Association for the Advancement of Artificial Intelligence

Decomposing and Fusing Intra- and Inter-Sensor Spatio-Temporal Signal for Multi-Sensor Wearable Human Activity Recognition

Haoyu Xie, Haoxuan Li, Chunyuan Zheng, Haonan Yuan, Guorui Liao, Jun Liao, Li Liu

Chongqing University, Peking University, Beihang University

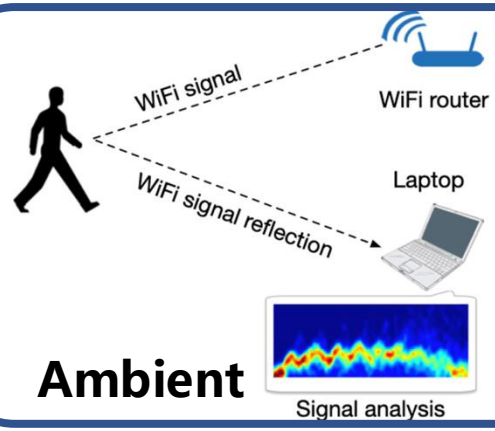
Contact: haoyuxie@stu.cqu.edu.cn

Introduction to Wearable Human Activity Recognition (WHAR)

What is HAR

•Human Activity Recognition (HAR) classifies human actions based on data sources.

Vision



Ambient

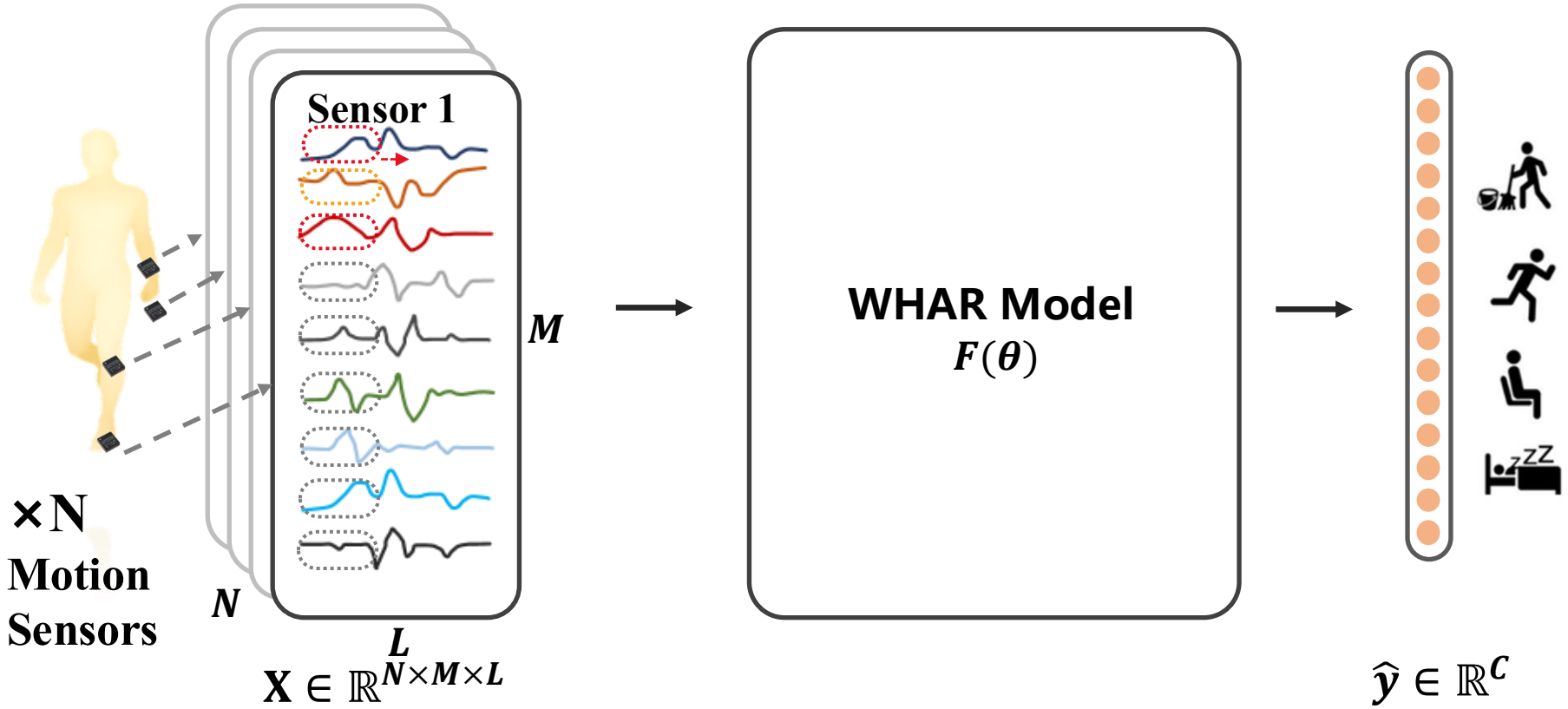
Wearable

Categories:

- **Vision-based:** Limited by lighting, occlusions, and privacy concerns.
- **Environment-based:** Relies on ambient sensors (WiFi, sound), sensitive to background noise.
- **Wearable Sensor-based:** Provides direct motion data, robust to environmental changes.
  - Uses **IMUs (accelerometers, gyroscopes, magnetometers)** placed on the body.
  - Applicable in **healthcare, sports, smart homes, and rehabilitation.**
  - Offers **continuous monitoring, low latency, and privacy advantages.**

Problem Definition

Wearable Human Activity Recognition **(WHAR)** is modeled as a **multivariate time series classification task.**



Input: Motion Data from Wearable Sensors

- N **wearable sensors**, each capturing
- M **modality variables**

(e.g., accelerometer, gyroscope, magnetometer)

- L **time steps of an activity window**

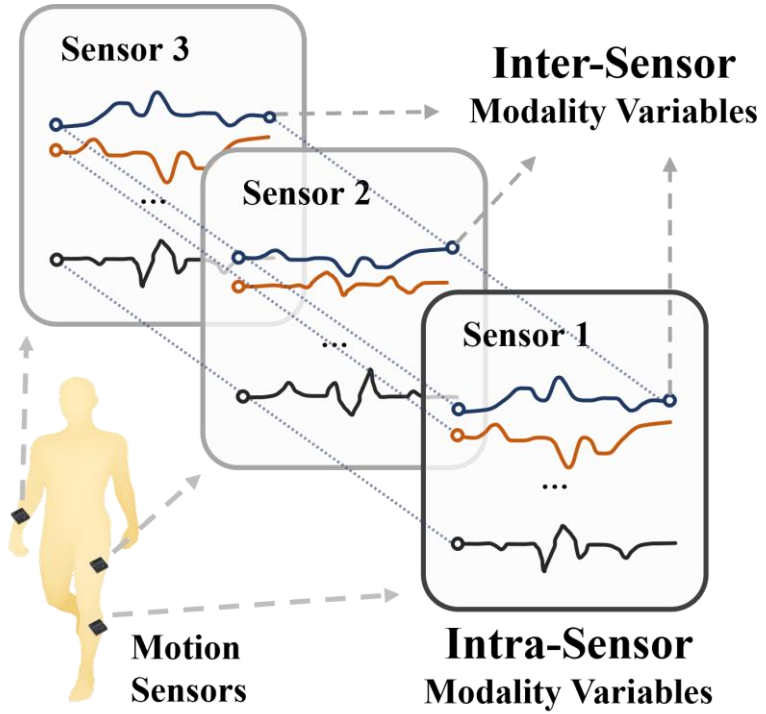
Objective: Activity Classification

- Train a model $F(\theta)$
- Minimize classification error between
  - **Predicted labels** $\hat{y}$
  - **True labels** y

Challenges in WHAR

Two Types of Variable Relationships in WHAR

- **Intra-Sensor Variables:** Different variables from the same sensor.
- **Inter-Sensor Variables:** Variables from sensors on different body parts.



- **Intra-Sensor Temporal Feature Extraction:** How to balance local and global temporal pattern extraction and cross-variable integration.

- **Inter-Sensor Spatio-Temporal Correlations:** How to capture coordinated sensor spatial interactions (e.g., arms & legs in activities like running & cycling).

Existing issues of WHAR Methods

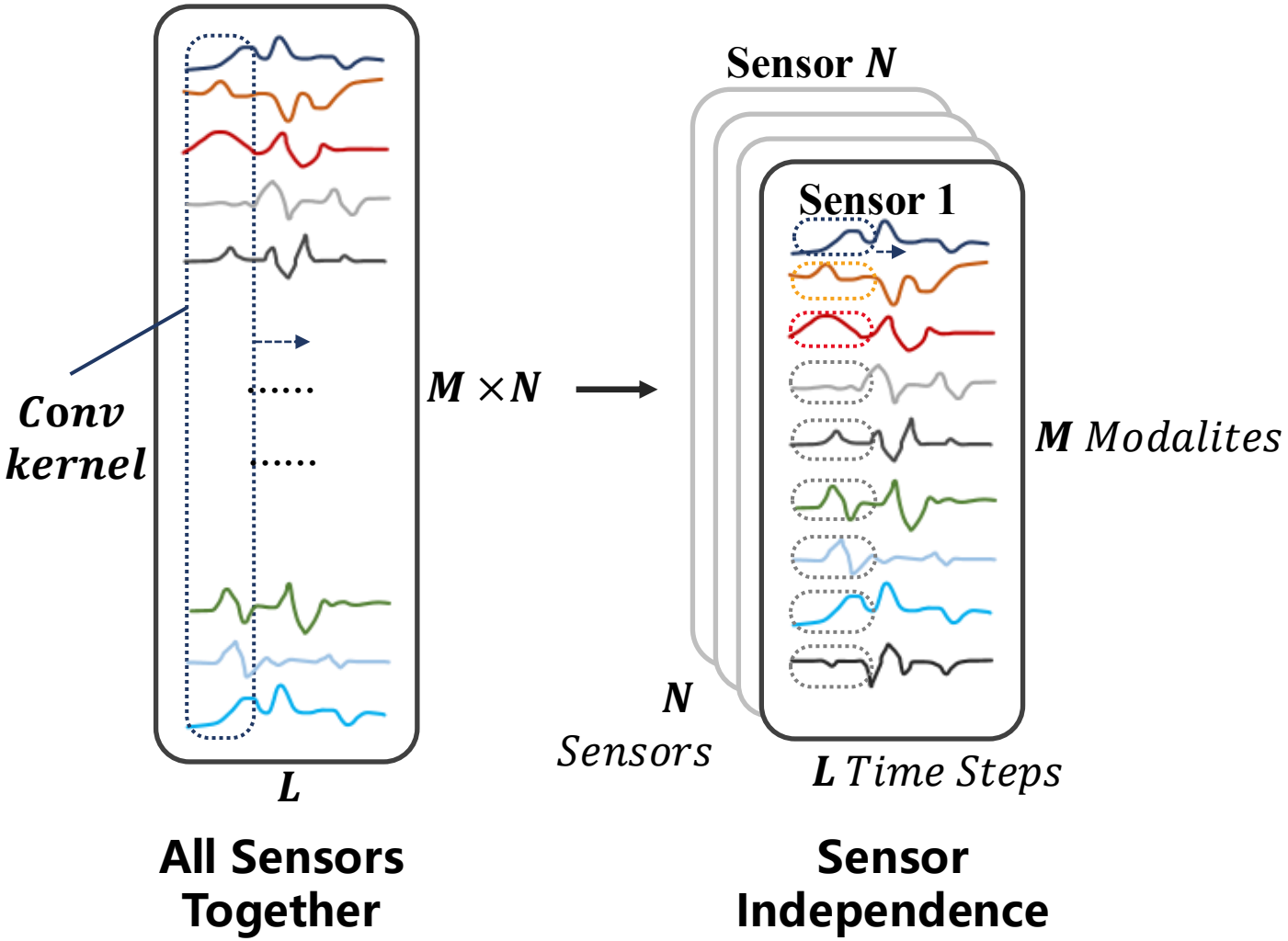
Capturing Intra-Sensor Temporal Features:

- Approach 1: Shared CNN kernels + LSTM (e.g., *DeepConvLSTM*, *Attend & Discriminate)*
  - **Pros:** Integrate local patterns & global dependencies
  - **Cons: Lose individual variable details, limiting cross-variable interactions**
- Approach 2: Conv1D fusion before temporal extraction (*DynamicWHAR*, *HyperHAR)*
  - **Pros:** Straightforward multimodal combination
  - **Cons: Lose high-level variable-specific temporal info**

Capturing Inter-Sensor Spatio-Temporal Correlations

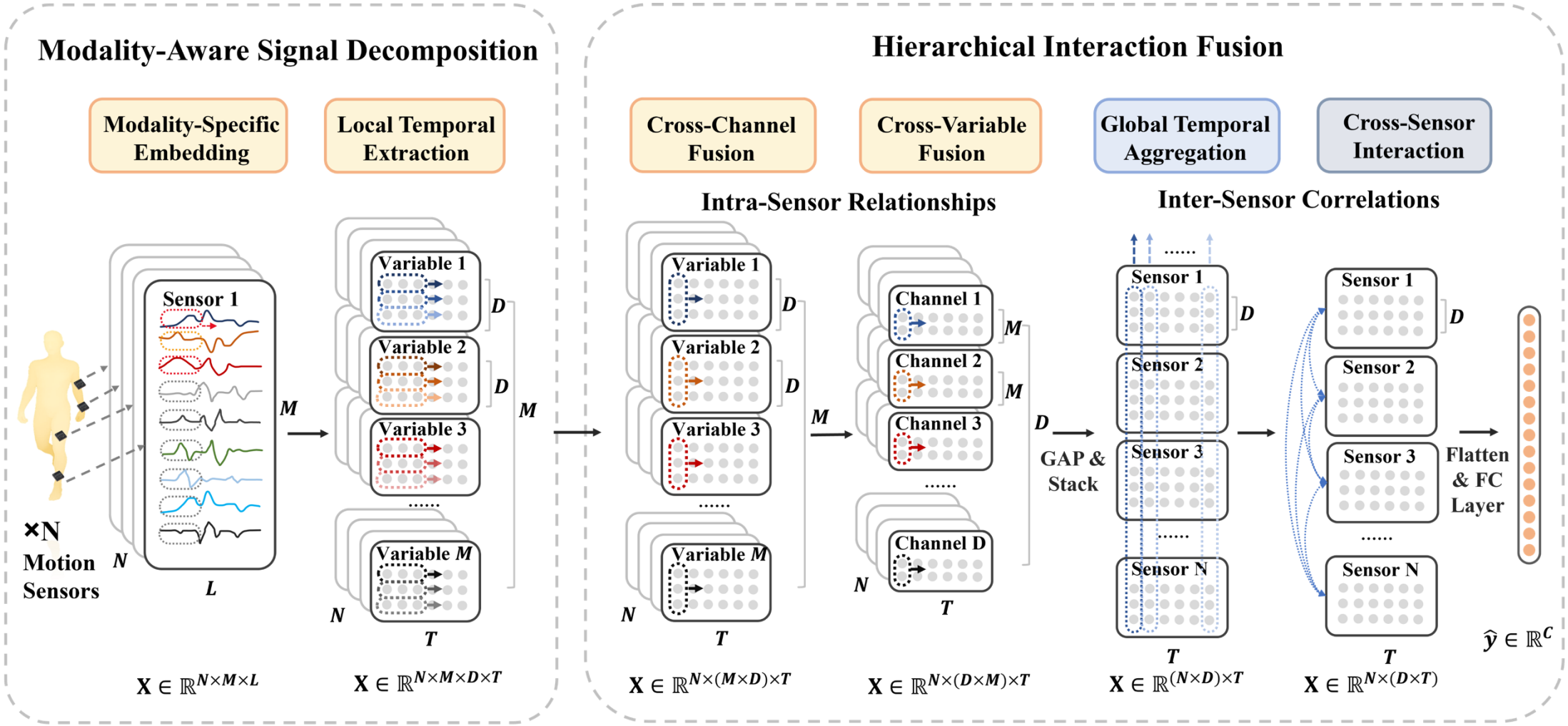
- Common methods (dense layers, RNNs) fuse all sensor variables indiscriminately, missing spatial relationships.
- *DynamicWHAR* : Separates sensors, then uses GCN for dynamic correlations:
  - **Depend on static or predefined graph structures**
  - **Struggle with directional correlations & scaling to large sensor networks**

Motivation: Decompose!

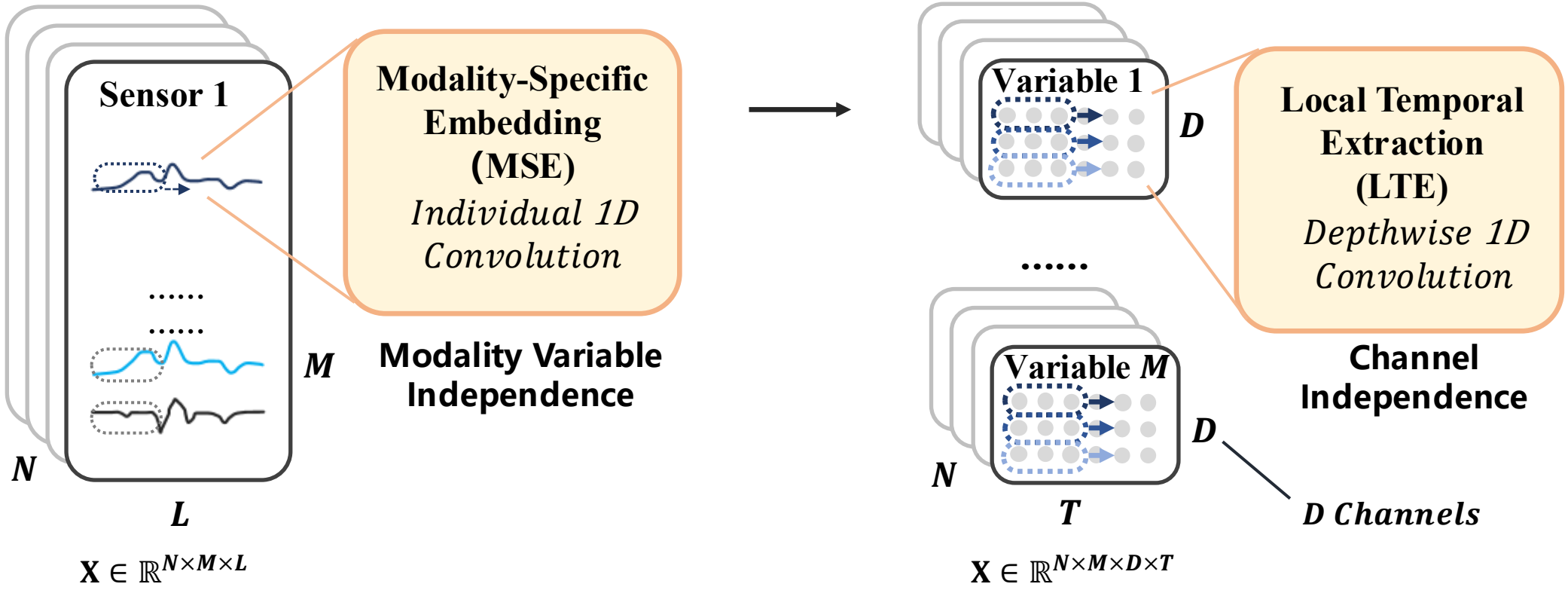


- **Sensor Independence (Inter-sensor):** Instead of fusing all sensor variables together, we first separate each sensor to extract features individually and later fuse them with spatial correlations.
- **Modality Independence (Intra-sensor):** Treat each modality variables independently and extract high-level temporal representations from each modality, mitigating interference between intra-sensor modalities.

Model Overview: DecomposeWHAR

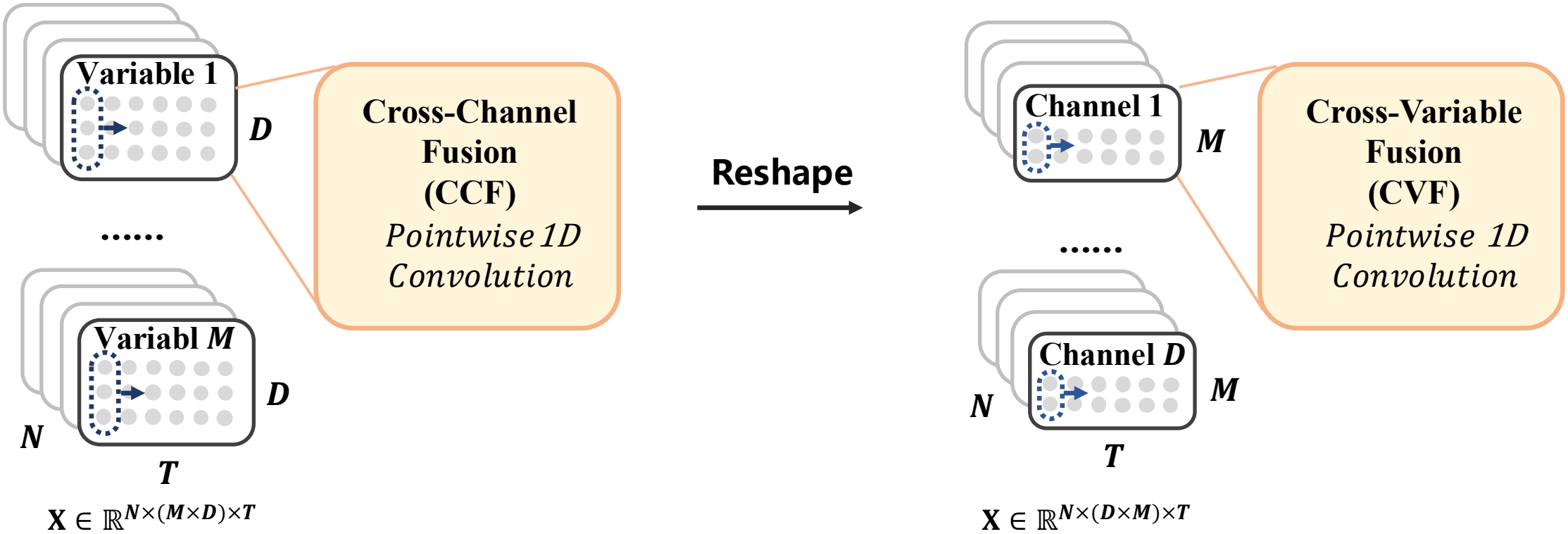


Modality-Aware Signal Decomposition



- **Modality-Specific Embedding (MSE):** Transform data into high-dimensional representations, independent capture of each modality variable's temporal dynamics.
- **Local Temporal Extraction (LTE):** Captures local temporal features maintaining the independence of channels of modality variables. Each variable channel undergoes separate convolution operations, preserving their distinct characteristics.

Hierarchical Interaction Fusion
Cross-Channel & Cross-Variable Fusion

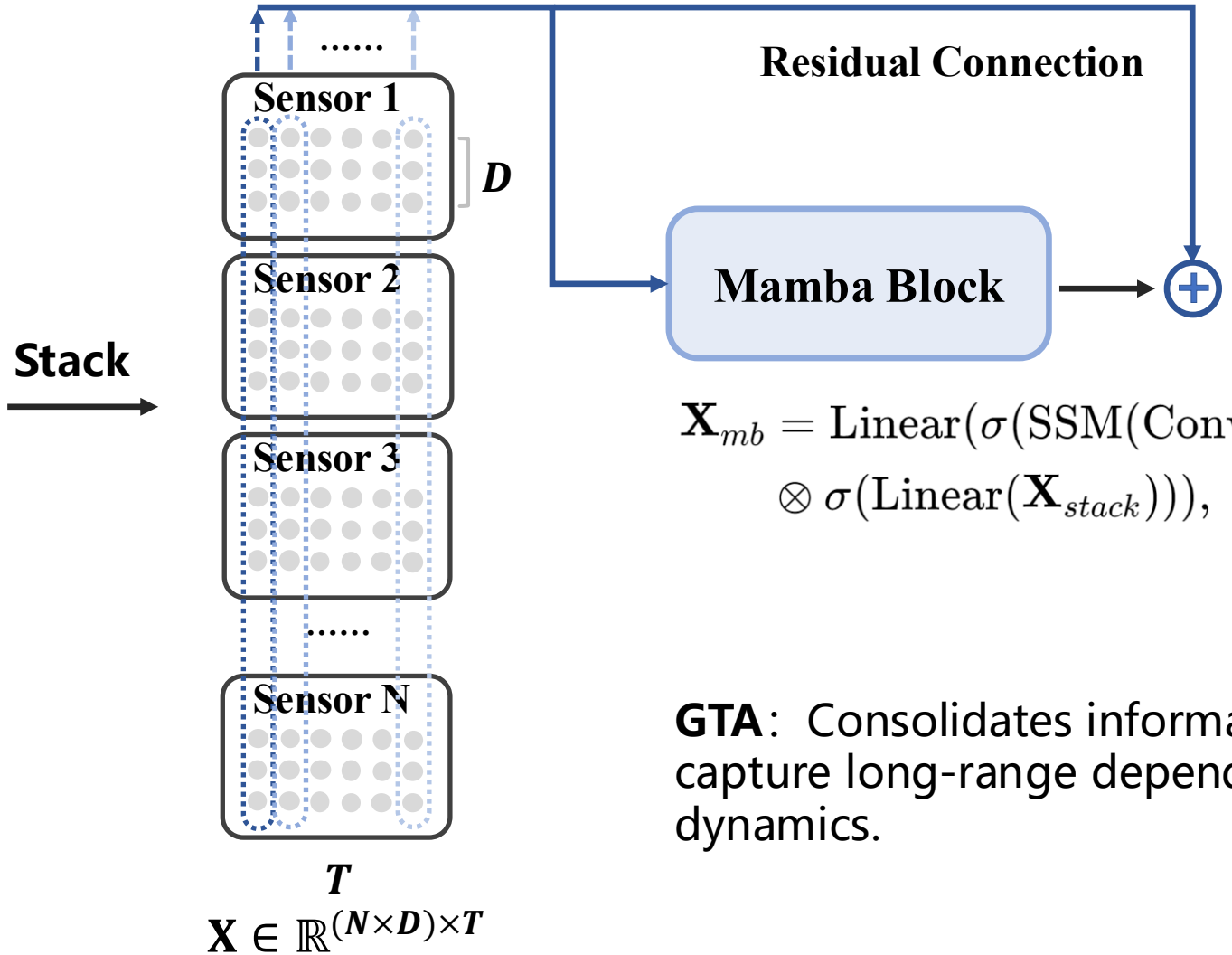


- **CCF:** Merge features across different sensor channels, capturing **inter-channel** dependencies within each variable.

- **CVF:** Merge different modality variables, capturing **cross-variable** dependencies within the same sensor.

Hierarchical Interaction Fusion

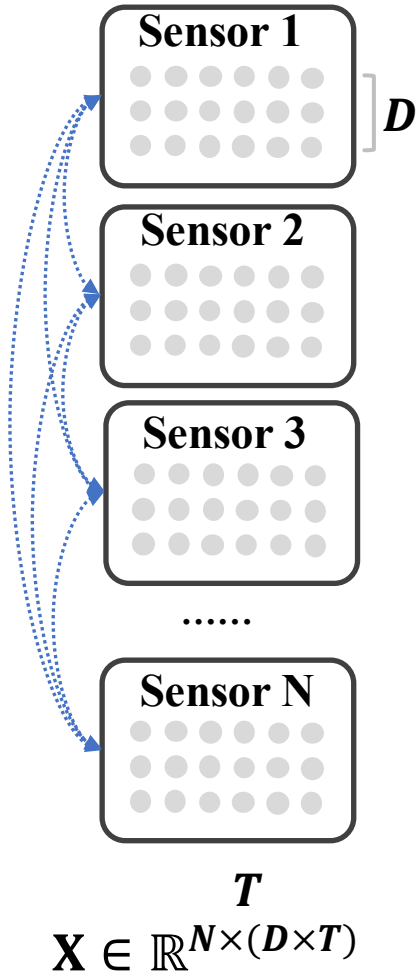
Global Temporal Aggregation (GTA)

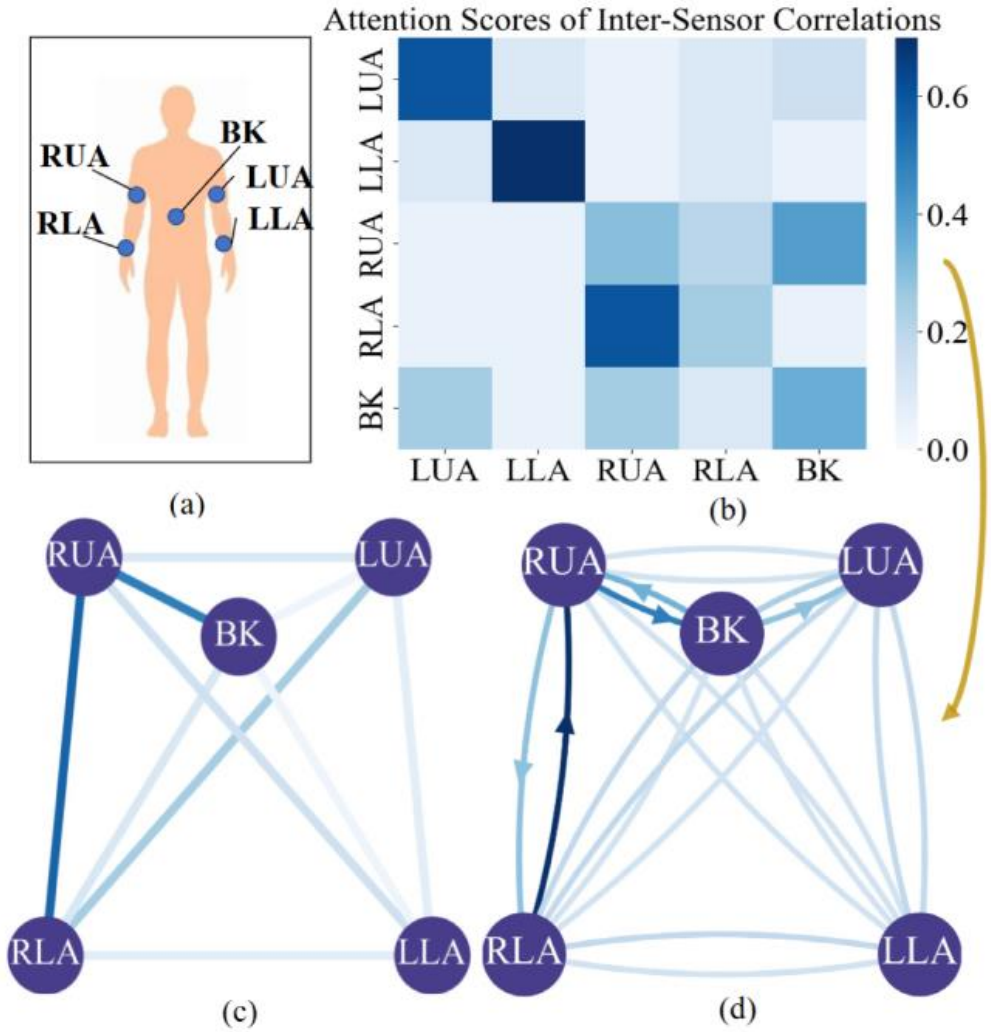


$$\mathbf{X}_{mb} = \text{Linear}(\sigma(\text{SSM}(\text{Conv}(\text{Linear}(\mathbf{X}_{stack})))) \otimes \sigma(\text{Linear}(\mathbf{X}_{stack}))),$$

GTA: Consolidates information across all time steps to capture long-range dependencies and overall temporal dynamics.

Hierarchical Interaction Fusion: Cross-Sensor Interaction (CSI)





"Drink from Cup" action

Visualization of Inter-Sensor Correlations:
Compared with *DynamicWHAR* (GCNs), which assume symmetric correlations, our method captures directional dependencies.

"**BK**" relies on "**RUA**" for posture, while "**RUA**" prioritizes arm movement. Similar asymmetries, like "**RUA**" ↔ "**RLA**", enhance inter-sensor spatial modeling.

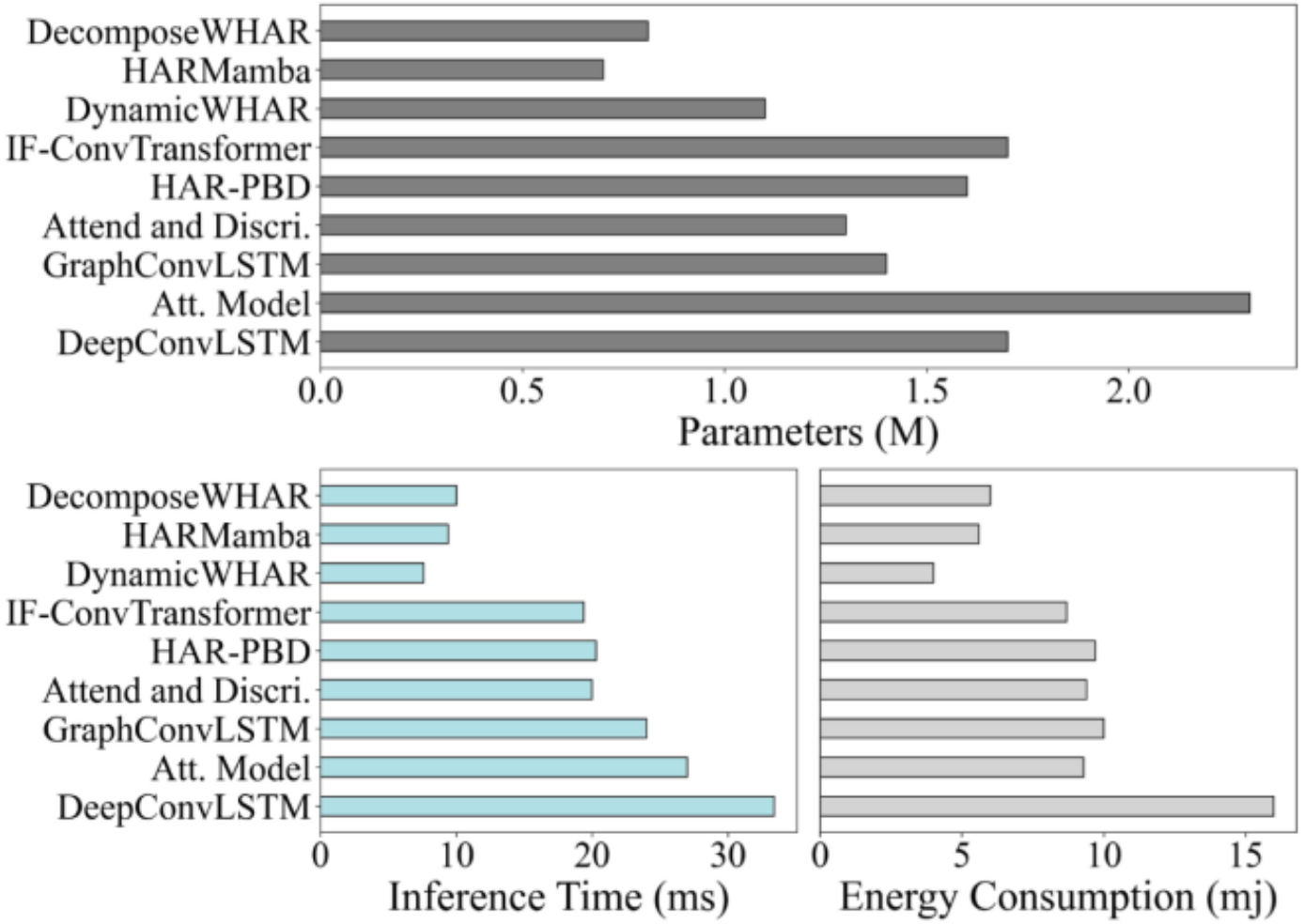
Evaluation: Recognition Performance

- **Recognition Performance:** Outperforms previous state-of-the-art models on Opportunity, Realdisp, and Skoda datasets. Significant improvement in accuracy and macro-F1 score.

Model	Opportunity		Realdisp		Skoda	
	Accuracy (%)	Macro-F1 (%)	Accuracy (%)	Macro-F1 (%)	Accuracy (%)	Macro-F1 (%)
DeepConvLSTM	69.30 ± 0.12	61.32 ± 0.75	85.61 ± 0.63	83.56 ± 0.72	90.31 ± 0.45	88.63 ± 0.39
Att. Model	71.64 ± 0.53	63.43 ± 0.81	88.37 ± 0.84	87.70 ± 1.06	91.55 ± 0.90	90.76 ± 0.83
GraphConvLSTM	70.38 ± 0.92	62.17 ± 1.05	89.94 ± 0.65	89.04 ± 0.41	91.19 ± 0.68	89.88 ± 0.91
HAR-PBD	71.82 ± 0.67	62.37 ± 0.66	91.41 ± 1.02	90.48 ± 1.14	88.37 ± 1.51	86.26 ± 1.87
Attend and Discriminate	72.75 ± 0.88	64.18 ± 0.43	90.46 ± 0.97	89.76 ± 0.53	92.16 ± 0.98	90.87 ± 0.77
IF-ConvTransformer	74.19 ± 0.94	65.92 ± 0.81	90.97 ± 0.88	90.57 ± 0.62	92.27 ± 0.83	90.41 ± 0.64
DynamicWHAR	74.27 ± 0.46	66.13 ± 0.32	92.58 ± 0.50	91.93 ± 0.57	93.80 ± 0.54	91.26 ± 0.51
HARMamba	75.13 ± 0.81	67.05 ± 0.69	91.29 ± 0.85	90.96 ± 0.73	93.96 ± 0.92	91.87 ± 0.85
(Improvement)	1.14%	1.37%	–1.41%	–1.07%	1.75%	0.43%
DecomposeWHAR	**78.28 ± 0.68**	**72.04 ± 0.51**	**96.64 ± 0.75**	**96.10 ± 0.63**	**97.61 ± 0.52**	**97.24 ± 0.47**
(Improvement)	**4.02%**	**6.93%**	**4.21%**	**4.34%**	**3.74%**	**5.52%**

Evaluation: Computation Efficiency

- **Computation Efficiency**: Reduce parameters and FLOPs while maintaining high accuracy. Optimized for real-world applications like smartwatches.



Deployed on Mi Watch running WearOS based on Android

Limitations & Future Work

Optimization Potential:

- **Computation Cost:** Model complexity may still pose deployment challenges in ultra-low-power devices.
- **Mamba Block Refinement:** The Mamba block within our framework requires further special refinement to enhance efficiency and overall performance.

Generalizability and Broader Applicability:

- Adaptability to **other multivariable classification** problems and time series tasks.
- Expand applicability across **diverse domains (**e.g.,**EEG、ECG)**, further validating its practical effectiveness.

Codes: https://github.com/Anakin2555/DecompseWHAR