

Appendix

This Appendix includes additional details for the paper, “MAESTRO : Adaptive Sparse Attention and Robust Learning for Multimodal Dynamic Time Series”, including the reproducibility statement, additional details on symbolic tokenization and theoretical proof of **Corollary 3.2** of the main paper (Section A), additional details for sparse multihead attention (Section B), additional experimental setup and training details of MAESTRO (Sections C and D) with detailed dataset introduction (Section D.2), more detailed results of the experiments shown in the main paper (Section E), additional experiments (Section F), and more discussion on broader impacts.

Reproducibility Statement

A minimal source-code has been provided in the Supplementary Materials. We use public datasets and provide implementation details in the following sections.

A Additional Details on the Symbolic Tokenization

Piecewise Aggregate Approximation (PAA) and Symbolic Aggregate approXimation (SAX) are sequential time-series compression methods that reduce temporal resolution by dividing the signal into fixed-size windows. PAA summarizes each window by its mean value, producing a smoothed, real-valued lower-dimensional representation. SAX builds on PAA by further discretizing these means into symbolic tokens using breakpoints derived from a standard Gaussian distribution, resulting in a compact and interpretable symbolic sequence. As detailed in Section 3.2.1 of the main paper, PAA serves as an intermediate step in the SAX transformation pipeline, bridging raw signal compression and symbolic quantization. While PAA captures coarse temporal structure, SAX enables symbolic reasoning, efficient indexing, and explicit missingness encoding via a reserved symbol. An illustration for their comparison is shown in Figure 8.

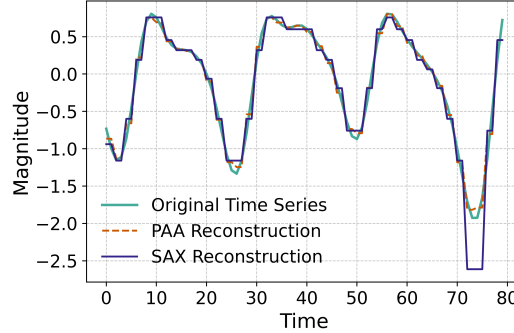


Figure 8: Comparison of PAA and SAX representations. The raw photoplethysmograph (PPG) signal $x[t]$ is first segmented and averaged into $\bar{x}[w]$ via PAA. SAX extends this by mapping each PAA segment to a discrete symbol $s[w] \in \{s_1, \dots, s_\alpha\}$, yielding a compact symbolic sequence. A reserved token s_0 can be used to indicate missing segments.

Proof Sketch for Corollary in Section 3.2.1 (Cross-Modal Relational Preservation)

Goal. Show that:

$$\left| \text{Dist}_{\text{sym}}^j(s_i^j, s_k^j) - \text{Dist}_{\text{sym}}^m(s_i^m, s_k^m) \right| \leq \left| \|x_i^j - x_k^j\|_2 - \|x_i^m - x_k^m\|_2 \right| + \epsilon_j + \epsilon_m.$$

Assumptions.

1. Lower-bound property of SAX (MINDIST):

$$\text{Dist}_{\text{sym}}^j(s_i^j, s_k^j) \leq \|x_i^j - x_k^j\|_2, \quad \text{Dist}_{\text{sym}}^m(s_i^m, s_k^m) \leq \|x_i^m - x_k^m\|_2.$$

2. Bounded symbolic approximation error:

$$\|x_i^j - x_k^j\|_2 - \text{Dist}_{\text{sym}}^j(s_i^j, s_k^j) \leq \epsilon_j, \quad \|x_i^m - x_k^m\|_2 - \text{Dist}_{\text{sym}}^m(s_i^m, s_k^m) \leq \epsilon_m.$$

Proof. We begin by applying the *triangle inequality*, which states that for any real numbers a, b, c , the following holds:

$$|a - c| \leq |a - b| + |b - c|.$$

We use this property iteratively to decompose the difference between the symbolic distances.

$$\begin{aligned} |\text{Dist}_{\text{sym}}^j - \text{Dist}_{\text{sym}}^m| &= \left| \text{Dist}_{\text{sym}}^j - \|x_i^j - x_k^j\|_2 + \|x_i^j - x_k^j\|_2 - \|x_i^m - x_k^m\|_2 + \|x_i^m - x_k^m\|_2 - \text{Dist}_{\text{sym}}^m \right| \\ &\leq \left| \text{Dist}_{\text{sym}}^j - \|x_i^j - x_k^j\|_2 \right| + \left| \|x_i^j - x_k^j\|_2 - \|x_i^m - x_k^m\|_2 \right| + \left| \|x_i^m - x_k^m\|_2 - \text{Dist}_{\text{sym}}^m \right|. \end{aligned}$$

By the assumption of bounded symbolic error, we have:

$$\left| \text{Dist}_{\text{sym}}^j - \|x_i^j - x_k^j\|_2 \right| \leq \epsilon_j, \quad \left| \text{Dist}_{\text{sym}}^m - \|x_i^m - x_k^m\|_2 \right| \leq \epsilon_m.$$

Substituting these bounds, we obtain:

$$\left| \text{Dist}_{\text{sym}}^j - \text{Dist}_{\text{sym}}^m \right| \leq \left| \|x_i^j - x_k^j\|_2 - \|x_i^m - x_k^m\|_2 \right| + \epsilon_j + \epsilon_m.$$

B Additional Details for Sparse Multihead Attention

We adopt the sparsity measurement proposed by Zhou et al. [63] to efficiently identify dominant queries without computing all query-key pairs. For each query vector $\mathbf{q} \in Q$, we compute its sparsity score $P(\mathbf{q}, K')$ over a sampled subset of keys $K' \subset K$ where $|K'| = u \log L_K$:

$$P(\mathbf{q}, K') = \max_{\mathbf{k} \in K'} \left(\frac{\mathbf{q}\mathbf{k}^\top}{\sqrt{d}} \right) - \frac{1}{|K'|} \sum_{\mathbf{k} \in K'} \frac{\mathbf{q}\mathbf{k}^\top}{\sqrt{d}}. \quad (2)$$

This max-mean measurement evaluates the query’s attention diversity by comparing its maximum alignment with the average alignment over the key subset. Queries with higher $P(\mathbf{q}, K')$ scores contain more distinctive information and are prioritized in our sparse attention mechanism.

The top- v queries with $v = u \log L_Q$ highest scores are selected for full attention computation, reducing the complexity from $O(L_Q L_K)$ to $O(u \log L_Q \cdot u \log L_K)$. This adaptive selection enables efficient processing while preserving the most informative query-key interactions.

C Training and Optimization Details

All experiments are performed on an Ubuntu OS server equipped with NVIDIA TITAN RTX GPU cards using PyTorch framework. Every experiment is carried out with 3 different seeds (2711, 2712, 2713). During model training, we use Adam optimizer with a learning rate from 1e-5 to 1e-3 and maximum number of epochs is set to 150 based on the suitability of each setting. We tune these optimization-related hyperparameters for each setting and save the best model checkpoint based on early exit based on the minimum value of the loss function achieved on the validation set.

Modality Dropout Scheme based on Curriculum Learning. As described in Section 3.2.5, the modality dropout is illustrated in Figure 9.

Hyperparameters. The key hyperparameters in MAESTRO are listed in Table 5 and are kept fixed across all datasets in this paper. However, a more comprehensive hyperparameter search could potentially yield further improvements in performance.

D Experimental Setup Details

D.1 Performance Metrics

Accuracy. Given N samples, Accuracy is defined as the proportion of correct predictions:

$$\text{Accuracy} = \frac{1}{N} \sum_{i=1}^N \mathbb{I}(\hat{y}_i = y_i)$$

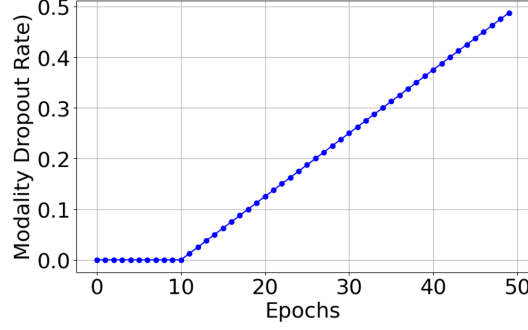


Figure 9: Modality Dropout Scheme based on Curriculum Learning (supporting illustration for Section 3.2.5).

Table 5: Model design components and hyperparameters.

Design Component	Hyperparameter	Value
Symbolic Tokenization	α – Number of alphabets	20
	W – Compression factor	2
Max Attention Budget	β	5
Sparse MoE	Ω – Number of experts	4
	k – Selected experts per token	1

975 **Macro-F1.** Let \mathcal{C} be the set of classes. For each class $c \in \mathcal{C}$, we compute precision P_c and recall
976 R_c :

$$F1_c = \frac{2P_cR_c}{P_c + R_c}, \quad \text{Macro-F1} = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} F1_c$$

977 **Relative Improvement.** We report relative measurements for all metrics as follows. Let $\text{Metric}_{\text{base}}$
978 be the baseline performance and $\text{Metric}_{\text{ours}}$ be our model’s performance. The relative improvement is:

$$\text{Relative Improvement} = \frac{\text{Metric}_{\text{ours}} - \text{Metric}_{\text{base}}}{\text{Metric}_{\text{base}}} \times 100\%$$

979 **Absolute Improvement.** We report absolute improvement in the case of accuracy-based metrics as
980 the direct difference between our model’s performance and the baseline performance, i.e.:

$$\text{Absolute Improvement} = \text{Metric}_{\text{ours}} - \text{Metric}_{\text{base}}$$

981 D.2 Dataset Details

982 In this section, we present detailed information about the datasets used for evaluation, including class
983 distributions, overall statistics, and preprocessing steps.

984 **WESAD.** We directly leverage the synchronized data from [41] for the WESAD dataset. The
985 modalities and their corresponding sampling rates are summarized in Table 6, and the class distribution
986 is shown in Figure 10.

987 **DaliaHAR.** We adapt the DaLiA dataset for activity recognition using multimodal sensor data,
988 addressing the scarcity of datasets that offer both diverse modalities and fine-grained activity labels.
989 Figure 11 illustrates the raw distribution of activity labels from a single subject recording. To
990 preprocess the data, we first segment the continuous recordings based on absolute timestamps
991 provided in the annotation files. Non-informative segments such as baseline and no-activity periods
992 are excluded. We then apply a sliding window approach with a window size of 8 seconds and a
993 2-second overlap. Since all sensor streams are temporally aligned, each modality is segmented
994 consistently, and each window is assigned the corresponding activity label for that subject.

995 The resulting processed data summary is given in Table 7 and the class distribution is shown in
996 Figure 12.

Table 6: WESAD dataset modality details.

Modality	Sampling Rate (Hz)	Variates
chest_ACC	700	3
chest_ECG	700	1
chest_EMG	700	1
chest_RESP	700	1
chest_EDA	700	1
chest_TEMP	700	1
wrist_ACC	32	3
wrist_BVP	64	1
wrist_EDA	4	1
wrist_TEMP	4	1
<i>Output Classes: Baseline, Stress, Amusement</i>		

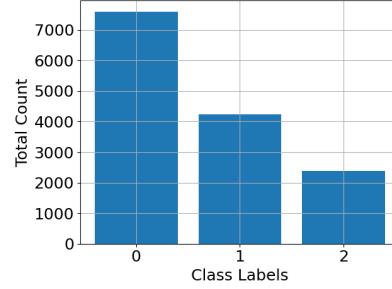


Figure 10: Distribution of the classes for WESAD dataset.

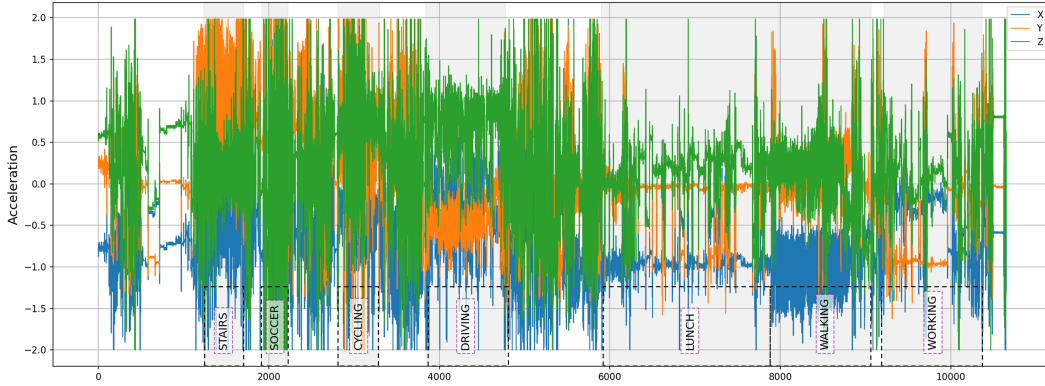


Figure 11: Visualizing the raw accelerometer data from wrist for the Dalia dataset.

Modality	Sampling Rate (Hz)	Variates
chest_ACC	700	3
wrist_ACC	32	3
wrist_BVP	64	1
wrist_EDA	4	1
wrist_TEMP	4	1
<i>Output Classes: STAIRS, SOCCER, CYCLING, DRIVING, LUNCH, WALKING, WORKING</i>		

Table 7: DaliaHAR dataset modality details.

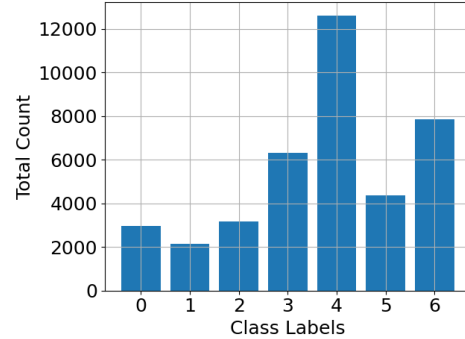


Figure 12: Distribution of activity classes in the DaliaHAR dataset.

997 **DSADS.** For the DSADS dataset, we follow the original preprocessing steps and report the cor-
 998 responding sensor modalities along with their specifications in Table 8. The class distribution is
 999 presented in Figure 13.

1000 **MIMIC.** We adopt the preprocessing for the MIMIC dataset as defined by the multimodal bench-
 1001 marking suite, MultiBench [25], to comply with standardized benchmarking practices. Overall, our
 1002 performance—shown in Table 2 in Section 4.1 of the main paper—aligns with the results reported in
 1003 the benchmarking suite. However, we observe a clear class imbalance, as illustrated in Figure 14.
 1004 Therefore, we also report the Macro-F1 score, which is particularly more informative than accuracy
 1005 for the MIMIC dataset.

Modality	Sampling Rate (Hz)	Variates
torso	25	9
right_arm	25	9
left_arm	25	9
right_leg	25	9
left_leg	25	9
<i>Output Classes: Sitting, Standing, Lying on back, Lying on right side, Ascending stairs, Descending stairs, Standing in an elevator (still), Moving around in an elevator, Walking in a parking lot, Walking on treadmill (flat), Walking on treadmill (inclined), Running on treadmill, Exercising on a stepper, Exercising on a cross trainer, Cycling on exercise bike (horizontal), Cycling on exercise bike (vertical), Rowing, Jumping, Playing basketball</i>		

Table 8: DSADS dataset modality details.

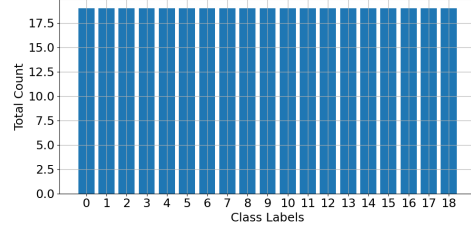


Figure 13: Class distribution of DSADS activity labels.

Modality	Sampling Rate (Hz)	Variates
glasgow	1	1
BP	1	1
HR	1	1
Temp	1	1
oxy	1	1
urine	1	1
urea	1	1
wbc	1	1
bdc2	1	1
Na	1	1
K	1	1
Bil	1	1
Age	1	1
icd9	1	1
hem_mal	1	1
cancer	1	1
adm_type	1	1
<i>Output Classes: 6 ICD-9 diagnostic categories (coarse-grained)</i>		

Table 9: MIMIC-III dataset modality details.

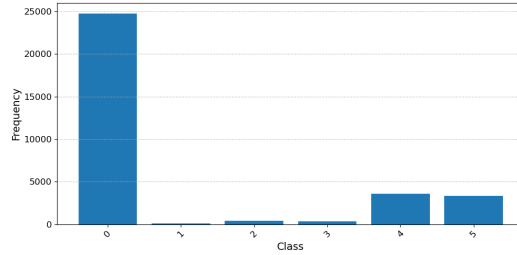


Figure 14: Distribution of ICD-9 class labels in the MIMIC-III dataset.

1006 D.3 Baseline Implementation Details

1007 **Multivariate Baselines.** We follow the MTS-Bakeoff [37] implementation for InceptionTime and
1008 ResNet1D. For the Transformer, we adopt the standard implementation with 8 heads and 2 layers,
1009 including positional encoding. In the case of the missingness analysis in Figure 5 of the main paper,
1010 we adapt the Transformer by retraining it using the same modality dropout scheme as employed
1011 in MAESTRO, to enable a fair comparison with a multivariate baseline that is not natively robust to
1012 missing data. This modification allows the adapted Transformer to serve as a competitive baseline, as
1013 evidenced by the performance trends shown in Figure 5.

1014 **Multimodal Baselines.** For the LRTF and MULT baselines, we implement them following the
1015 MultiBench [25] framework. For the remaining baselines, we use their original implementations and
1016 open-source resources for reproduction.

1017 In the robustness study in Section 4.1 in the main paper, we include only those baselines that
1018 natively support missingness, along with the adapted Transformer. We train these models using their
1019 original settings and evaluate them on samples with dynamically missing modalities. Specifically,
1020 we randomly drop out modalities with increasing severity, ranging from 10% to 40% and report the
1021 performance.

1022 E More Detailed Results of the Main Paper Experiments

1023 This section includes more detailed results from those experiments discussed in Section 4 of the main
1024 paper.

1025 E.1 Robustness Results

1026 The primary results using the Macro-F1 performance metric for varying levels of missingness across
 1027 all datasets are presented in Figure 5 in Section 4.1 of the main paper. Tables 10, 11, 8, and 13
 1028 provide the complete statistics for accuracy and F1 scores for all datasets—WESAD, DaliaHAR,
 1029 DSADS, and MIMIC-III, respectively.

Table 10: Accuracy and F1-score (mean_{std}) across different missingness levels for the WESAD dataset (supporting results for Figure 5 in Section 4.1 in the main paper).

Model	0%		10%		20%		30%		40%	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
Transformer	0.67 _{0.04}	0.53 _{0.06}	0.63 _{0.09}	0.45 _{0.06}	0.62 _{0.00}	0.49 _{0.05}	0.61 _{0.04}	0.48 _{0.09}	0.61 _{0.08}	0.46 _{0.01}
FlexMoE	0.71 _{0.09}	0.64 _{0.11}	0.68 _{0.08}	0.61 _{0.08}	0.65 _{0.06}	0.58 _{0.09}	0.62 _{0.08}	0.56 _{0.08}	0.59 _{0.07}	0.52 _{0.07}
FuseMoE	0.48 _{0.13}	0.40 _{0.06}	0.46 _{0.01}	0.41 _{0.03}	0.45 _{0.01}	0.40 _{0.03}	0.41 _{0.03}	0.38 _{0.04}	0.39 _{0.03}	0.37 _{0.02}
ShaSpec	0.65 _{0.56}	0.54 _{0.44}	0.55 _{0.51}	0.49 _{0.44}	0.53 _{0.50}	0.47 _{0.44}	0.48 _{0.46}	0.44 _{0.41}	0.43 _{0.42}	0.41 _{0.39}
Ours(w/o SAX)	0.69 _{0.13}	0.55 _{0.06}	0.68 _{0.11}	0.54 _{0.05}	0.67 _{0.09}	0.53 _{0.04}	0.65 _{0.08}	0.52 _{0.04}	0.66 _{0.08}	0.53 _{0.06}
Ours	0.77 _{0.07}	0.66 _{0.03}	0.77 _{0.06}	0.68 _{0.04}	0.74 _{0.07}	0.54 _{0.04}	0.74 _{0.07}	0.64 _{0.06}	0.71 _{0.07}	0.61 _{0.06}

Table 11: Accuracy and F1-score (mean_{std}) across different missingness levels for the DaliaHAR dataset (supporting results for Figure 5 in Section 4.1 in the main paper).

Model	0%		10%		20%		30%		40%	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
Transformer	0.72 _{0.04}	0.70 _{0.08}	0.69 _{0.04}	0.67 _{0.07}	0.67 _{0.04}	0.65 _{0.07}	0.64 _{0.06}	0.62 _{0.08}	0.61 _{0.06}	0.59 _{0.08}
FlexMoE	0.70 _{0.06}	0.70 _{0.05}	0.63 _{0.04}	0.61 _{0.05}	0.53 _{0.04}	0.51 _{0.06}	0.45 _{0.01}	0.43 _{0.02}	0.37 _{0.03}	0.35 _{0.03}
FuseMoE	0.78 _{0.01}	0.79 _{0.03}	0.67 _{0.01}	0.68 _{0.03}	0.58 _{0.01}	0.58 _{0.03}	0.48 _{0.03}	0.46 _{0.03}	0.40 _{0.04}	0.36 _{0.03}
ShaSpec	0.74 _{0.00}	0.77 _{0.02}	0.64 _{0.02}	0.67 _{0.03}	0.55 _{0.02}	0.57 _{0.02}	0.48 _{0.02}	0.47 _{0.01}	0.44 _{0.02}	0.39 _{0.01}
Ours(w/o SAX)	0.82 _{0.03}	0.84 _{0.03}	0.80 _{0.05}	0.83 _{0.02}	0.77 _{0.04}	0.79 _{0.03}	0.73 _{0.03}	0.76 _{0.02}	0.71 _{0.02}	0.73 _{0.01}
Ours	0.83 _{0.01}	0.84 _{0.01}	0.83 _{0.04}	0.85 _{0.03}	0.81 _{0.03}	0.82 _{0.03}	0.78 _{0.03}	0.79 _{0.03}	0.74 _{0.03}	0.75 _{0.03}

Table 12: Accuracy and F1-score (mean_{std}) across different missingness levels for the DSADS dataset (supporting results for Figure 5 in Section 4.1 in the main paper).

Model	0%		10%		20%		30%		50%	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
Transformer	0.88 _{0.04}	0.88 _{0.04}	0.78 _{0.13}	0.77 _{0.15}	0.74 _{0.16}	0.73 _{0.14}	0.71 _{0.13}	0.70 _{0.15}	0.66 _{0.10}	0.68 _{0.09}
FlexMoE	0.67 _{0.01}	0.63 _{0.03}	0.46 _{0.05}	0.44 _{0.05}	0.35 _{0.09}	0.34 _{0.08}	0.25 _{0.00}	0.23 _{0.01}	0.19 _{0.06}	0.17 _{0.07}
FuseMoE	0.84 _{0.03}	0.85 _{0.03}	0.59 _{0.00}	0.61 _{0.01}	0.46 _{0.04}	0.46 _{0.04}	0.43 _{0.02}	0.44 _{0.03}	0.37 _{0.03}	0.38 _{0.03}
ShaSpec	0.81 _{0.01}	0.79 _{0.00}	0.58 _{0.03}	0.58 _{0.03}	0.45 _{0.02}	0.45 _{0.02}	0.29 _{0.01}	0.27 _{0.02}	0.22 _{0.01}	0.20 _{0.02}
Ours(w/o SAX)	0.78 _{0.02}	0.77 _{0.02}	0.79 _{0.00}	0.77 _{0.01}	0.76 _{0.01}	0.75 _{0.01}	0.72 _{0.01}	0.71 _{0.01}	0.66 _{0.05}	0.65 _{0.04}
Ours	0.89 _{0.01}	0.88 _{0.01}	0.86 _{0.01}	0.86 _{0.01}	0.84 _{0.00}	0.84 _{0.01}	0.83 _{0.01}	0.82 _{0.01}	0.79 _{0.02}	0.79 _{0.02}

Table 13: Accuracy and F1-score (mean_{std}) across different missingness levels in MIMIC dataset (Supporting results for Figure 5 in Section 4.1 in the main paper).

Model	0%		10%		20%		30%		40%	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
Transformer	0.78 _{0.02}	0.22 _{0.01}	0.78 _{0.04}	0.21 _{0.03}	0.78 _{0.01}	0.22 _{0.02}	0.77 _{0.01}	0.22 _{0.01}	0.77 _{0.04}	0.21 _{0.03}
FlexMoE	0.79 _{0.02}	0.27 _{0.03}	0.78 _{0.02}	0.25 _{0.02}	0.79 _{0.01}	0.25 _{0.01}	0.78 _{0.01}	0.24 _{0.02}	0.78 _{0.01}	0.21 _{0.01}
FuseMoE	0.76 _{0.01}	0.26 _{0.00}	0.75 _{0.01}	0.27 _{0.01}	0.73 _{0.02}	0.23 _{0.00}	0.72 _{0.03}	0.23 _{0.01}	0.72 _{0.02}	0.22 _{0.01}
ShaSpec	0.76 _{0.01}	0.25 _{0.01}	0.74 _{0.03}	0.25 _{0.01}	0.74 _{0.01}	0.23 _{0.00}	0.74 _{0.00}	0.20 _{0.02}	0.70 _{0.05}	0.21 _{0.00}
Ours(w/o SAX)	0.80 _{0.01}	0.31 _{0.01}	0.80 _{0.02}	0.31 _{0.02}	0.77 _{0.01}	0.28 _{0.01}	0.79 _{0.02}	0.27 _{0.01}	0.77 _{0.02}	0.26 _{0.02}
Ours	0.79 _{0.01}	0.30 _{0.02}	0.78 _{0.02}	0.30 _{0.03}	0.76 _{0.01}	0.29 _{0.03}	0.78 _{0.01}	0.27 _{0.01}	0.77 _{0.02}	0.25 _{0.02}

1030 E.2 Ablation Results

1031 The complete statistics of the Figure 6 in Section 4.2 of the main paper is given in Table 14.

1032 F Additional Experiments

1033 In addition to the experiments shown in the main paper and Section E, we also conducted additional
 1034 experiments to further evaluate our approach, as shown below.

Table 14: Accuracy (mean_{std}) on the WESAD dataset under full data and 40% missingness conditions for an ablation analysis (supporting results for Figure 6 in Section 4.2 of the main paper).

Method	Full	40% Missing
Ours	0.78 _{0.07}	0.71 _{0.06}
w/o Symbolic Transformation	0.69 _{0.13}	0.66 _{0.07}
w/o Modality Embedding	0.70 _{0.10}	0.55 _{0.04}
w/o Modality Dropout	0.76 _{0.09}	0.62 _{0.07}
w/o Adaptive Attn Budget	0.71 _{0.11}	0.67 _{0.08}
w/o Cross-modal SparseMoE	0.70 _{0.08}	0.50 _{0.01}

F.1 Unimodal Sweep Results

This section presents the unimodal results for WESAD, DSADS, DALIAHAR, and MIMIC, as shown in Table 15. For unimodal training, we train individual models for each modality using the Transformer backbone. These results indicate that in some applications, the modalities contain redundant information—as seen in DSADS—where unimodal performance is not significantly lower than multimodal performance. In contrast, in datasets like Dalia, certain modalities perform close to random guessing (e.g., `wrist_TEMP` with 0.38 accuracy for a 3-class classification task). Since MIMIC shows overall lower performance, we report the top five modalities and additionally provide the F1-score to highlight the performance boost achieved by using MAESTRO (with an F1-score of 0.30), compared to unimodal models which typically achieve around 0.15 F1 in most cases.

Table 15: Unimodal performance across all datasets (supporting results for Section 4.2).

Dataset	Modality	ACC	STDEV
DSADS	Torso	0.63	0.01
	Right Arm	0.74	0.02
	Left Arm	0.85	0.03
	Right Leg	0.81	0.02
	Left Leg	0.83	0.01
Dalia	wrist_ACC	0.81	0.04
	wrist_BVP	0.50	0.04
	wrist_EDA	0.45	0.03
	wrist_TEMP	0.35	0.03
	chest_ACC	0.85	0.01
WESAD	chest_ACC	0.67	0.04
	chest_ECG	0.75	0.13
	chest_EMG	0.75	0.01
	chest_RESP	0.63	0.04
	chest_EDA	0.60	0.04
	chest_TEMP	0.71	0.02
	wrist_ACC	0.66	0.12
	wrist_BVP	0.71	0.08
	wrist_EDA	0.60	0.04
	wrist_TEMP	0.38	0.12
MIMIC	glasgow	0.77	0.01
	BP	0.72	0.08
	HR	0.72	0.07
	Temp	0.77	0.01
	oxy	0.76	0.01

F.2 Supporting Results for Case Study on DaliaHAR from Section 4.2

Based on a known *a priori*, we evaluate three scenarios: 1) using the two best modalities—`wrist_ACC` and `chest_ACC`—we train a bimodal model, 2) we run inference on the adapted Transformer using

Table 16: F1 score and standard deviation for 5 best MIMIC modalities.

Modality	F1 Score	STD
glasgow	0.17	0.02
BP	0.17	0.02
HR	0.16	0.02
Temp	0.15	0.01
oxy	0.16	0.01

only these two modalities while dropping the rest, and 3) we evaluate MAESTRO in the presence of only these two modalities.

Our results, shown in Table 17, highlight that MAESTRO with *a priori* knowledge at run-time performs competitively (with an absolute improvement of 3%) compared to an end-to-end model trained on these pre-selected modalities. This demonstrates that MAESTRO is capable of effectively modeling multimodal interactions and learning task-dependent semantics from multimodal data.

Table 17: Accuracy with *a priori* modality mask at inference: [wrist_ACC, wrist_BVP, wrist_EDA, wrist_TEMP, chest_ACC] = [1, 0, 0, 0, 1] (supporting results for Section 4.2).

Method	Accuracy
Bimodal (wrist + chest)	0.82 \pm 0.02
Transformer (adapted)	0.71 \pm 0.01
MAESTRO	0.85 \pm 0.01

F.3 Visualization of Unimodal Representations for DaliaHAR

To further support the results in Table 4, we plot the t-SNE projections of the unimodal models' latent representations for each modality in DaliaHAR, as shown in Figure 15.

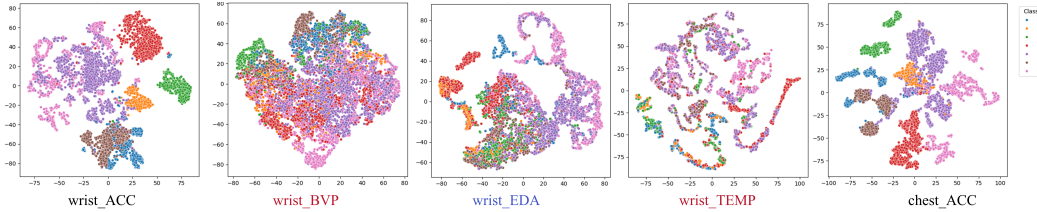


Figure 15: t-SNE projections of the unimodal models' latent representations for each modality in DaliaHAR (supporting results for Table 4 in Section 4.2 in the main paper).

G Broader Impacts

MAESTRO paves the way for more efficient and practical handling of heterogeneous sensing data. It has the potential to enhance analytics and, in turn, the performance of ubiquitous sensing applications across diverse domains—including smart home monitoring, daily living assistance, fitness and wellness interventions, elderly care, healthcare, and environmental monitoring. These applications rely on rich, continuous streams of sensory data, where sensor reliability often cannot be guaranteed. For example, in environmental monitoring within remote or inaccessible locations, sensors may fail due to power loss or harsh conditions. In such scenarios, maintaining robust performance with only a subset of available modalities is critical.

Currently, MAESTRO addresses complete modality-level missingness, demonstrating its ability to sustain model effectiveness even under challenging sensing conditions. In future, we aim to explore more advanced symbolic encoding strategies and extend the framework to address irregularly sampled and asynchronous sensing modalities and