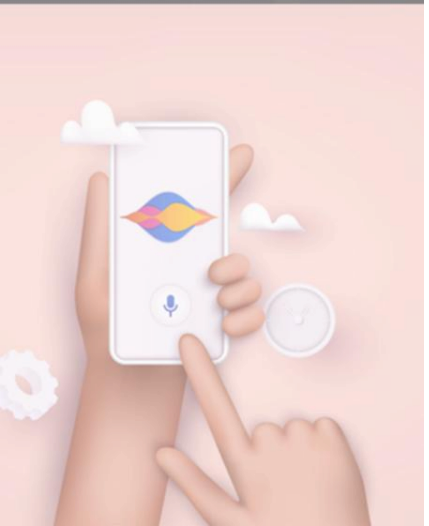


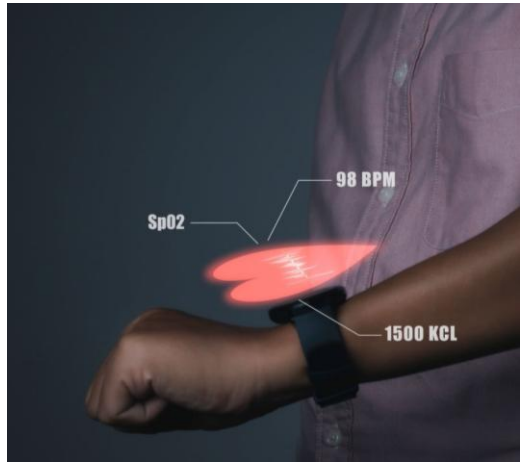
Robust Machine Learning for Real-world Time-series Applications

Payal Mohapatra, PhD Candidate (Final Year)
Department of Electrical and Computer Engineering, Northwestern University
October 8, 2025

Time Series Data are Ubiquitous

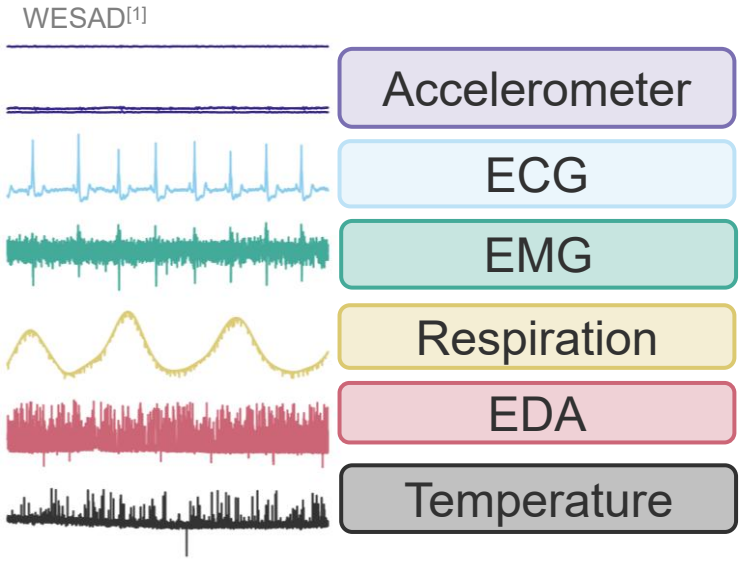
Clinical Applications

Environment Monitoring

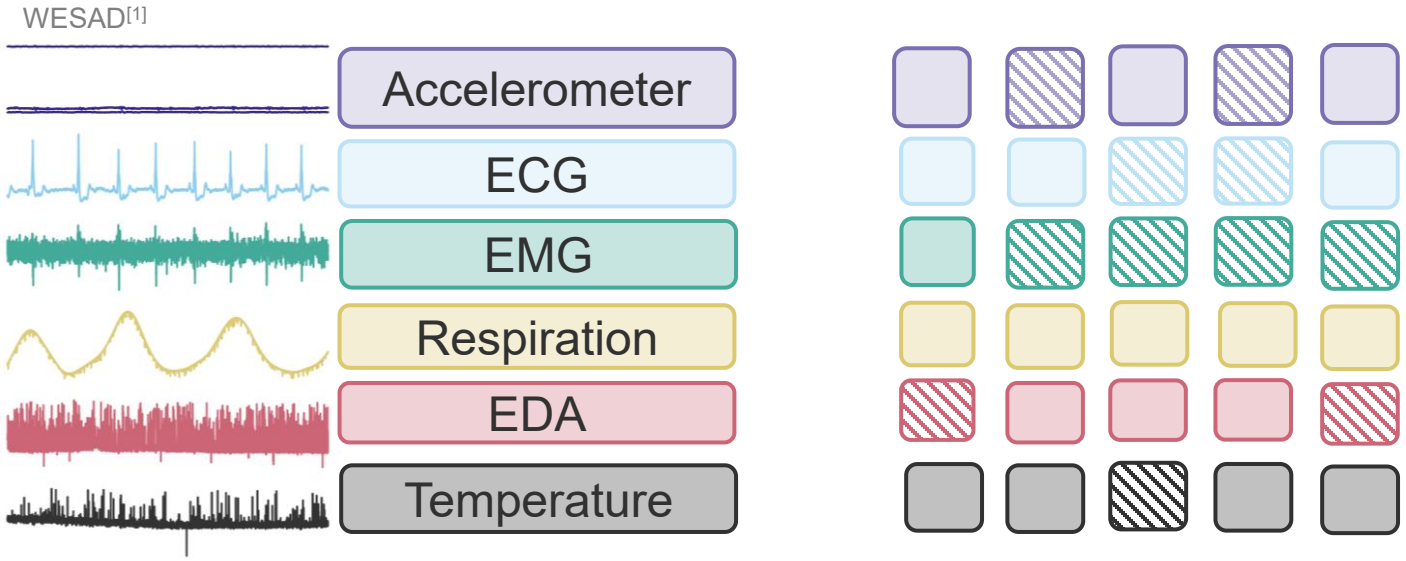


Lifestyle Enhancement

Challenges in Time Series Data Analyses



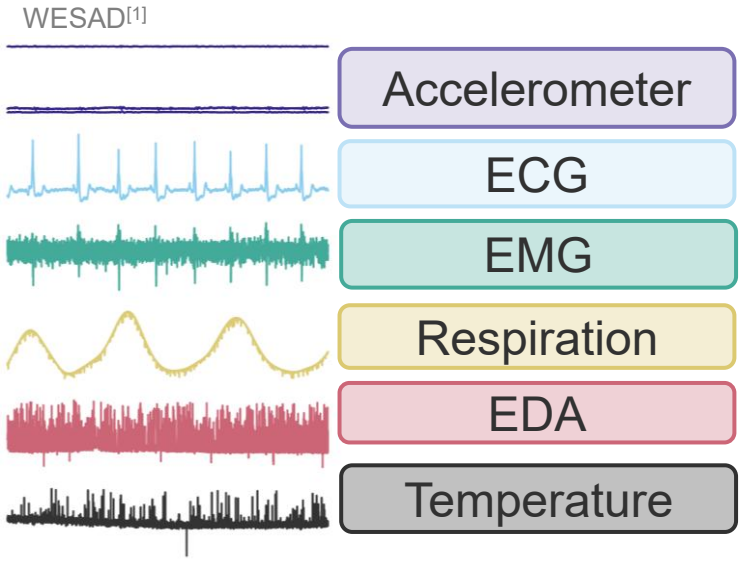
Challenges in Time Series Data Analyses



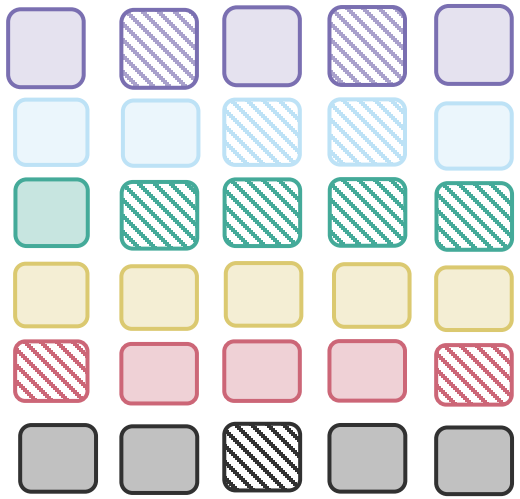
Heterogeneity

Missingness

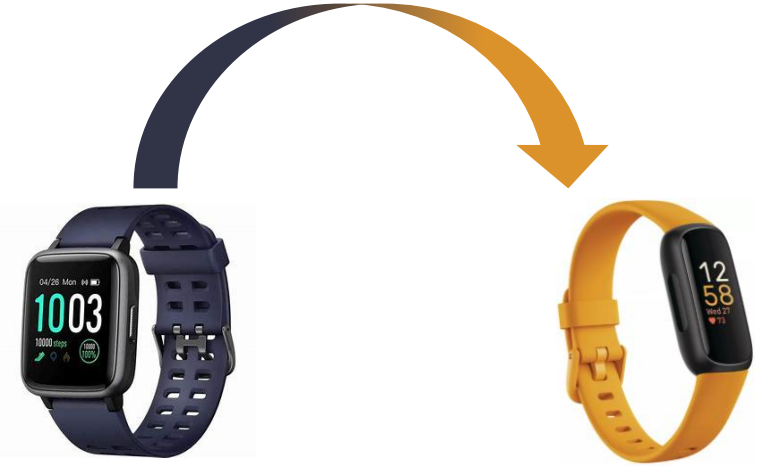
Challenges in Time Series Data Analyses



Heterogeneity



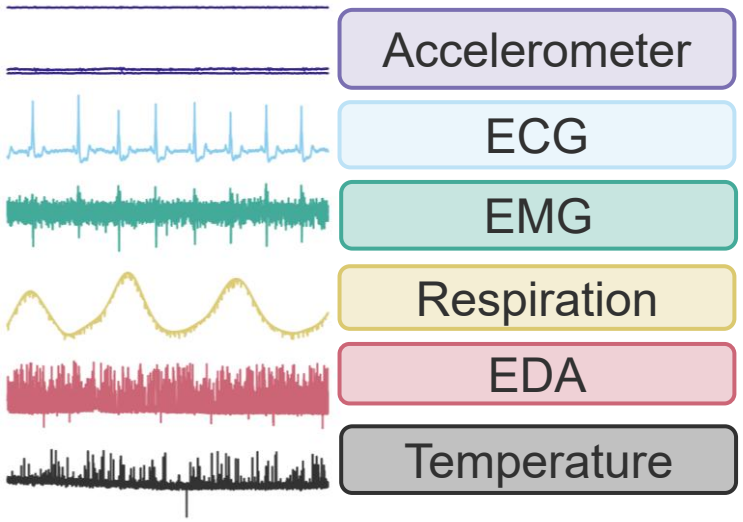
Missingness



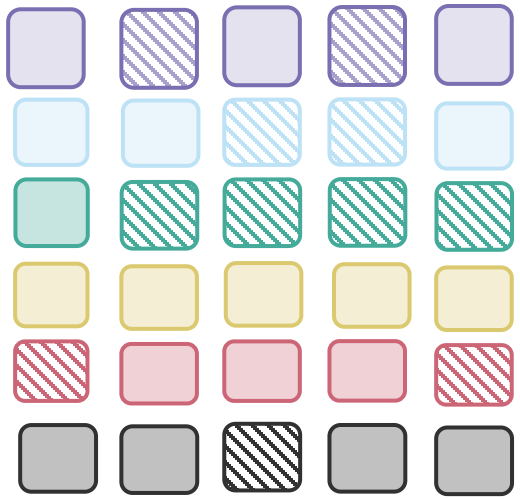
Device Change/Distribution Shift

Challenges in Time Series Data Analyses

WESAD[1]



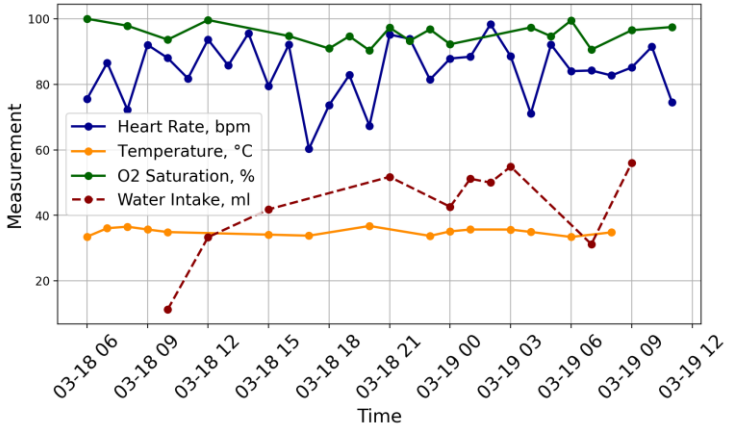
Heterogeneity



Missingness

Device Change/Distribution Shift

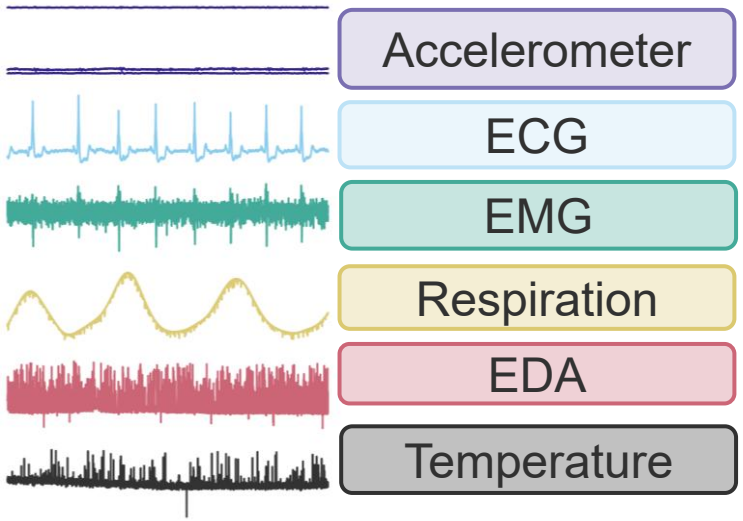
MIMIC III[2]



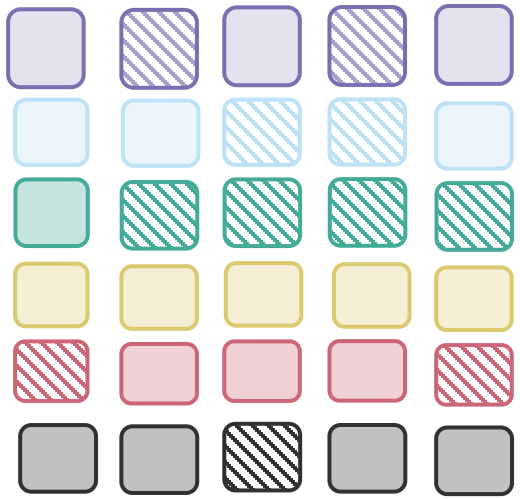
Irregularity

Challenges in Time Series Data Analyses

WESAD[1]



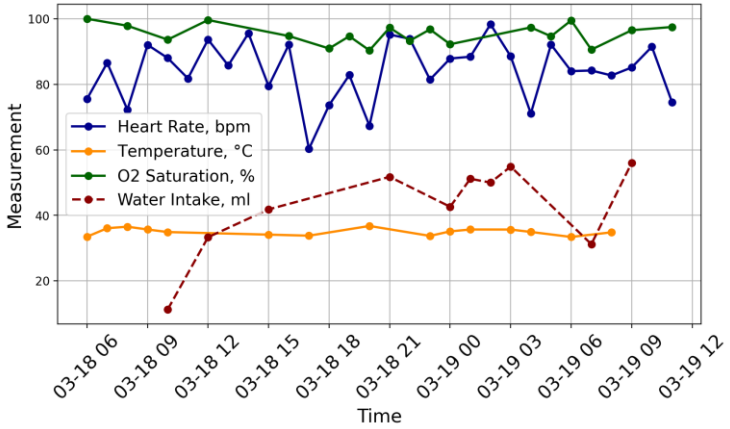
Heterogeneity



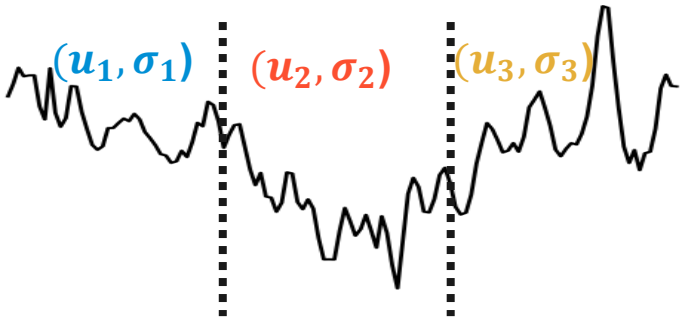
Missingness

Device Change/Distribution Shift

MIMIC III[2]



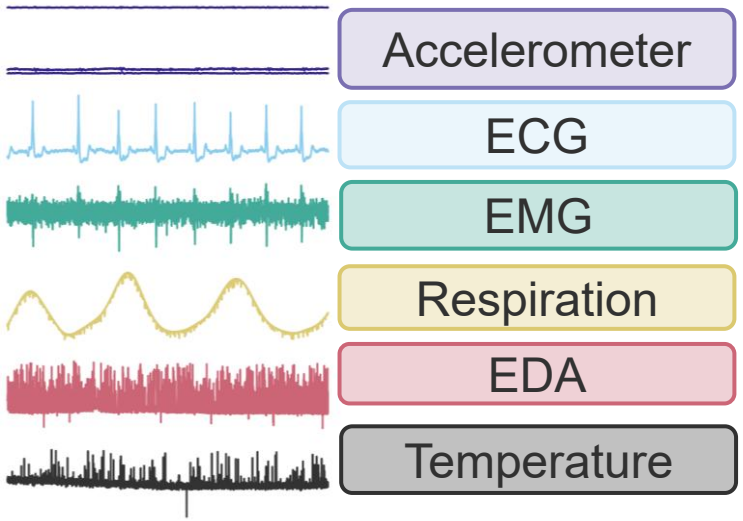
Irregularity



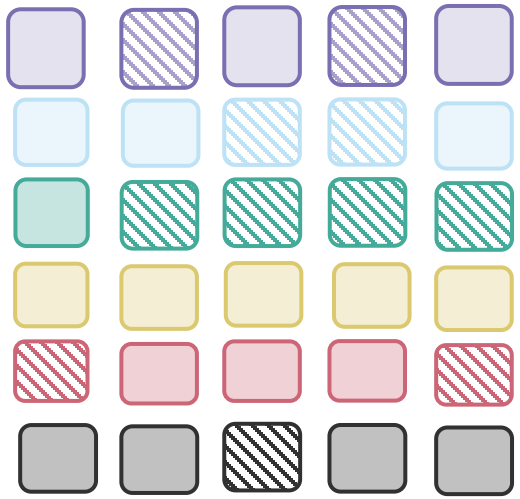
Nonstationarity

Challenges in Time Series Data Analyses

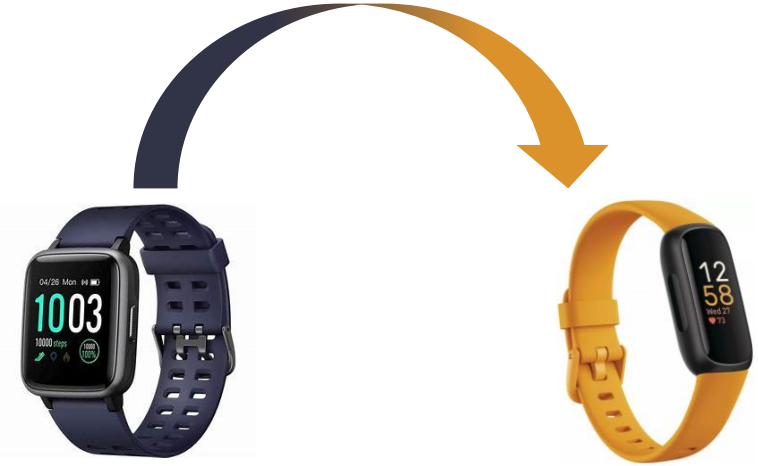
WESAD[1]



Heterogeneity

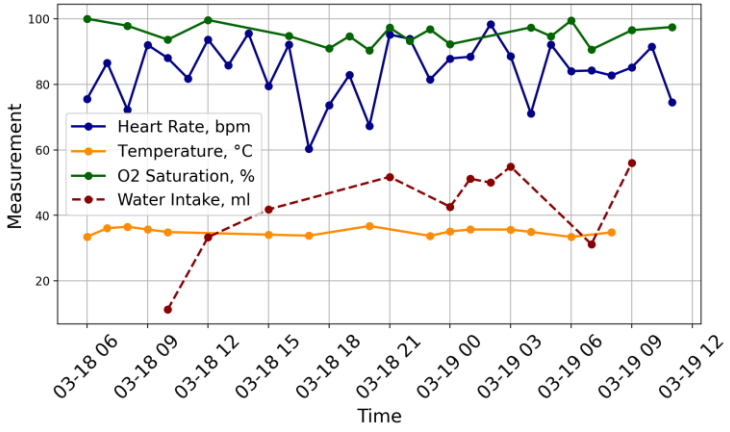


Missingness

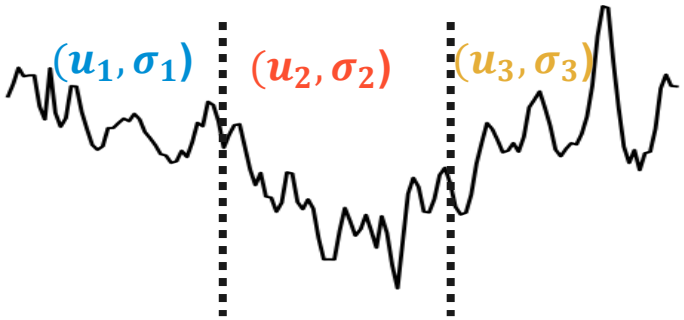


Device Change/Distribution Shift

MIMIC III[2]



Irregularity

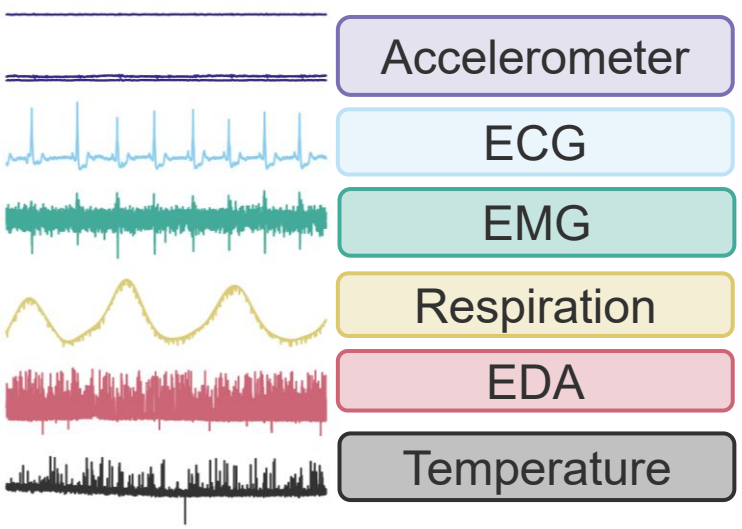


Nonstationarity

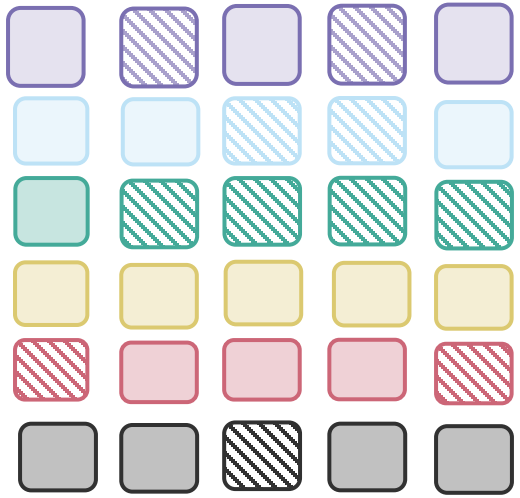
Subjectivity in labels

(Human-centered applications)

Agenda



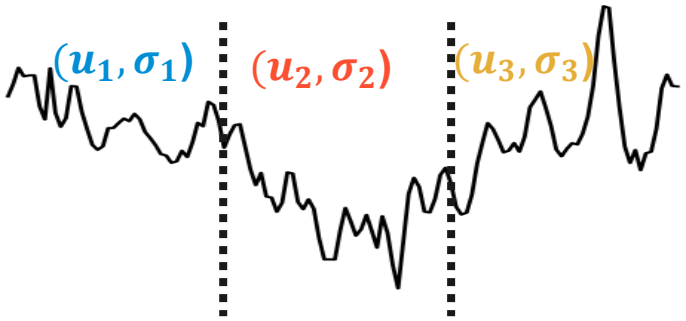
Heterogeneity



Missingness

I. Robustness to Missing Modalities.

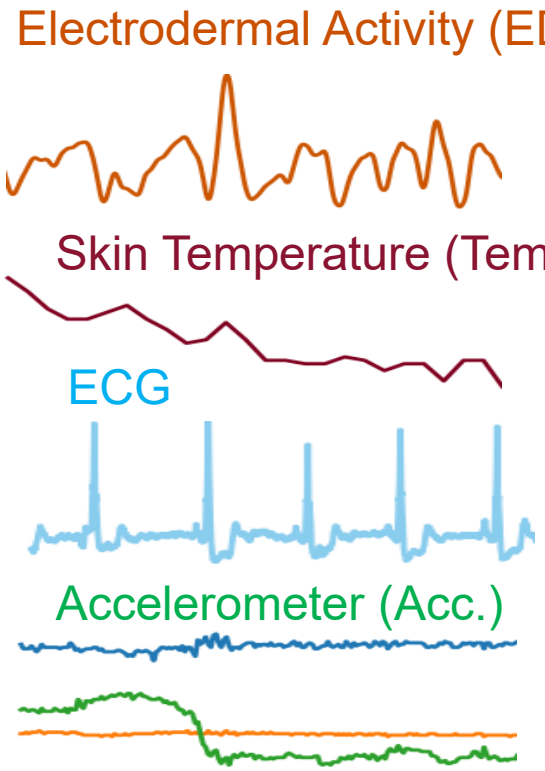
Device Change



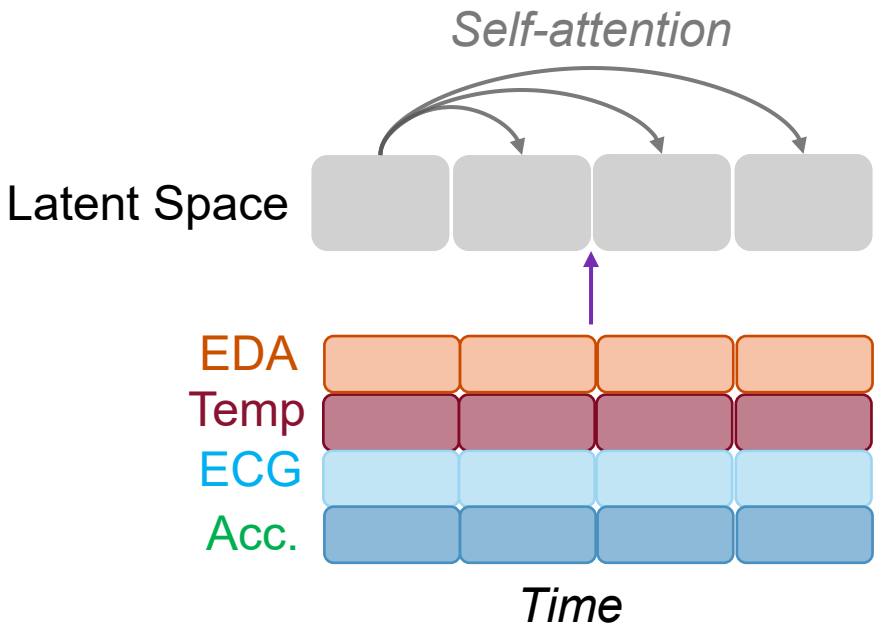
Nonstationarity

II. Robustness to Distribution Shifts.

Traditional Multivariate Treatment of Time-series



Many applications like stress monitoring have many heterogenous sensors but mostly treated **multivariate** because the data structure is same.



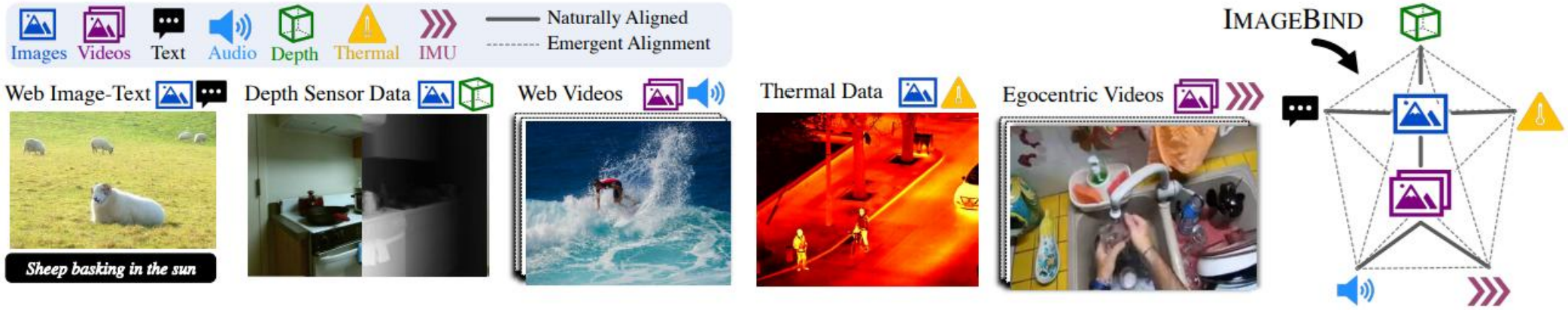
While effective in simple tasks :

- 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.

Need to view as multimodal learning.

Common multimodal learning paradigms

I. Binding to an anchor modality[1]



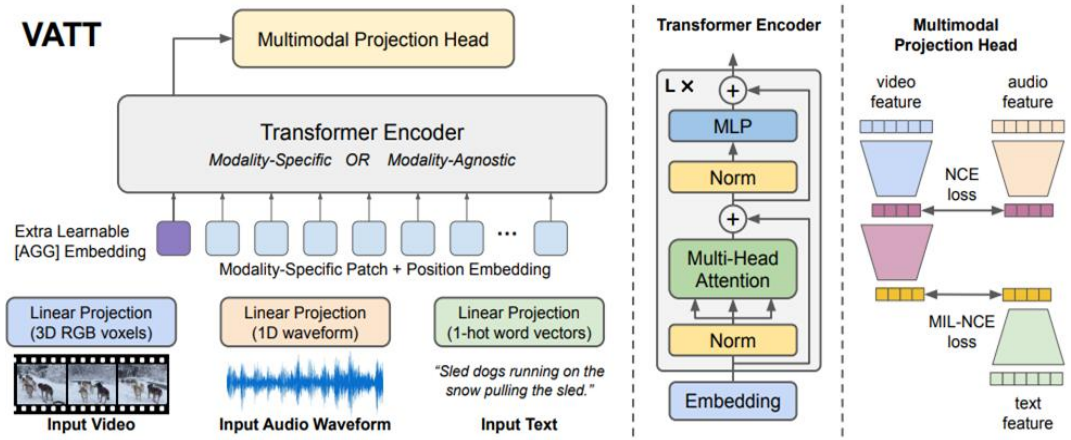
- Need to identify primary modality.
- Primary modality should be always available.

[1] IMAGEBIND, CVPR 2023 (highlight paper)

Common multimodal learning paradigms

II. Pairwise similarity matching[2]

- Structurally contrastive losses work for a pair of inputs (ongoing works on factorized contrastive losses[3]).
- Assumption of high mutual information among modalities.



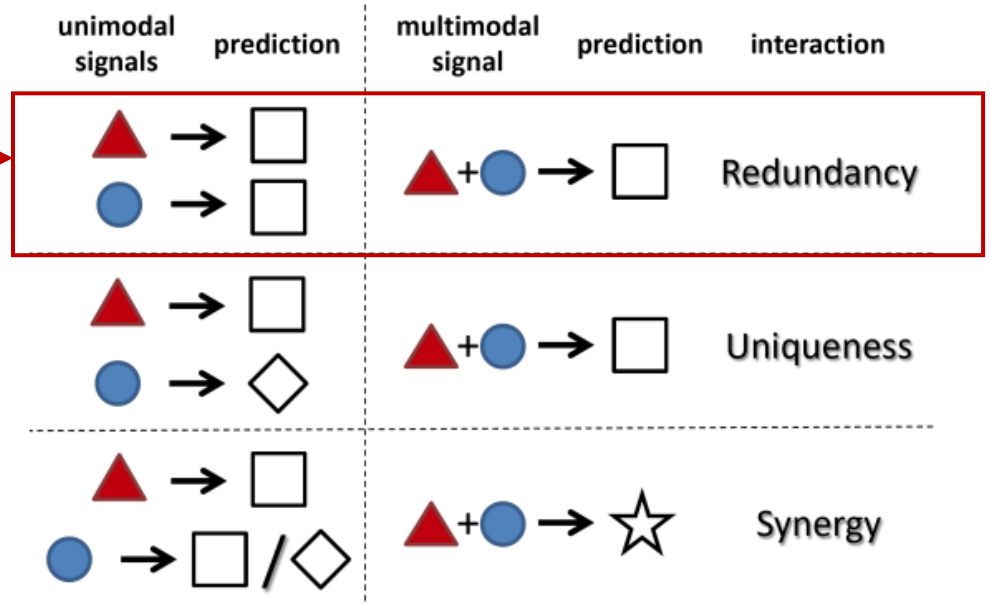
[2] VATT, Neurips 2021
[3] Factorized Contrastive Learning, Neurips 2023

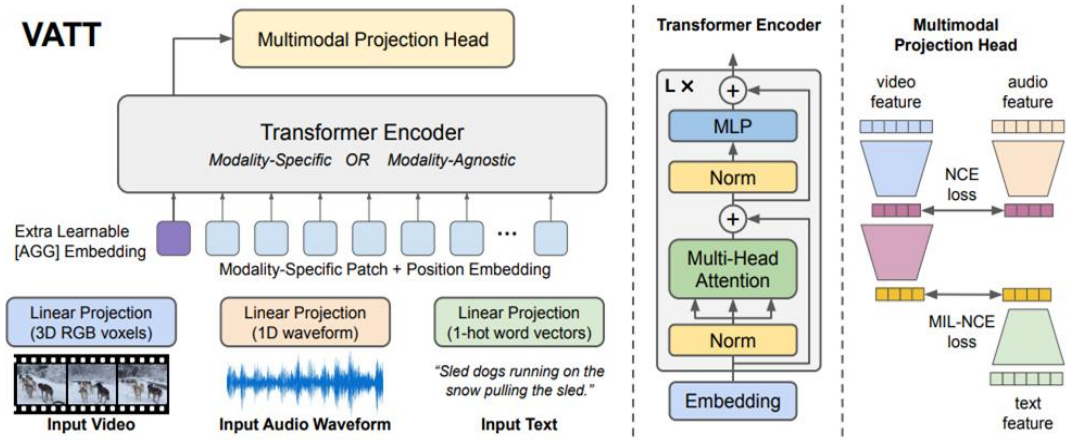
Common multimodal learning paradigms

II. Pairwise similarity matching[2]

- Structurally contrastive losses work for a pair of inputs (ongoing works on factorized contrastive losses[3]).
- Assumption of high mutual information among modalities.

[4]



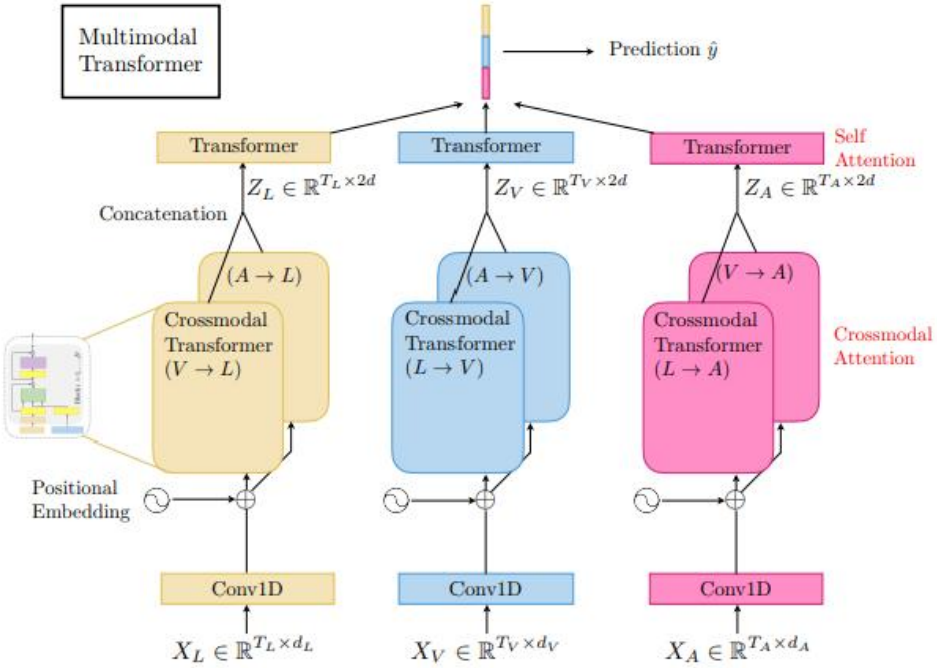


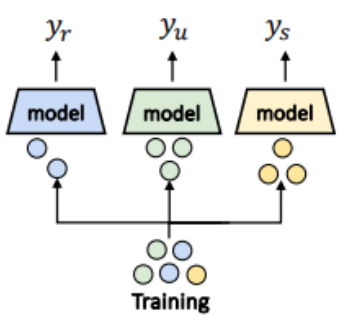
[2] VATT, Neurips 2021
[3] Factorized Contrastive Learning, Neurips 2023
[4] Multimodal Fusion Interactions, ICMI 2023

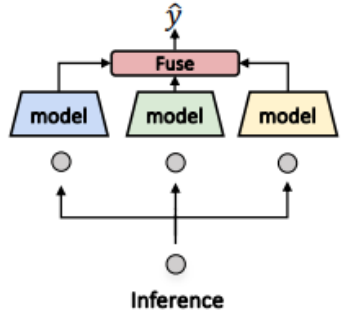
Common multimodal learning paradigms

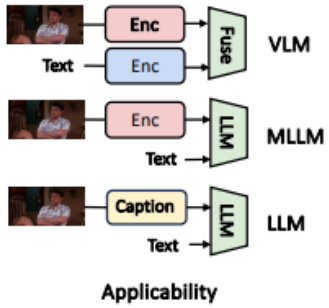
II. Explicit cross-modal interaction modeling[5, 6]

- 🗝 Suitable for small number of modalities to conduct exhaustive pair-wise analysis.









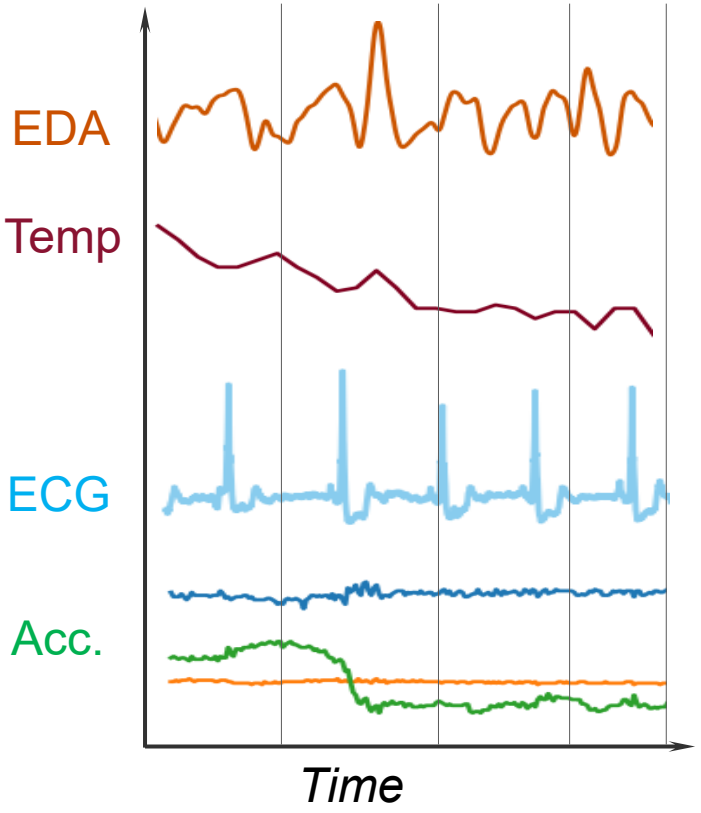
[5] MULT, ACL 2020
[6] MMOE, EMNLP 2024

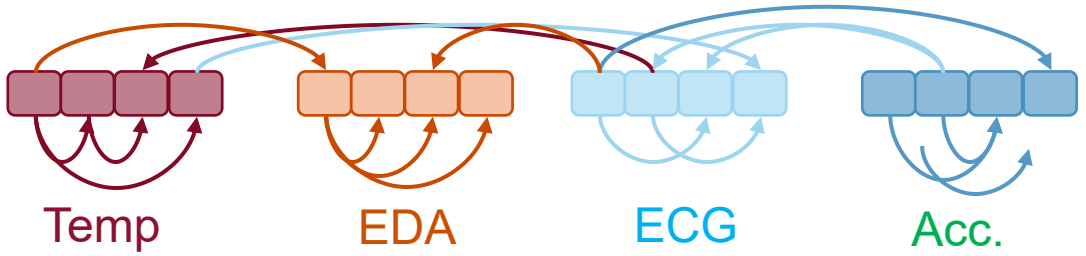
Considerations for Multimodal Real-world Time-series

- Need to identify primary modality.
- Assumption of high mutual information among modalities.
- Pairwise interaction modeling.

- *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.

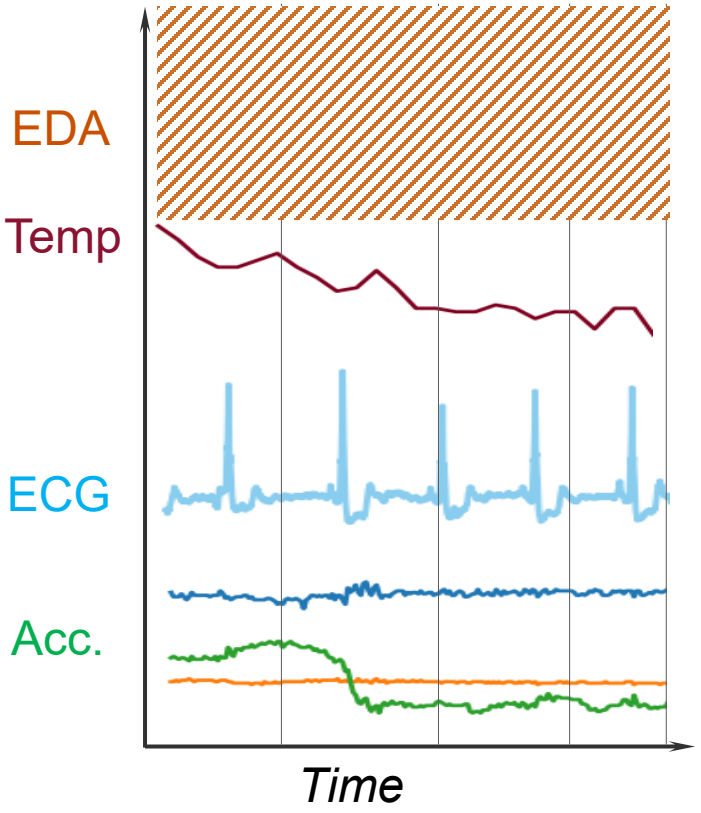
Cross-modal-attention for Multimodal Time-series

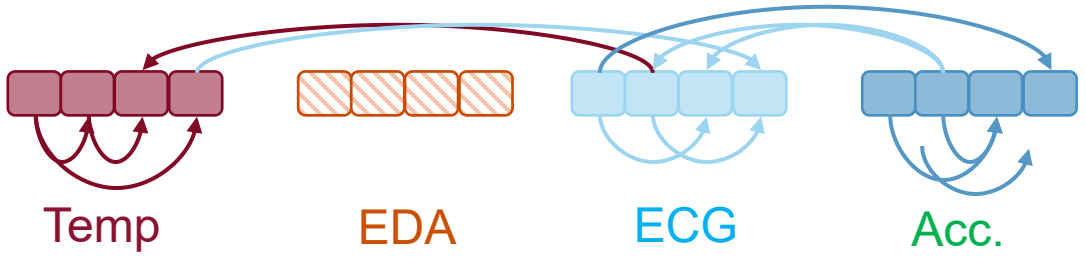




Cross-attention can allow learning task-relevant modality interaction.

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]

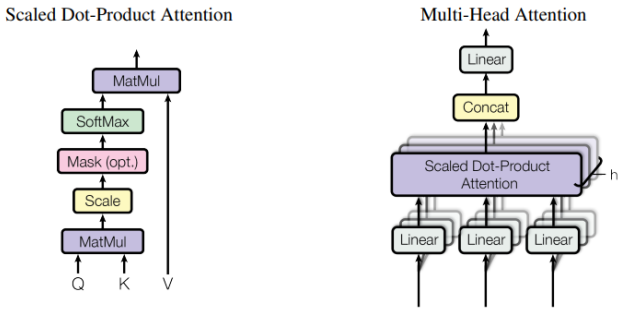


Figure 2: (left) Scaled Dot-Product Attention. (right) Multi-Head Attention consists of several attention layers running in parallel.

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

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

[1] Attention Is All You Need, Neurips 2017

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

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

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

Canonical Self-Attention[1]

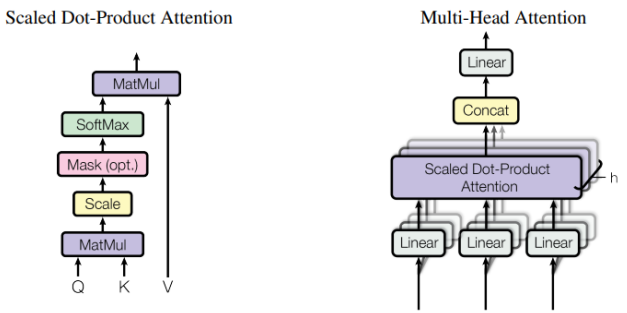


Figure 2: (left) Scaled Dot-Product Attention. (right) Multi-Head Attention consists of several attention layers running in parallel.

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

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

$$\mathcal{O}(Llog(L))$$

[1] Attention Is All You Need, Neurips 2017

Handling long time-series through sparse attention

- 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.

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

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

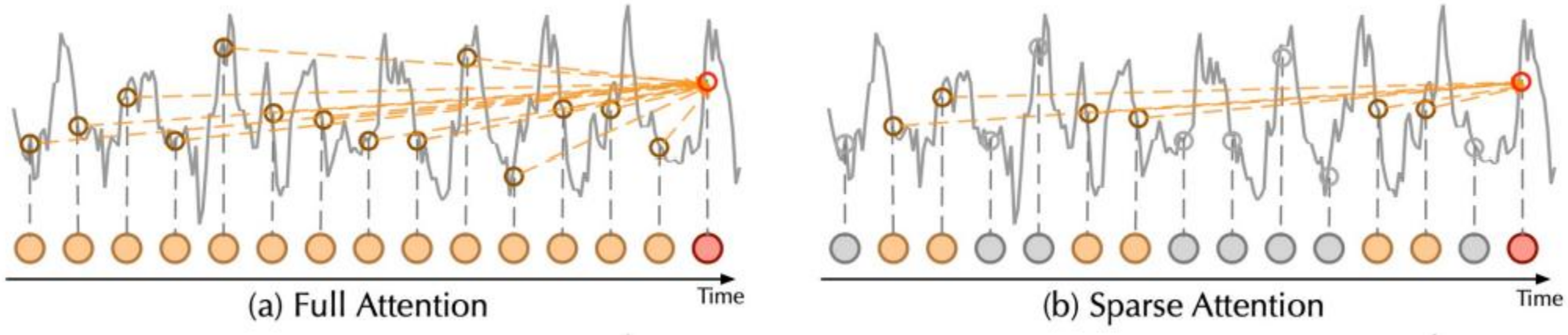
$$\dot{s} = \bar{s} + \hat{s}$$

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

Handling long time-series through sparse attention

Key Design Motivation :

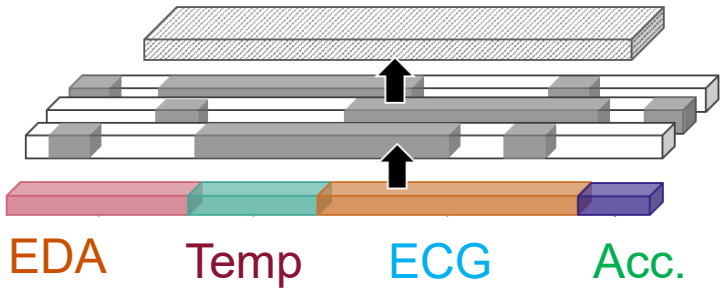
- Self-attention is long-tailed. So only a few dot-product pairs contribute to the major attention and others can be ignored[1].
- Using a sparsity metric[1], we can sample the more informative query-key pairs.
- Instead of L queries, we now use **top-v** where $\mathrm{v} = u * logL$ queries → $\mathcal{O}(Llog(L))$



[1] Informer, AAAI 2021

Figure source : https://wuhaixu2016.github.io/pdf/NeurIPS2021_Autoformer.pdf

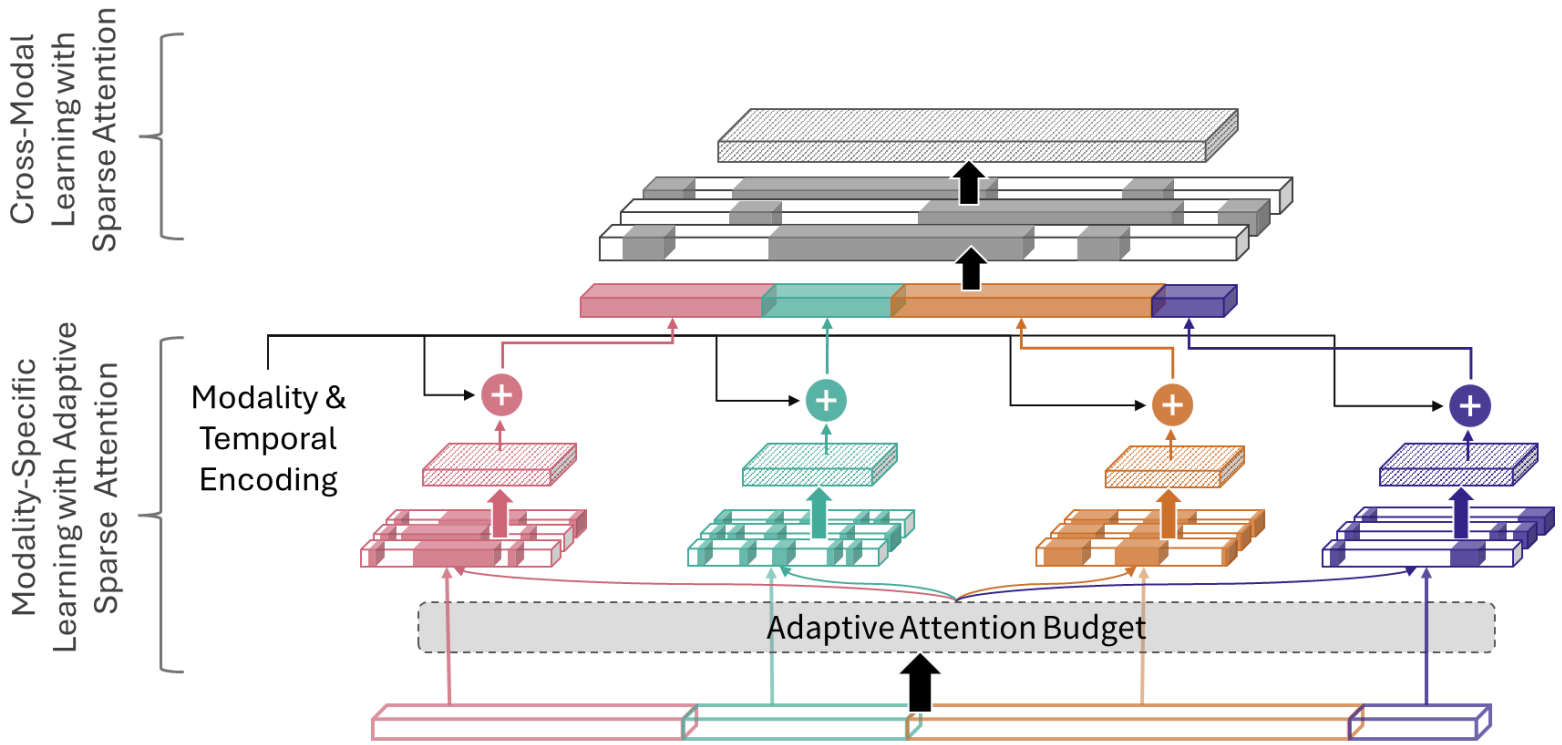
Handling long multimodal time-series through sparse attention



Sparse cross-modal learning

1. Adaptive Attention Budget per modality

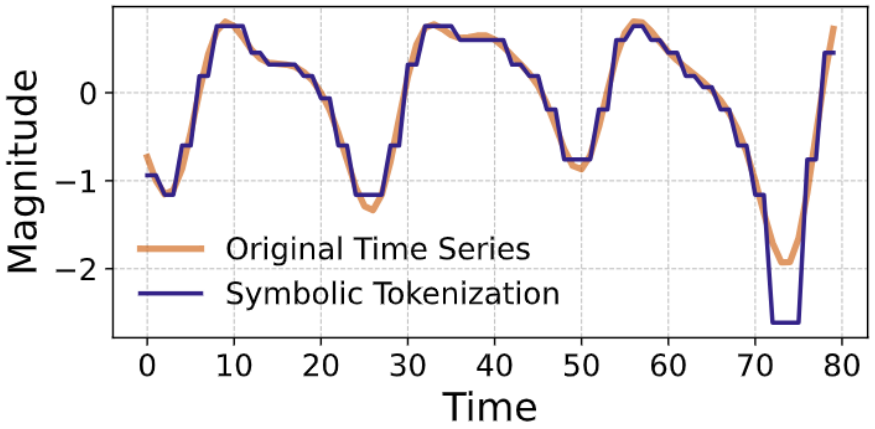
Top-v queries in ProbSparse Computation : v is modulated by modality's relevance and availability



2. Symbolic Tokenization

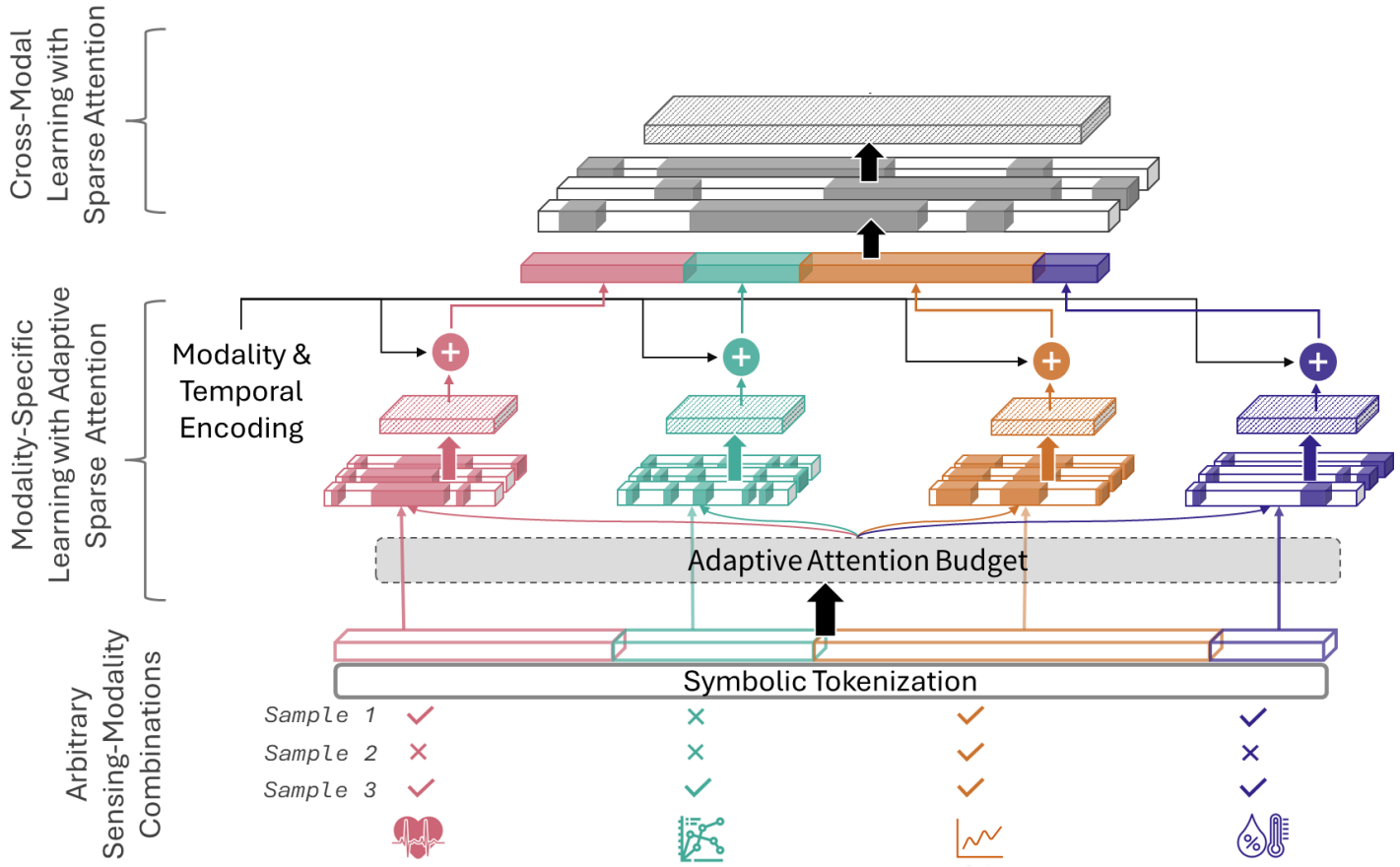
1. Converts time-series to discrete *symbols*[1].
   - Has some nice properties – guarantees a lower bound Euclidean distance between the symbolic time-series and the original time-series.
   - We extend it under some assumptions that this tokenization preserves multimodal relational structure.

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



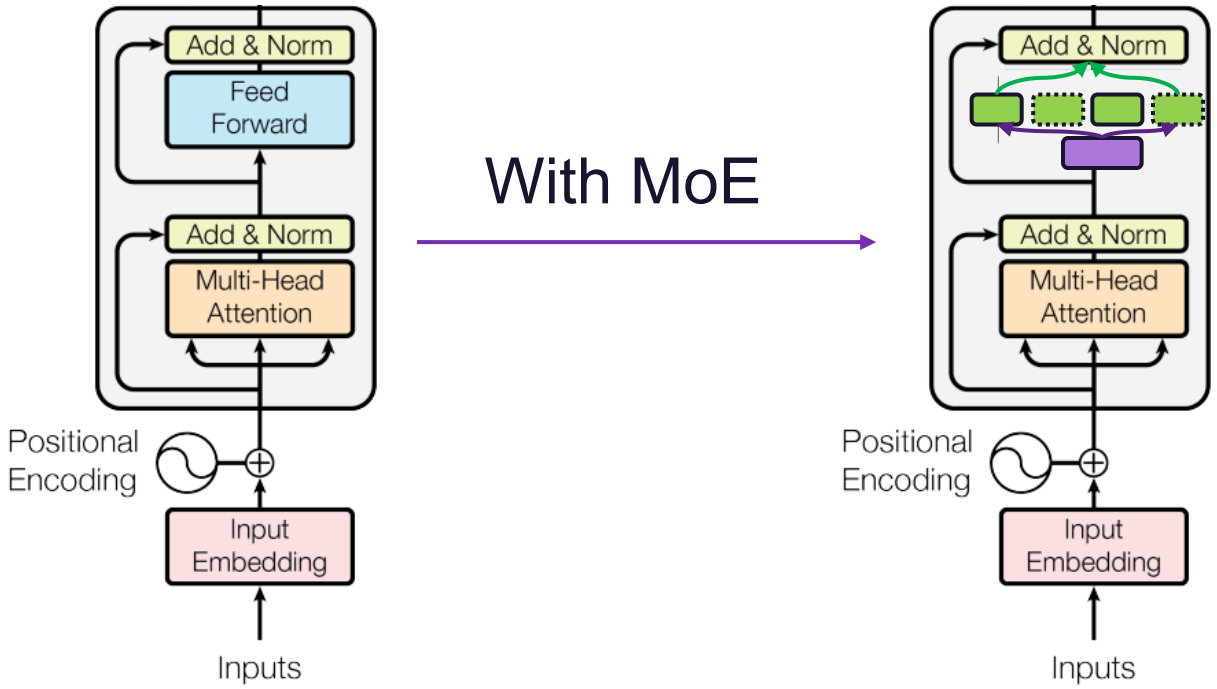
[1] Experiencing SAX: a Novel Symbolic Representation of Time Series, Data Mining and Knowledge Discovery, 2007

2. Symbolic Tokenization



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.

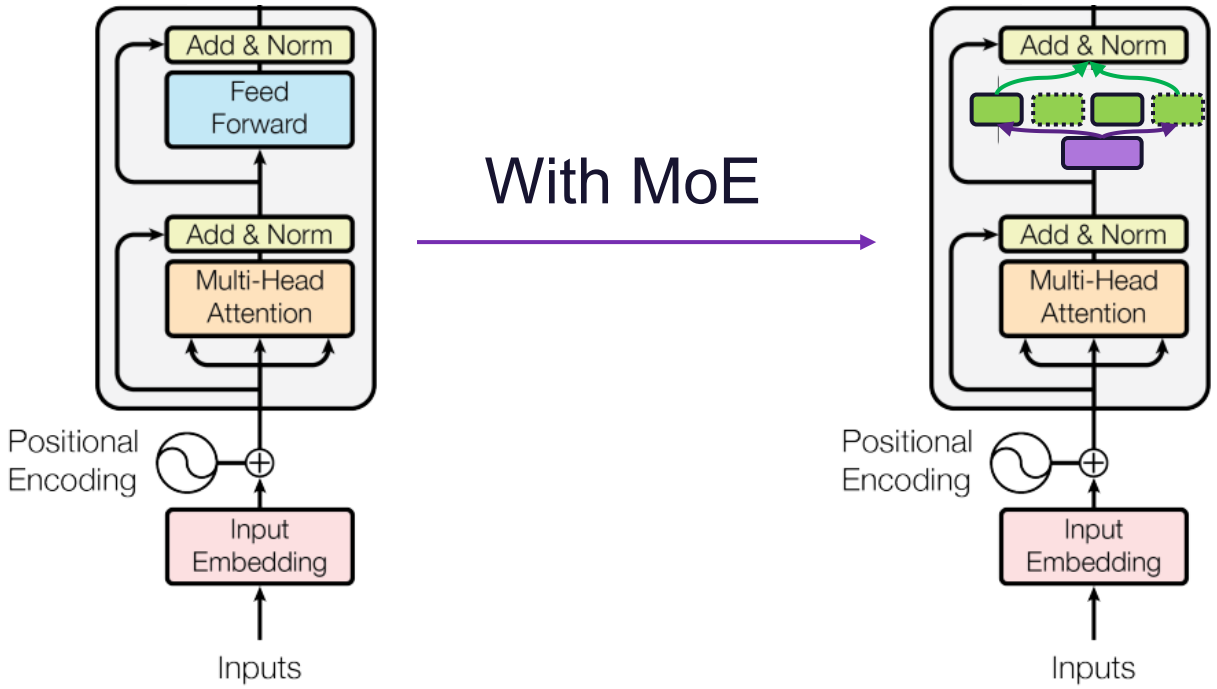


Vanilla Transformer[1]

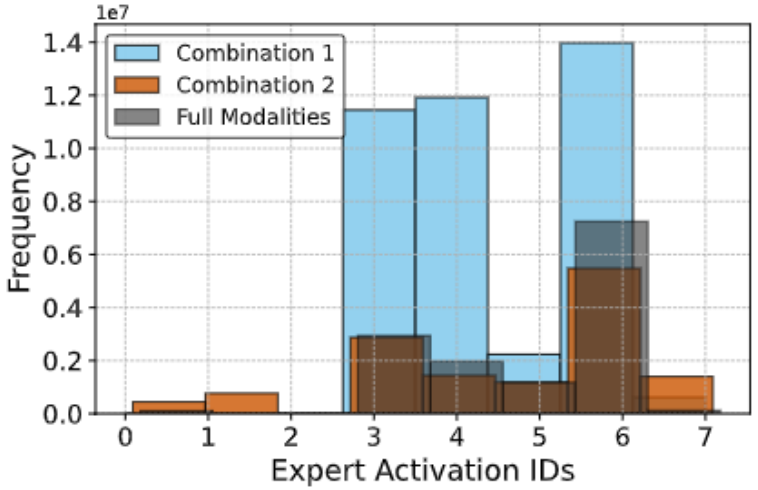
[1] Attention Is All You Need, Neurips 2017

3. Handling Missingness through Mixture-of-Experts

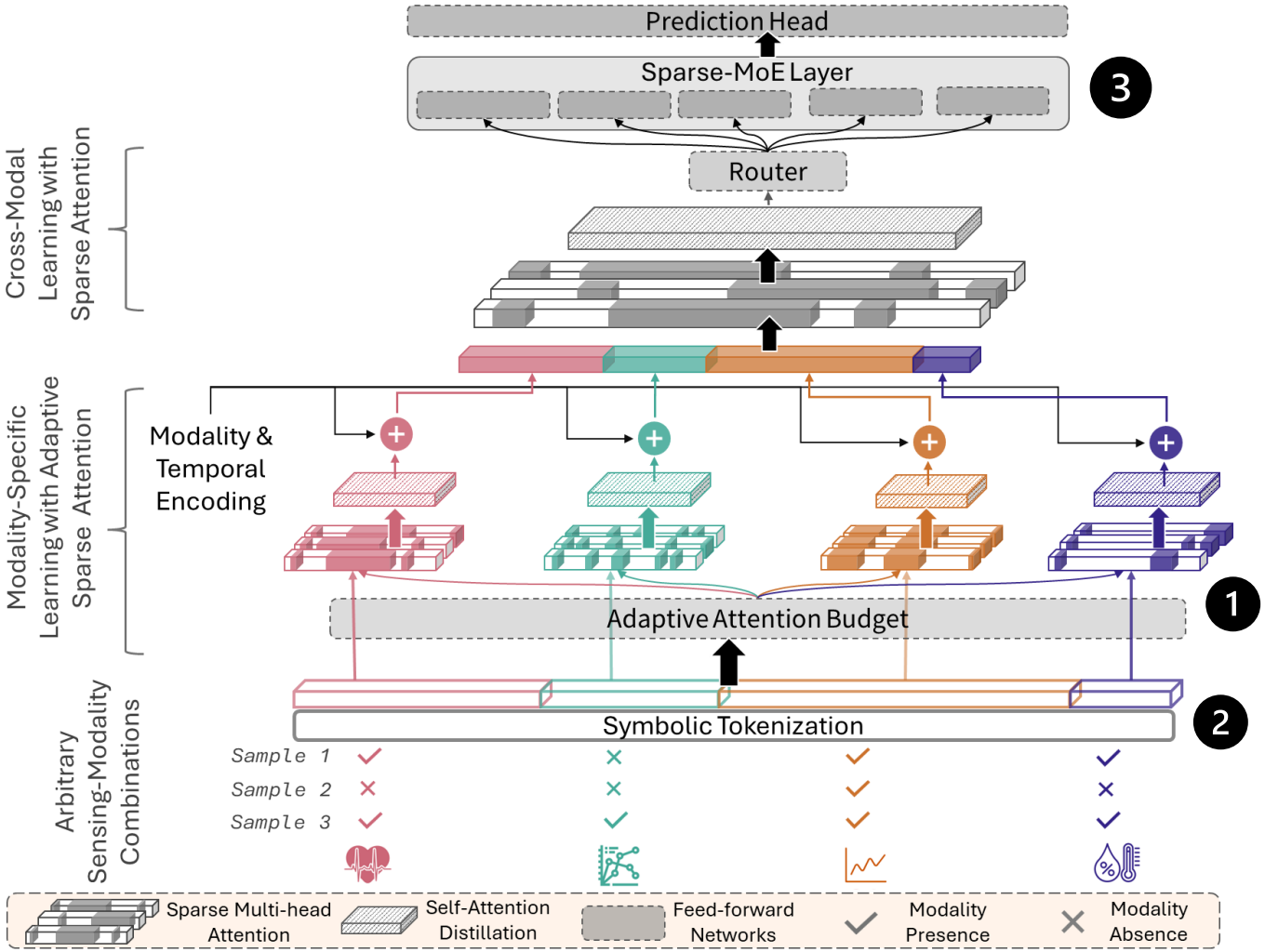
Instead of the fixed Feed-forward layer of the transformer, we can use a mixture of experts(MoE) for implicit modality specialization.

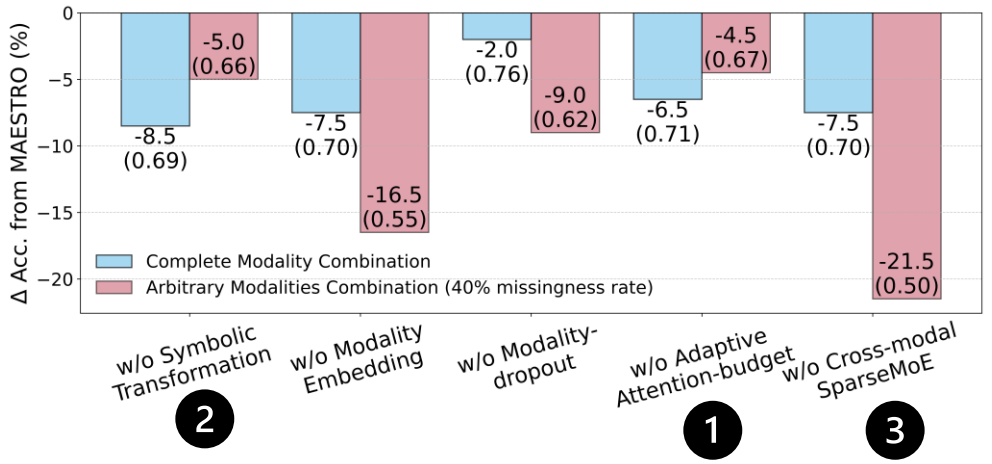


Vanilla Transformer[1]



[1] Attention Is All You Need, Neurips 2017

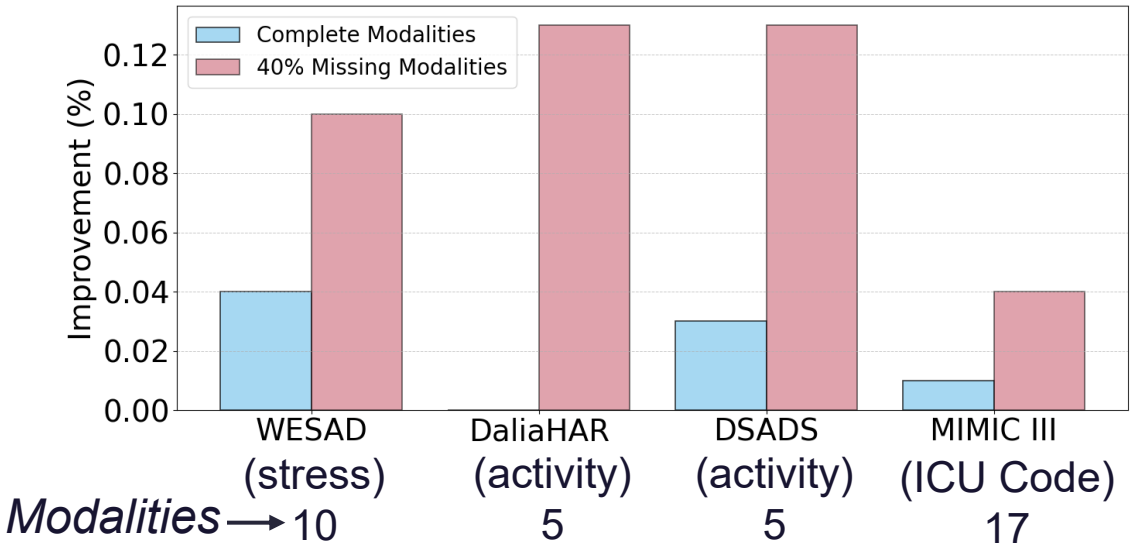




MAESTRO : Adaptive Sparse Attention and Robust Learning for Multimodal Dynamic Time Series, Neurips 2025 (Spotlight)

Key Results : MAESTRO

Performance with complete and 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

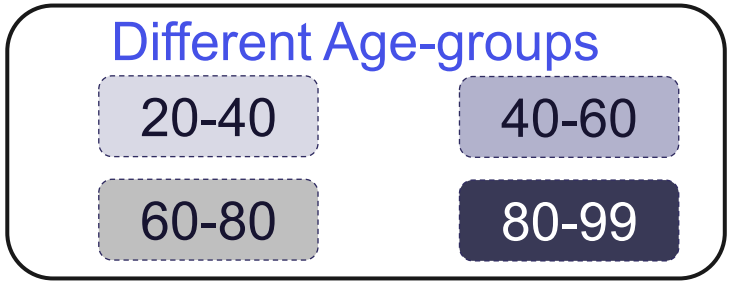
Payal Mohapatra and Yueyuan Sui and Akash Pandey and Stephen Xia and Qi Zhu, MAESTRO : Adaptive Sparse Attention and Robust Learning for Multimodal Dynamic Time Series

Need for Generalization in Sensing Applications

Different Devices[1, 2]

Domains

Subpopulation Shift [1, 2]



Human Activity Recognition (HAR) using Inertial Measurement Units' recordings.

Sleep-Stage Classification (SSC) using Electroencephalogram (EEG) Recordings.

[1] WOODS: Benchmarks for Out-of-Distribution Generalization in Time Series
[2] ADATIME: A Benchmarking Suite for Domain Adaptation on Time Series Data

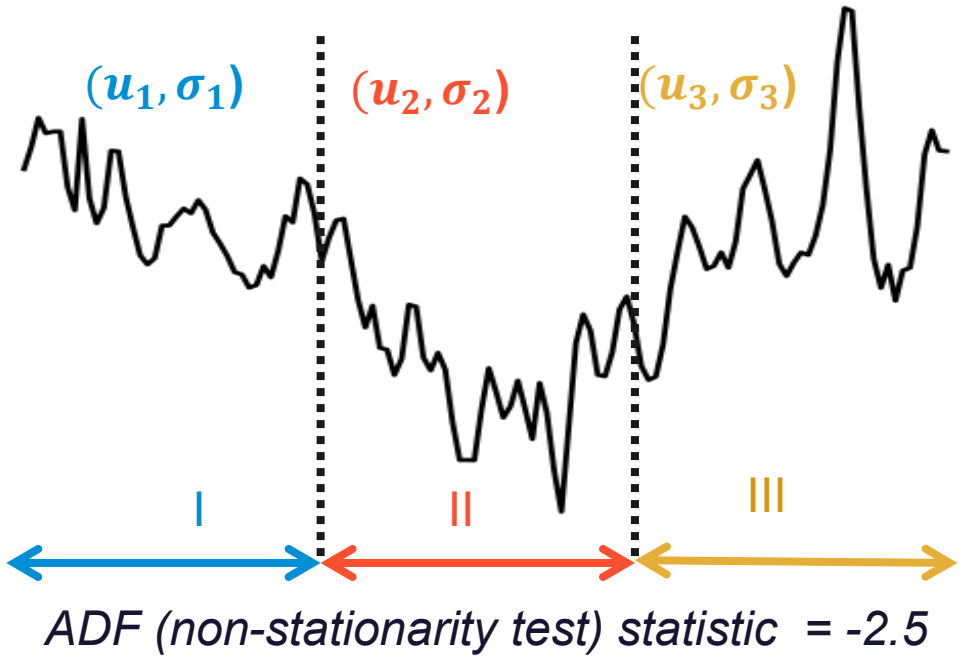
Challenges for Time-Series Domain Generalization

Not only is there a domain-gap[1]

	HAR				
	UCIHAR	WISDM	HHAR	SSC	MFD
Same Domain (Target-only)	100.00	98.02	98.55	72.09	99.39
Cross-Domain (Source-only)	65.94	48.60	63.07	51.67	72.51
Gap (δ)	37.32	49.44	33.86	18.38	26.88

But also intra-domain nonstationarity[2, 3].

- varying statistical and spectral properties over time.



[1] ADATIME: A Benchmarking Suite for Domain Adaptation on Time Series Data
[2] Out-of-Distribution Representation Learning for Time-Series Classification
[3] AdaRNN: Adaptive Learning and Forecasting of Time Series

Motivation

- 🗝 Applying standard Domain Generalization (DG) algorithms to time-series is ineffective[1, 2].
- 🗝 Time-Series Domain Adaption techniques[3].
  - Need access to some target domain data for alignment.

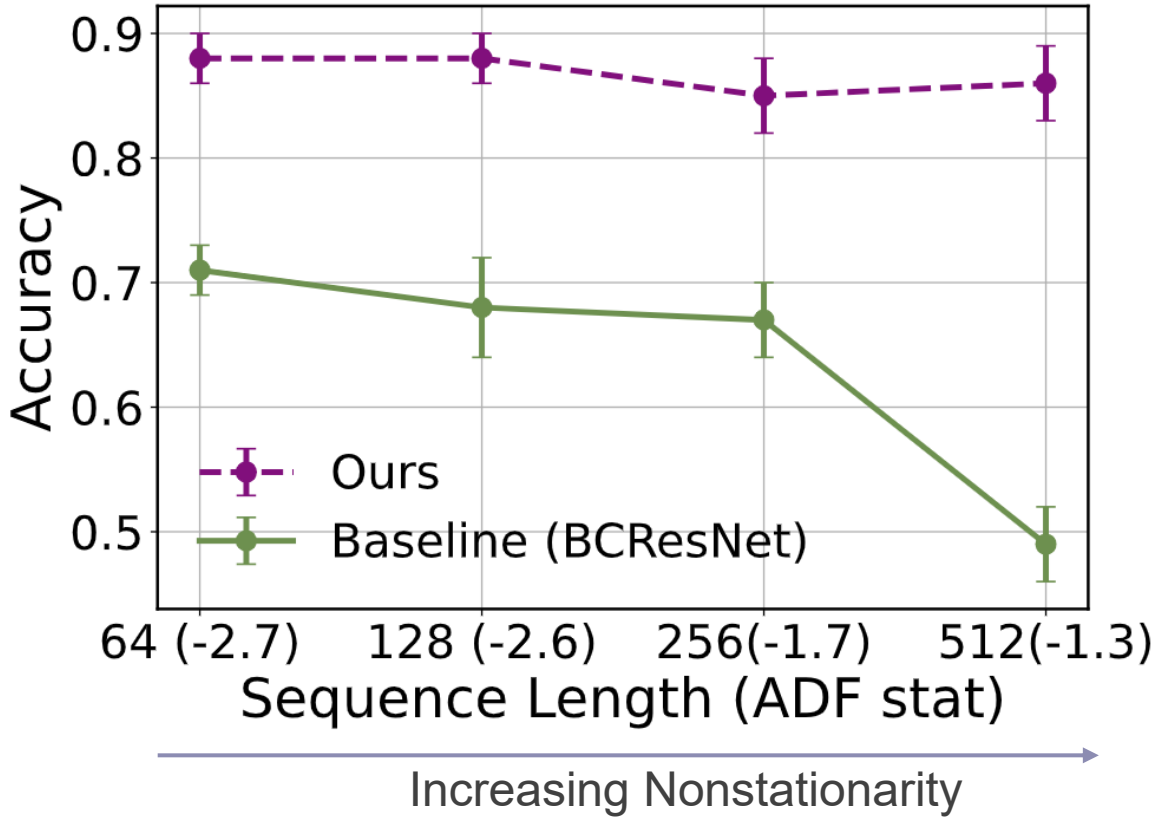
❑ Need for time-series classification method that

- ✓ Learns generalizable representations.
- ✓ Doesn't access target domain data.
- ✓ Doesn't require explicit domain labels.

[1] GLOBEM Dataset: Multi-Year Datasets for Longitudinal Human Behavior Modeling Generalization
[2] WOODS: Benchmarks for Out-of-Distribution Generalization in Time Series
[3] Source-Free Domain Adaptation with Temporal Imputation for Time Series Data

Our Approach : Phase-driven Domain-Generalization for Time-series Classification (PhASER)

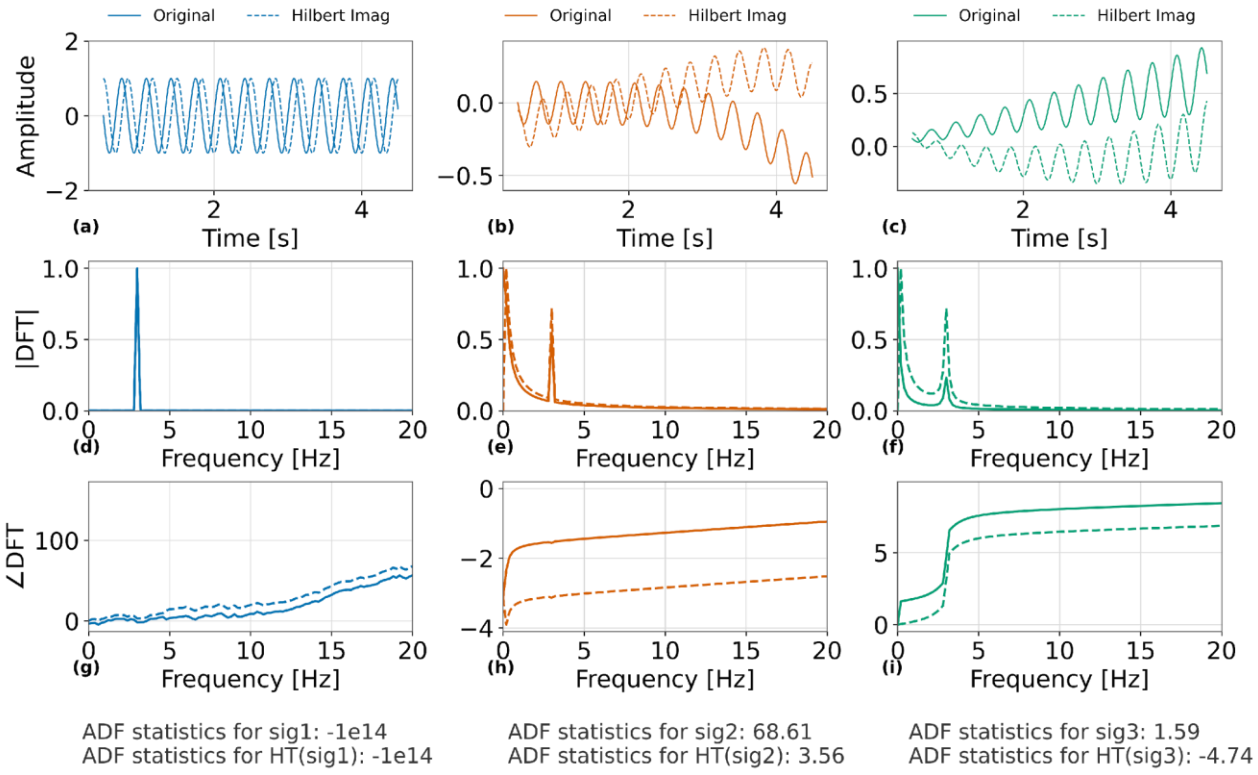
1. Nonstationarity impacts generalization.



Mohapatra, Payal, Lixu Wang, and Qi Zhu. "Phase-driven Domain Generalizable Learning for Nonstationary Time Series." **TMLR 2025**.

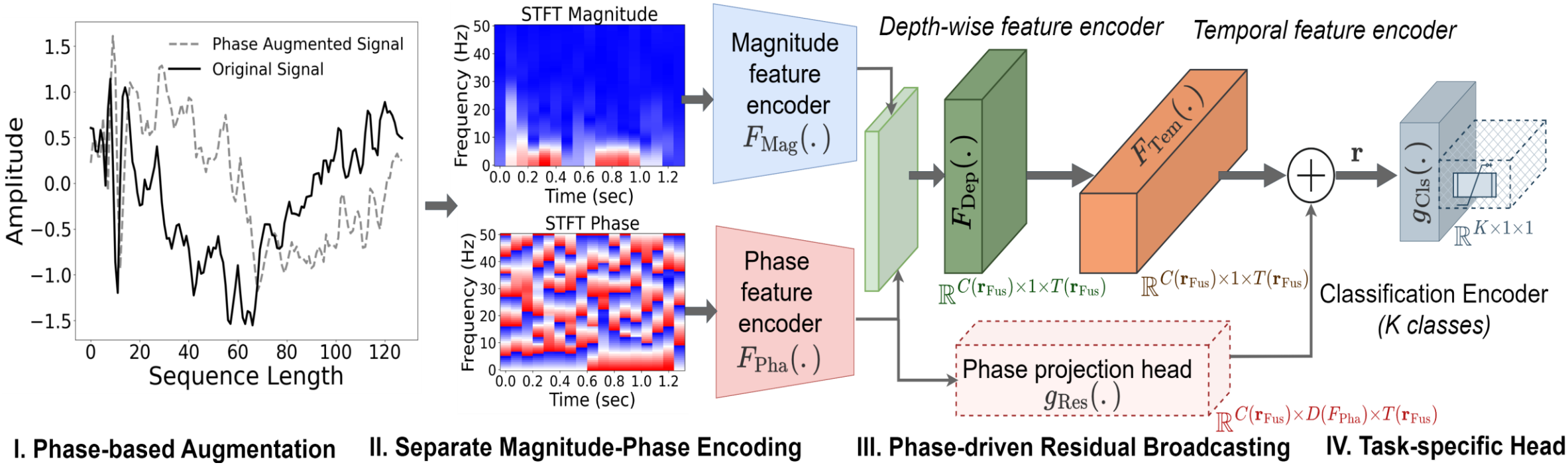
Our Approach : Phase-driven Domain-Generalization for Time-series Classification (PhASER)

1. Nonstationarity impacts generalization.
2. Phase information of a signal encodes nonstationarity.
3. Changing Phase (carefully) can diversify nonstationarity.
   - which can aid in generalization.



Mohapatra, Payal, Lixu Wang, and Qi Zhu. "Phase-driven Domain Generalizable Learning for Nonstationary Time Series." **TMLR 2025**.

Our Approach : Phase-driven Domain-Generalization for Time-series Classification (PhASER)



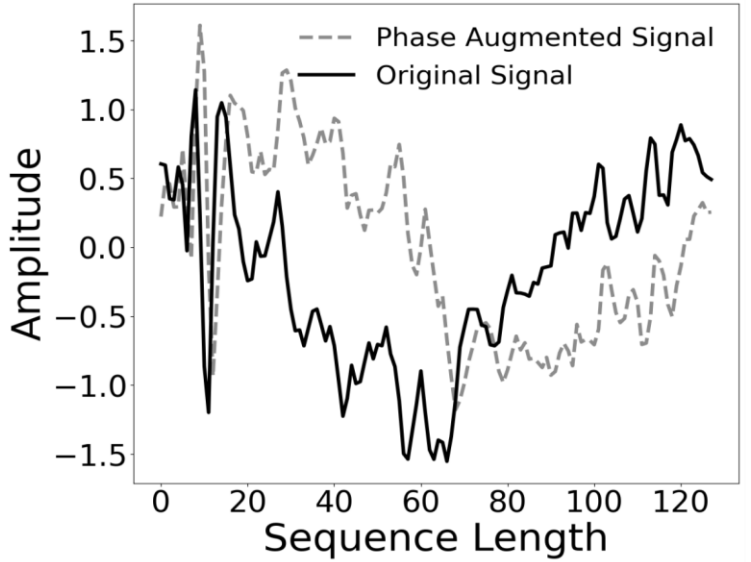
[Approach]: PhASER leverages phase augmentation, separate magnitude-phase encoding, and phase-residual broadcasting for distribution-invariant learning, achieving up to 11% performance improvement across five datasets against 13 baselines.

Mohapatra, Payal, Lixu Wang, and Qi Zhu. "Phase-driven Domain Generalizable Learning for Nonstationary Time Series." **TMLR 2025**.

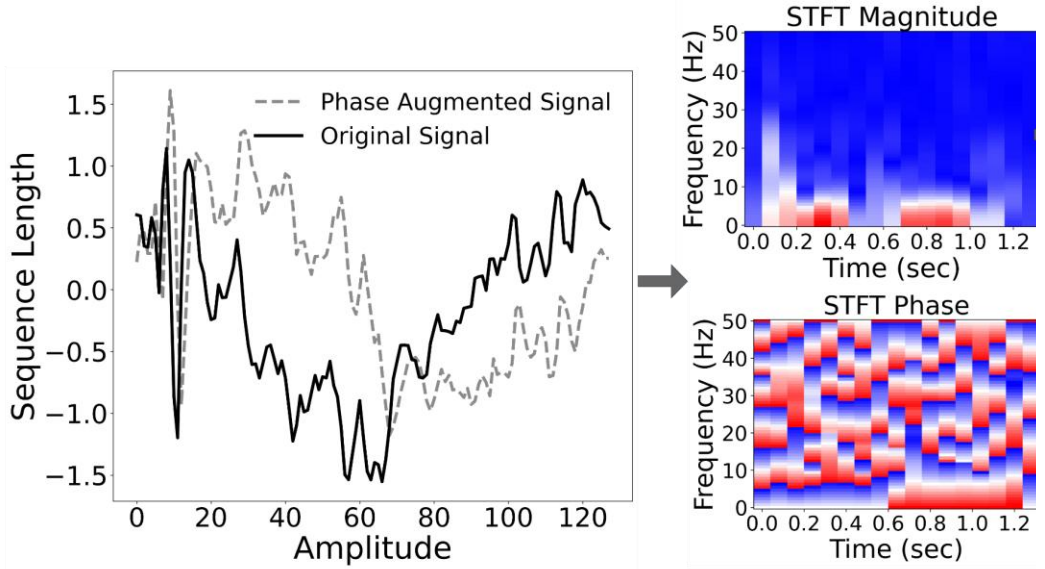
1. Hilbert-transform based augmentation

Interesting properties of Hilbert Transform for us

- Capable of handling nonstationary signals.
- Retains magnitude response.
- In-sample augmentation (unlike mixup and other strategies)
  - No need to incorporate any application-specific characteristics to achieve this.
  - Simple.

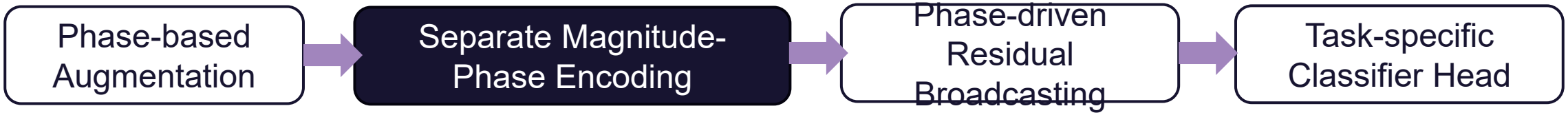


2. Separate Magnitude and Phase Encoding



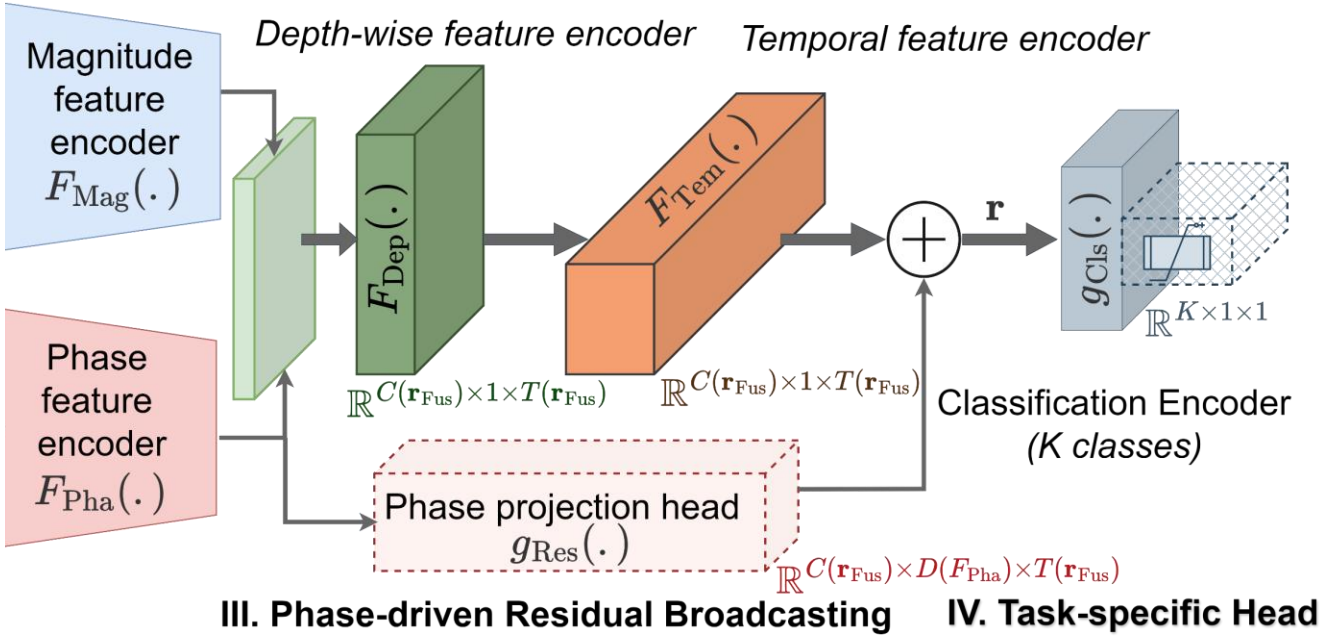
I. Phase-based Augmentation **II. Separate Magnitude-Phase Encoding**

- Phase modification diversifies the nonstationarity statistics.
- Separately encode phase and magnitude better design choice.
  - Short-term Fourier Transform feature encoding.

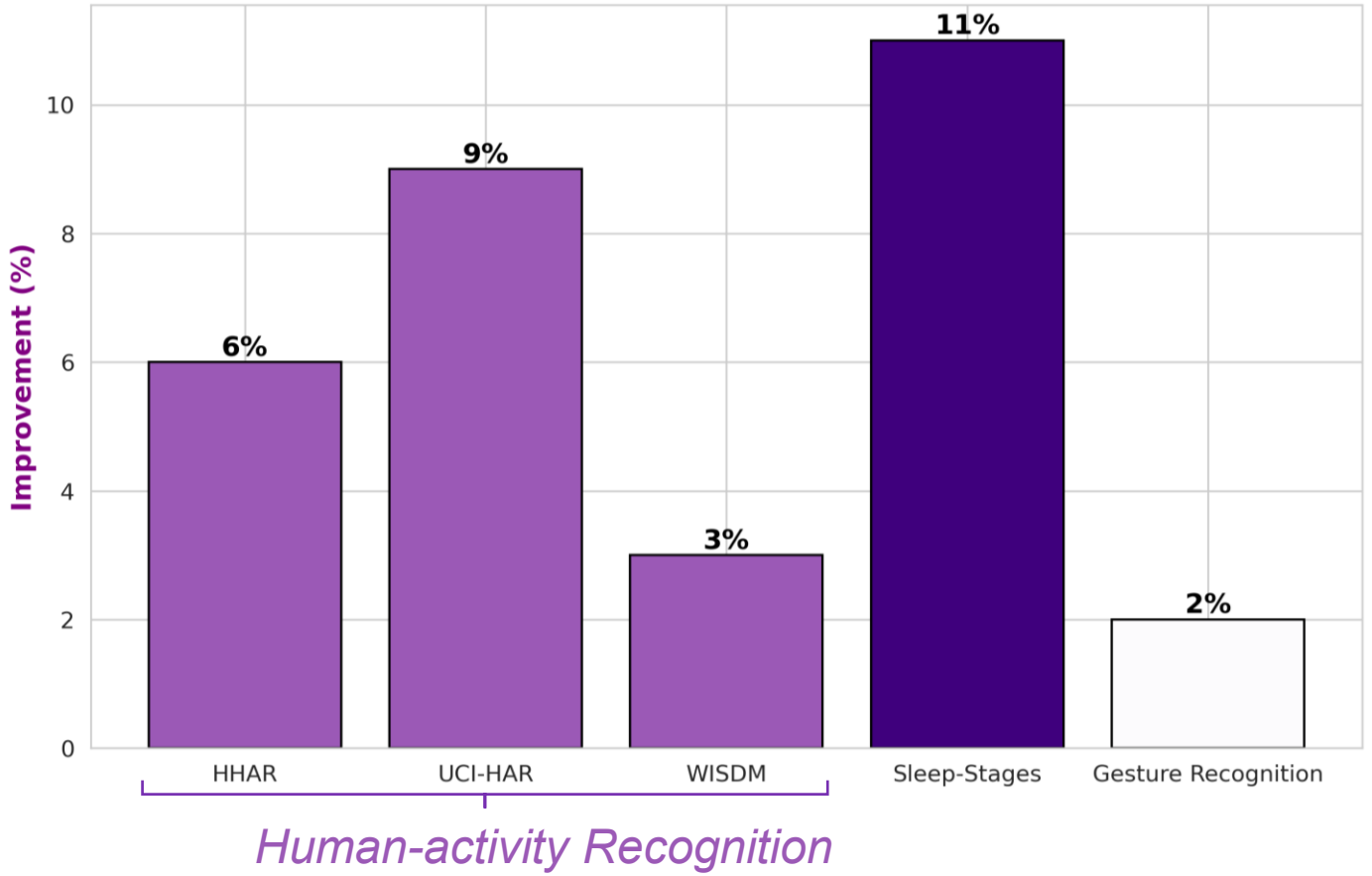


3. Phase Residual-Broadcasting Network

- Residual-Broadcasting Network
  - Collapse to 1D then broadcast using Phase-embeddings to 2D.
- Use Phase-embedding as a residual.



Key Result : PhASER



Mohapatra, Payal, Lixu Wang, and Qi Zhu. "Phase-driven Domain Generalizable Learning for Nonstationary Time Series." **TMLR 2025**.

Summary

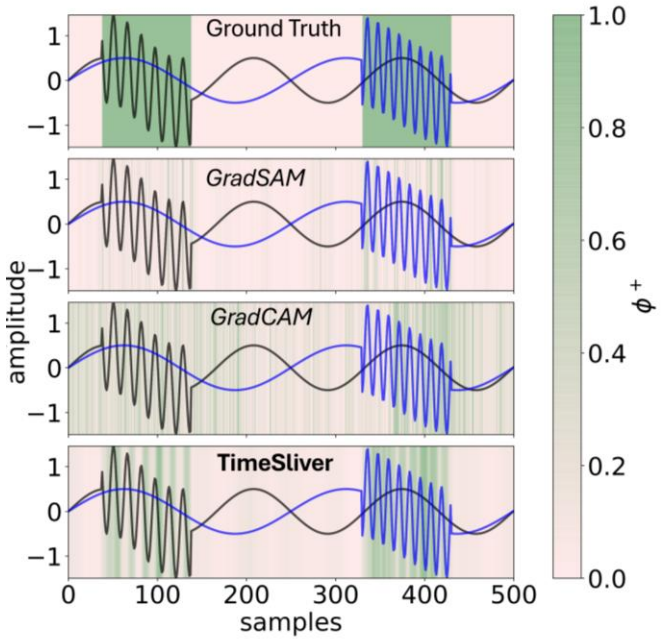
- MAESTRO**:** Multimodal handling of time series using symbolic tokenization and a Mixture-of-Experts framework for **robustness against** arbitrary **sensor missingness.**

- PhASER**:** Anchoring design in phase information through augmentation, separate feature encoding, and residual broadcasting to enable **generalizable representations for nonstationary time-series** classification.

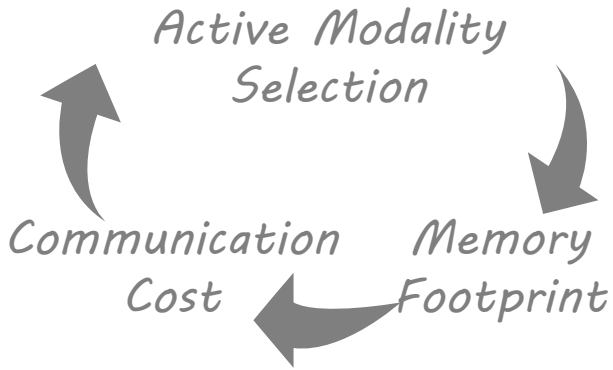
Addressing the Challenges in Time Series Data Analyses

Heterogeneity	P.Mohapatra et.al, Can LLMs Understand Unvoiced Speech? Exploring EMG-to-Text Conversion with LLMs, **ACL 2025** P.Mohapatra et.al, MAESTRO : Adaptive Sparse Attention and Robust Learning for Multimodal Dynamic Time Series, **Neurips 2025** *(Spotlight)*
Missingness	P.Mohapatra et.al, Missingness-resilient video-enhanced multimodal disfluency detection, **Interspeech 2024** P.Mohapatra et.al, Person identification with wearable sensing using missing feature encoding and multi-stage modality fusion, **ICASSP 2023** *(top performer in Signal Processing Grand Challenge)*
Distribution Shift	P.Mohapatra et.al, Phase-driven domain generalizable learning for nonstationary time series, **TMLR 2025**
Subjectivity in labels (Human-centered applications)	P.Mohapatra et.al, Wearable network for multilevel physical fatigue prediction in manufacturing workers, **PNAS Nexus 2024** P.Mohapatra et.al, Effect of attention and self-supervised speech embeddings on non-semantic speech tasks, **ACM Multimedia 2023** *(top performer in ComPaRe challenge)*
Data-Efficient Learning	P.Mohapatra et.al, Efficient stuttering event detection using siamese networks, **ICASSP 2023** P.Mohapatra et.al, Speech disfluency detection with contextual representation and data distillation, IASA@**Mobisys 2022**

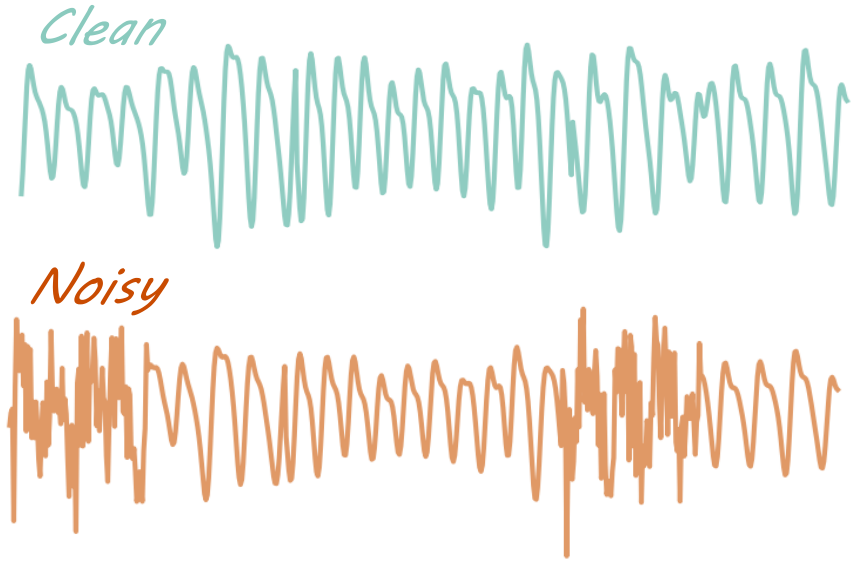
Ongoing Efforts and Future works



Interpretability

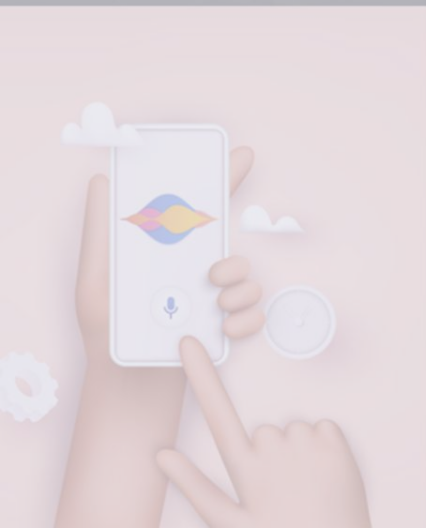


System Efficiency



Noisy Data Streams

…and many more ☺



Thank You.

Payal Mohapatra

Designing technical enablers to improve analytics frameworks
for real world time-series applications.

Connect with me here!

Learn more about my work here!