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

Northwestern University

Openreview Link

Payal Mohapatra Yueyuan Sui Akash Pandey Stephen Xia Qi Zhu

Northwestern University, USA

Connect with me

NEURAL INFORMATION PROCESSING SYSTEMS

Motivation

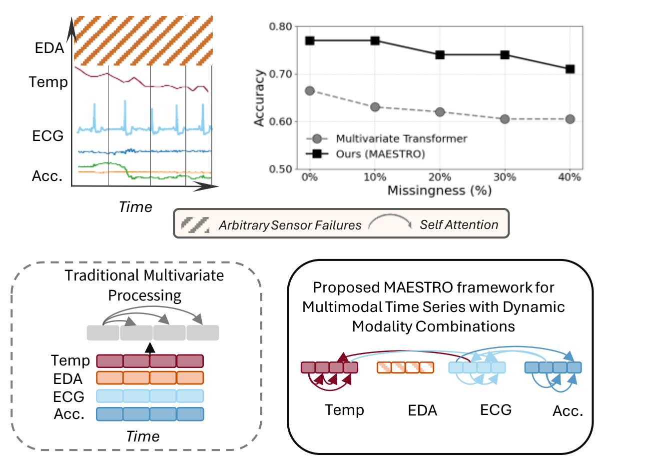
- Traditional approaches treat multisensor data as *multivariate* time-series, ignoring heterogeneous modalities and their interactions.
- Existing *sensor-fusion* approaches are often application-specific and heuristic.

We present a novel perspective on modeling time-series data from multiple diverse sensing modalities as **multimodal learning**, enabling the discovery of task-relevant, modality-specific and cross-modal interactions.

Common multimodal paradigms cannot be directly applied to real-world sensing applications:

- Reliance on a *single anchor modality* is not feasible when the primary modality is not known *a priori*.
- *Contrastive learning* assumes high mutual information among modalities, which is not always guaranteed in heterogeneous sensing modalities.
- *Pairwise interaction modeling* grows combinatorially as the number of modalities increases ($M \geq 4$).

Another important consideration in designing multimodal learning frameworks for time-series from real-world sensing applications is supporting learning from **arbitrary combinations of sensing modalities**, as sensor malfunction is common.



We introduce MAESTRO, a robust framework that overcomes key limitations of existing multimodal and multivariate learning approaches while handling samples with arbitrary sensor combinations. It constructs **long multimodal sequences** to facilitate dynamic intra- and cross-modal interactions based on task relevance. We use **sparse attention, symbolic tokenization, adaptive attention budgeting, and Mixture-of-Experts** to efficiently realize MAESTRO.

Handling Long Multimodal Sequences

Although self-attention's computational complexity increases quadratically with sequence length, it exhibits a long-tailed distribution. We can sample only the most informative queries from the long cross-modal sequence to improve computational efficiency.

Sparse Attention: For each query $\mathbf{q} \in \mathbf{Q}$, we first sample a random subset of keys $\mathbf{K}' \subset \mathbf{K}$ where $|\mathbf{K}'| = u \log L_K$. We compute a sparsity metric $\mathcal{P}(\mathbf{q}, \mathbf{K}')$ that evaluates the query's attention diversity.

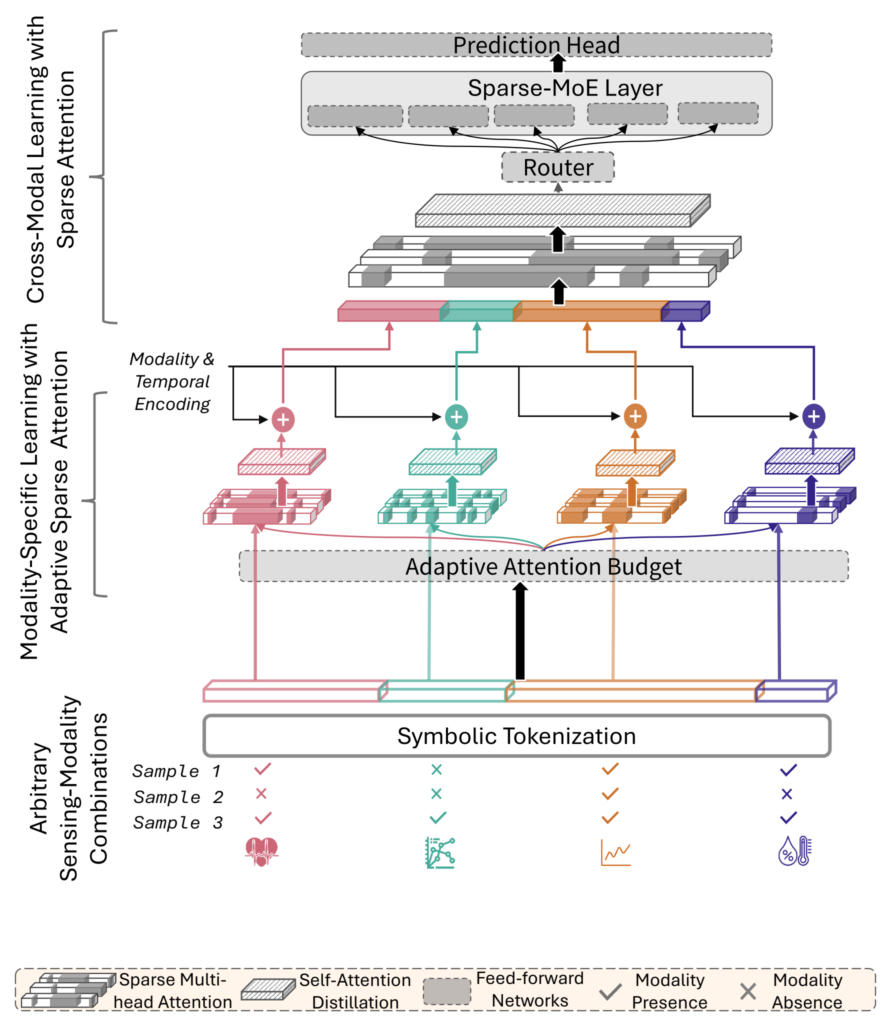
Queries with higher $\mathcal{P}(\mathbf{q}, \mathbf{K}')$ scores contain more distinctive information. We select the top-v queries (where $v = u \log L_Q$) with the highest sparsity scores to favor diverse queries. This selection is performed independently for each attention head to avoid excessive information loss.

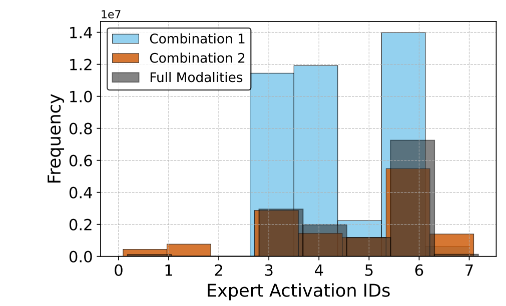
Complexity comparison. $\hat{L}$: multimodal length, M: modalities, $L_{\max}$: longest sequence.

Method	Time	Space
Dense	$\mathcal{O}(\hat{L}^2)$	$\mathcal{O}(\hat{L}^2)$
Pairwise	$\mathcal{O}(M^2 L_{\max}^2)$	$\mathcal{O}(M^2 L_{\max}^2)$
Sparse (Ours)	$\mathcal{O}(\hat{L} \log \hat{L})$	$\mathcal{O}(\hat{L} \log \hat{L})$

Overall Model Architecture

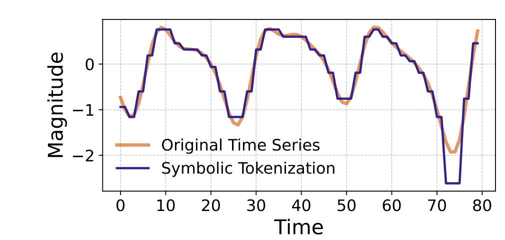
MAESTRO integrates four key components: ❶ symbolic tokenization with reserved tokens for missing data, ❷ adaptive attention budgeting guided by modality availability and relevance, ❸ sparse attention over long multimodal sequences to capture rich cross-modal context, and ❹ loss-free, MoE-based dynamic routing that adapts to varying modality observations.





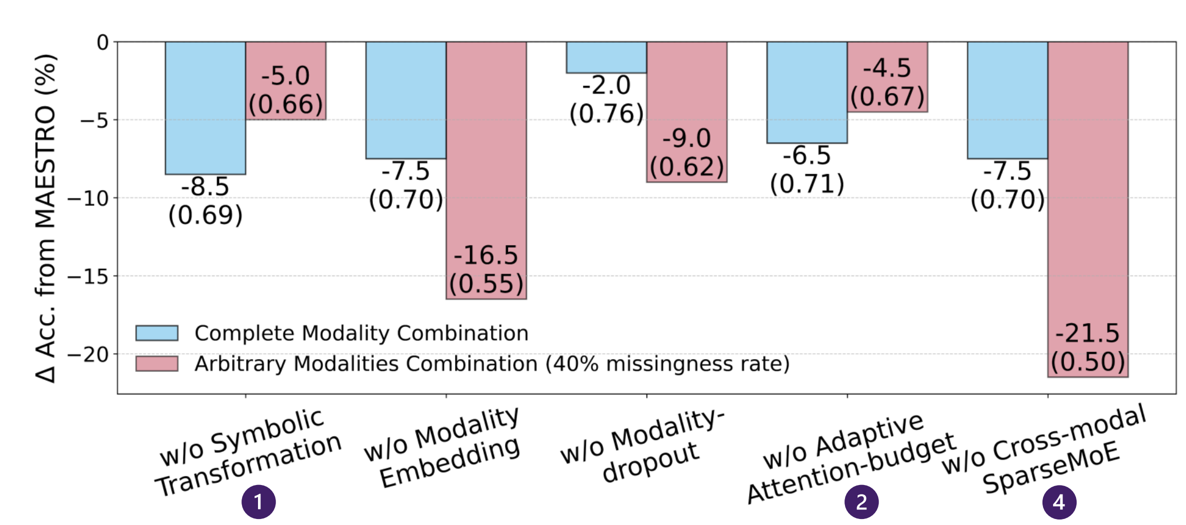
❹ Sparse MoE supports input-dependent dynamism based on modality combinations.

❷ Adaptive Attention Budget parameterizes v for selecting top-v dominant queries for each modality.

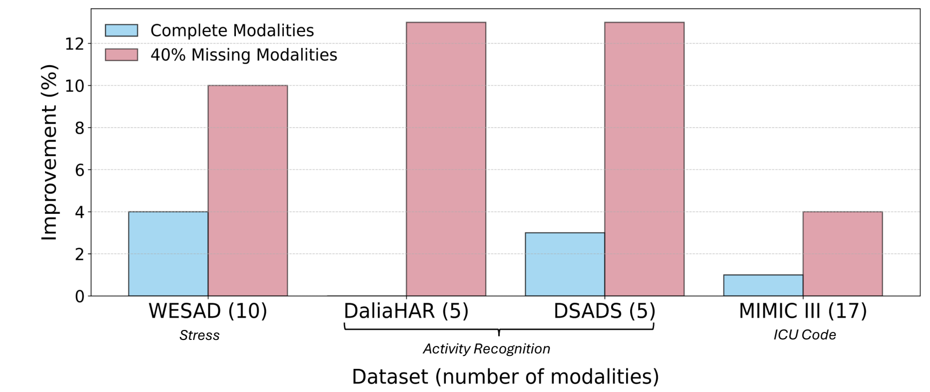


❶ Time-series, $x[t]$, is converted to discrete tokens $s[w] \in \{s_0, s_1, \ldots, s_\alpha\}$, where s_0 is reserved to denote missingness.

Ablation Study



Key Experimental Results



Overall MAESTRO offers an average improvement of 8% over top multivariate and 4% over top multimodal baselines—across all benchmarks.

Under dynamic conditions with arbitrary number of missing modalities, MAESTRO achieves increasing accuracy improvements over the best baseline—7.6% improvement under 10% missingness and 9.4% under 40%.

Computational Complexity

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

References

[1] Informer: Beyond efficient transformer for long sequence time-series forecasting. *AAAI Conference on Artificial Intelligence, 2021.*
[2] Quantifying & modeling multimodal interactions: An information decomposition framework. *Advances in Neural Information Processing Systems, 2023.*
[3] Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023.*
[4] Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the Association for Computational Linguistics, 2019.*
[5] Flex-MoE: Modeling arbitrary modality combination via the flexible mixture-of-experts. *Advances in Neural Information Processing Systems, 2024.*