concat	✗	✗	0.6731	0.6699	0.5392	0.5281	0.4454	0.4366	0.7247	0.7217	0.6584	0.6543
	CMC [62]	✗	✓	0.6792	0.6702	0.4313	0.4356	0.4173	0.4032	0.6919	0.6877	0.6497	0.6335
	FOCAL [40]	✗	✓	0.7462	0.7432	0.6249	0.6249	0.5613	0.5579	0.7549	0.7527	0.7194	0.7160
Unimodal: 100% Multimodal: 5%	GMC [54]	✗	✓	0.7354	0.7317	0.6591	0.6523	0.4756	0.4720	0.8044	0.8053	0.7247	0.7211
	SimCLR [6]	✗	✓	0.3061	0.2742	0.2873	0.2609	0.2974	0.2758	0.2981	0.2698	0.2800	0.2308
	TNC [63]	✗	✓	0.1969	0.0815	0.1788	0.1312	0.1855	0.1021	0.1929	0.0896	0.1949	0.1041
	TSTCC [14]	✗	✓	0.3001	0.2706	0.2639	0.2393	0.2867	0.2432	0.3048	0.2842	0.2860	0.2337
	InfoMAE	✗	✓	**0.7950**	**0.7929**	**0.6986**	**0.7007**	**0.5928**	**0.5908**	**0.8326**	**0.8324**	**0.7636**	**0.7537**
Unimodal: 0% Multimodal: 5%	Joint Pretrain	✓	✗	Acc: 0.3329				F1: 0.3039					

Cross-Modal Alignment (MOD)

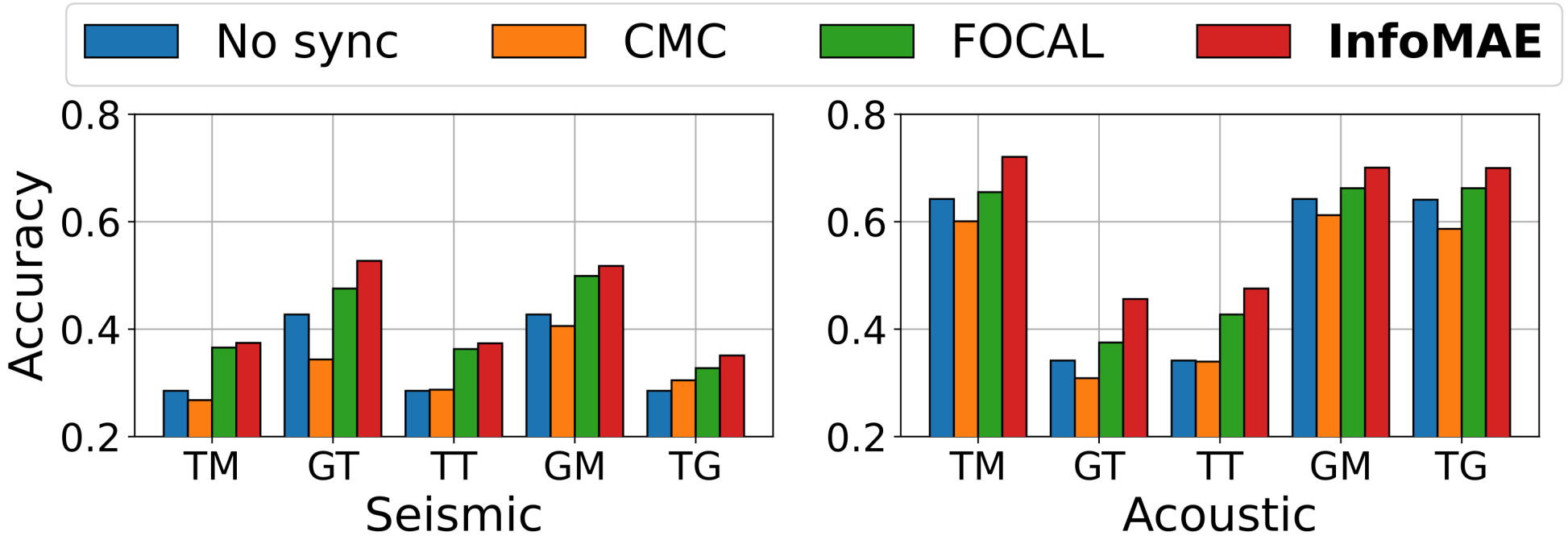
Alignment with varying paired data from the same domain

Unimodal Pretrain domain == Multimodal Alignment domain

Multimodal Data	Supervised		Joint Pretrain		CMC		GMC		FOCAL		**InfoMAE**	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
5%	0.5740	0.5663	0.3329	0.3039	0.7087	0.6989	0.8614	0.8616	0.8694	0.8668	**0.8828**	**0.8808**
15%			0.6142	0.6104	0.8111	0.8062	0.8781	0.8753	0.8727	0.8703	**0.9049**	**0.9028**
25%			0.7071	0.7938	0.8433	0.8372	0.8774	0.8759	0.8848	0.8831	**0.9290**	**0.9270**
50%			0.8942	0.8920	0.8754	0.8724	0.8948	0.8938	0.9009	0.8994	**0.9377**	**0.9367**

Cross-Modal Alignment (MOD)

Unimodal enhancement after cross-modal alignment on domain M



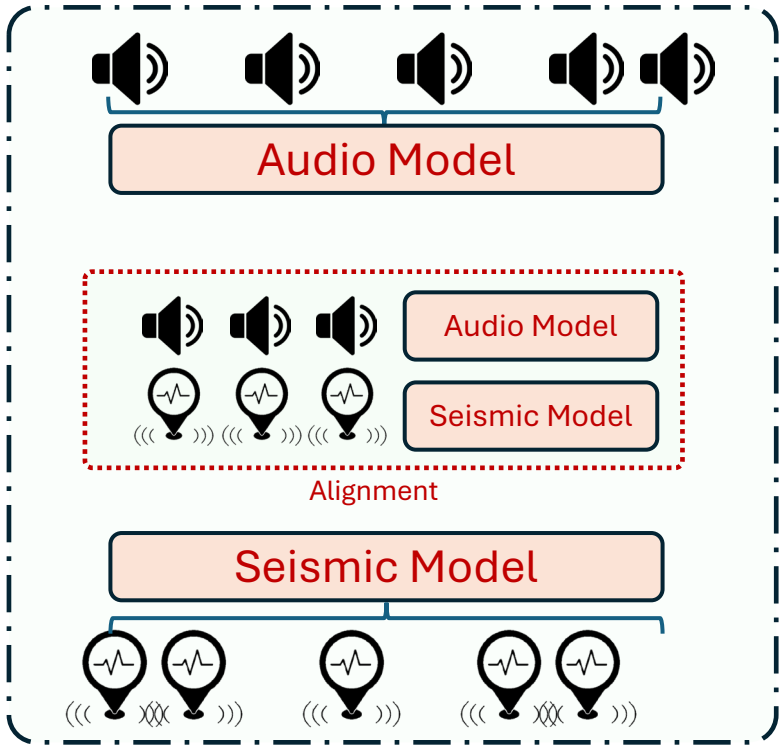
Full-Pair Multimodal Pretraining

- Multimodal Pretraining on Domain M
- Finetuning on Domain G and Domain T

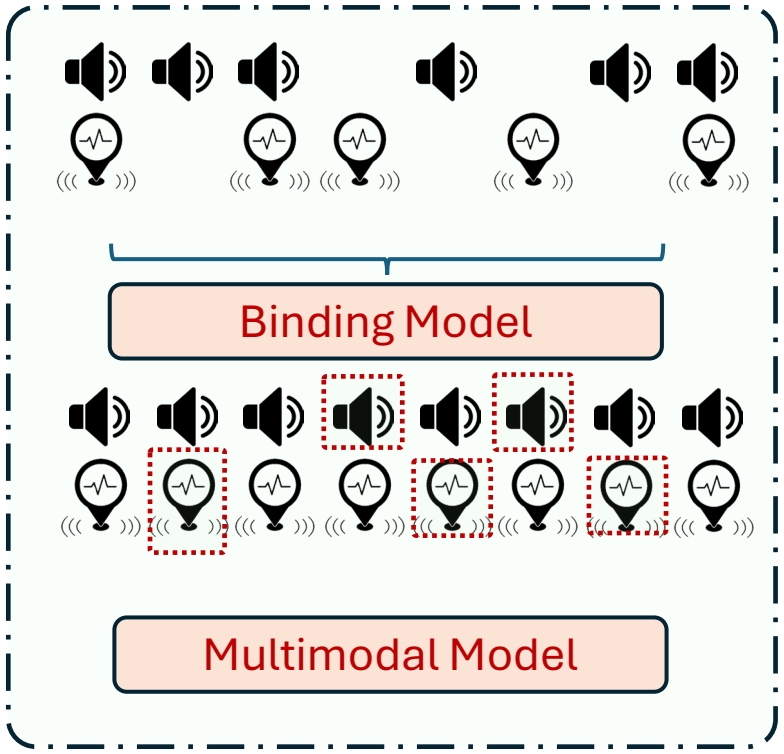
- InfoMAE shows flexibility as a
  - Multimodal Pretraining SSL
  - Pair-Efficient Pretraining SSL

Domain	**Domain G**		**Domain T**	
Frameworks	Acc	F1	Acc	F1
CMC [8]	0.7924	0.7897	0.6791	0.6776
FOCAL [40]	0.9137	0.9111	0.8156	0.8130
GMC [55]	0.7986	0.7947	0.3457	0.3387
MoCo [8]	0.8719	0.8688	0.7500	0.7483
SimCLR [6]	0.8418	0.8386	0.7288	0.7207
TNC [64]	0.6916	0.6797	0.5680	0.5625
TSTCC [15]	0.7080	0.7004	0.5804	0.5766
MAE [24]	0.6708	0.6642	0.4421	0.4365
CAV-MAE [18]	0.5507	0.5282	0.3457	0.3387
InfoMAE	**0.9196**	**0.9186**	**0.8546**	**0.8535**

Binding Between Multimodal Data



Model Binding

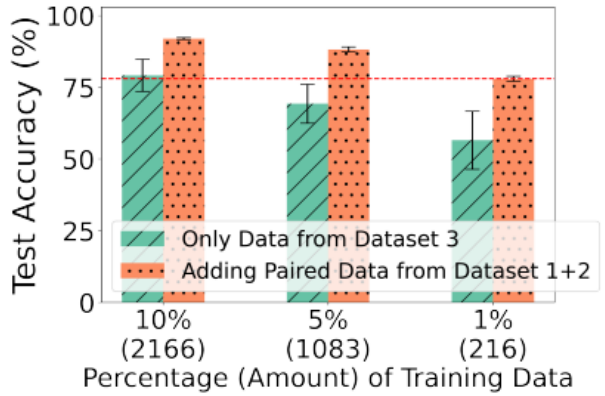


Data Binding

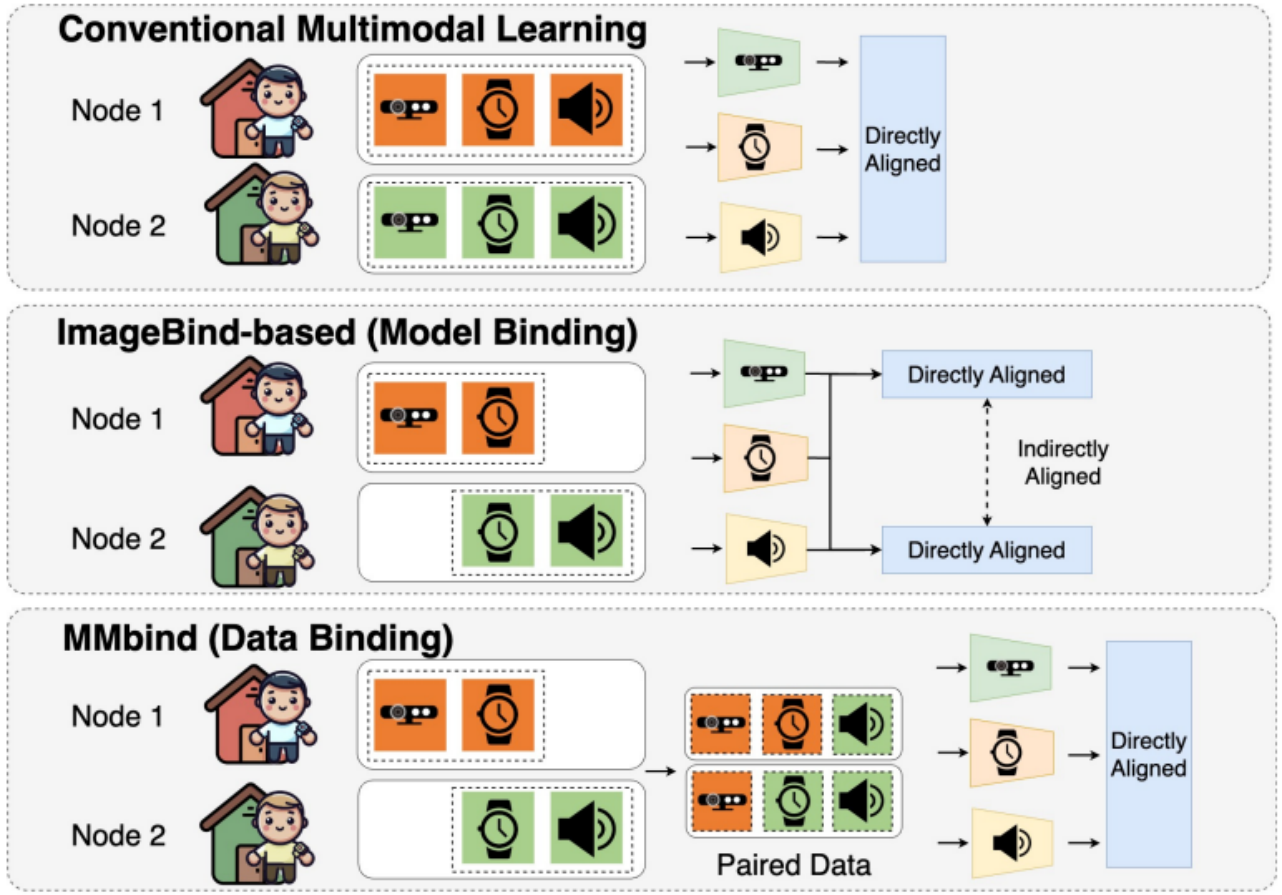
MMBind: Incomplete Sensor Data Binding

Distributed IoT Settings

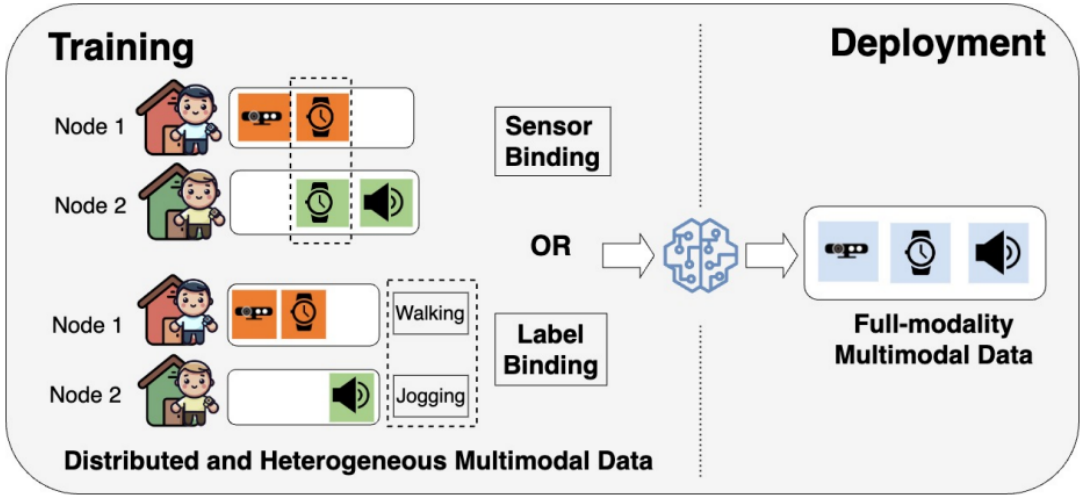
- Nodes are separate
- Different modalities are available

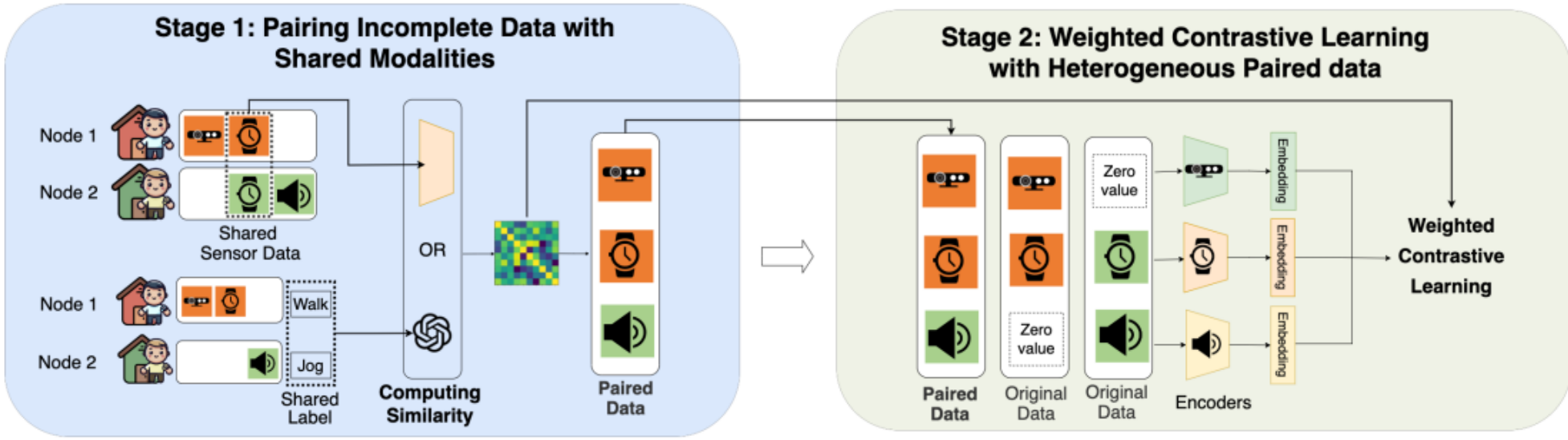


Dataset	Name	Sensor Modality	Samples
1	MotionSense	Acc, Gyro	12,636
2	Shoaib	Acc, Mag	4,500
3	RealWorld	Acc, Gyro, Mag	21,663 (1%~10% training)

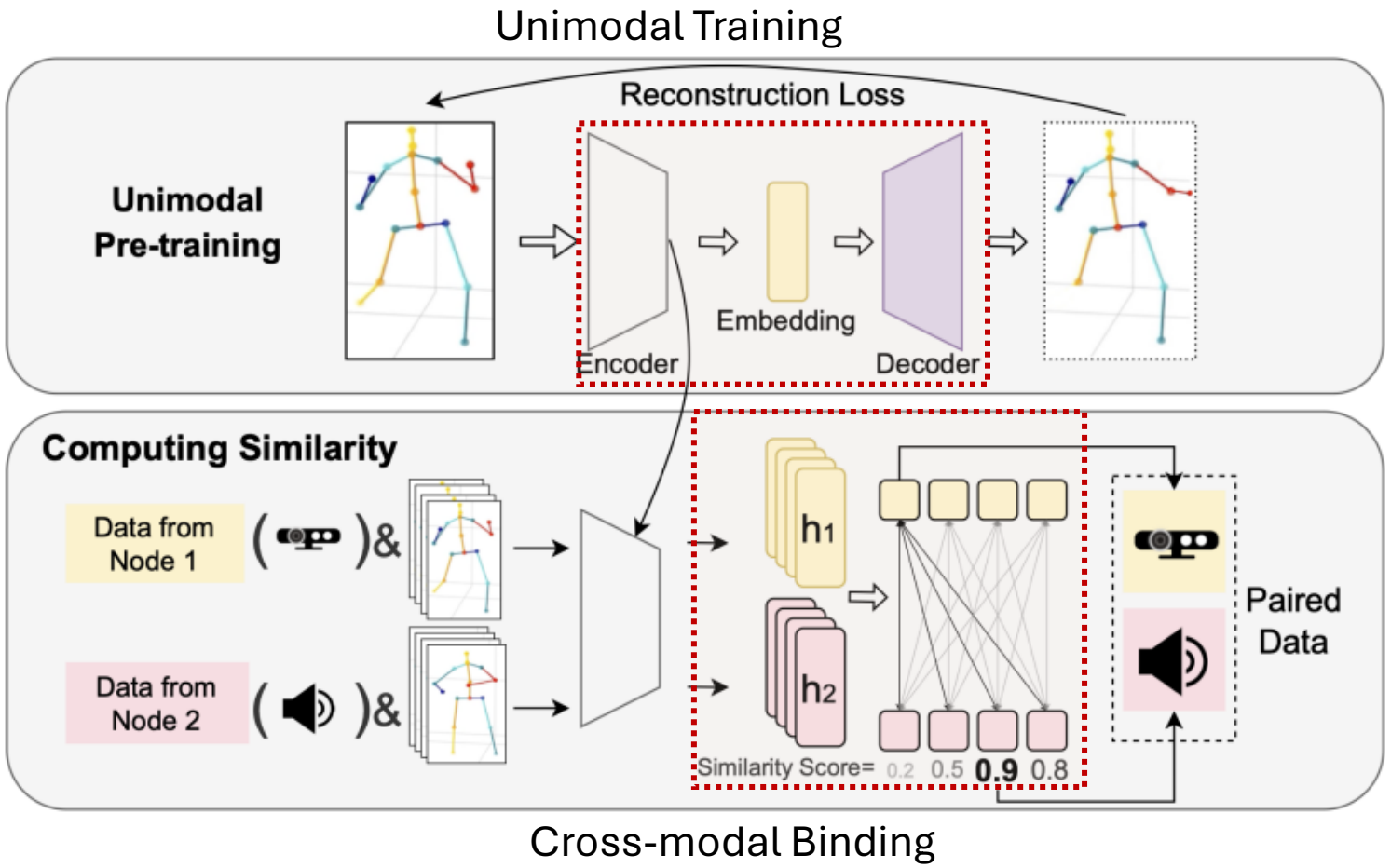


Two-stage Binding and Training

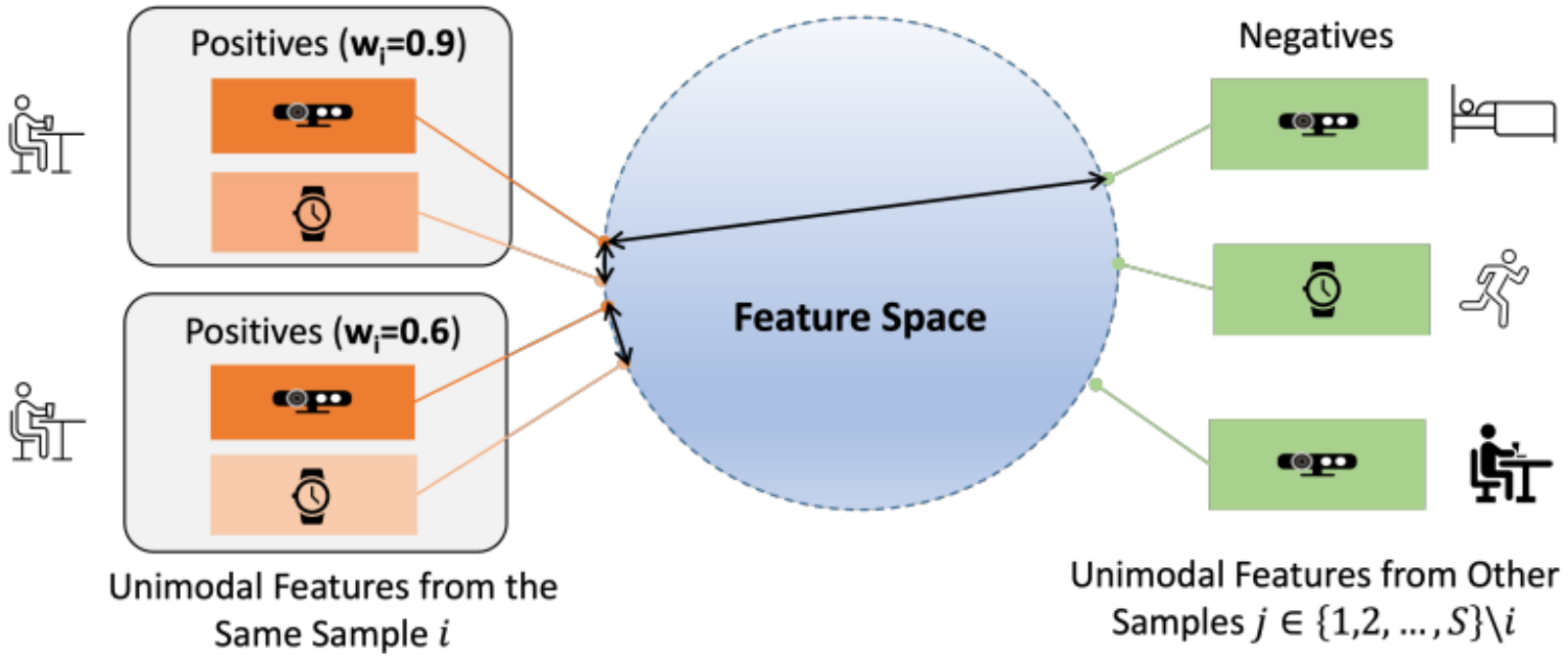




MAE-based Data-Binding through Sensors



Weighted Contrastive Learning



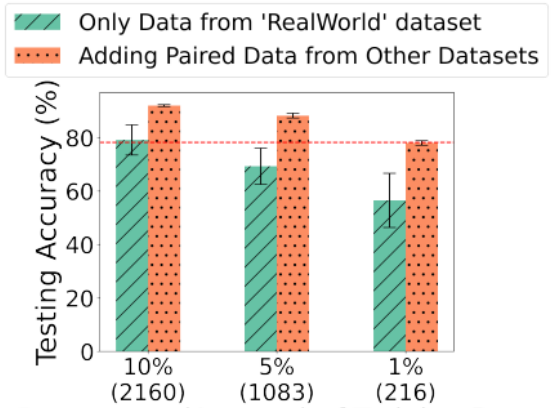
Cross-Dataset Binding

Datasets	UTD-MHAD		MM-FI		PAMAP2		SUN-RGBD	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1
Lower Bound	40.41	0.380	65.74	0.654	64.51	0.609	34.80	0.331
Unimodal	69.04	0.646	53.91	0.532	59.44	0.528	33.23	0.323
MIM	62.23	0.590	68.31	0.676	63.38	0.567	40.33	0.390
MPM	69.74	0.666	70.71	0.701	64.15	0.592	41.10	0.394
CMG	61.69	0.592	72.17	0.722	61.62	0.577	35.96	0.342
DCM	59.25	0.563	68.26	0.678	64.43	0.597	40.14	0.374
ImageBind	62.99	0.612	70.04	0.698	61.58	0.568	44.08	0.418
MMBind	**78.86**	**0.763**	**77.72**	**0.775**	**69.08**	**0.654**	**44.56**	**0.419**
Upper Bound	78.68	0.768	72.45	0.720	68.87	0.636	42.61	0.398

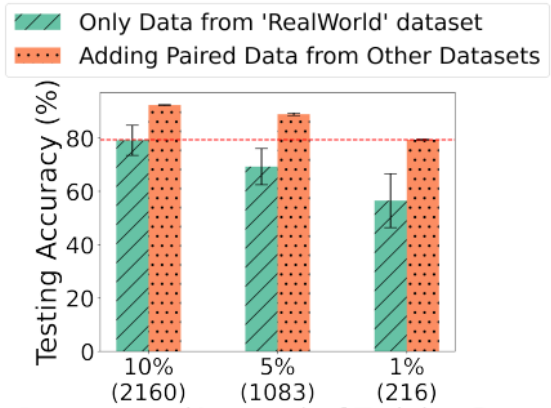
Table 3: Performance on cross-node data binding with a shared sensor modality.

Datasets	UTD-MHAD		MM-FI		PAMAP2		SUN-RGBD	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1
Lower Bound	40.09	0.378	65.74	0.654	63.52	0.581	30.420	0.290
Unimodal	69.54	0.668	44.35	0.432	67.68	0.637	32.68	0.247
MIM	75.84	0.729	61.80	0.611	68.45	0.657	53.52	0.533
MPM	73.95	0.726	63.44	0.623	69.23	0.678	52.18	0.506
MMBind	**83.04**	**0.813**	**77.09**	**0.775**	**74.44**	**0.723**	**55.54**	**0.549**
Upper Bound	85.63	0.844	84.12	0.841	73.65	0.722	63.91	0.637

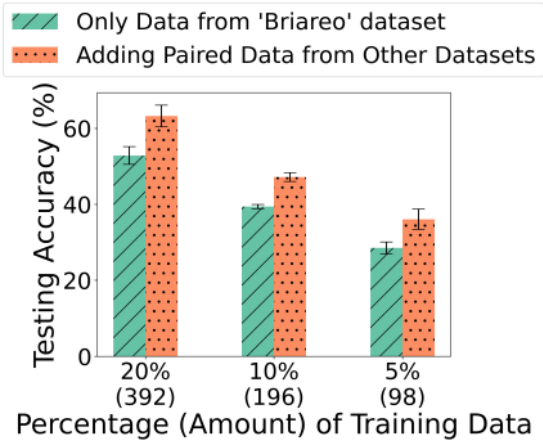
Table 4: Performance on cross-node data binding with shared data labels.



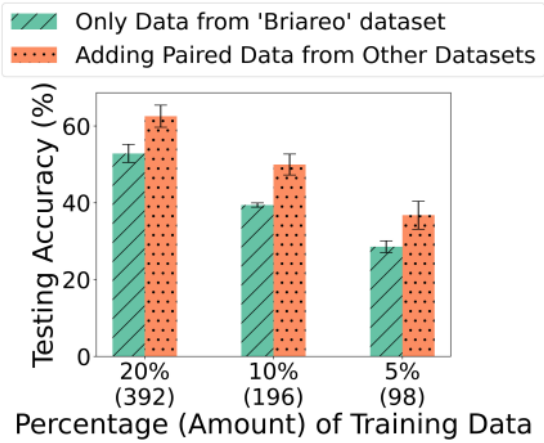
(a) Label binding.



(b) Acc binding.

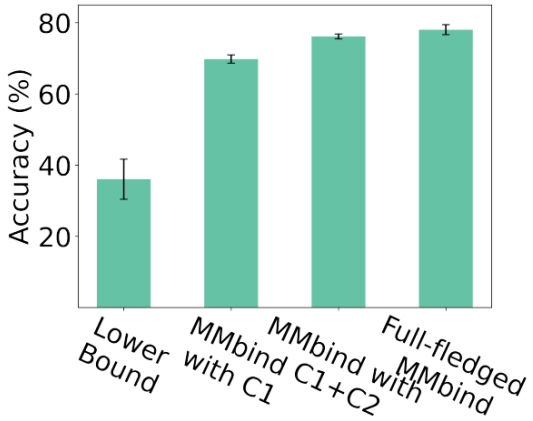


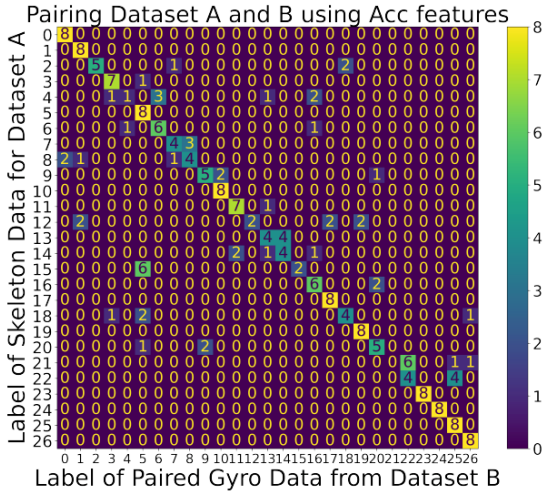
(a) Label binding.



(b) Skeleton binding.

Understanding MMBind's Performance





(a) Ablation study. (b) Data pairing.

Figure 11: Understanding MMBind's Performance.

C1: Leveraging paired data

C2: Adaptive learning with heterogeneous data

C3: Weighted contrastive

Binding Modality	Pairing Accuracy	MMBind Performance
Skeleton	**77.55%**	**77.72%**
WiFi CSI	4.2%	68.79%
Randomly	2.31%	68.96%

Table 7: Different binding modalities.

Amount of Paired Data	Pairing Accuracy	MMBind Performance
169 samples	82%	72.04%
432 samples	77%	76.32%
2,610 samples	40%	75.89%

Table 8: Impact of paired dataset size.

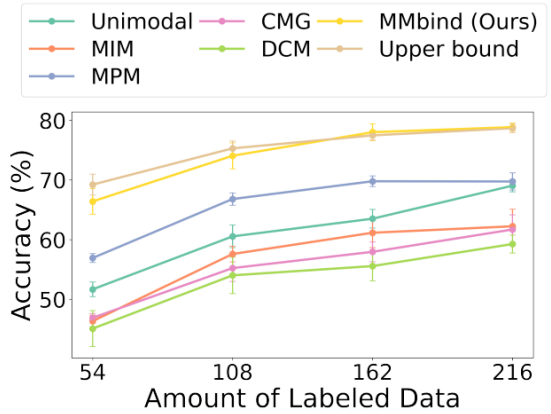


Figure 12: Different amounts of naturally paired data.

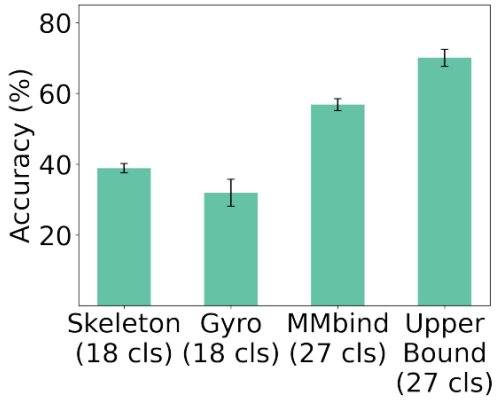
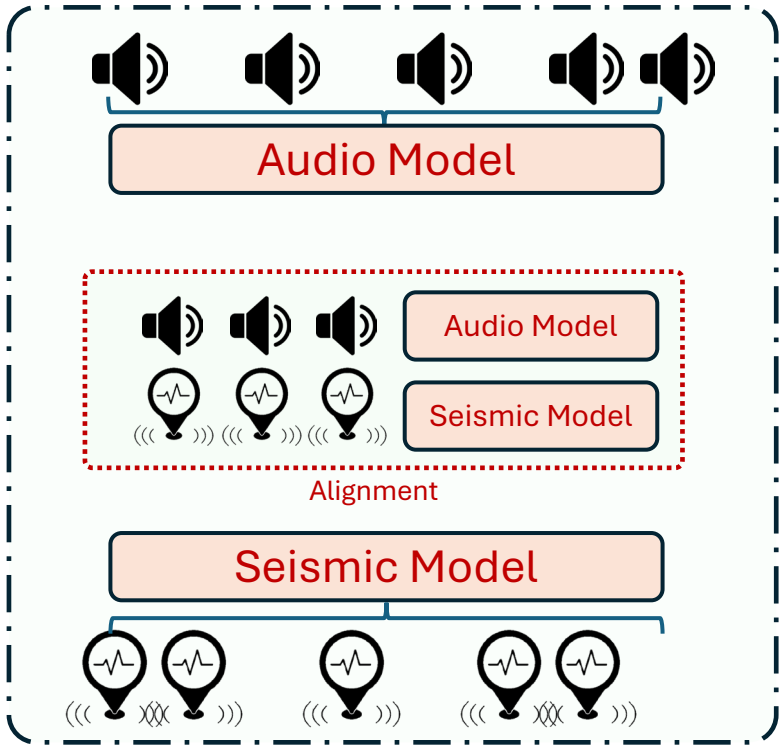
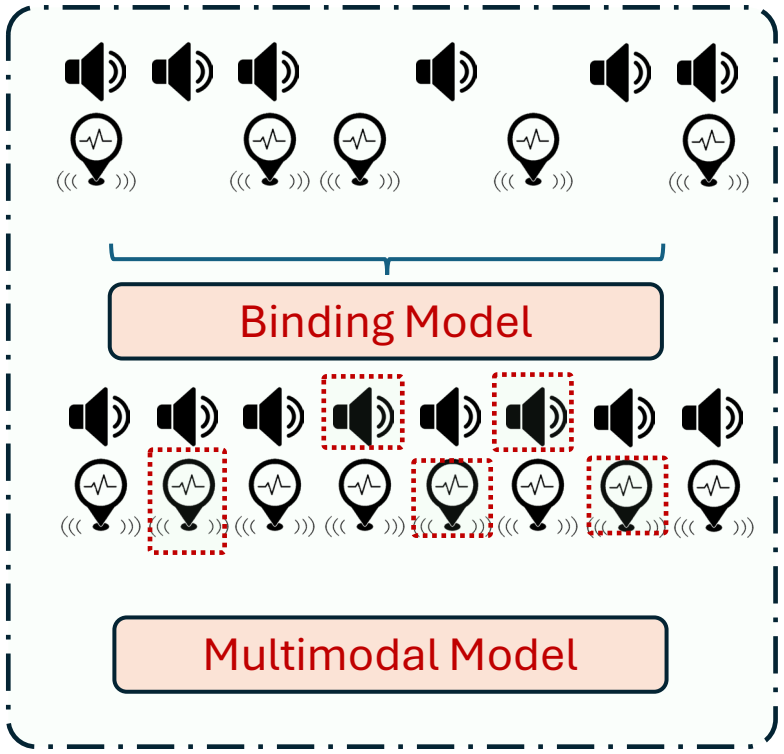


Figure 13: Partial class overlap.

Binding Between Multimodal Data



Model Binding



Data Binding

Agenda

1. Overview and Background
2. Self-Supervised Representation Learning for IoT Signals
3. Multimodal Representation with Incomplete Sensor Signals
4. **Multimodal Sensing Applications and Age of LLM**
5. Conclusion + Q & A

From Representations to Multimodal LLMs

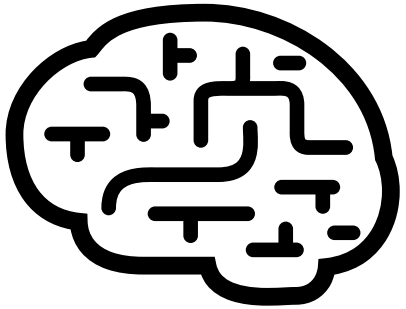
- **Multimodal Representation for Sensing Applications:**
  - Encode sensing data from different sensors to one representation
  - Use the representation for downstream tasks – **classification**
  - **Key Issue:**
    - Only works for fixed categories
    - Cannot extract insights from sensor data

- **Goal: Interpret sensing signals**
  - Extract richer information from sensing data
  - Reason over the extracted information for downstream tasks

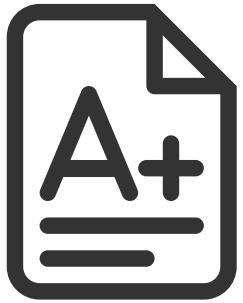
- **Solution: Multimodal LLMs**

Two Key Questions

How to implement Multimodal LLMs for Sensing applications?

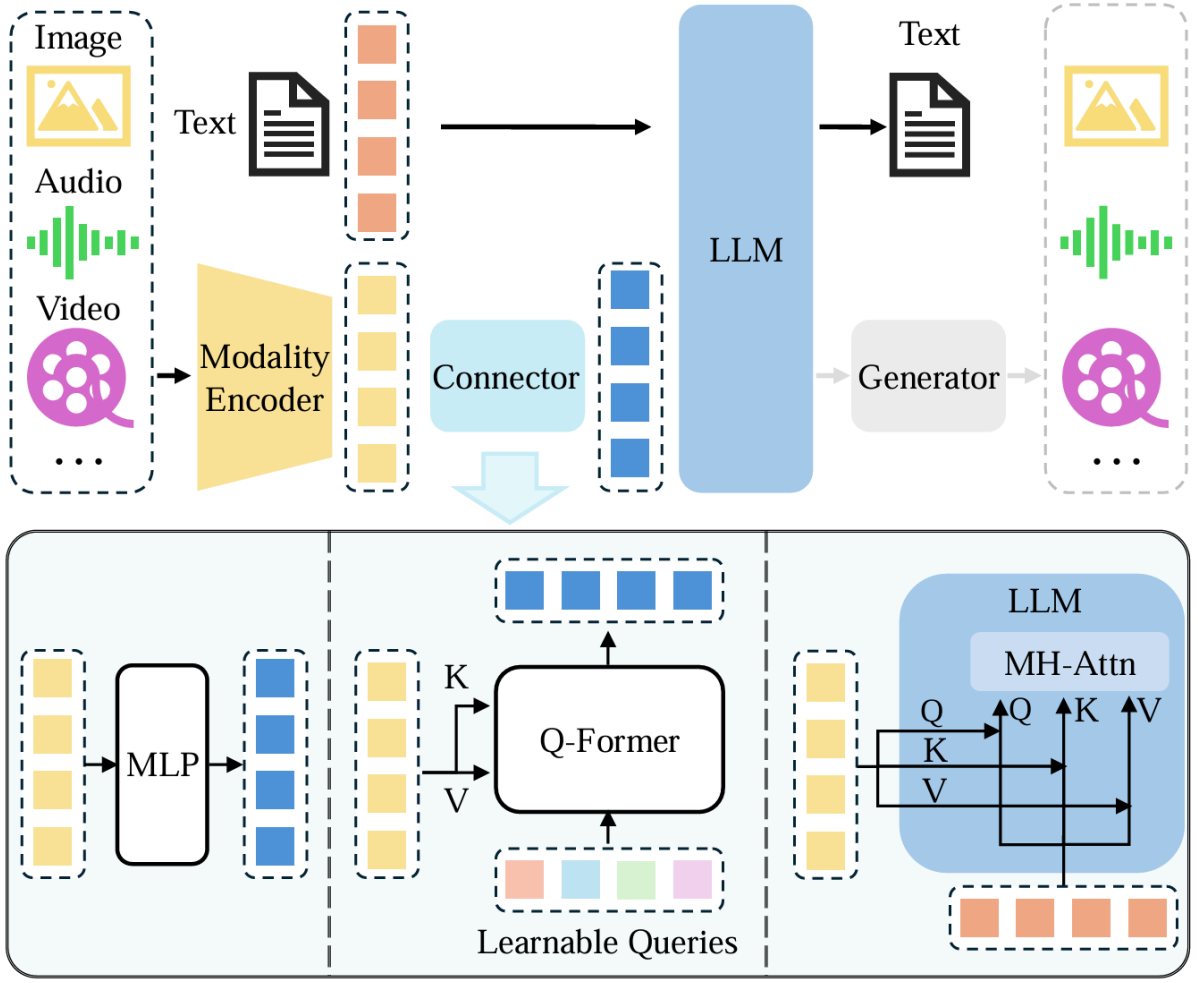


How to evaluate Multimodal LLMs in Sensing Applications?



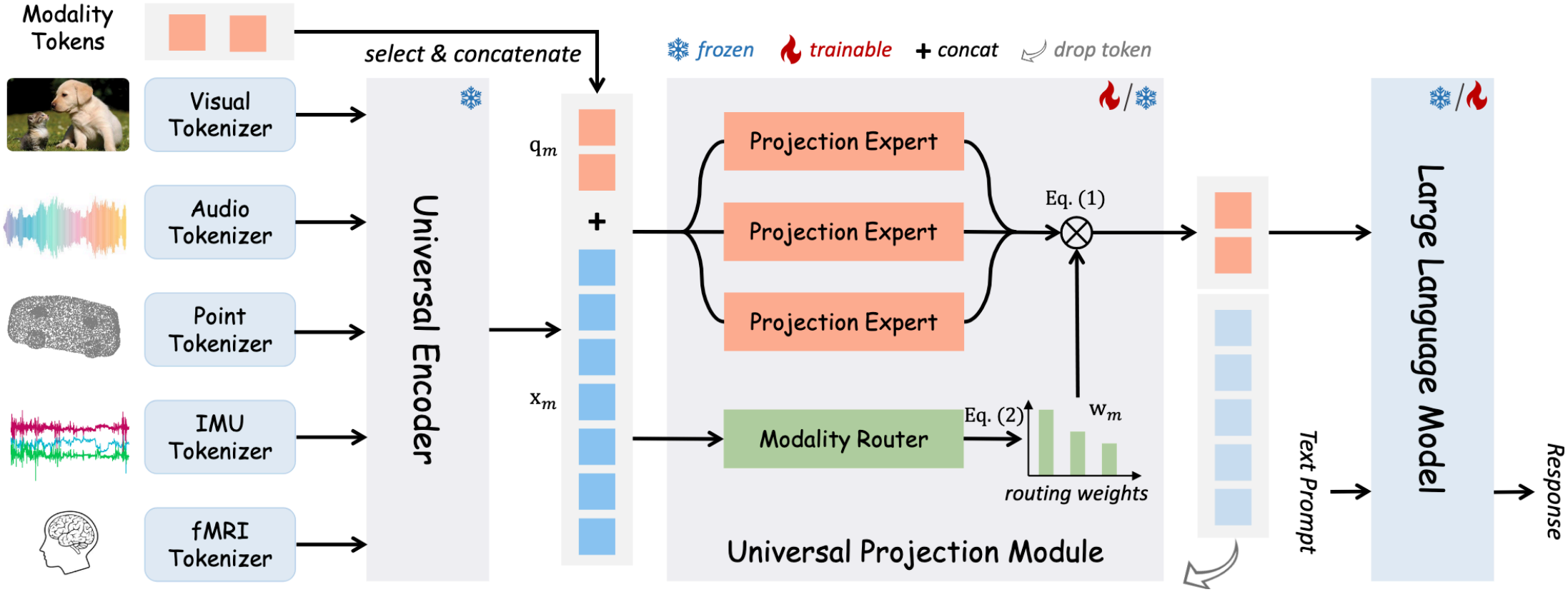
Incorporating Modalities into LLMs

- **Modules:**
  - Pre-trained Modality Encoders
  - Pre-trained LLMs
  - Connector for Two Modules
- **Types of Connectors:**
  - Projection-based
  - Query-based
  - Fusion-based



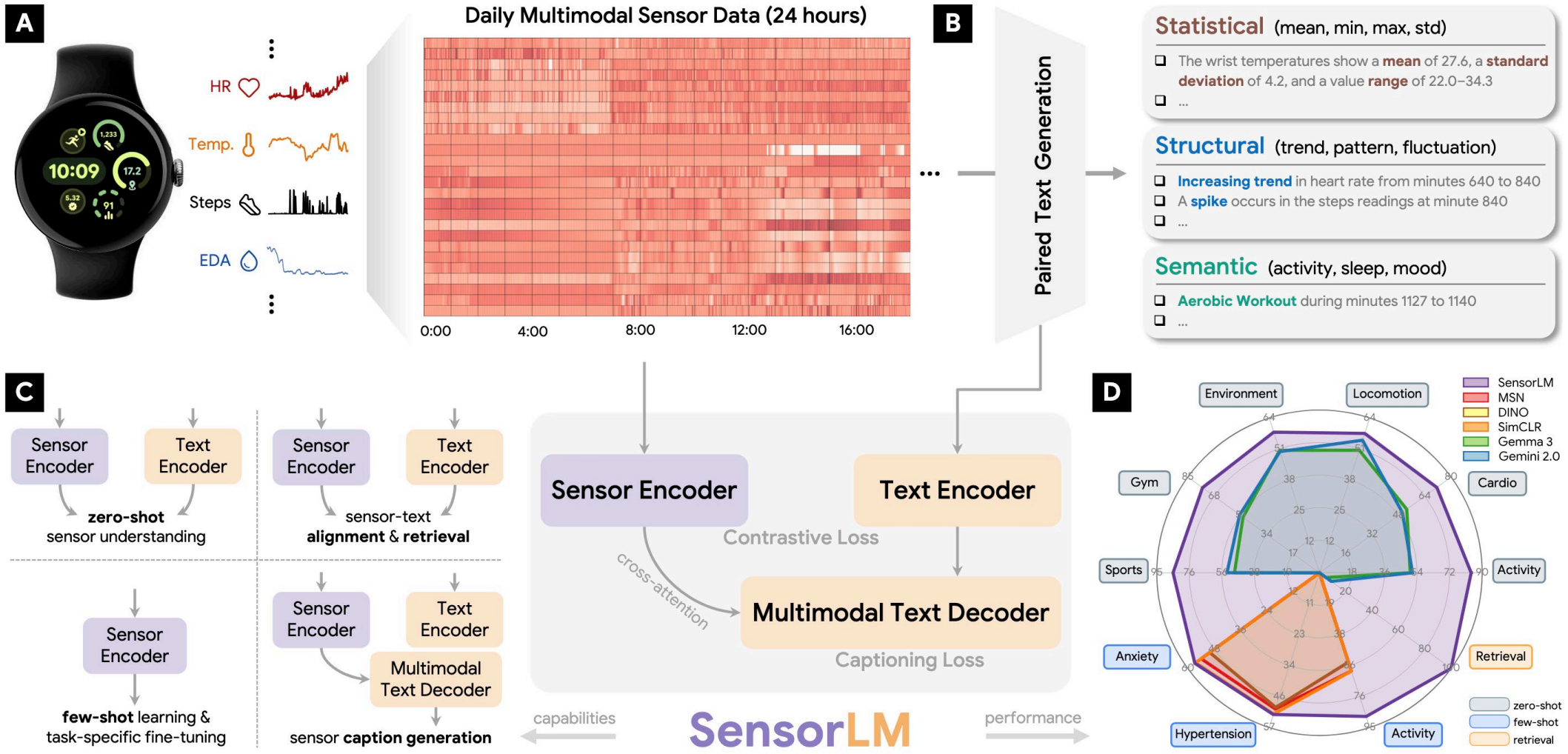
Yin et al. “A Survey on Multimodal Large Language Models” arXiv 2024

Example: OneLLM



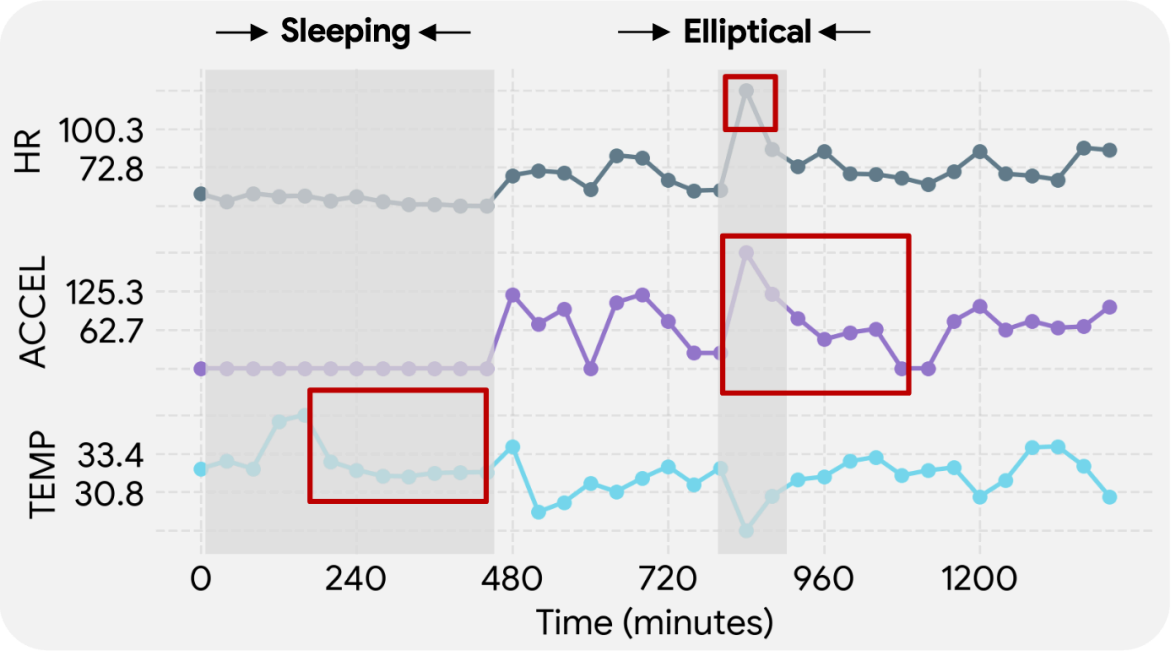
Han et al. "OneLLM: One Framework to Align All Modalities with Language" CVPR 2024

SensorLM – Aligning Sensor Data & Language



SensorLM – Sensor-Language Dataset

- **Statistical Captions:** quantitative summary of the sensor data
- **Structural Captions:** dynamic characteristics & patterns
- **Semantic Captions:** high-level meaning & context



Statistical

- ❑ The heart rate readings show a mean of 65.3, a standard deviation of 14.7 ...
- ❑ Accelerometer log energy features a mean of 49.2, a maximum of 194.0, a minimum of 0.0, and a standard deviation of 46.9 ...

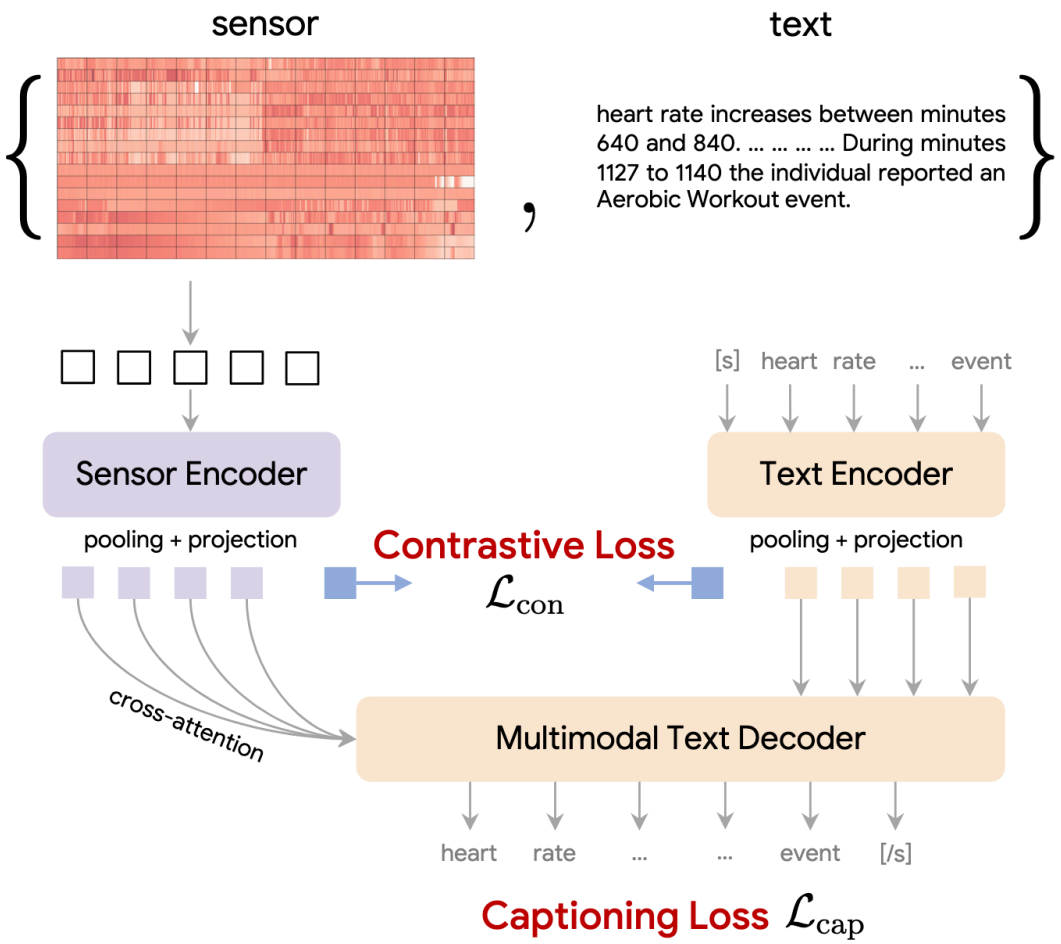
Structural

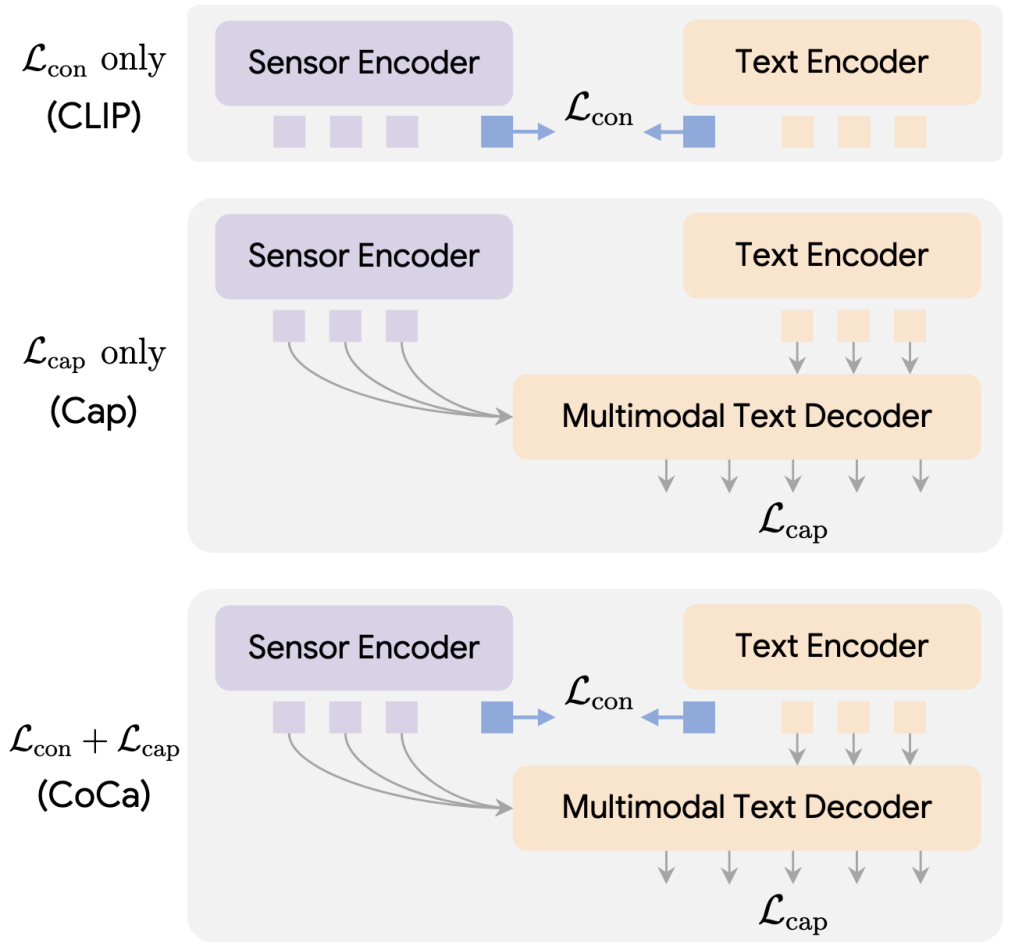
- ❑ A peak heart rate is detected at minute 840 ...
- ❑ From minute 840 to 1120, accel log energy exhibits an decreasing trend ...
- ❑ Skin temp slope indicates a decreasing trend spanning minute 40 to 480 ....

Semantic

- ❑ Elliptical episode occurred between minute 830 and 850 ...
- ❑ Sleep took place during the minutes 5 through 449 ...

SensorLM – Architectures & Pretraining Objectives





SensorQA – Benchmark for Sensing Applications

- **Motivation:**
  - Early benchmarks focused on **classifying sensor data** – cannot extract insights
  - QA benchmarks is mainly studied in **language & vision domains**
  - LLMs for sensor data currently constrained on **short-term sensor data**

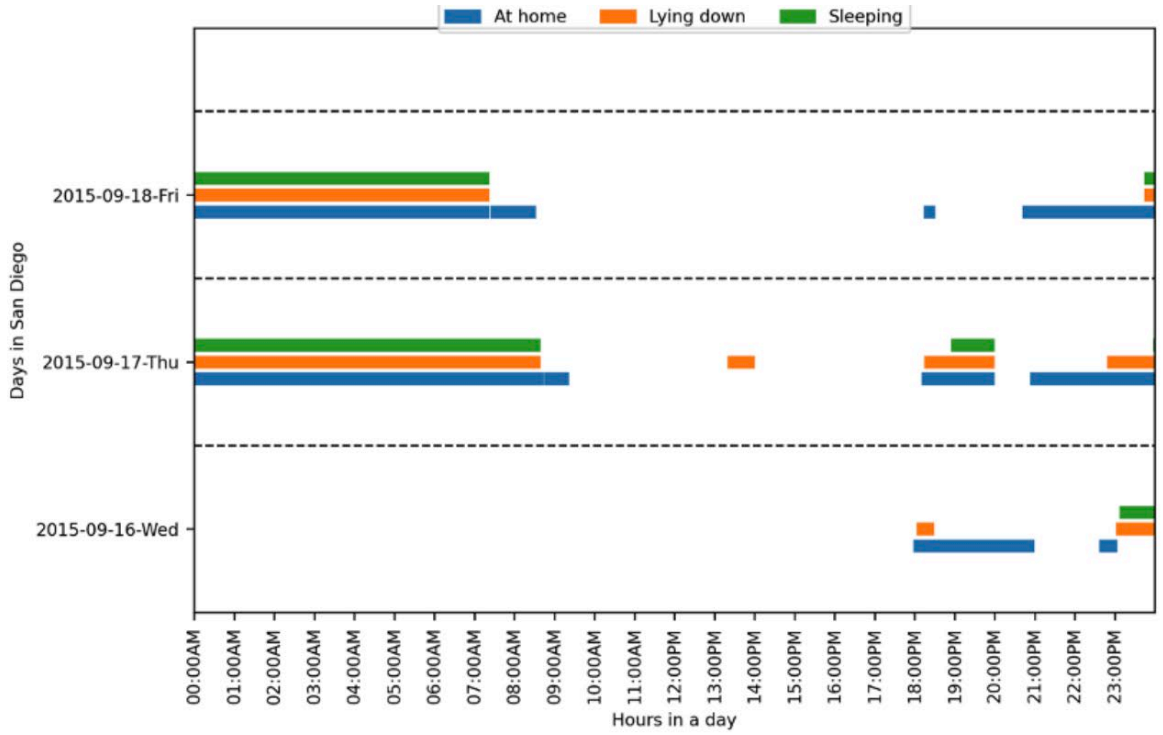
- **Sensor Data Collection: ExtraSensory**
  - In-the-wild sensing data collection – emphasizing **real-life settings**
  - Sensor measurements from 60 subjects & more than 300K min of data

- **QA Data Collection: Amazon Mechanical Turk**
  - **Human-created contents** – reflecting human interests & needs
  - Sensor data are visualized like **Gantt charts** to let workers generate QAs

SensorQA – Benchmark for Sensing Applications

On which day did I spend the most time with co-workers?

You spent the most time with co-workers on Friday.



How did I spend my time at home on Friday morning?

You spent your time at home on Friday lying down and sleeping.

SensorQA – Benchmark Results

- Best baseline accuracy is only **28%**
- Sensor + Text baselines underperform even Text-only baselines

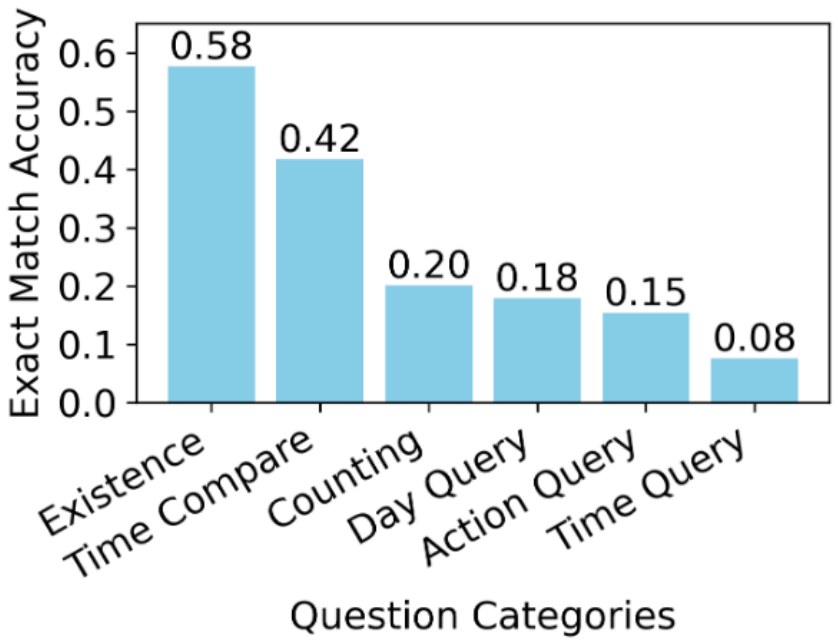
Modalities	Backbone Model	FSL/FT[1]	Full Answers					Short Answers
			Rouge-1 (↑)	Rouge-2 (↑)	Rouge-L (↑)	Meteor (↑)	Bleu (↑)	Accuracy (↑)
Text	GPT-3.5-Turbo	FSL	0.35	0.23	0.32	0.43	0.16	3.0%
Text	GPT-4	FSL	0.66	0.51	0.64	0.66	<u>0.39</u>	16.0%
Text	T5-Base	FT	0.71	0.55	0.69	<u>0.70</u>	**0.43**	25.4%
Text	Llama	FT	<u>0.72</u>	**0.62**	**0.72**	**0.72**	0.38	26.5%
Vision+Text	GPT-4-Turbo	FSL	0.38	0.28	0.36	0.51	0.15	14.0%
Vision+Text	GPT-4o	FSL	0.39	0.28	0.37	0.61	0.25	7.0%
Vision+Text	Llama-Adapter	FT	**0.73**	<u>0.57</u>	<u>0.71</u>	**0.72**	**0.43**	**28.0%**
Vision+Text	Llava-1.5	FT	0.62	0.46	0.60	0.58	0.35	21.5%
Sensor+Text	IMU2CLIP-GPT4	FSL	0.44	0.28	0.40	0.53	0.16	13.0%
Sensor+Text	DeepSQA	FT	0.34	0.05	0.34	0.18	0.0	<u>27.4%</u>
Sensor+Text	OneLLM	FT	0.12	0.04	0.12	0.04	0.0	5.0%

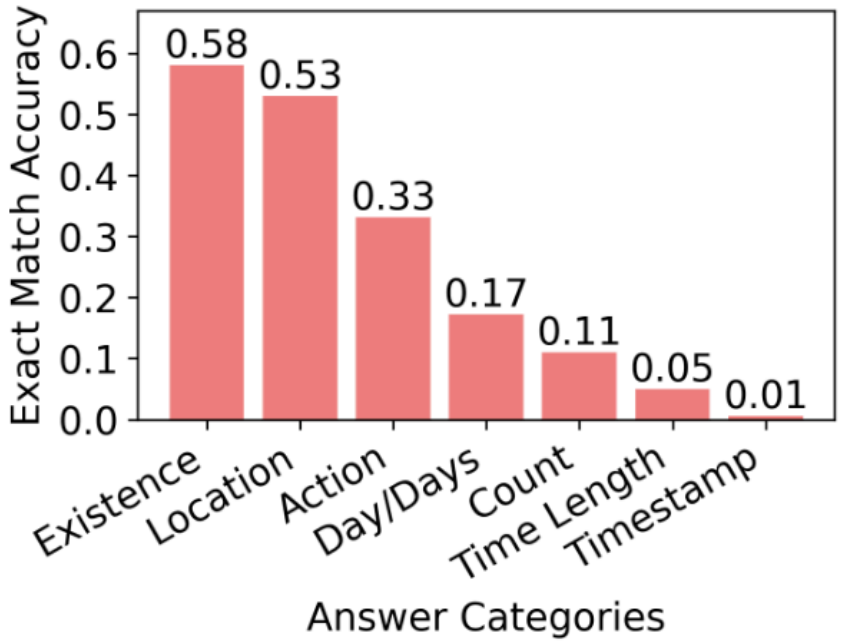
[1]FS: Few-Shot Learning. FT: Finetuning.

Table 4: Benchmark results of baselines on SensorQA. Bold and underlined values show the best and second-best results.

SensorQA – Benchmark Results with Llama

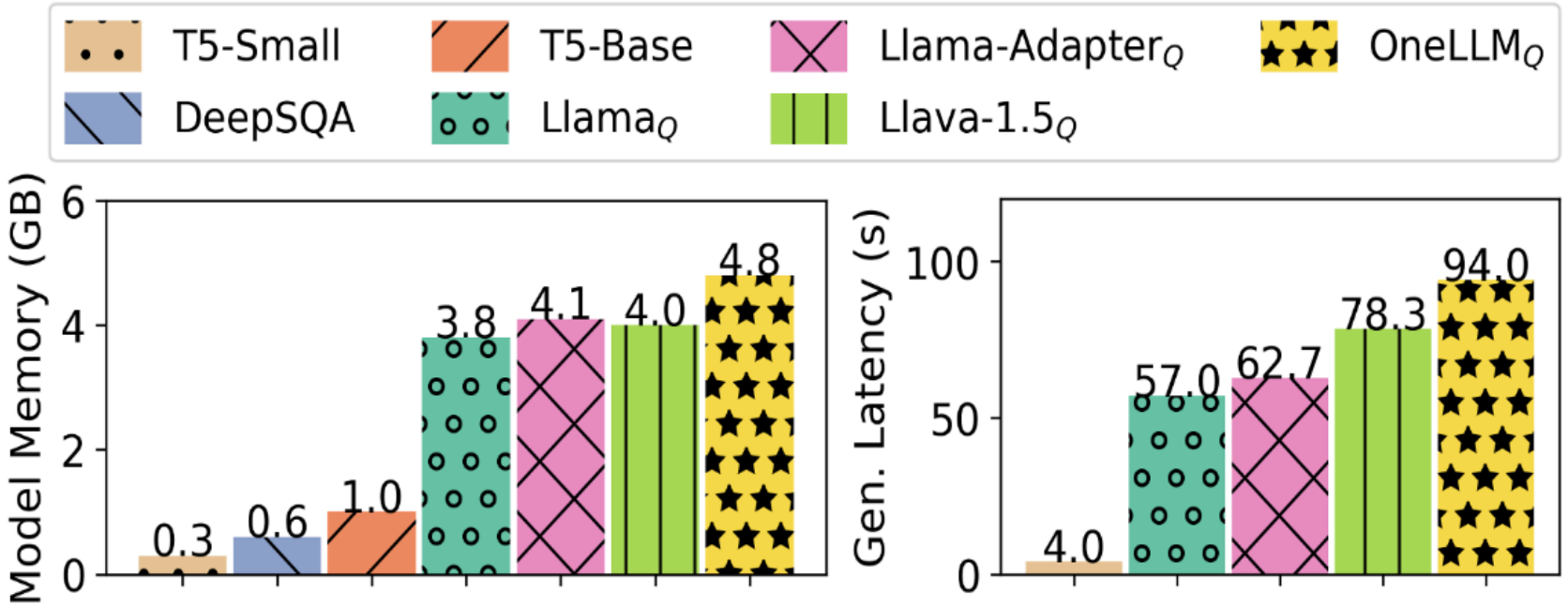
- **Time-related queries** are especially hard
- Even the accuracy for **queries about existence** is only **58%**





SensorQA – Benchmark Results

- **Out-of-memory** can happen on some baselines
- On efficiency, LLM approaches are very **slow** on edge devices (Jetson TX2)



Discussion

- Sensing Multimodal LLMs largely mirrors **vision language models (VLMs)**
  - Use similar pipeline to incorporate new sensing modalities

- **Key Ch**
  - It
  - Bu

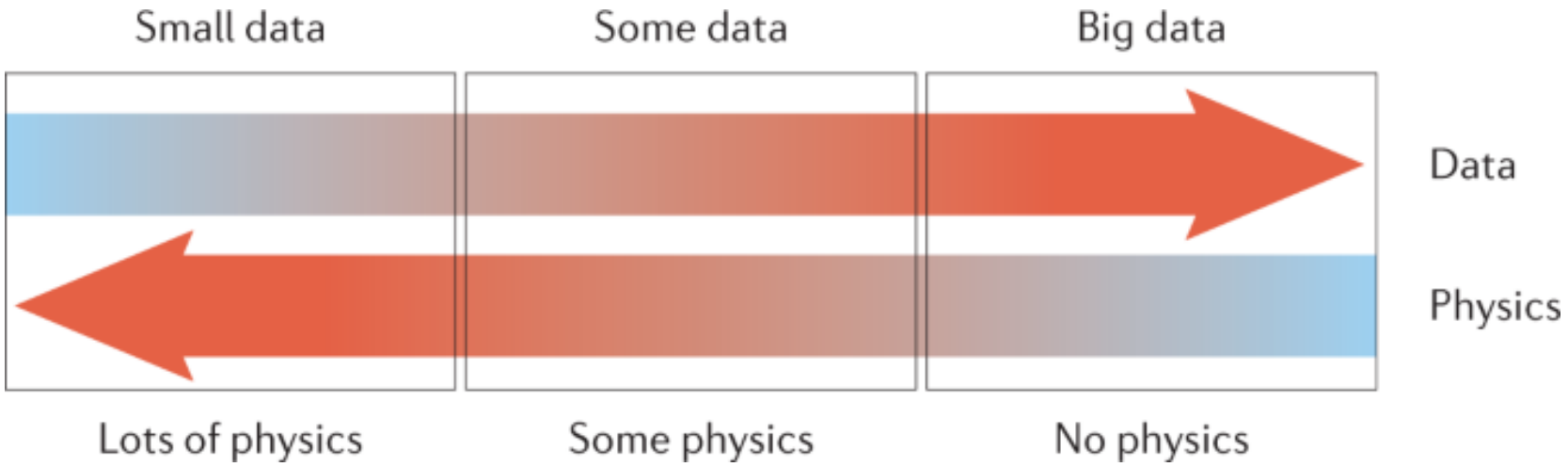
- **Why**
  - Or
  - **Ha**
  - A lot of **noise** exists in sensing data

We need new techniques tailored for Sensing Multimodal LLMs!

Discussion

What is not considered in VLMs – Physics & Prior Knowledge

- Sensing data is sampled from **Cyber Physical Systems**
  - Interaction with real physical world implies potential physical knowledge
- Physics & prior knowledge can work as **constraints**, reducing the requirements on data

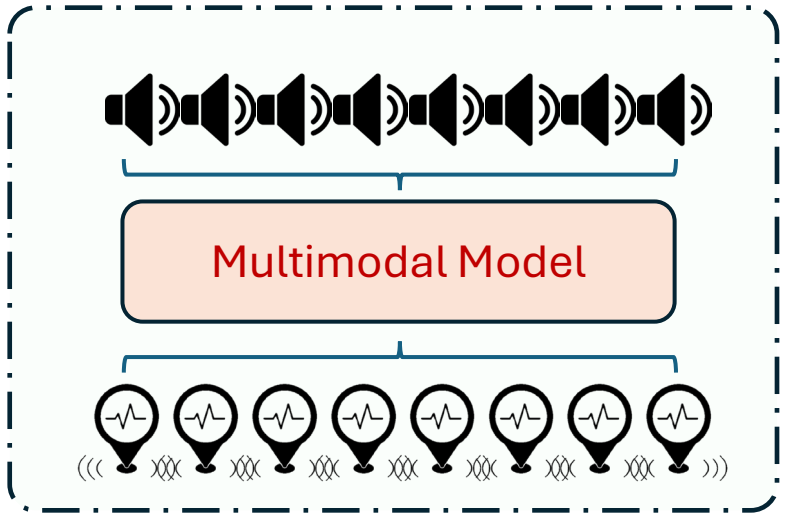


Karniadakis et al. “Physics-informed Machine Learning”, Nature Review Physics

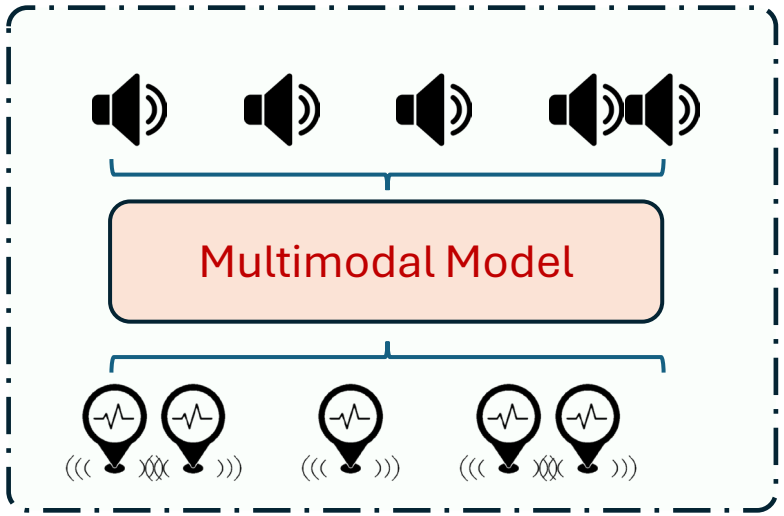
Agenda

1. Overview and Background
2. Self-Supervised Representation Learning for IoT Signals
3. Multimodal Representation with Incomplete Sensor Signals
4. Multimodal Sensing Applications and Age of LLM
5. **Conclusion + Q & A**

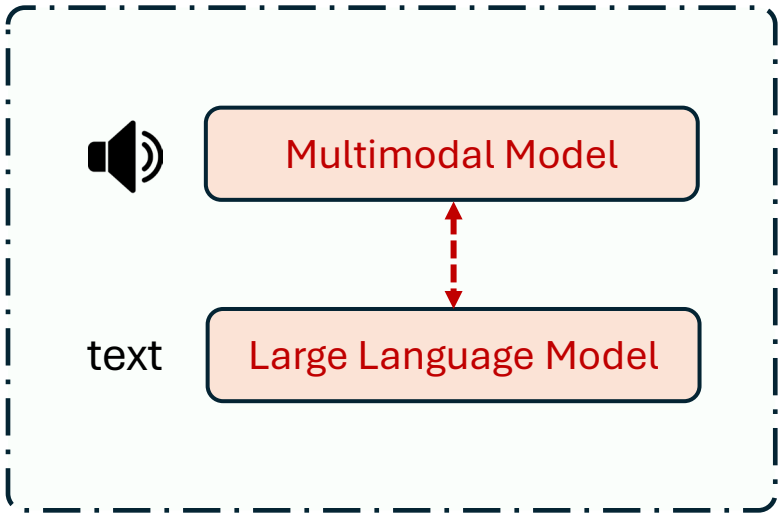
Conclusion



Multimodal Representation Learning



Multimodal SSL with Incomplete Multimodal Pairs



Multimodal SSL with LLM and Benchmarking

Thank you – Q & A

{tkimura4, yuhengp2, hongjue2}@Illinois.edu