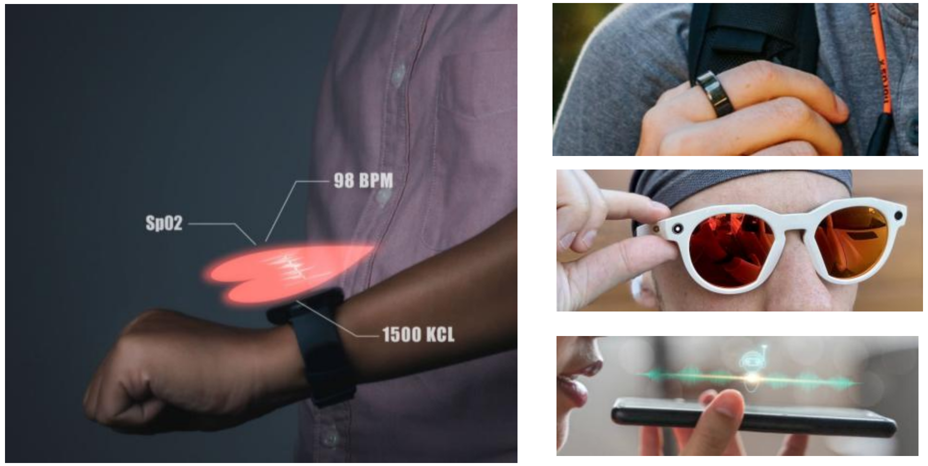
MAESTRO : Adaptive Sparse Attention and Robust Learning for Multimodal Dynamic Time Series

Payal Mohapatra Yueyuan Sui Akash Pandey Stephen Xia Qi Zhu
Northwestern University

Time Series Data are Ubiquitous

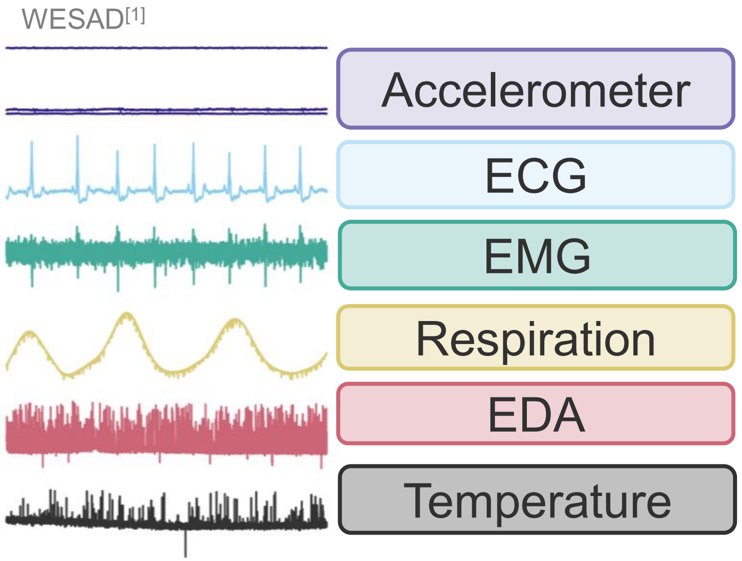
Clinical Applications

Environment Monitoring

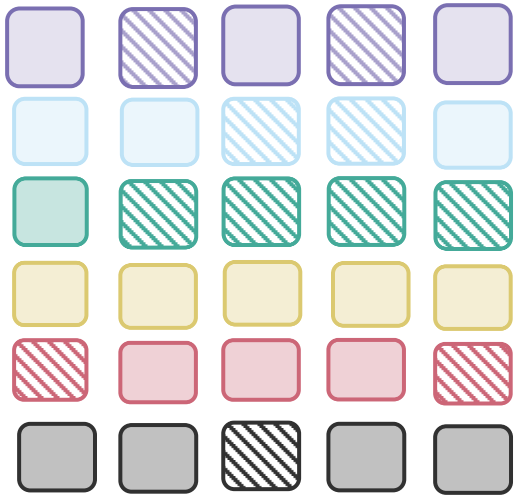


Lifestyle Enhancement

Challenges in Time Series Data Analyses



Heterogeneity

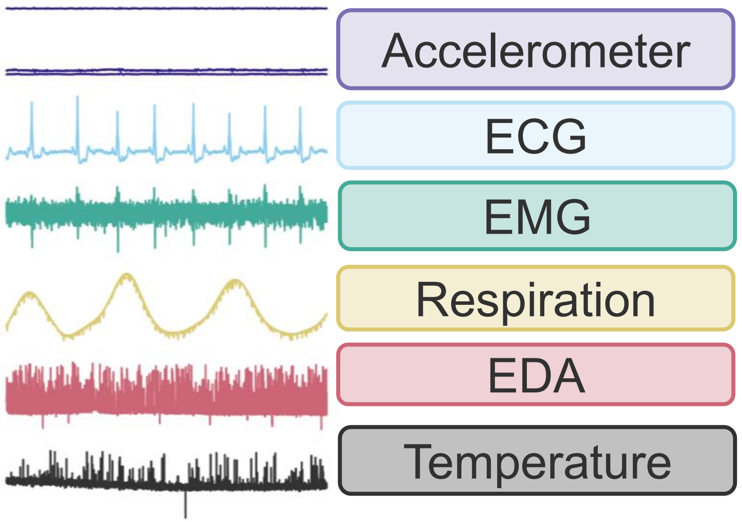


Missingness

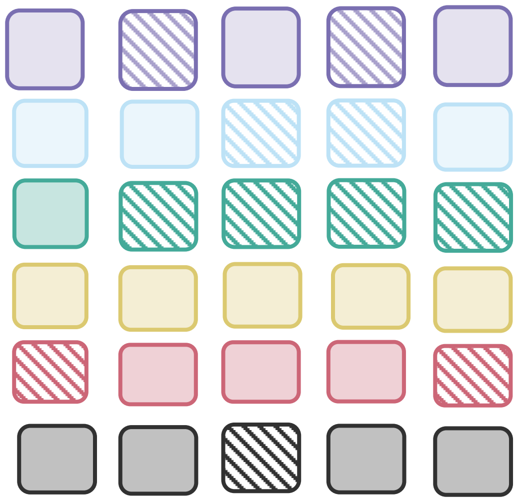
[1] Wearable Stress and Affect Detection, 2018

Challenges in Time Series Data Analyses

WESAD[1]

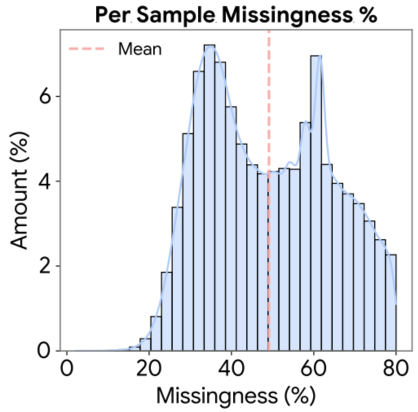


Heterogeneity



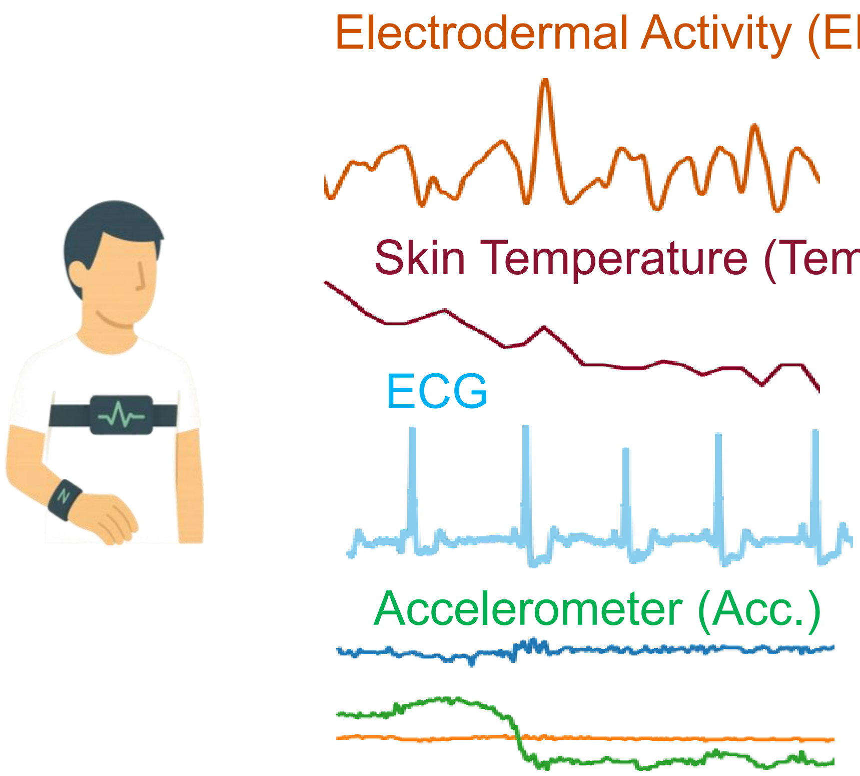
Missingness

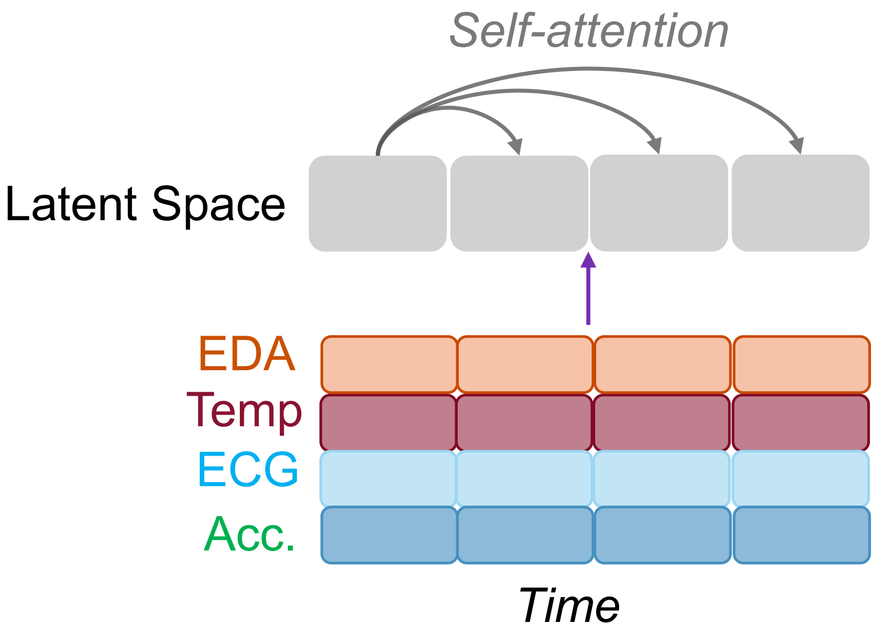
40M hours of day-long multimodal sensor data from LSM-2[2]



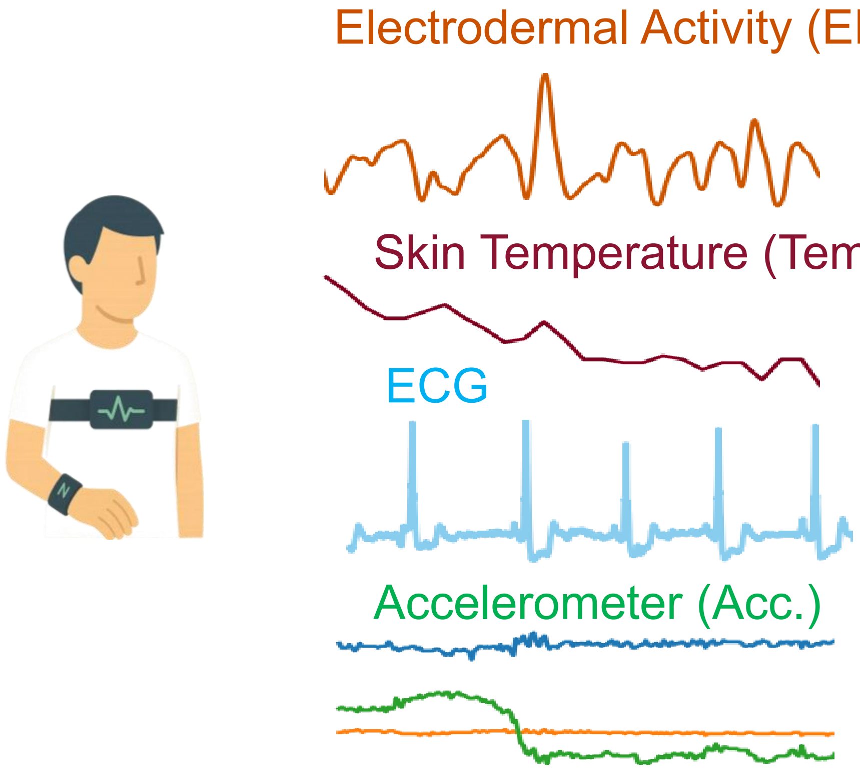
[1] Wearable Stress and Affect Detection, 2018
[2] LSM-2: Learning from Incomplete Wearable Sensor Data, 2025

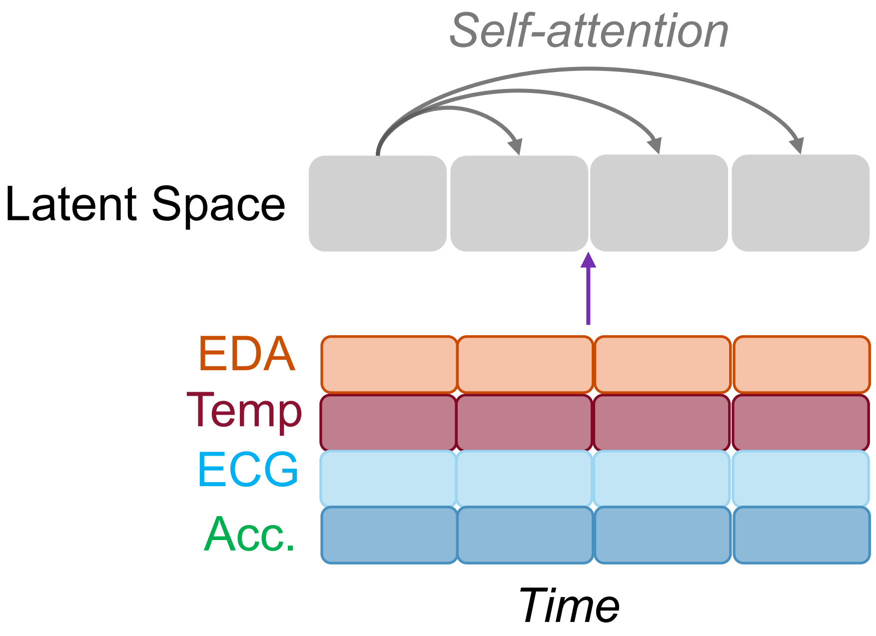
Traditional Treatment of Time-series





Traditional Treatment of Time-series

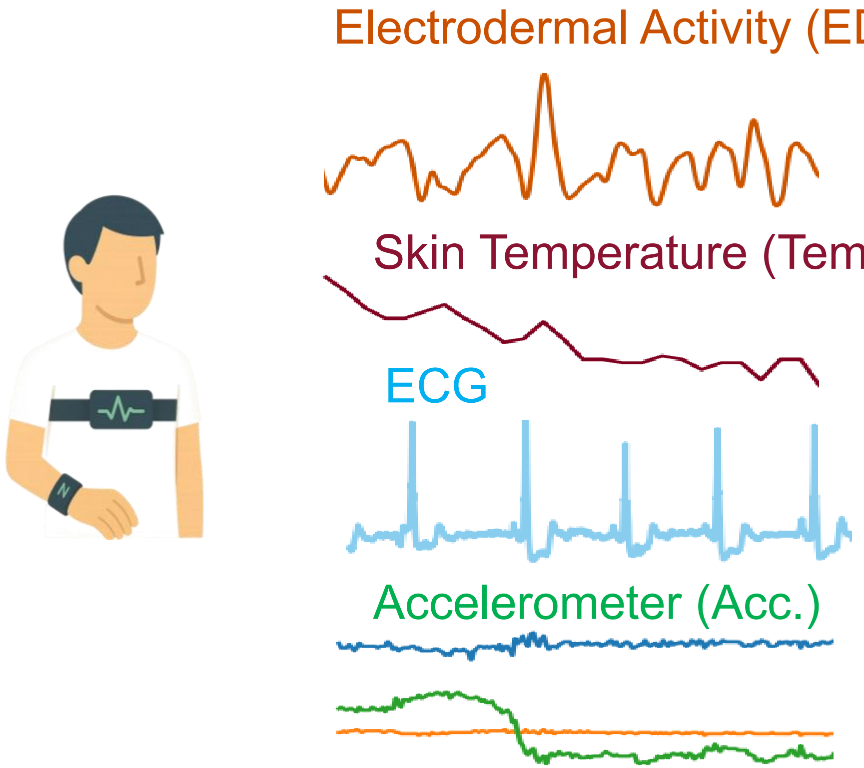


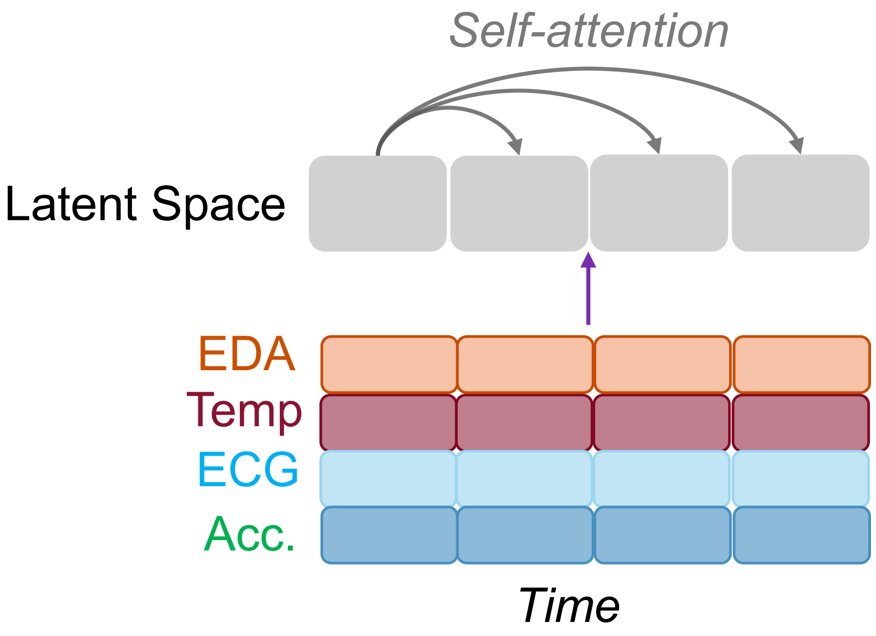


While effective in simple tasks :

- Does not model inherent heterogeneity
- Misses modeling inter-modal interactions effectively
- Cannot disentangle representation in case one modality is missing/corrupted.
- Cannot handle different sequence length across modalities.

Traditional Treatment of Time-series





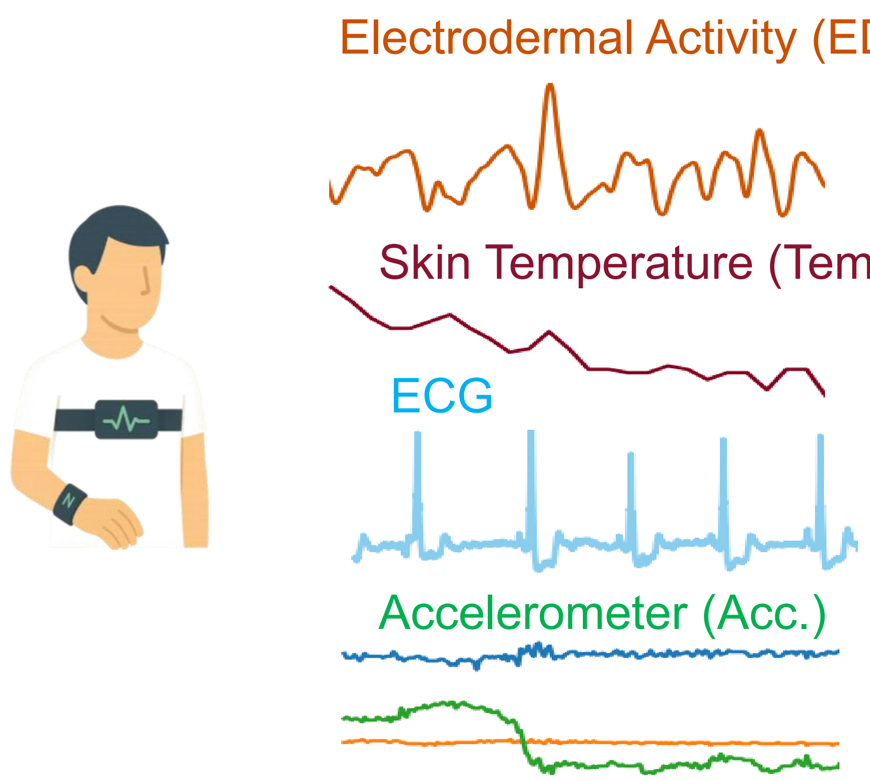
Another approach is sensor fusion, but it is highly application-specific and often heuristic.

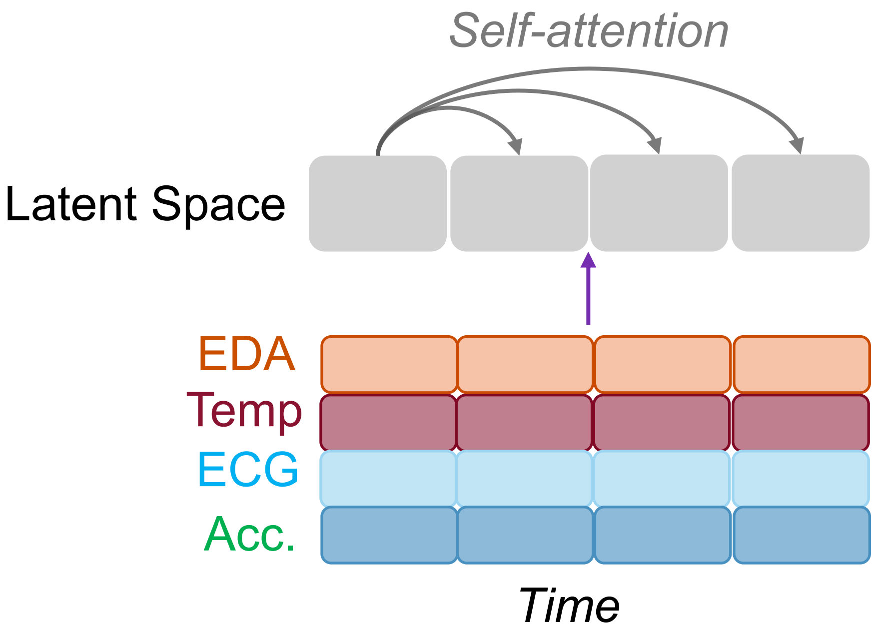
For example, in heart rate monitoring from wearables, several proposals have been made in the literature, such as:

- *early fusion using multi-wavelength photoplethysmography (PPG)[1],*
- *multi-site PPG[2], and*
- *late fusion with temperature[3].*

[1] Meier & Holz (2024). *Effect of wavelength on PPG reliability outdoors — CISS.*
[2] Meier & Holz (2024). *PPG accuracy across body locations and motion — CISS.*
[3] Meier, Demirel & Holz (2024). *WildPPG: Real-world long-duration PPG dataset — NeurIPS.*

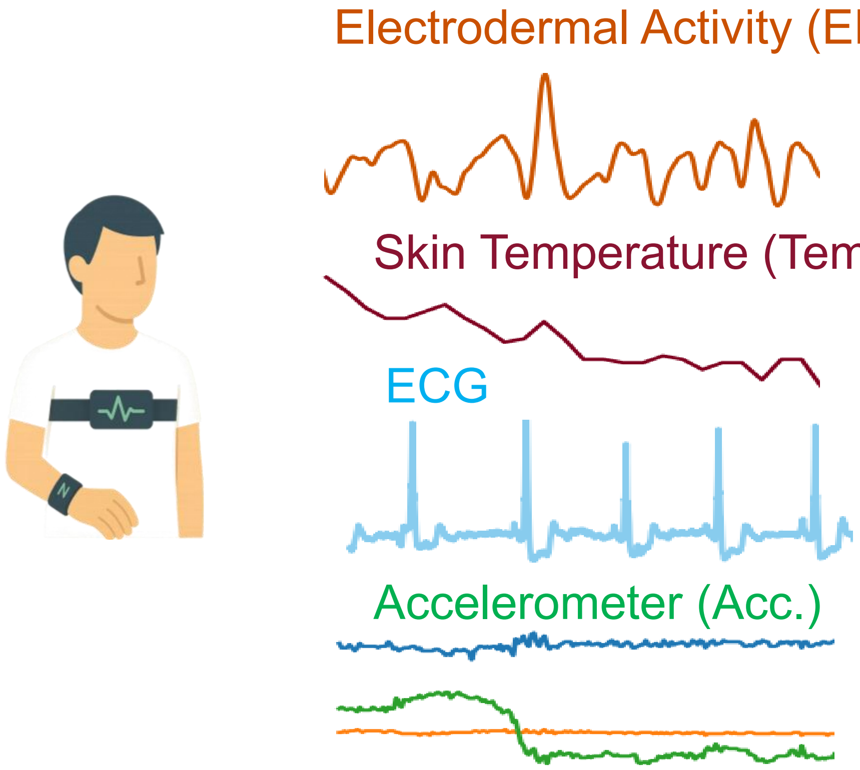
Traditional Treatment of Time-series





Need to view as multimodal data instead of multivariate-time-series!

Traditional Treatment of Time-series



Generally, video, audio, and text representations are considered multimodal. However, even though time-series data are represented in the same numerical form, they can be highly heterogeneous.

Need to view as multimodal data instead of multivariate-time-series!

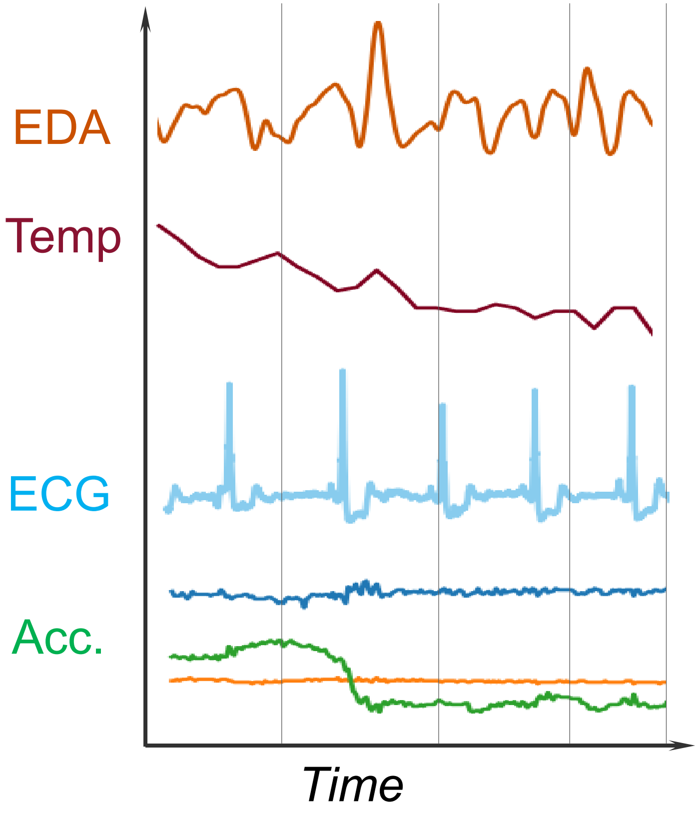
Considerations for Multimodal Real-world Time-series

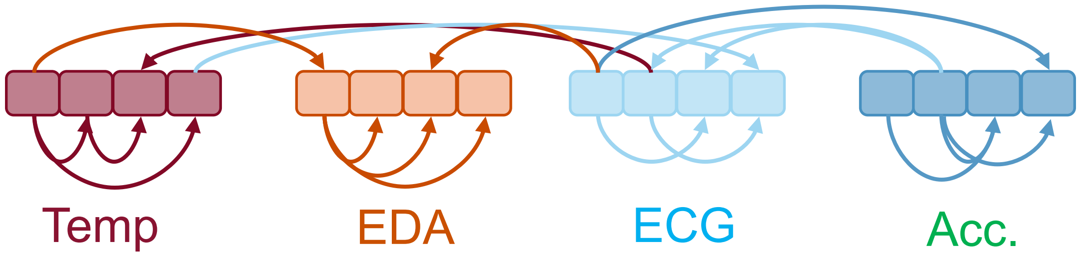
- Need to identify primary modality[1].
- Assumption of high mutual information among modalities[2, 3, 4].
- Pairwise interaction modeling [5, 6].

- *Apriori* of primary modality is not always guaranteed.
- Heterogenous modalities.
- Number of modalities can be greater than 10. Combinatorially expansive!
- Random missingness due to sensor malfunction.

[1] IMAGEBIND, CVPR 2023 (highlight paper)
[2] VATT, Neurips 2021
[3] Factorized Contrastive Learning, Neurips 2023
[4] Multimodal Fusion Interactions, ICMI 2023
[5] MULT, ACL 2020
[6] MMOE, EMNLP 2024

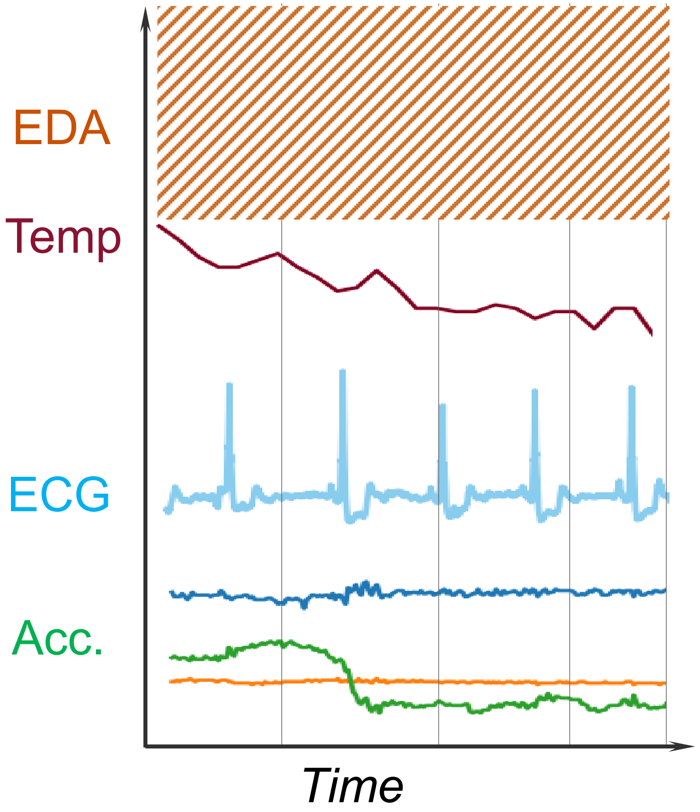
Cross-modal-attention for Multimodal Time-series

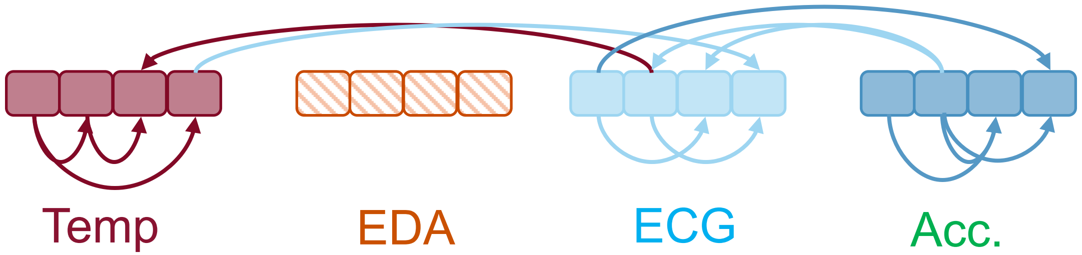




Cross-attention can allow learning task-relevant <u>modality interaction</u>.

Cross-modal-attention for Multimodal Time-series





Cross-attention can allow learning from arbitrary modality combinations.

But applying Cross-modal-attention through Long Multimodal Time-series increases the computational complexity!

Canonical Self-Attention[1]

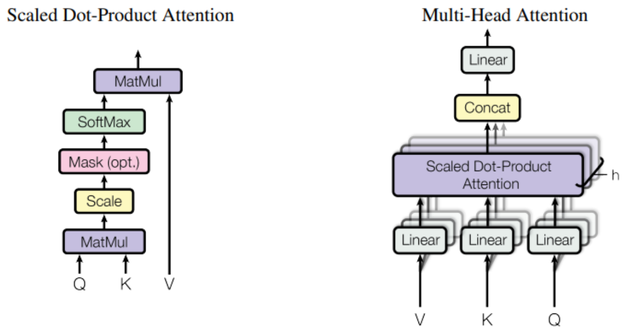


Figure 2: (left) Scaled Dot-Product Attention. (right) Multi-Head Attention consists of several attention layers running in parallel.

$$\mathcal{A}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathtt{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^{\top}}{\sqrt{d}}\right)\mathbf{V}$$

Point-wise self-attention for a sequence length of L, has a quadratic computational complexity→$\mathcal{O}(L^2)$

Consider M modalities each of sequence length L, then the computational complexity increases,

$$\mathcal{O}(M^2 L^2)$$

[1] Attention Is All You Need, Neurips 2017

But applying Cross-modal-attention through Long Multimodal Time-series increases the computational complexity!

Canonical Self-Attention[1]

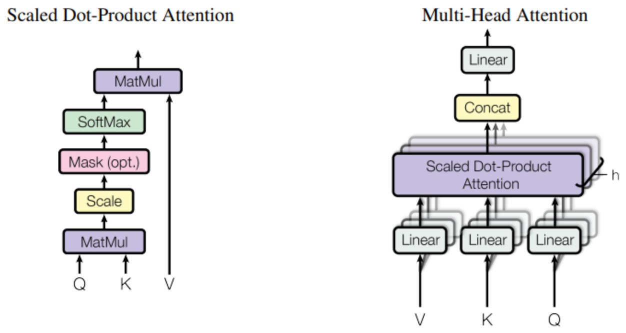


Figure 2: (left) Scaled Dot-Product Attention. (right) Multi-Head Attention consists of several attention layers running in parallel.

$$\mathcal{A}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \texttt{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}}\right)\mathbf{V}$$

Point-wise self-attention for a sequence length of L, has a quadratic computational complexity→$\mathcal{O}(L^2)$

Consider M modalities each of sequence length L, then the computational complexity increases,

$$\mathcal{O}(M^2L^2)$$

Sparse Attention

$$\mathcal{O}(MLlog(ML))$$

[1] Attention Is All You Need, Neurips 2017

Handling long time-series through sparse attention (Overview)

- Point-wise self-attention for a sequence length of L, has a quadratic computational complexity.

$$\mathcal{A}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \texttt{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}}\right)\mathbf{V}$$

ProbSparse Attention - $\mathcal{A}_s(\mathbf{Q},\mathbf{K},\mathbf{V}) = \texttt{Softmax}\left(\frac{\bar{\mathbf{Q}}\mathbf{K}^\top}{\sqrt{d}}\right)\mathbf{V}$

- Stacking N encoder layers further increases the memory consumption and computational complexity.

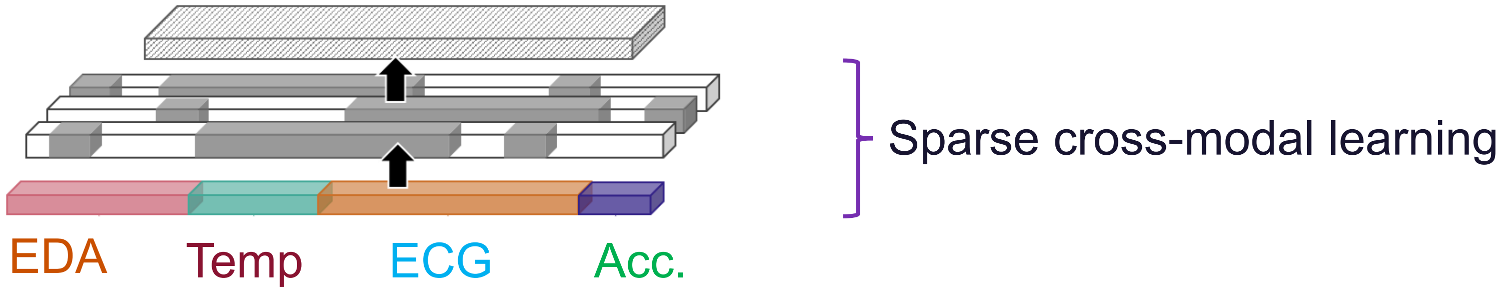
$$\hat{s} = s + \mathrm{PE}_{\sin}(s)$$

$$\bar{s} = \mathcal{A}_s(\mathbf{Q},\mathbf{K},\mathbf{V}) = \texttt{Softmax}\left(\frac{\bar{\mathbf{Q}}\mathbf{K}^\top}{\sqrt{d}}\right)\mathbf{V}$$

$$\dot{s} = \bar{s} + \hat{s}$$

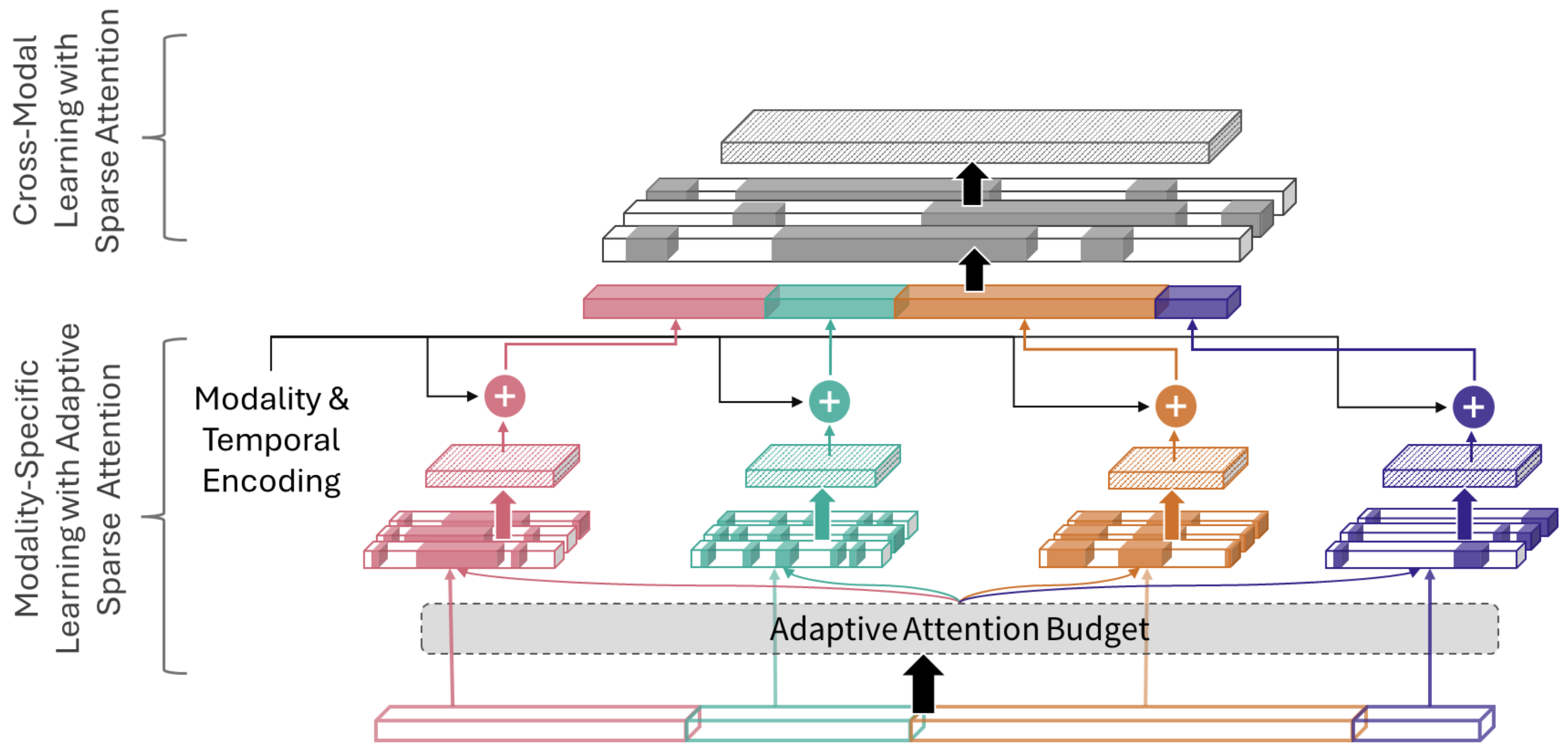
$$z = \texttt{distil}(\dot{s}) + \texttt{maxpool}(\hat{s})$$

Handling long multimodal time-series through sparse attention



1. Adaptive Attention Budget per modality

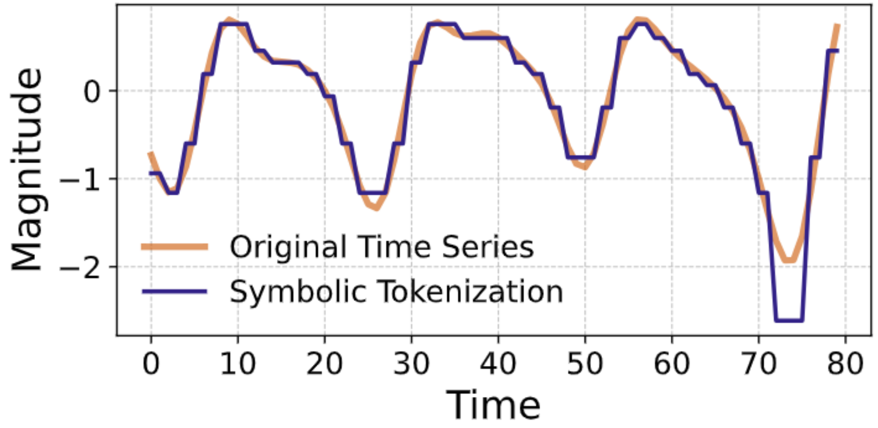
Top-**v** queries in ProbSparse Computation : **v** is modulated by modality's relevance and availability



2. Symbolic Tokenization

1. Converts time-series to discrete *symbols*[1].
    - Has some nice properties – guarantees a lower bound Euclidean distance between the symbolic time-series and the original time-series.
    - We extend it under some assumptions that this tokenization preserves multimodal relational structure.

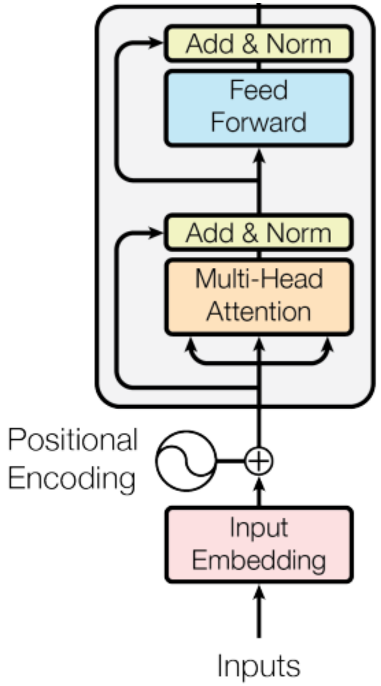
2. We can reserve a symbol for *missing* data naturally.
3. We can compress the signal further reducing the sequence length.



[1] Experiencing SAX: a Novel Symbolic Representation of Time Series, Data Mining and Knowledge Discovery, 2007

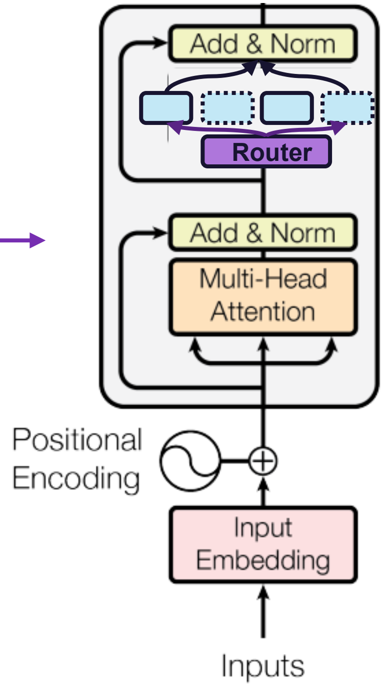
3. Handling Missingness through Mixture-of-Experts

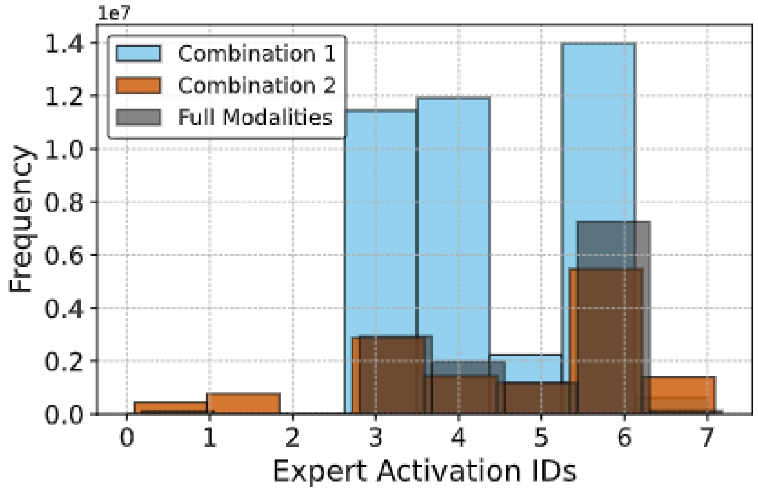
Instead of the fixed Feed-forward layer of the transformer, we can use a mixture of experts(MoE) for implicit modality specialization.



Vanilla Transformer[1]

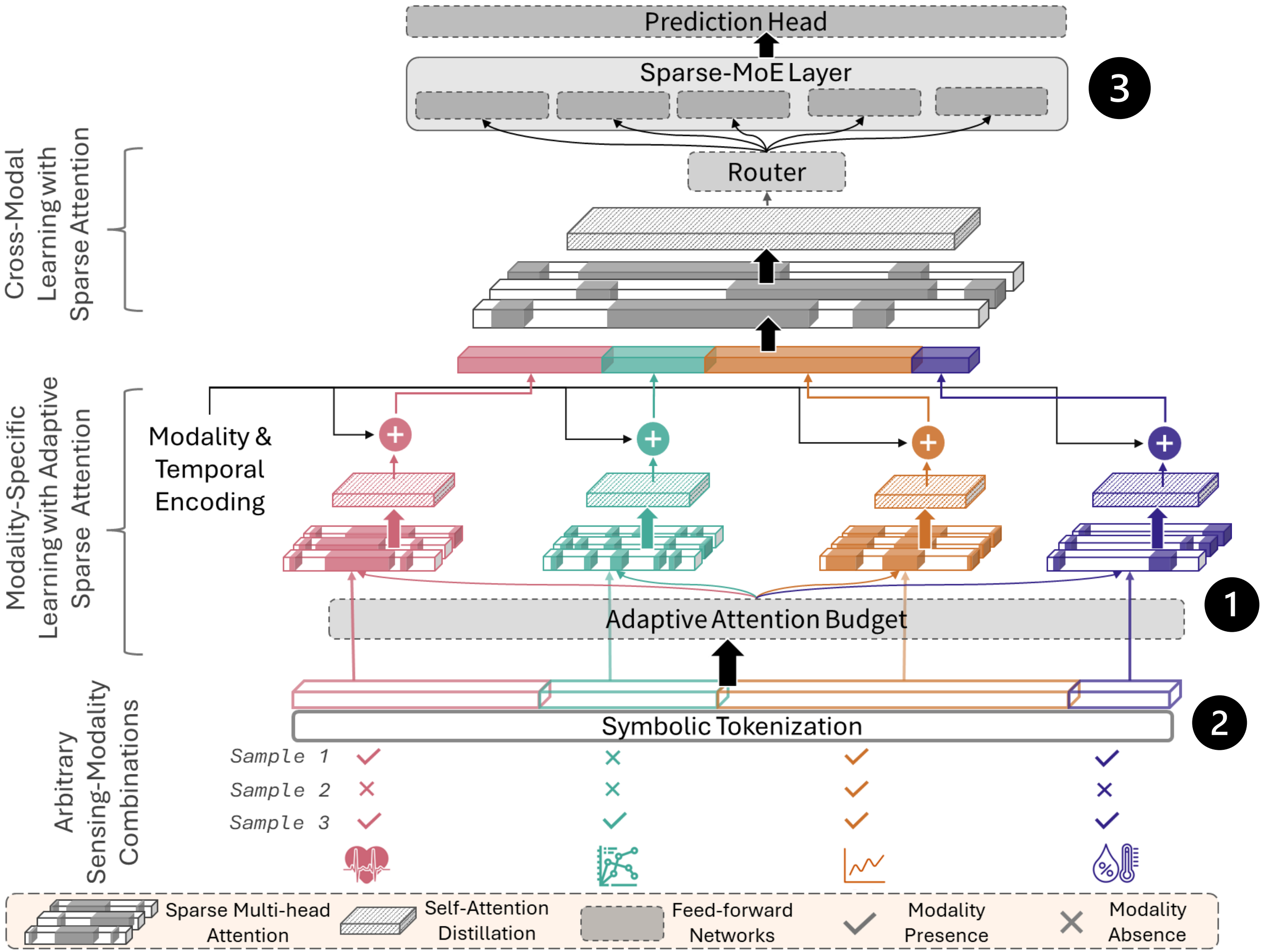
With MoE

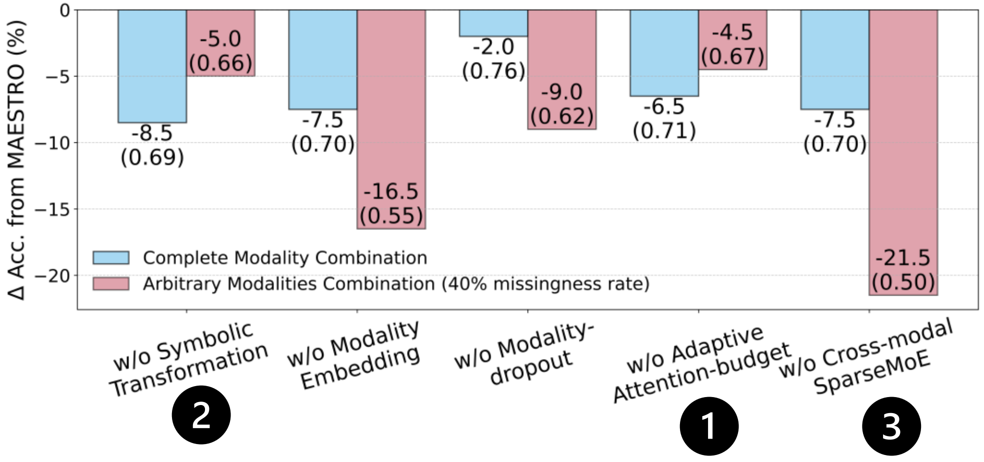




[1] Attention Is All You Need, Neurips 2017

Overall MAESTRO Framework

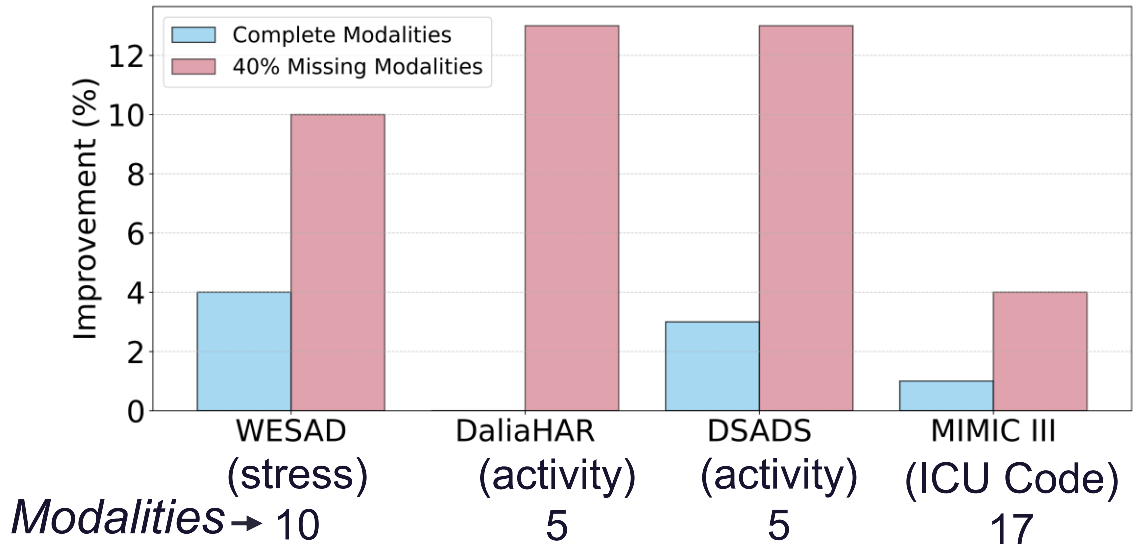




Ablation of Key Components

Key Results : MAESTRO

Average of 4% performance improvement with complete and 8% with arbitrary set of modalities.



Computational Efficiency

Model	Acc. ↑	MMAC ↓	GFLOPs ↓	Params (M)
Multivariate Models				
iTransformer	$0.67_{\pm 0.05}$	2833	5.73	12.82
Transformer	$0.63_{\pm 0.02}$	4331	8.66	1.68
Multimodal Models				
FuseMoE	$0.47_{\pm 0.41}$	6524	13.05	0.67
MULT	$0.60_{\pm 0.42}$	13324	26.65	3.71
ShaSpec	$0.62_{\pm 0.51}$	4556	9.11	216
MAESTRO	$0.77_{\pm 0.04}$	3066	6.13	1.39
– Full-Attn (Per-Modal)	$0.80_{\pm 0.03}$	3769	7.54	1.40
– Full-Attn (Cross-Modal)	$0.77_{\pm 0.07}$	3496	6.99	1.39
– All Full-Attention	$0.75_{\pm 0.05}$	4205	8.42	1.39
– All Full-Attention (no MoE)	$0.78_{\pm 0.04}$	4392	8.78	1.39

Thank you.