SOURCE-LEARNED RELIANCE FOR SELECTIVE TEST-TIME ADAPTATION OF MULTIMODAL TIME SERIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal wearable systems must remain reliable when sensor streams become noisy or unavailable. Existing multimodal test-time adaptation (TTA) methods often assess reliability online, but cross-modal agreement can be misleading when sensors measure different physical processes, and evaluating alternative modality configurations adds inference cost. We propose CARAT, which decouples model reliance from runtime corruption detection to guide omission or attenuation, amortizing reliance estimation through source training. An asymmetric modality-dropout curriculum prepares a missingness-resilient backbone for omission and derives a frozen, backbone-specific reliance proxy from windowed input-projection gradient norms. At deployment, a lightweight one-class detector flags suspect streams, and the proxy guides a joint choice between replacing the suspect set with the backbone's trained missingness symbol and attenuating its representations before fusion, without candidate-subset evaluation. Across four wearable datasets, five corruption types, three backbones, and eight TTA baselines, CARAT achieves the highest overall macro-F1 and best mean rank (2.42), exceeding EATA, the strongest baseline, by 1.58 F1 points across 12 equally weighted dataset–backbone settings. Across five profiled configurations, CARAT uses 9.49% fewer GFLOPs and updates 47.82% fewer parameters than EATA. A pattern also emerges across sensing regimes: multimodal TTA methods such as PTA are competitive on IMU-dominated homogeneous datasets, whereas unimodal TTA methods like TENT and EATA match or exceed it on heterogeneous datasets. These results position CARAT as a practical default to wearable TTA, offering competitive robustness with modest computational requirements and benefits that vary across backbones and dataset regimes.

1 INTRODUCTION

Multimodal wearable systems combine several sensor streams to support applications such as human activity recognition and stress monitoring (Altun et al., 2010; Reiss et al., 2019; Schmidt et al., 2018). These applications can involve numerous motion and physiological sensing streams, each with distinct physical characteristics and sources of noise. At deployment, missingness, sensor failure, motion artifacts, drift, and nonstationary noise can affect individual streams or several streams simultaneously (Zhang et al., 2014; Sharma et al., 2007). Maintaining reliable predictions therefore requires handling changes in both the availability and quality of these inputs.

Test-time adaptation (TTA) addresses deployment shifts by updating pretrained models using unlabeled test data (Wang et al., 2020a; Niu et al., 2022), with multimodal extensions targeting reliability imbalance and selective adaptation (Yang et al., 2024; Guo & Jin, 2025; Chen et al., 2025). These methods are evaluated on two or three modalities of one kind, such as audio, video, and optical flow (Yang et al., 2024; Dong et al., 2025), and decide which stream is unreliable either by comparing streams against each other, which fails when sensors measure different physical processes (an accelerometer and a skin-temperature sensor disagree by design), or by trying alternative modality configurations at inference, which costs extra forward passes. Wearable systems carry many streams of different kinds (ten in WESAD (Schmidt et al., 2018)), which makes both problems worse.

Missingness-resilient multimodal backbones offer a complementary opportunity (Wang et al., 2023; Yun et al., 2024; Mohapatra et al., 2025). Because they are trained to predict from incomplete inputs,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

their missing-input operation can be reused to deliberately omit suspect streams during adaptation. **However, just supporting missing inputs does not establish when omission is beneficial.** Removing a corrupted stream may suppress harmful inputs, but it can also discard useful information that the remaining modalities cannot replace. Choosing between omission and attenuation (keeping a stream but downweighting it before fusion) therefore requires considering the backbone's reliance on both the suspect and remaining streams, that is, how much its predictions depend on each.

The next challenge is to make this choice without the costly deployment-time probing. With K suspect streams, unrestricted retain/remove choices admit 2^K subsets; assessing these alternatives requires repeated predictions, and even leave-one-modality-out tests add inference work. **Source training can provide an opportunity to estimate modality reliance before deployment.** A source-derived reliance proxy can summarize how the backbone learned to use its modalities, while a detector assesses current input abnormality. Combining these signals offers a way to guide omission and attenuation without evaluating candidate subsets online.

Our Approach. We propose **CARAT** (**C**onditional modality composition **a**nd **r**eliance-guided **at**tenuation for efficient test-time adaptation). The main idea is: during source training, the backbone learns to predict with missing streams and, as a by-product, how much it relies on each one. At test time, a detector flags suspect streams; if the backbone relies mostly on the healthy streams, the suspect ones are dropped, otherwise they are kept but turned down (Figure 1). An asymmetric modality-dropout curriculum uses windowed gradient norms at modality-specific input projections to prepare the backbone for missing inputs and derive a frozen, backbone-specific reliance proxy. For each sample, the share of the reliance profile assigned to unflagged modalities determines whether to replace the entire suspect set with the trained missingness symbol or attenuate its encoder outputs through learnable scalars before fusion. Only normalization and attenuation parameters are updated using a sample-weighted entropy objective. The key contributions are:

- **Reliance estimated at training time, corruption detected at test time.** CARAT estimates a backbone-specific reliance prior during source training and reuses it alongside runtime abnormality flags to guide omission or attenuation. This amortizes reliance estimation across deployment samples, avoiding explicit cross-modal agreement estimation and candidate-subset evaluation at test time (Section 3.2.2).
- **An asymmetric curriculum for missingness robustness and reliance estimation.** We modulate modality-dropout probabilities using windowed input-projection gradient norms, preparing the backbone for omission while deriving a frozen reliance proxy for adaptation. Section 4.2 tests the curriculum's robustness benefit and the proxy's usefulness.
- **A multimodal time-series TTA benchmark with robustness and efficiency insights.** CARAT achieves the best mean rank (2.42) across 12 dataset–backbone settings. The benchmark reveals a pattern across sensing regimes: PTA is competitive on IMU-dominated homogeneous datasets, whereas TENT and EATA match or exceed it on heterogeneous datasets. Across five profiled configurations, CARAT uses 9.49% fewer GFLOPs and updates 47.82% fewer parameters than EATA, with substantial latency and CPU energy savings over AeO and PTA. Design ablations further examine the roles of source training, reliance-guided routing, and corruption detection (Sections 4.1 and 4.2).

2 Related Work

Multimodal test-time adaptation. TTA updates pretrained models using unlabeled deployment data (Wang et al., 2020a; Niu et al., 2022; Gong et al., 2023). READ (Yang et al., 2024) combines self-adaptive attention with a confidence-aware objective; SuMi (Guo & Jin, 2025) uses sample selection, unimodal assistance, and information sharing; PTA (Wang et al.) combines prediction-bias reweighting with attention alignment; and AEO (Dong et al., 2025) optimizes entropy and modality-prediction discrepancies for open-set adaptation. Closest to our intervention design, TSA (Chen et al., 2025) learns modality-specific adapters and a router at test time for single-modality shifts, evaluating alternative adapter configurations at inference. Recently proposed, DASP (He et al., 2026) uses test-batch feature redundancy to identify shifted modalities and selectively update stable or plastic adapters. CARAT combines a frozen source-derived reliance proxy with current abnormality flags to choose jointly between removing the suspect set and attenuating its representations before fusion, without candidate-subset predictions during routing.

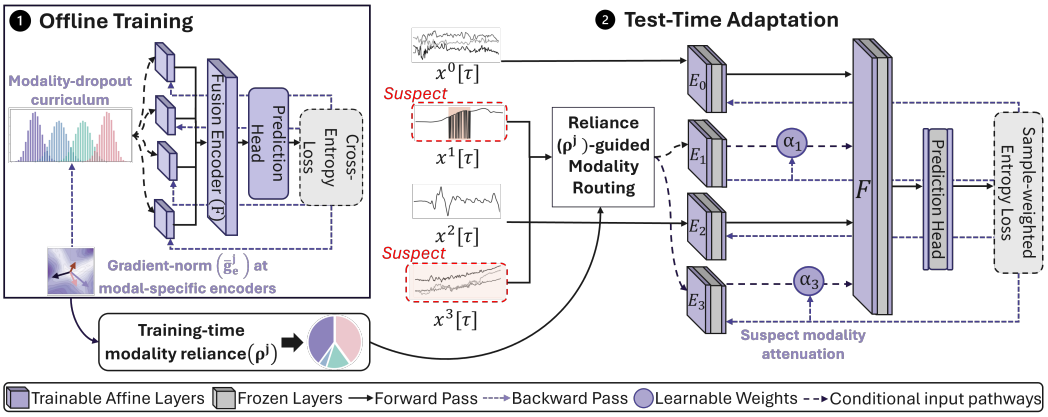


Figure 1: **Overview of CARAT.** During offline training, an asymmetric modality-dropout curriculum adjusts per-modality dropout rates using windowed gradient norms $\bar{g}_e^j$, yielding a training-time reliance profile ρ^j (Section 3.2.1). At test time, a lightweight detector identifies suspect modalities (red; here, x^1 and x^3), and the frozen reliance profile selects one intervention for the entire suspect set (Section 3.2.2): *modality composition* replaces that set with the missingness symbol s_0 when the unflagged modalities retain sufficient reliance mass; otherwise, *latent attenuation* downweights the suspect representations through learnable scalars a^j before fusion. Adaptation updates only normalization and attenuation parameters using a sample-weighted entropy objective.

Source training for missingness and modality imbalance. Missingness-resilient backbones such as ShaSpec (Wang et al., 2023), Flex-MoE (Yun et al., 2024), and MAESTRO (Mohapatra et al., 2025) support prediction with incomplete inputs, providing trained mechanisms for modality omission. Meanwhile, joint training can favor dominant modalities and under-optimize others (Wang et al., 2020b; Peng et al., 2022). OGM (Peng et al., 2022) modulates gradients according to relative modality contributions, while OPM (Wei et al., 2024) dynamically drops modality features based on discriminative-performance discrepancies. MIAM (Zbinden et al., 2026) adjusts token-masking distributions using modality-specific performance and learning dynamics, supporting incomplete inputs within and across modalities. OPM and MIAM thus already connect adaptive masking with missing-input robustness. CARAT builds on these capabilities to address a further decision: whether a currently abnormal stream can be discarded. Its windowed gradient norms guide missingness exposure and produce a frozen, backbone-specific reliance proxy, which combines with deployment-time abnormality flags to select joint removal or pre-fusion attenuation of the suspect set.

Time-series shift detection and adaptation. Time-series methods detect anomalies through reconstruction-based scoring (Xu et al., 2023; Sun et al.) or association discrepancy (Xu et al., 2021), and adapt through spectral alignment and temporal normalization (Gnassounou et al., 2024; 2025). Forecasting methods address nonstationarity through test-time calibration or shift-aware updates (Kim et al., 2025; Grover & Etemad, 2025). CARAT combines reconstruction error and power spectral density distance relative to clean source statistics to flag suspect streams, complementing its backbone-dependence proxy with observation-level abnormality information. Detector outputs enter routing as per-modality flags, permitting alternative detectors; Section 4.2 examines sensitivity to imperfect flags.

3 Methodology

3.1 Problem Formulation

We denote each multimodal time-series sample as $\mathcal{M}_i = \left(\{x_i^1[\tau]\}_{\tau=1}^{T^1}, \ldots, \{x_i^M[\tau]\}_{\tau=1}^{T^M}\right)$, where $x_i^j[\tau] \in \mathbb{R}^{D^j}$ is the feature vector at temporal index τ for modality j, and each modality forms a multivariate time series $x_i^j \in \mathbb{R}^{D^j \times T^j}$. Let f_θ denote a multimodal predictor trained on labeled source data $\mathcal{D}_s = \{(\mathcal{M}_i, y_i)\}_{i=1}^N \sim p_s(\mathcal{M}, y)$. At test time, the model receives a sequence of unlabeled batches $\{\mathcal{B}_t\}_{t\geq 1}$ in an episodic online adaptation setting. We study corruption-induced input distribution shifts generated by unknown processes $\mathcal{T}_t^j$ acting independently or jointly on sensor observations. These processes introduce additive noise, structured sensor degradation, or missingness,

changing the distribution of observed inputs relative to source data while leaving each sample's underlying task label unchanged. Multiple modalities may be corrupted simultaneously; the affected subset and corruption severity are not supplied to the adaptation method. We consider architectures of the form

$$f_\theta(\mathcal{M}_i) = h\big(F(E_1(x_i^1), \ldots, E_M(x_i^M))\big)\,,$$

where E_j denotes a modality-specific encoder, F a multimodal fusion module, and h the prediction head. The encoders and fusion module may incorporate transformer layers, gating, routing, or mixture-of-experts mechanisms. Our objective is to control modality participation and update a small subset of model parameters using only unlabeled test batches $\{\mathcal{B}_t\}$, without access to raw source samples during adaptation, to improve robustness under multimodal corruption.

3.2 Our Approach

We propose CARAT (**C**onditional modality composition **a**nd **r**eliance-guided **at**tenuation for efficient test-time adaptation), a multimodal test-time adaptation framework that combines current input abnormality with source-learned modality reliance. A missingness-resilient backbone supplies a trained operation for omitting inputs; the reliance profile helps decide when to use it. CARAT routes each sample between removal and attenuation without evaluating candidate modality subsets. During training, we replace uniform modality dropout with an asymmetric curriculum that modulates each modality's dropout rate using its windowed input-projection gradient norm, yielding a reliance profile $\{\rho^j\}_{j=1}^M$ at convergence (Section 3.2.1). At test time, a lightweight per-modality one-class detector identifies suspect inputs (Section 3.2.2), and the learnt reliance prior determines the intervention. If the unflagged modalities retain sufficient reliance mass, all suspect inputs to E_j are replaced with the missingness symbol; otherwise, the suspect modalities are propagated but attenuated prior to fusion in F. Adaptation proceeds by updating only normalization and attenuation parameters using a sample-weighted entropy objective over unlabeled test batches. An overview is shown in Figure 1.

3.2.1 Modality Reliance Estimation during Training

For the evaluated missingness-resilient backbones (Wang et al., 2023; Yun et al., 2024; Mohapatra et al., 2025), we start from a shared modality-dropout curriculum, exposing the model to easy, complete samples first and progressing to harder, incomplete samples. After a warm-up phase, existing schedules apply a scalar dropout rate $p_{\text{base}}(e)$ uniformly across all modalities at epoch e (full schedule in Appendix C). We introduce gradient-guided modulation, drawing on prior observations of imbalanced multimodal learning (Peng et al., 2022; Wang et al., 2020b): modalities with larger windowed gradient norms receive higher dropout probabilities, exposing the backbone more often to their absence. As a by-product, the resulting per-modality dropout rate at convergence serves as a proxy for the model's learned reliance on each modality, which we use directly at test time in Section 3.2.2. We follow the symbolic aggregate approximation of Mohapatra et al. (2025) for tokenization, with a reserved symbol s_0 indicating missing modalities.

Concretely, we use $p_{\text{base}}(e)$ as the target average dropout rate before clipping and let training dynamics redistribute it across modalities. Let $\theta_j \subset \theta$ denote the parameters of modality j's input projection. We measure the gradient norm $g_k^j = \|\nabla_{\theta_j} \mathcal{L}_k(\theta)\|_2$ of the supervised loss $\mathcal{L}_k$ at step k as a local optimization signal for this modality-specific parameter block. At the start of each epoch, we form a per-modality reliance ratio by normalizing windowed gradient norms across modalities.

$$r_e^j = \frac{\bar{g}_e^j}{\sum_{\ell=1}^M \bar{g}_e^\ell}, \qquad \bar{g}_e^j = \frac{1}{W} \sum_{u=k_e-W}^{k_e-1} g_u^j, \tag{1}$$

with $r_e^j \in [0,1]$ and $\sum_j r_e^j = 1$. The Bernoulli dropout probability for modality j during epoch e is then, $p_e^j = \min\big(p_{\text{cap}},\ M \cdot p_{\text{base}}(e) \cdot r_e^j\big)$, held fixed throughout the epoch. The factor M preserves $p_{\text{base}}(e)$ as the mean Bernoulli rate before clipping. We use the normalized gradient magnitudes as

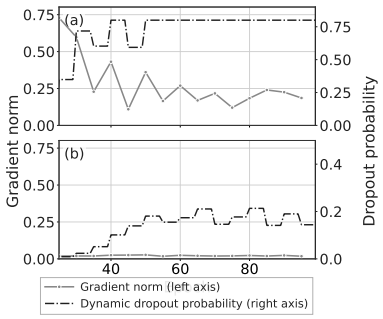


Figure 2: Per-modality gradient norm $\bar{g}_e^j$ and corresponding dropout probability p_e^j on activity recognition (DaliaHAR) dataset for (a) wrist accelerometer and (b) wrist temperature.

an optimization-based proxy for modality reliance. Figure 2 illustrates this on DaliaHAR (modalities: accelerometer, temperature, blood volume pulse, electrodermal activity): the gradient norm $\bar{g}_e^j$ is consistently larger for the accelerometer than for temperature, and the schedule translates this gap directly into a higher dropout probability for the accelerometer during training. We log the dropout probability at the trained model $\theta^\star$ as the model's *reliance profile* $\rho^j = p_{e^\star}^j$, which is a clipped, scaled version of the windowed gradient share recorded for that checkpoint. Source preparation therefore supplies both exposure to missing inputs and a frozen, backbone-specific prior for adaptation (Section 3.2.2); their empirical contributions are examined in Section 4.2 and Appendix E.

Optimization-based motivation. Gradient flow provides a local motivation for this signal (Lubana & Dick, 2020). For a differentiable loss under $\dot{\theta} = -\nabla_\theta \mathcal{L}$, the contribution of parameter block θ_j to $\dot{\mathcal{L}}$ is $-\|\nabla_{\theta_j}\mathcal{L}\|_2^2$. This identity motivates measuring gradient magnitude at modality-specific projections. Our windowed, clipped profile uses that signal as a reliance proxy; its connection to predictive dependence is assessed empirically rather than implied by the identity.

Figure 3 (left) connects the proxy to the intervention problem: removing the high-reliance accelerometers causes a larger F1 decline than removing low-reliance temperature on DaliaHAR–ShaSpec. This motivates retaining some information from suspect streams when the backbone relies strongly on them. Input removal reuses the missingness symbol seen during training; attenuation instead preserves a downweighted representation when removal may be costly. Figure 3 (right) shows that lower entropy accompanies higher F1 for CARAT in the illustrated evaluation. The prior perturbations in Section 4.2 directly test the utility of the learned profile, while Appendix A.1 details the gradient-flow interpretation and its limits.

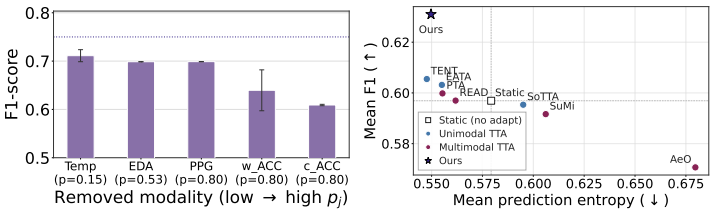


Figure 3: **Left:** Leave-one-modality-out F1 on DaliaHAR–ShaSpec, ordered by terminal dropout probability ρ^j. **Right:** Mean prediction entropy and F1 under Diverse corruption (10 scenarios, 5 corruption types).

3.2.2 Reliance-Guided Test-Time Adaptation

For each test sample $\mathcal{M}_i$, a per-modality detector $\mathcal{D}^j(\cdot)$ flags suspect modalities. The detector assesses current input abnormality; the source-derived profile guides the intervention. The frozen reliance profile $\{\rho^j\}$ then determines the intervention: if the surviving (unflagged) modalities account for a sufficient share of the reliance budget, the suspect inputs are replaced with the missingness symbol s_0 at the input to E_j (dropped); otherwise, they propagate but their pre-fusion contribution is dampened by a learnable attenuator before F.

Suspect modality detection. For each sample $\mathcal{M}_i$ and modality j, a lightweight one-class detector $\mathcal{D}^j(\cdot)$ computes a corruption score using reconstruction and spectral-distance features relative to compact clean-reference statistics computed during source training (Xu et al., 2023; Gnassounou et al., 2024; 2025). No raw source data is accessed during adaptation. Modalities whose score exceeds a fixed threshold are flagged as suspect, producing the routing set $\mathcal{S}_i$. Since CARAT consumes only this binary routing signal, the framework remains agnostic to the specific detector choice. The detectors are trained offline and reused at deployment. Appendix D details their formulation and sensitivity analyses, while Appendix L.1 reports their training cost, detector-only inference latency, memory footprint, and stored artifacts.

Reliance-guided routing. Given $\mathcal{S}_i$, the reliance profile $\{\rho^j\}_{j=1}^M$ guides the adaptation action on the suspect modalities. The share of the training-time reliance budget assigned to the unflagged modalities of sample i is

$$\beta_i = \frac{\sum_{j \notin \mathcal{S}_i} \rho^j}{\sum_{\ell=1}^{M} \rho^\ell} \in [0, 1]. \tag{2}$$

Sample i is routed to modality-composition adaptation when $\beta_i \geq \nu$, and to latent-attenuation adaptation otherwise. This selects one action for the entire suspect set using the relative reliance mass, without evaluating alternative subsets. Both ordering and relative magnitudes can affect β_i; Table 4 tests the sensitivity to reversing the ordering and to order-preserving perturbations.

Modality-composition adaptation. When $\beta_i \geq \nu$, the unflagged modalities meet the retained-reliance threshold, and we use the backbone's trained missing-input operation. We replace suspect modalities $j \in \mathcal{S}_i$ with the missingness symbol s_0 at the encoder input. The substitution is applied to each flagged modality and introduces no additional parameters; it reuses the missingness symbol employed by the modality-dropout curriculum (Section 3.2.1).

Suspect modality attenuation. When $\beta_i < \nu$, the prior assigns a larger share of reliance to the suspect set, motivating attenuation to retain potentially useful signal while limiting its contribution to fusion. We attenuate each $j \in \mathcal{S}_i$ using a learnable scalar $a^j = \sigma(\alpha^j) \in (0, 1)$ applied to its encoder output:

$$E'_j(\tilde{x}_i^j) = \begin{cases} a^j \cdot E_j(\tilde{x}_i^j), & j \in \mathcal{S}_i, \\ E_j(\tilde{x}_i^j), & \text{otherwise.} \end{cases} \tag{3}$$

The scalar a^j acts as a per-modality gain on the latent stream entering fusion. Since each α^j must adapt within only a few episodic steps, we use a substantially larger learning rate than for LayerNorm affine parameters ($\eta_\alpha \gg \eta_{\text{LN}}$). Appendix Figure 12 shows that small learning rates keep the gates near the sigmoid midpoint, while larger values enable meaningful adaptation within the same update budget.

Overall test-time optimization. A test batch $\mathcal{B}_t$ may contain both samples with flagged modalities ($\mathcal{S}_i \neq \emptyset$) and samples without flagged modalities ($\mathcal{S}_i = \emptyset$). Following Wang et al. (2020a), we minimize a weighted entropy objective:

$$\mathcal{L}_{\text{TTA}}(\theta; \mathcal{B}_t) = \frac{\sum_{i=1}^{B} w_i \, \mathcal{H}(p_\theta(y \mid \mathcal{M}_i))}{\sum_{i=1}^{B} w_i}, \qquad w_i = 1 + \lambda \mathbb{1}[\mathcal{S}_i = \emptyset]. \tag{4}$$

Here, $\mathcal{H}(\cdot)$ denotes prediction entropy and $\lambda \geq 0$ upweights samples without flagged modalities, to emphasize inputs judged less affected by the detector. We update only normalization parameters and attenuation logits α^j, keeping the source reliance profile fixed and resetting the adaptive parameters between test batches.

4 Results and Discussion

We evaluate whether source-learned reliance helps a missingness-resilient backbone decide when to remove suspect inputs and when to retain them through attenuation. The experiments test overall robustness and deployment cost (Section 4.1), then examine the contributions of source preparation, reliance ordering, detection, and online updates (Section 4.2).

Benchmark. We use four wearable sensing datasets: DSADS (Altun et al., 2010) (19-class activity recognition from five body-worn IMUs), PAMAP2 (Reiss & Stricker, 2012; de Frana, n.d.) (12-class activity recognition from three IMUs and heart rate), DaliaHAR (Reiss et al., 2019) (7-class activity recognition from wrist motion and physiological signals plus chest acceleration), and WESAD (Schmidt et al., 2018) (3-class stress classification from ten chest and wrist streams). The benchmark spans IMU-dominated and heterogeneous sensing configurations. We apply five synthetic sensor corruptions to selected modality subsets: discrete missingness, waveform anomalies, nonlinear degradation, frequency-selective attenuation, and nonstationary noise. The complete grids contain 50, 50, 45, and 95 conditions per backbone on DSADS, DaliaHAR, PAMAP2, and WESAD, respectively. Appendix B details the sensors and splits; Appendix A specifies corruption models and parameters, with examples in Figure 7.[1]

Backbones and comparators. We evaluate MAESTRO (Mohapatra et al., 2025), ShaSpec (Wang et al., 2023), and Flex-MoE (Yun et al., 2024), which differ in their encoders and fusion mechanisms but support missing inputs and modality-specific attenuation before fusion. Comparators comprise Static (no adaptation) and eight TTA methods: TENT (Wang et al., 2020a), EATA (Niu et al., 2022), and SoTTA (Gong et al., 2023); multimodal methods READ (Yang et al., 2024), SuMi (Guo & Jin, 2025), AeO (Dong et al., 2025), and PTA (Wang et al.); and time-series-specific PSDNorm (Gnassounou et al., 2025). In the main benchmark, all methods share the same source checkpoint trained

[1] A minimal implementation is included in the supplementary materials; we will release the corruption scenarios and baseline implementations.

Table 1: **Robustness across datasets and backbones.** Mean macro-F1 (↑, 0–1) over 50/50/45/95 modality-subset × corruption conditions per backbone on DSADS/DaliaHAR/PAMAP2/WESAD. Overall F1 equally weights the 12 settings; Rank (↓) averages ranks of their means before rounding, with average ranks for exact ties. **Bold** / underline: best / second-best displayed values; rounded ties are co-marked. MAE: MAESTRO; Sha: ShaSpec; Flex: Flex-MoE.

	DSADS			DaliaHAR			PAMAP2			WESAD			Overall	
Method	MAE	Sha	Flex	MAE	Sha	Flex	MAE	Sha	Flex	MAE	Sha	Flex	**F1 ↑**	**Rank ↓**
Static	0.69	0.44	0.53	0.74	0.60	0.64	**0.92**	0.70	0.75	**0.65**	0.53	0.53	0.64	5.42
TENT	0.69	0.42	0.54	**0.76**	0.61	**0.68**	**0.92**	0.69	**0.77**	**0.65**	0.53	**0.54**	0.65	3.92
EATA	0.69	0.42	0.54	**0.76**	0.60	**0.68**	**0.92**	0.69	**0.77**	**0.65**	0.53	**0.54**	0.65	3.50
READ	0.69	0.39	0.54	0.74	0.60	0.65	0.91	0.66	0.75	0.64	0.53	0.53	0.64	6.83
SuMi	0.68	0.43	0.54	0.74	0.59	0.64	0.91	0.70	0.76	0.64	0.53	0.51	0.64	6.58
SoTTA	0.69	0.43	0.54	0.75	0.60	0.66	**0.92**	0.69	0.76	**0.65**	0.53	0.52	0.64	4.92
AeO	0.69	0.44	0.53	0.72	0.57	0.61	0.91	0.69	0.76	0.64	0.52	0.49	0.63	7.83
PTA	0.70	0.43	0.54	0.74	0.60	0.60	**0.92**	0.70	**0.77**	0.63	0.52	0.50	0.64	5.42
PSDNorm	0.60	0.44	0.46	0.68	0.57	0.61	0.85	0.71	0.72	0.64	0.48	0.51	0.61	8.17
CARAT (Ours)	**0.72**	**0.52**	**0.55**	**0.76**	**0.63**	0.67	**0.92**	**0.72**	**0.77**	0.63	**0.55**	**0.54**	**0.66**	**2.42**

with our asymmetric curriculum and the same corrupted test batches. Thus, Table 1 compares adaptation on a common source model. Source-training effects are assessed separately in Appendix Table 8, and replacement curricula are examined in Section 4.2. Baseline implementations and update schedules are in Appendix F.

Evaluation and reporting. We use episodic TTA (Wang et al., 2020a): each method adapts on an unlabeled test batch, predicts that batch, and resets before the next. Table 1 averages macro-F1 within each dataset–backbone setting and weights the 12 settings equally; its descriptive mean rank is computed from the setting means before rounding. In the DaliaHAR ablations, Diverse (Div.) averages ten modality-subset scenarios and five corruption types. One F1 point denotes 0.01 on the 0–1 scale; numerical differences are computed before display rounding when finer source values are available. Appendix N retains the paired Wilcoxon summaries and five corruption-redraw trials from the repeated evaluation. We profile updated parameters, GFLOPs, latency, and CPU energy for batches of 64, including adaptation, inference, and detector overhead (Appendix L).

4.1 Effectiveness under Multimodal Test-Time Corruption

Table 1 yields three key observations. **(i) CARAT achieves the highest overall macro-F1 and best mean rank.** It improves on EATA, the strongest baseline overall, by 1.58 F1 points on average, with a median paired improvement of 0.75 points across the 12 settings. Its mean rank is 2.42, compared with 3.50 for EATA. **(ii) The largest gains occur on ShaSpec across all four datasets.** CARAT leads in every ShaSpec setting, averaging 4.39 points above EATA, compared with 0.15 on MAESTRO and 0.19 on Flex-MoE. **(iii) CARAT leads or remains close to the best method in 11 of the 12 settings.** It ranks first in seven settings and third in four, with the latter trailing the best method by only 0.18–0.56 F1 points. Appendices H–K provide every modality-subset and corruption result. Appendix N reports paired effects, intervals, and condition-level ranks; its separate five-redraw DSADS–MAESTRO experiment reports positive paired intervals against all eight included comparators, assessing corruption variability for a fixed checkpoint.

The benchmark reveals a structural trend: the relative strength of TTA methods changes with the sensing configuration. On IMU-dominated DSADS and PAMAP2, multimodal methods such as PTA are competitive; PTA is the strongest baseline on DSADS–MAESTRO and DSADS–Flex-MoE. On heterogeneous DaliaHAR and WESAD, however, TENT and EATA outperform PTA across all six dataset–backbone settings. With Flex-MoE, both obtain approximately 0.68 on DaliaHAR and 0.54 on WESAD, compared with PTA's 0.60 and 0.50. A plausible explanation is that disagreement between sensors measuring different physical processes can reflect complementary information, making it an ambiguous cue for corruption. This observation motivates CARAT's separation of sensor abnormality from backbone reliance: current detector flags identify suspect inputs, while the source-learned prior guides their omission or attenuation. Appendix Table 32 provides the corresponding condition-level breakdown. The comparison remains observational, since the datasets also differ in task, sensor count, and corruption difficulty.

CARAT combines modest adaptation compute with substantial runtime savings over multimodal TTA. Across five profiled configurations, CARAT uses 9.49% fewer GFLOPs and updates 47.82% fewer parameters than EATA, measured by median within-configuration ratios. Relative to AeO and PTA, it reduces median latency by 71.03% and 73.97%, respectively. These computational savings coexist with implementation overhead: the reference implementation has 44.23% higher latency and 37.37% higher CPU energy than EATA. A separate DSADS–MAESTRO study demonstrates that this overhead can be substantially reduced through implementation optimization: batching and reusing detector computations lowers end-to-end latency by 26.3%, from 1,713 to 1,262 ms, while preserving detector features, suspect flags, and predictions (Appendix Table 27). Together, these results support CARAT as a practical default when adaptation FLOPs and the number of updated parameters are primary constraints, with further runtime savings demonstrated through detector optimization.

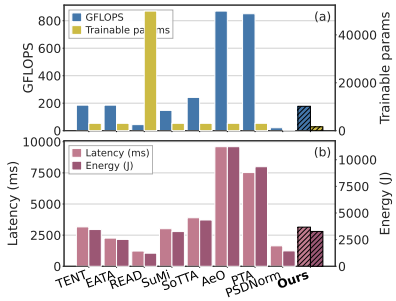


Figure 4: PAMAP2–MAESTRO efficiency for batches of 64, including detector overhead: GFLOPs, updated parameters, latency, and CPU energy. Full measurements: Appendix Table 25.

Detecting corruption is insufficient to justify removing a modality. Unconditional dropping replaces every suspect stream with the missingness symbol, bypassing the reliance-guided choice between removal and attenuation. In the MAESTRO comparison (Figure 5), PAMAP2 F1 falls from 0.92 for Static to 0.68 under unconditional dropping, while CARAT retains 0.92. On DaliaHAR, the corresponding scores are 0.74, 0.67, and 0.76 (Appendix Table 28). These losses show that a missingness-resilient backbone alone does not make omission a safe response to every flagged input, supporting intervention that accounts for the information retained by suspect streams. Across Table 1, CARAT improves on Static in 11 of 12 settings. The complete grids also include all-modality corruption, where unconditional removal would discard every stream (Appendix Tables 12–14).

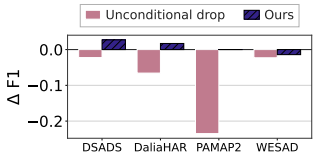


Figure 5: Absolute ΔF1 relative to Static on MAESTRO, averaged across the evaluated modality-subset and corruption conditions. Full results: Appendix Table 28.

4.2 Ablation and Further Design Analyses

We examine DaliaHAR–ShaSpec, whose non-uniform reliance profile distinguishes high-reliance wrist acceleration (w_acc) from low-reliance temperature (w_temp); all_wrist tests joint wrist-sensor corruption, and Div. aggregates 50 conditions. Full reliance profiles are in Appendix G.1.

Selective intervention and online updates make complementary contributions. Table 2 links the framework's design choices to performance. Always dropping (R1) loses 7 F1 points on w_acc and 5 on Div.; always attenuating (R2) ties CARAT on these two settings but trails on w_temp and all_wrist. Freezing attenuation and disabling the optimizer (R8) costs 4 points on w_acc and 3 on Div., measuring the joint benefit of normalization and attenuation updates. Removing attenuation alone (R6) lowers all four scores. Detection matters most under joint wrist corruption: removing it (R4) reduces F1 from 0.67 to 0.60, whereas oracle flags (R5) raise it to 0.68. Noisy flags (R3) preserve Div. F1 but lose 3 points on w_acc. The uniform-training control (R9) reduces Div. F1 to 0.60; it changes both source weights and the prior, so it measures their combined contribution.

Table 2: **Component ablations.** Macro-F1 (↑); all_w: all_wrist. R8 disables online updates; R9 changes source training and its prior. **Bold**/underline: best/second displayed score, including ties.

ID	Variant	w_acc	w_temp	all_w	Div.
R1	Always drop	0.63	0.70	0.61	0.58
R2	Always attenuate	**0.70**	0.73	0.66	**0.63**
R3	Noisy flags	0.67	0.73	0.65	**0.63**
R4	No detector	0.66	**0.75**	0.60	0.61
R5	Oracle flags	**0.70**	**0.75**	**0.68**	**0.63**
R6	No attenuation	0.67	0.74	0.66	0.62
R8	Frozen $a^j = \rho^j$	0.66	0.74	0.64	0.60
R9	Uniform training	0.59	0.59	0.59	0.60
R7	**CARAT**	**0.70**	**0.75**	0.67	**0.63**

Asymmetric training improves robustness to missing inputs. Averaging the reported F1 means across the 12 settings in Appendix Table 8, asymmetric training improves on each backbone's original regime by 9.58 and 14.50 points at 20% and 40% missingness. Complete-input effects vary: DaliaHAR–ShaSpec changes from 0.77 to 0.74, while its 40%-missing F1 rises from 0.39 to 0.61.

Table 3: **TTA across source-training regimes.** DaliaHAR–ShaSpec, Div. macro-F1 (↑).

TTA	OGM	OPM	MIAM
Static	0.57	0.59	0.58
EATA	0.62	0.63	0.61
CARAT	**0.63**	**0.65**	**0.63**

CARAT remains effective across alternative source-training regimes. On DaliaHAR–ShaSpec, CARAT exceeds the strongest evaluated TTA baseline by 1, 2, and 2 Div. F1 points with OGM-, OPM-, and MIAM-trained checkpoints, respectively (Table 3); EATA matches the best baseline in each case. OPM-trained CARAT reaches 0.65, compared with 0.63 under the asymmetric curriculum (Table 2). These results support the portability of reliance-guided adaptation across the tested source-training regimes. The asymmetric curriculum's benefits for missing-input robustness are evaluated separately in Appendix Tables 8 and 34; Appendix Table 35 gives the full TTA comparison.

Preserving the learned reliance ordering is more effective than reversing it. With the source checkpoint fixed, reversing the prior lowers w_acc F1 from 0.70 to 0.64 and Div. F1 from 0.63 to 0.61 (Table 4). Order-preserving jitter matches CARAT at two-decimal precision in all four settings; random priors reduce the three named-scenario scores but tie on Div. These interventions support the usefulness of the learned profile and tolerance to the tested magnitude perturbations, without implying that routing is invariant to prior values.

Table 4: **Prior perturbations.** Rev.: reversed ordering; Rand.: random; Jit.: order-preserving jitter.

Prior	w_acc	w_temp	all_w	Div.
Rev.	0.64	0.71	0.62	0.61
Rand.	0.67	0.73	0.65	**0.63**
Jit.	**0.70**	**0.75**	**0.67**	**0.63**
CARAT	**0.70**	**0.75**	**0.67**	**0.63**

CARAT remains ahead of Static and TENT across detector-quality bins. Its margin grows from low to high detector quality in Figure 6, while absolute F1 falls for all methods. The bins contain different corruption instances, so this trend does not measure the causal effect of improving detector accuracy; the controlled flag ablations in Table 2 isolate detector changes. Appendix G.2 reports routing-threshold and batch-size sensitivity (Figures 11 and 10).

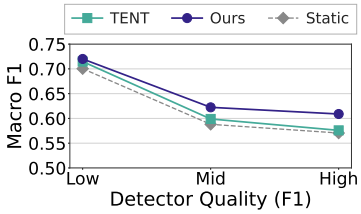


Figure 6: **Detector-quality bins on Div.**

Additional diagnostics probe uninformative inputs and modality projection width. In Table 5A, replacing one to three training modalities with variance-matched white noise yields terminal dropout probabilities ρ^j of 0.00–0.01 for those streams and F1 of 80.40–82.50 points. Panel B instead reports normalized gradient-based reliance r^j: widening the chest or temperature projection eightfold increases its share by 2.90 or 4.80 percentage points, with F1 changes of -0.10 or -0.80 points. The profile responds to width, but the widened modality's share does not increase eightfold. These controls characterize the proxy under the tested training changes; they do not establish parameterization invariance. Appendix Tables 36 and 37 give complete results.

Table 5: **DaliaHAR–MAESTRO diagnostics.** T/E/B: wrist temp./EDA/BVP; C: chest acc. ρ: terminal dropout; r: normalized reliance.

A. Noise	Max. ρ (noised)		F1 pts.
None	–		$80.20_{\pm 1.40}$
T	0.01		81.40
T, E	0.00		82.50
T, E, B	0.00		80.40
B. Width	$100r^{\mathrm{C}}$	$100r^{\mathrm{T}}$	F1 pts.
None	$23.70_{\pm 13.00}$	$9.80_{\pm 4.10}$	$80.30_{\pm 0.70}$
C ×8	$\mathbf{26.60}_{\pm 11.60}$	$21.30_{\pm 6.20}$	$80.20_{\pm 0.70}$
T ×8	$23.40_{\pm 3.80}$	$\mathbf{14.60}_{\pm 11.50}$	$79.50_{\pm 2.10}$

5 Conclusion and Future Work

CARAT amortizes reliance estimation through source training, separating it from runtime corruption detection to guide omission or attenuation without subset search. Across 12 equally weighted dataset–backbone settings, it leads overall macro-F1 by 1.58 points over EATA and achieves the best mean rank (2.42), with the largest gains on ShaSpec. PTA is competitive on IMU-dominated datasets, whereas TENT and EATA are stronger on heterogeneous datasets. These findings support CARAT as a practical option in the evaluated setting. Evaluation covers synthetic wearable-classification corruptions and backbones supporting trained omission and separable pre-fusion representations. The frozen, parameterization-dependent prior may become stale, and attenuation cannot ensure reliability when all streams are uninformative. Future work targets slow-timescale prior refresh, uncertainty-aware adaptation and abstention, finer suspect-set selection, and validation under natural failures across broader tasks and architectures.

Ethics Statement

This work studies robustness and deployment cost in multimodal test-time adaptation using the publicly available DSADS, DaliaHAR, PAMAP2, and WESAD datasets, with no new participant recruitment or data collection. These datasets contain human motion and physiological recordings, making privacy and potential dataset biases relevant considerations. Errors in suspect-modality detection or reliance-guided removal and attenuation may suppress useful signals and affect predictive performance differently across classes and user groups. Deployment in sensitive applications therefore requires appropriate privacy safeguards and application-specific evaluation of reliability and subgroup performance.

Reproducibility Statement

A minimal source-code implementation is included in the supplementary materials. Appendix B describes or references dataset preprocessing and subject-level split protocols, while Appendix A specifies the corruption processes and their parameters. Source-training seeds and hardware are reported at the start of the appendix. Appendices C, G, D, and F document the curriculum, backbone implementations, detector, and episodic baseline adaptation, respectively. Appendices H–K provide complete results across corruption conditions, and Appendix L specifies computational-cost measurement protocols and results. The appendix also provides sensitivity analyses and additional statistical comparisons. Appendix A.1 presents the gradient-flow interpretation of the reliance profile, including its derivation, assumptions, and limitations.

AI Use Statement

For manuscript writing, we used OpenAI Codex only to polish the authors' original text. We also used Codex to assist with generating figures using Python. The authors manually verified all content presented in the paper and take full responsibility for the accuracy and originality of the manuscript and accompanying materials.

References

John Allen. Photoplethysmography and its application in clinical physiological measurement. *Physiological measurement*, 28(3):R1–R39, 2007.

Kerem Altun, Billur Barshan, and Orkun Tunçel. Comparative study on classifying human activities with miniature inertial and magnetic sensors. *Pattern Recognition*, 43(10):3605–3620, 2010.

Md Abul Bashar and Richi Nayak. Tanogan: Time series anomaly detection with generative adversarial networks. In *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 1778–1785. IEEE, 2020.

Mohammad Braei and Sebastian Wagner. Anomaly detection in univariate time-series: A survey on the state-of-the-art. *arXiv preprint arXiv:2004.00433*, 2020.

MingCai Chen, Baoming Zhang, Zongbo Han, Wenyu Jiang, Yanmeng Wang, Shuai Feng, Yuntao Du, and Bingkun Bao. Test-time selective adaptation for uni-modal distribution shift in multi-modal data. In *Forty-second International Conference on Machine Learning*, 2025.

Diego Silva de Frana. Physical activity prediction dataset. https://www.kaggle.com/datasets/diegosilvadefrana/fisical-activity-dataset, n.d. Accessed: 2026-05-05.

Hao Dong, Eleni Chatzi, and Olga Fink. Towards robust multimodal open-set test-time adaptation via adaptive entropy-aware optimization. *arXiv preprint arXiv:2501.13924*, 2025.

Théo Gnassounou, Antoine Collas, Rémi Flamary, Karim Lounici, and Alexandre Gramfort. Multi-source and test-time domain adaptation on multivariate signals using spatio-temporal monge alignment. *arXiv preprint arXiv:2407.14303*, 2024.

Théo Gnassounou, Antoine Collas, Rémi Flamary, and Alexandre Gramfort. Psdnorm: Test-time temporal normalization for deep learning in sleep staging. *arXiv preprint arXiv:2503.04582*, 2025.

Taesik Gong, Yewon Kim, Taeckyung Lee, Sorn Chottananurak, and Sung-Ju Lee. Sotta: Robust test-time adaptation on noisy data streams. *Advances in Neural Information Processing Systems*, 36:14070–14093, 2023.

Shivam Grover and Ali Etemad. Shift-aware test time adaptation and benchmarking for time-series forecasting. In *Second Workshop on Test-Time Adaptation: Putting Updates to the Test! at ICML 2025*, 2025.

Zirun Guo and Tao Jin. Smoothing the shift: Towards stable test-time adaptation under complex multimodal noises. *arXiv preprint arXiv:2503.02616*, 2025.

Yongbo He, Zirun Guo, and Tao Jin. Decoupling stability and plasticity for multi-modal test-time adaptation. *arXiv preprint arXiv:2603.00574*, 2026.

HyunGi Kim, Siwon Kim, Jisoo Mok, and Sungroh Yoon. Battling the non-stationarity in time series forecasting via test-time adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 17868–17876, 2025.

Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*. John Wiley & Sons, 2019.

Ekdeep Singh Lubana and Robert P Dick. A gradient flow framework for analyzing network pruning. *arXiv preprint arXiv:2009.11839*, 2020.

Payal Mohapatra, Yueyuan Sui, Akash Pandey, Stephen Xia, and Qi Zhu. Maestro: Adaptive sparse attention and robust learning for multimodal dynamic time series. *arXiv preprint arXiv:2509.25278*, 2025.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient test-time model adaptation without forgetting. In *International conference on machine learning*, pp. 16888–16905. PMLR, 2022.

Alan V Oppenheim. *Discrete-time signal processing*. Pearson Education India, 1999.

Christina Orphanidou. Signal quality assessment in physiological monitoring: state of the art and practical considerations. 2017.

Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. Balanced multimodal learning via on-the-fly gradient modulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8238–8247, 2022.

Attila Reiss and Didier Stricker. Creating and benchmarking a new dataset for physical activity monitoring. In *Proceedings of the 5th International Conference on PErvasive Technologies Related to Assistive Environments*, PETRA '12, New York, NY, USA, 2012. Association for Computing Machinery. ISBN 9781450313001. doi: 10.1145/2413097.2413148. URL https://doi.org/10.1145/2413097.2413148.

Attila Reiss, Ina Indlekofer, Philip Schmidt, and Kristof Van Laerhoven. Deep ppg: Large-scale heart rate estimation with convolutional neural networks. *Sensors*, 19(14):3079, 2019.

Philip Schmidt, Attila Reiss, Robert Duerichen, Claus Marberger, and Kristof Van Laerhoven. Introducing wesad, a multimodal dataset for wearable stress and affect detection. In *Proceedings of the 20th ACM international conference on multimodal interaction*, pp. 400–408, 2018.

Abhishek Sharma, Leana Golubchik, and Ramesh Govindan. On the prevalence of sensor faults in real-world deployments. In *2007 4th Annual IEEE Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks*, pp. 213–222. IEEE, 2007.

Xin Sun, Heng Zhou, and Chao Li. Multivariate time series anomaly detection with idempotent reconstruction. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726*, 2020a.

Guowei Wang, Fan Lyu, and Changxing Ding. Partition-then-adapt: Combating prediction bias for reliable multi-modal test-time adaptation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Hu Wang, Yuanhong Chen, Congbo Ma, Jodie Avery, Louise Hull, and Gustavo Carneiro. Multi-modal learning with missing modality via shared-specific feature modelling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15878–15887, 2023.

Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12695–12705, 2020b.

Yake Wei, Di Hu, Henghui Du, and Ji-Rong Wen. On-the-fly modulation for balanced multimodal learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(1):469–485, 2024.

Jiehui Xu, Haixu Wu, Jianmin Wang, and Mingsheng Long. Anomaly transformer: Time series anomaly detection with association discrepancy. *arXiv preprint arXiv:2110.02642*, 2021.

Zhijian Xu, Ailing Zeng, and Qiang Xu. Fits: Modeling time series with $10k$ parameters. *arXiv preprint arXiv:2307.03756*, 2023.

Mouxing Yang, Yunfan Li, Changqing Zhang, Peng Hu, and Xi Peng. Test-time adaptation against multi-modal reliability bias. In *The twelfth international conference on learning representations*, 2024.

Sukwon Yun, Inyoung Choi, Jie Peng, Yangfan Wu, Jingxuan Bao, Qiyiwen Zhang, Jiayi Xin, Qi Long, and Tianlong Chen. Flex-moe: Modeling arbitrary modality combination via the flexible mixture-of-experts. In *Advances in Neural Information Processing Systems*, volume 37, pp. 98782–98805. Curran Associates, Inc., 2024.

Robin Zbinden, Wesley Monteith-Finas, Gencer Sumbul, Nina van Tiel, Chiara Vanalli, and Devis Tuia. Miam: Modality imbalance-aware masking for multimodal ecological applications. In *The Fourteenth International Conference on Learning Representations*, 2026.

Zhilin Zhang, Zhouyue Pi, and Benyuan Liu. Troika: A general framework for heart rate monitoring using wrist-type photoplethysmographic signals during intensive physical exercise. *IEEE Transactions on biomedical engineering*, 62(2):522–531, 2014.

APPENDIX

This Appendix provides additional details supporting the main paper, including corruption-generation procedures and detector construction (Sections A and D), dataset and implementation details for all evaluated backbones and TTA baselines (Sections B, F, and G), analyses of the asymmetric modality-dropout curriculum and learned reliance profiles (Sections C, A.1, and G.1), additional sensitivity studies and ablations (Section G.2), and complete per-condition experimental results beyond the aggregates reported in the main paper.

All experiments are implemented in PyTorch and conducted on Ubuntu servers equipped with NVIDIA TITAN RTX GPUs. Source-training experiments are repeated across three random seeds $\{2711, 2712, 2713\}$, and results are reported as mean $\pm$ standard deviation where applicable.

For test-time adaptation evaluation, each method is evaluated across at least 10 modality-corruption scenarios and 5 corruption types per dataset/backbone combination under the episodic adaptation protocol.

A TEST-TIME CORRUPTION SIMULATION DETAILS

We describe below each noise model used in this work. Each corruption is applied independently per channel to a univariate signal $x[t] \in \mathbb{R}$ for $t = 1, \dots, T$, producing a corrupted signal $\tilde{x}[t]$. The set $\mathcal{P}$ is locally redefined per corruption type.

Discrete Missingness (Little & Rubin, 2019). A subset of $\lceil \rho T \rceil$ indices $\mathcal{P} \subset \{1, \dots, T\}$ is sampled uniformly at random, and each selected sample is replaced with zero:

$$\tilde{x}[t] = \begin{cases} 0, & t \in \mathcal{P}, \\ x[t], & \text{otherwise.} \end{cases} \tag{5}$$

Zero-replacement (rather than masking) is used to remain consistent with downstream backbones that handle missing inputs via zero-imputation.

Waveform Anomaly (Braei & Wagner, 2020; Bashar & Nayak, 2020). A subset of segments is randomly selected, and within each segment the original signal is replaced by one of two synthetic waveforms — a square wave or a sinusoid — chosen with equal probability:

$$\tilde{x}[t] = \begin{cases} w[t], & t \in \mathcal{P}, \\ x[t], & \text{otherwise,} \end{cases} \qquad w[t] = \begin{cases} c \sum_{k=0}^{L-1} \frac{1}{2k+1} \sin\big(2\pi(2k+1) f\, t\big), & \text{(square wave)} \\ a \sin(2\pi f t), & \text{(sinusoid)} \end{cases} \tag{6}$$

where $\mathcal{P}$ is the union of the selected segments, each of radius r centered at a random index. The constant c is a free amplitude parameter (not the Fourier coefficient $4/\pi$ of an idealized square wave); L controls the harmonic order of the truncated expansion.

Nonlinear Sensor Degradation (Allen, 2007). The signal is amplitude-clipped at the q-th percentile of $|x[t]|$ and then compressed via a power law:

$$\tilde{x}[t] = \operatorname{sign}\big(\bar{x}[t]\big) \cdot \big|\bar{x}[t]\big|^{\beta}, \qquad \bar{x}[t] = \operatorname{clip}(x[t],\ -\tau_q,\ +\tau_q)\,, \tag{7}$$

where τ_q is the q-th percentile of $|x[t]|$ and $\beta \in (0, 1)$ controls compression. This models the saturation-then-compression behavior of degraded transducers.

Frequency-Selective Attenuation (Oppenheim, 1999). A low-pass filter is applied in the frequency domain by attenuating components above a cutoff frequency f_c by A dB:

$$\tilde{x}[t] = \mathcal{F}^{-1}\Big\{ M(f) \cdot \mathcal{F}\{x[t]\} \Big\}, \qquad M(f) = \begin{cases} 1, & f \le f_c, \\ 10^{-A/20}, & f > f_c, \end{cases} \tag{8}$$

where $\mathcal{F}$ denotes the discrete Fourier transform and f_s is the modality's sampling rate.

Nonstationary Noise (Orphanidou, 2017). Additive Gaussian noise is injected with a time-varying standard deviation, such that the local SNR degrades linearly across the observation window:

$$\tilde{x}[t] = x[t] + \sigma[t]\, \eta[t], \qquad \eta[t] \sim \mathcal{N}(0, 1), \qquad \sigma^2[t] = \frac{\bar{P}_x}{10^{\mathrm{SNR}[t]/10}}, \tag{9}$$

with $\mathrm{SNR}[t] = \mathrm{SNR}_0 + (\mathrm{SNR}_T - \mathrm{SNR}_0)\, \frac{t-1}{T-1}$ and $\bar{P}_x = \frac{1}{T} \sum_t x[t]^2$.

Default parameters. Table 6 summarizes the default values used across all experiments unless otherwise stated.

Table 6: Default parameters for each test-time corruption.

Corruption	Default parameters
Discrete missingness	$\rho = 0.05$
Waveform anomaly	$\rho = 0.05$, $r = 5$, $c = 3.0$, $L = 5$, $f = 0.04$, $a = 2.0$
Nonlinear degradation	$q = 60$, $\beta = 0.3$
Frequency attenuation	$f_c = 0.3 \cdot f_s/2$, $A = 20\,\text{dB}$
Nonstationary noise	$\text{SNR}_0 = 30\,\text{dB}$, $\text{SNR}_T = 5\,\text{dB}$

A.1 Gradient-Flow Interpretation of the Reliance Profile

We relate the reliance profile ρ^j in Section 3.2.1 to the gradient-flow framework of Lubana & Dick (2020). The decomposition motivates the gradient signal used by the curriculum; the connection between the resulting proxy and predictive dependence is evaluated empirically.

Setup. Let f_θ have a disjoint parameter partition $\theta = (\theta_1, \ldots, \theta_M, \theta_{\text{rest}})$, where θ_j is modality j's input-projection block and θ_{rest} collects the remaining encoder, fusion, and prediction parameters. For a fixed differentiable loss $\mathcal{L}(\theta)$, we consider Euclidean gradient flow. The selected blocks identify where gradients are measured; shared downstream parameters can also carry information from multiple modalities.

Gradient-flow decomposition. Under gradient flow,

$$\dot{\theta} = -\nabla_\theta \mathcal{L}(\theta),$$

the chain rule gives

$$\dot{\mathcal{L}}(\theta) = \langle \nabla_\theta \mathcal{L}, \dot{\theta} \rangle = -\|\nabla_\theta \mathcal{L}\|_2^2.$$

Partitioning the parameters into these blocks yields

$$\dot{\mathcal{L}}(\theta) = -\sum_{j=1}^{M} \|\nabla_{\theta_j} \mathcal{L}\|_2^2 - \|\nabla_{\theta_{\text{rest}}} \mathcal{L}\|_2^2.$$

When the selected blocks have nonzero total squared-gradient norm, their normalized contributions to instantaneous loss reduction are

$$\sigma^j(\theta) = \frac{\|\nabla_{\theta_j} \mathcal{L}\|_2^2}{\sum_{\ell=1}^{M} \|\nabla_{\theta_\ell} \mathcal{L}\|_2^2}, \qquad \sum_{j=1}^{M} \sigma^j(\theta) = 1.$$

Connection to the implemented profile. The main method first averages gradient norms over the training window, then normalizes them and applies the dropout schedule:

$$\bar{g}_e^j = \frac{1}{W} \sum_{u=k_e-W}^{k_e-1} \|\nabla_{\theta_j} \mathcal{L}_u(\theta_u)\|_2, \qquad r_e^j = \frac{\bar{g}_e^j}{\sum_{\ell=1}^{M} \bar{g}_e^\ell},$$

$$\rho^j = p_{e^\star}^j = \min\left(p_{\text{cap}},\, M p_{\text{base}}(e^\star) r_{e^\star}^j\right).$$

Thus, r_e^j is a normalized gradient share, whereas ρ^j is the terminal dropout probability and need not sum to one. At a fixed parameter state, nonnegative gradient norms and their squares have the same ordering. Window averaging need not preserve the instantaneous ordering or the ordering of averaged squared norms, and clipping can introduce ties. Consequently, the implemented profile is an optimization-based proxy, rather than the exact loss-reduction share σ^j.

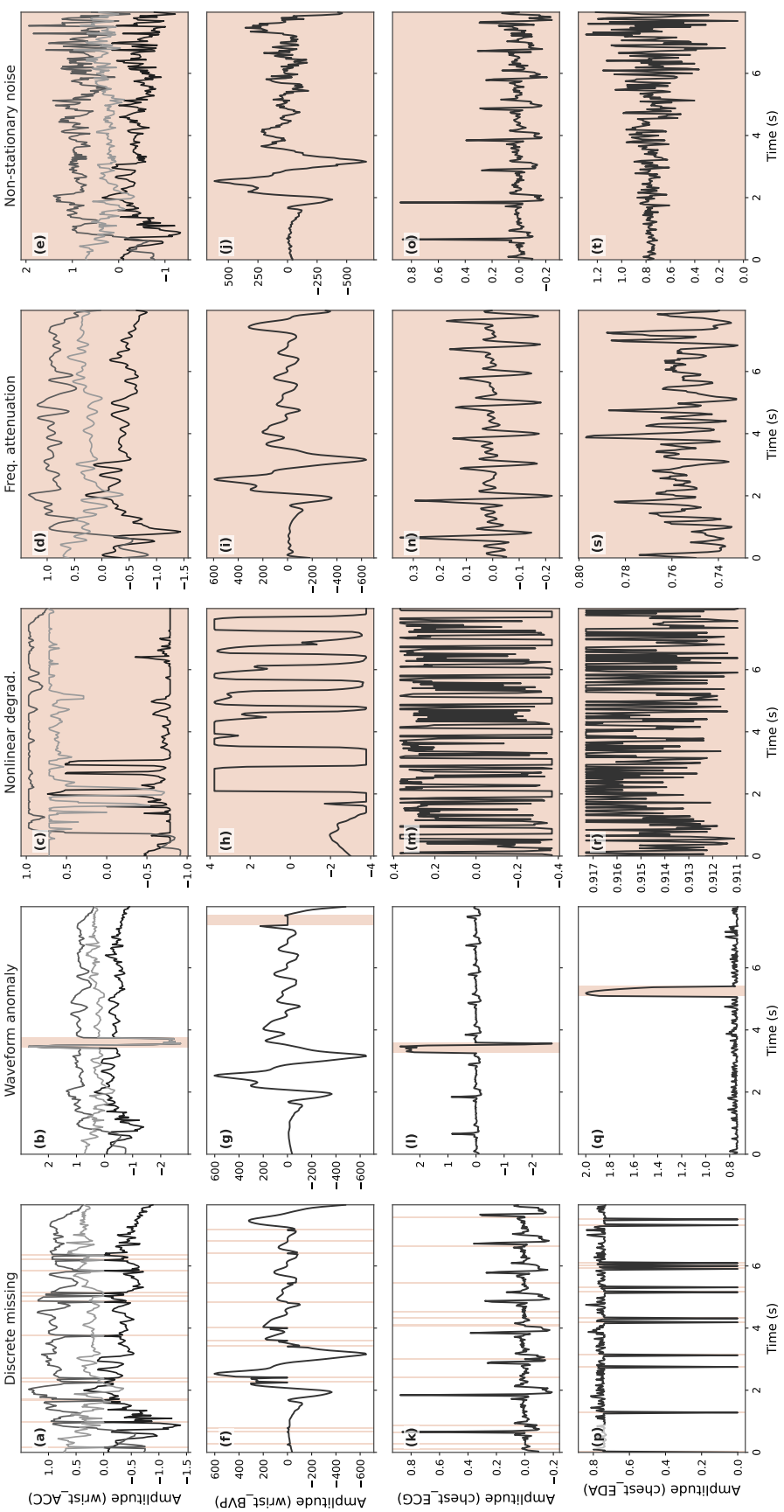


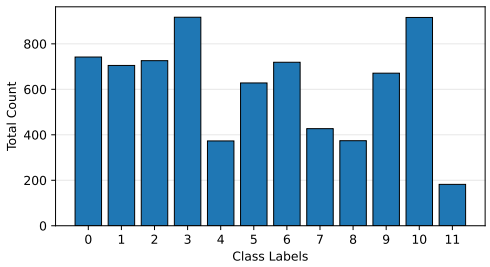
Figure 7: **Test-time noise injections across heterogeneous wearable sensors.** Rows show four modalities — wrist_ACC (motion), wrist_BVP (photoplethysmography), chest_ECG (electrocardiogram), and chest_EDA (electrodermal activity). Columns show the five noise models from Appendix A: *discrete missingness* (sparse zero replacements), *waveform anomaly* (segment substitution by a square wave or sinusoid), *nonlinear sensor degradation* (amplitude clipping with power-law compression), *frequency-selective attenuation* (high-frequency rolloff), and *nonstationary noise* (linearly increasing noise variance). Orange highlights mark localized perturbations; shaded panels indicate window-wide corruption. The same mechanism produces qualitatively different distortions across modalities.

Table 7: PAMAP2 dataset modality details.

Modality	Rate (Hz)	Variates
IMU_hand	100	17
IMU_chest	100	17
IMU_ankle	100	17
heart_rate	9	1

Sampling rates denote the original sensor rates before resampling to 64 Hz.

Figure 8: Class distribution of the PAMAP2 activity labels.



Connection to reliance-guided adaptation. The profile connects two stages of CARAT. During training, higher epoch-wise dropout probabilities increase exposure to inputs in which the corresponding modalities are absent. At deployment, the frozen terminal profile supplies the relative reliance mass used by the routing rule. Removal invokes the trained missingness operation, while attenuation retains downweighted suspect representations when the retained-reliance threshold is not met. These interventions act at different locations in the model; the construction does not establish equivalence of their representations or proximity to the source distribution. Their utility is assessed through the omission and prior-perturbation experiments in Section 4.

Scope of the interpretation. The gradient-flow decomposition holds for the specified parameterization and loss and requires no stationarity assumption. Finite-step optimization, adaptive preconditioning, stochastic batches, and changing dropout masks differ from this idealization. Parameter scaling and projection width can also change the measured norms. We therefore use ρ^j as a backbone-specific routing prior, without interpreting it as a causal attribution of predictive information or a guarantee that a modality can be removed safely. The leave-one-modality-out comparison, prior perturbations, and width diagnostics provide empirical checks of this use.

B Details on Datasets

We evaluate on WESAD (Schmidt et al., 2018), DaliaHAR (Reiss et al., 2019), DSADS(Altun et al., 2010), and PAMAP2 (Reiss & Stricker, 2012). For WESAD, DaliaHAR, and DSADS, we follow the preprocessing protocol of MAESTRO (Mohapatra et al., 2025), including the modality definitions, temporal alignment, resampling, windowing, label filtering, and subject-level split construction. We therefore do not repeat the full dataset specifications for these datasets here. Below, we summarize the PAMAP2 preprocessing used in our experiments, which follows the same window-based setup.

PAMAP2. PAMAP2 is a wearable activity-recognition dataset collected from inertial sensors and a heart-rate monitor. We treat the hand, chest, and ankle IMUs as three separate modalities, and use the heart-rate signal as an additional modality. Each IMU contains 17 channels, including temperature, two accelerometer streams, gyroscope, magnetometer, and orientation signals. We remove transient activities, segment the synchronized recordings into 8-second windows, and resample each window to a common base rate of 64 Hz, resulting in 512 timesteps per modality. We use 12 activity classes and subject-wise three-fold cross-validation. Table 7 and Figure 8 summarize the processed PAMAP2 setup.

C More Details on Curriculum Dropout Schedule

Following (Wang et al., 2023; Yun et al., 2024; Mohapatra et al., 2025), we abstract the three-phase modality-dropout curriculum as a single scalar rate:

$$p_{\text{base}}(e) = \begin{cases} 0 & e < 0.2E \\ p_{\max} \frac{e-0.2E}{0.3E} & 0.2E \le e < 0.5E \\ p_{\max} & e \ge 0.5E \end{cases}$$

where $p_{\max} \in [0, 1]$ is the target missingness rate and E the total number of training epochs. Existing methods treat $p_{\text{base}}(e)$ as the per-modality Bernoulli dropout probability, applied uniformly across all M modalities.

D Detector Analyses

Overview. Our suspect-modality detector is intentionally lightweight and serves only as a coarse routing signal for downstream adaptation rather than a standalone corruption classifier. For each sample $\mathcal{M}_i$ and modality j, the detector computes a corruption score s_i^j from two complementary feature families: FITS reconstruction-error statistics and log-spectral-distance (LSD) features relative to clean source data. Modalities whose score exceeds a clean-reference threshold are flagged as suspect and routed accordingly during adaptation.

FITS reconstruction features. For each fold, we train a frequency-domain FITS reconstructor $\mathcal{F}_\theta : \mathbb{R}^T \to \mathbb{R}^T$ (Xu et al., 2023) on clean source data using a standard MSE objective for 30 epochs (d=64, learning rate 10^{-3}, batch size 64). FITS operates in the DFT domain on the lowest $T/2$ frequency components, capturing low-rank periodic structure while leaving higher-frequency corruption in the residual. For each channel c, we compute residuals

$$\epsilon_t = |\hat{x}_{i,t}^c - x_{i,t}^c|,$$

and standardize them using clean-train statistics $(\bar{\epsilon}^c, \sigma_\epsilon^c)$. The resulting FITS feature block is

$$\phi_{\text{FITS}}^c(x_i) = \Big[\mu(\epsilon), \max(\epsilon), \mu(z), \max(z), \tfrac{1}{T}\sum_t \mathbb{1}[z_t > 2], \overline{\text{top}_5}(z)\Big],$$

where

$$z_t = (\epsilon_t - \bar{\epsilon}^c)/\sigma_\epsilon^c.$$

These statistics capture residual magnitude, deviation from the clean distribution, and sparsity/extremity of anomalies.

Log-spectral-distance features. We estimate the per-channel power spectral density $\text{PSD}(x_i^c)$ using Welch's method ($n_{\text{perseg}} = \min(T, 64)$, 50% overlap), and compute a clean PSD barycenter $\overline{\text{PSD}}_{\text{src}}^c$ over the clean training set following (Gnassounou et al., 2024; 2025). The log-spectral distance between PSDs p and $\bar{p}$ is

$$\text{LSD}(p, \bar{p}) = \left[\frac{1}{|\mathcal{F}|}\sum_{f\in\mathcal{F}} \big(10\log_{10} p(f) - 10\log_{10}\bar{p}(f)\big)^2\right]^{1/2}.$$

We additionally separate low/high-frequency bands using f_{cut}=4 Hz to capture localized spectral shifts. The resulting LSD feature block is

$$\phi_{\text{LSD}}^c(x_i) = \Big[\text{LSD}_{\text{tot}}, \text{LSD}_{\text{low}}, \text{LSD}_{\text{high}}, \Delta\arg\max, \Delta H_{\text{spec}}\Big],$$

where $\Delta\arg\max$ measures peak-frequency drift and ΔH_{spec} measures spectral-entropy deviation relative to the clean PSD barycenter.

Per-modality detector formulation. For modality j with channel set C_j, we concatenate the FITS and LSD feature blocks across channels:

$$\phi_i^j = \text{concat}_{c\in C_j}\big[\phi_{\text{FITS}}^c(x_i), \phi_{\text{LSD}}^c(x_i)\big] \in \mathbb{R}^{11|C_j|}.$$

Features are standardized using clean-train statistics, and an empirical Gaussian reference (μ^j, Σ^j) is fit on the resulting vectors. The corruption score is computed using the squared Mahalanobis distance

$$s_i^j = (\phi_i^j - \mu^j)^\top (\Sigma^j)^{-1}(\phi_i^j - \mu^j).$$

A modality is flagged as suspect when

$$s_i^j > \delta^j,$$

where δ^j is set to the 95th percentile of clean-train scores. The resulting binary routing indicator is

$$m_i^j = \mathbb{1}[s_i^j > \delta^j],$$

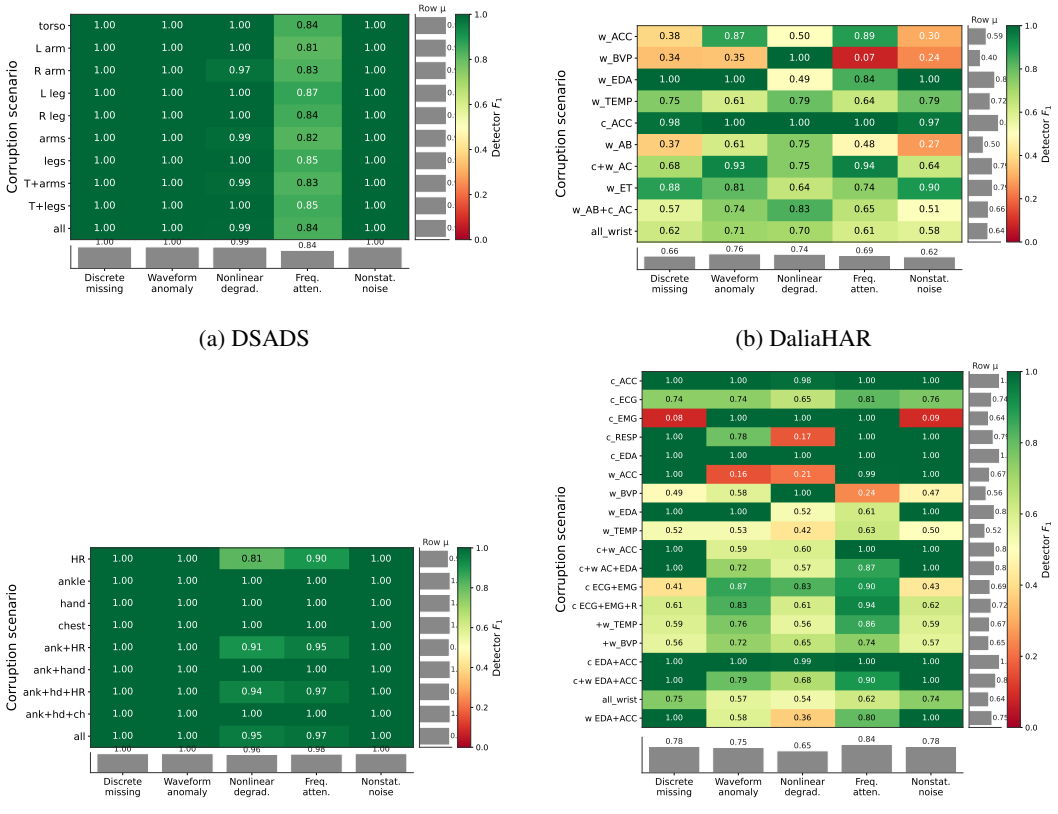
and the suspect-modality set used by CARAT is

$$\mathcal{S}_i = \{j : m_i^j = 1\}.$$

All detector hyperparameters are held fixed across datasets (z-threshold $= 2$, f_{cut}=4 Hz, percentile $= 95$, FITS d=64 / 30 epochs).

Why both FITS and LSD? The two feature families capture complementary corruption signatures. FITS residuals are sensitive to transient or localized deviations from the clean source distribution, but may absorb broad spectral shifts such as smooth attenuation. In contrast, LSD features capture global spectral-shape deviations but are less sensitive to localized temporal anomalies that average out in the PSD estimate. Empirically, combining both yields more robust suspect-modality detection across diverse corruption types.

Figure 9 reports detector F1 scores across (modality, corruption) combinations for all datasets. Performance is generally strongest for structured corruption affecting spectral content or waveform morphology, while localized corruption and mild missingness are intentionally harder to separate from clean data under the coarse one-class setting. Importantly, CARAT remains robust under moderate detector errors because routing depends only on the binary suspect indicator rather than calibrated corruption estimates.



(a) DSADS

(b) DaliaHAR

(c) PAMAP2

(d) WESAD

Figure 9: Detector quality (F1 at flagging suspect modalities) across (modality, noise) combinations on each dataset.

Table 8: Performance under complete and missing modalities. Accuracy and F1-score (↑) on DSADS, DaliaHAR, WESAD, and PAMAP2 across three backbones (MAESTRO, ShaSpec, FlexMoE) under `Complete`, `20% missing`, and `40% missing` conditions under their recommended training regime and our proposed asymmetric training regime.

			Accuracy			F1		
Dataset	**Backbone**	**Variant**	Complete	20% miss.	40% miss.	Complete	20% miss.	40% miss.
DSADS	MAESTRO	Original	$0.86_{\pm.00}$	$0.82_{\pm.01}$	$0.78_{\pm.01}$	$0.86_{\pm.00}$	$0.82_{\pm.01}$	$0.77_{\pm.01}$
		Ours	$0.87_{\pm.00}$	$0.85_{\pm.01}$	$0.79_{\pm.01}$	$0.86_{\pm.00}$	$0.83_{\pm.01}$	$0.78_{\pm.01}$
	ShaSpec	Original	$0.81_{\pm.01}$	$0.45_{\pm.02}$	$0.22_{\pm.01}$	$0.79_{\pm.00}$	$0.45_{\pm.02}$	$0.20_{\pm.02}$
		Ours	$0.78_{\pm.00}$	$0.75_{\pm.00}$	$0.64_{\pm.03}$	$0.77_{\pm.00}$	$0.74_{\pm.00}$	$0.64_{\pm.03}$
	FlexMoE	Original	$0.67_{\pm.01}$	$0.35_{\pm.09}$	$0.19_{\pm.06}$	$0.63_{\pm.03}$	$0.34_{\pm.08}$	$0.17_{\pm.07}$
		Ours	$0.74_{\pm.00}$	$0.73_{\pm.01}$	$0.60_{\pm.04}$	$0.74_{\pm.00}$	$0.73_{\pm.01}$	$0.60_{\pm.05}$
DaliaHAR	MAESTRO	Original	$0.74_{\pm.00}$	$0.69_{\pm.01}$	$0.60_{\pm.01}$	$0.80_{\pm.00}$	$0.74_{\pm.01}$	$0.63_{\pm.01}$
		Ours	$0.74_{\pm.00}$	$0.71_{\pm.01}$	$0.66_{\pm.01}$	$0.79_{\pm.00}$	$0.74_{\pm.02}$	$0.67_{\pm.02}$
	ShaSpec	Original	$0.74_{\pm.00}$	$0.55_{\pm.02}$	$0.44_{\pm.02}$	$0.77_{\pm.02}$	$0.57_{\pm.02}$	$0.39_{\pm.01}$
		Ours	$0.72_{\pm.00}$	$0.66_{\pm.02}$	$0.60_{\pm.02}$	$0.74_{\pm.00}$	$0.68_{\pm.02}$	$0.61_{\pm.02}$
	FlexMoE	Original	$0.70_{\pm.06}$	$0.53_{\pm.04}$	$0.37_{\pm.03}$	$0.70_{\pm.05}$	$0.51_{\pm.06}$	$0.35_{\pm.03}$
		Ours	$0.73_{\pm.00}$	$0.68_{\pm.00}$	$0.62_{\pm.01}$	$0.77_{\pm.00}$	$0.70_{\pm.00}$	$0.64_{\pm.02}$
WESAD	MAESTRO	Original	$0.67_{\pm.00}$	$0.64_{\pm.02}$	$0.56_{\pm.05}$	$0.60_{\pm.00}$	$0.57_{\pm.02}$	$0.47_{\pm.05}$
		Ours	$0.77_{\pm.00}$	$0.74_{\pm.01}$	$0.73_{\pm.01}$	$0.66_{\pm.00}$	$0.64_{\pm.01}$	$0.63_{\pm.01}$
	ShaSpec	Original	$0.65_{\pm.06}$	$0.53_{\pm.05}$	$0.43_{\pm.04}$	$0.54_{\pm.04}$	$0.47_{\pm.04}$	$0.41_{\pm.03}$
		Ours	$0.74_{\pm.00}$	$0.70_{\pm.01}$	$0.70_{\pm.02}$	$0.56_{\pm.00}$	$0.54_{\pm.00}$	$0.54_{\pm.02}$
	FlexMoE	Original	$0.69_{\pm.00}$	$0.69_{\pm.01}$	$0.65_{\pm.03}$	$0.55_{\pm.00}$	$0.55_{\pm.01}$	$0.50_{\pm.05}$
		Ours	$0.65_{\pm.00}$	$0.66_{\pm.01}$	$0.62_{\pm.02}$	$0.55_{\pm.00}$	$0.56_{\pm.01}$	$0.51_{\pm.03}$
PAMAP2	MAESTRO	Original	$0.95_{\pm.00}$	$0.93_{\pm.00}$	$0.84_{\pm.01}$	$0.95_{\pm.00}$	$0.92_{\pm.00}$	$0.85_{\pm.00}$
		Ours	$0.95_{\pm.00}$	$0.93_{\pm.00}$	$0.86_{\pm.01}$	$0.95_{\pm.00}$	$0.93_{\pm.00}$	$0.86_{\pm.00}$
	ShaSpec	Original	$0.86_{\pm.00}$	$0.80_{\pm.02}$	$0.71_{\pm.03}$	$0.86_{\pm.00}$	$0.81_{\pm.02}$	$0.73_{\pm.03}$
		Ours	$0.85_{\pm.00}$	$0.80_{\pm.02}$	$0.72_{\pm.03}$	$0.86_{\pm.00}$	$0.81_{\pm.02}$	$0.73_{\pm.03}$
	FlexMoE	Original	$0.87_{\pm.00}$	$0.83_{\pm.01}$	$0.78_{\pm.03}$	$0.88_{\pm.00}$	$0.84_{\pm.01}$	$0.79_{\pm.03}$
		Ours	$0.90_{\pm.00}$	$0.84_{\pm.00}$	$0.79_{\pm.05}$	$0.89_{\pm.00}$	$0.84_{\pm.00}$	$0.79_{\pm.05}$

E Results on Recommended and Asymmetric Curriculum Dropout Schedules

Table 8 supports the discussion in Section 4 by comparing the source checkpoints trained with each backbone's recommended regime and with our asymmetric modality-dropout curriculum under complete and missing-modality inputs with three seeds. We follow the same experimental setup as Mohapatra et al. (2025).

F Implementation Details on Baseline TTA Methods

We evaluate the eight TTA baselines described in Section 4 under the same episodic protocol. In addition to these baselines, we report Static as the no-adaptation reference. For each test batch, the adaptation model is restored to the source checkpoint, adapted using only the current batch, and then used to predict the same batch. All baselines are evaluated on the same corrupted batches and modality-corruption scenarios.

Unless otherwise specified, we follow the recommended hyperparameter settings of each baseline and use Adam with learning rate 10^{-3} for test-time updates. Following the common TTA evaluation

setting, all iterative adaptation methods are run for five steps unless the baseline recommends a single-update procedure. In our implementation, TENT (Wang et al., 2020a), EATA (Niu et al., 2022), SoTTA (Gong et al., 2023), AeO (Dong et al., 2025), and PTA (Wang et al.) use five adaptation steps. READ (Yang et al., 2024) uses one update step, and SuMi (Guo & Jin, 2025) performs one update after its sample-selection stage, following their recommended adaptation procedures. Static and PSDNorm (Gnassounou et al., 2025) do not update model parameters. Here, B denotes the batch size, C the number of classes, $H(\cdot)$ the prediction entropy, and $p_i^{\max}$ the maximum predicted probability for sample i.

Static. Static is the frozen source model evaluated directly on the corrupted test batch. It provides the no-adaptation reference.

TENT. TENT updates the affine parameters of all LayerNorm modules by minimizing prediction entropy:

$$\mathcal{L}_{\text{Tent}} = -\frac{1}{B}\sum_{i=1}^{B}\sum_{c} p_{ic}\log p_{ic},$$

where $p_i = \text{softmax}(z_i)$. All modalities remain active with unit weights throughout adaptation.

EATA. EATA follows the same LayerNorm adaptation target as TENT, but applies an entropy-based reliability filter before the update. We use the entropy threshold $0.4\log C$, where C is the number of classes. Adaptation is performed only when the filtered predictions provide a valid entropy-minimization signal; otherwise, the update is skipped.

READ. READ updates fusion-stage parameters with a reliability-aware objective. For MAESTRO and Flex-MoE, the trainable parameters are the query and key/value projections in the fusion attention module. For ShaSpec, which does not expose an explicit fusion-attention block, we adapt the Linear and LayerNorm parameters in its composition layer, classifier, and domain classifier. The loss is

$$\mathcal{L}_{\text{READ}} = \mathcal{L}_{\text{ra}} - \mathcal{L}_{\text{bal}},$$

with

$$\mathcal{L}_{\text{ra}} = \frac{1}{B}\sum_{i} p_i^{\max}\left(1 - \log p_i^{\max} + \log\gamma\right), \qquad \gamma = e^{-1}.$$

The balance term discourages degenerate batch-level predictions.

SuMi. SuMi identifies reliable samples using unimodal and multimodal predictions. We obtain unimodal predictions by activating one modality at a time and masking the remaining modalities. Samples are retained if they pass an interquartile-range feature filter and satisfy

$$H_{\text{uni}}(x_i) \geq \tau_{\text{uni}}, \qquad H_{\text{multi}}(x_i) \leq \tau_{\text{multi}},$$

where $\tau_{\text{uni}} = e/10$ and $\tau_{\text{multi}} = \log(C)/2 - 0.5$. On the selected samples, SuMi performs one LayerNorm update using entropy minimization together with a mutual-information sharing term that aligns unimodal and multimodal predictions via KL divergence.

SoTTA. SoTTA uses high-confidence uniform-class sampling followed by sharpness-aware entropy minimization. We use confidence threshold 0.99 and first retain samples whose maximum softmax probability exceeds this threshold. The retained samples are capped per pseudo-class to form a class-balanced memory for the batch. SoTTA then applies a SAM-style entropy objective with perturbation radius $\rho = 0.05$. If the high-confidence memory is empty, adaptation is skipped for that batch.

AeO. AeO applies an adaptive entropy objective coupled with cross-modal prediction discrepancy. For each sample, the normalized entropy defines the adaptive weight

$$w_i = \tanh\left(\beta\left(H(p_i)/\log C - \alpha\right)\right),$$

with $\alpha = 0.5$ and $\beta = 5.0$. The final objective combines adaptive entropy, unimodal entropy, pairwise cross-modal prediction discrepancy, and a class-diversity regularizer. We use $\gamma_1 = 0.5$ and $\gamma_2 = 0.1$, and update LayerNorm parameters.

PTA. PTA partitions the batch by combining confidence with predicted-class frequency. For sample i, we compute

$$r_i = \max_c p_{ic} \left(1 - f_{\hat{y}_i}\right),$$

where $\hat{y}_i = \arg\max_c p_{ic}$ and $f_{\hat{y}_i}$ is the batch frequency of the predicted class. Samples above the 75th percentile of r_i receive weight 2.0, while samples below the 25th percentile receive weight 0.1. PTA minimizes the resulting weighted entropy loss and adds a cross-modal alignment term that pulls unimodal predictions toward the detached multimodal prediction.

PSDNorm. PSDNorm is a non-parametric frequency-domain correction. We estimate clean source PSD barycenters from the source training set. At test time, when the detector flags a modality as corrupted, each channel in that modality is corrected before SAX tokenization using

$$G(\omega) = \sqrt{\frac{\mathrm{PSD}_{\mathrm{src}}(\omega)}{\mathrm{PSD}_{\mathrm{test}}(\omega) + \epsilon}}.$$

Unflagged modalities are left unchanged, and no model parameters are updated.

G Implementation Details of different backbones

We evaluate CARAT on three missingness-resilient backbones: MAESTRO (Mohapatra et al., 2025), ShaSpec (Wang et al., 2023), and Flex-MoE (Yun et al., 2024). To keep the comparison across backbones fair, we preserve the original backbone architectures and supervised training objectives, while applying the asymmetric modality-dropout curriculum introduced in Section 3.2.1 during source training. This curriculum changes the modality masking schedule but does not modify the backbone loss or add CARAT-specific modules to the source model. It produces the backbone-specific reliance profile used by CARAT at test time. In our implementation, all source backbones are trained on SAX-tokenized source data. The checkpoint with the best validation accuracy is used as the source checkpoint for CARAT and all baseline TTA methods.

MAESTRO. MAESTRO follows the original MAESTRO training objective and is optimized with cross-entropy loss over activity labels. During source training, we apply the asymmetric modality-dropout curriculum to estimate the MAESTRO-specific reliance profile. In the curriculum, the base modality-dropout probability is zero during the initial warm-up stage, increases linearly after warm-up, and is then held fixed. The per-modality dropout probabilities are assigned according to running gradient norms of the modality input projections, so modalities with larger estimated reliance are masked more often during training.

Flex-MoE. Flex-MoE follows the original Flex-MoE supervised objective. The training loss combines cross-entropy classification loss with the gate load-balancing loss from its fusion module:

$$\mathcal{L} = \mathcal{L}_{\mathrm{ce}} + \lambda_{\mathrm{gate}} \mathcal{L}_{\mathrm{gate}}, \qquad \lambda_{\mathrm{gate}} = 10^{-2}.$$

During source training, we apply the same asymmetric modality-dropout curriculum to estimate the Flex-MoE-specific reliance profile. The curriculum is applied to the modality inputs while keeping the Flex-MoE architecture and fusion objective unchanged.

ShaSpec. ShaSpec follows the original ShaSpec shared-specific training objective, which combines task classification, shared-feature alignment, and modality-specific domain classification:

$$\mathcal{L} = \mathcal{L}_{\mathrm{ce}} + \alpha \mathcal{L}_{\mathrm{shared}} + \beta \mathcal{L}_{\mathrm{specific}}, \qquad \alpha = 0.1, \quad \beta = 0.02.$$

The shared loss aligns shared modality features, and the specific loss trains modality-domain predictions. During source training, we apply the same asymmetric modality-dropout curriculum to estimate the ShaSpec-specific reliance profile, while keeping the ShaSpec architecture and shared-specific objective unchanged.

G.1 Modality Reliance Profiles Across Backbones and Datasets

To better understand how different multimodal backbones distribute optimization importance across sensing streams, Tables 9–11 report the learned modality reliance for each dataset and backbone combination. Following Section 3.2.1, these weights are derived from the normalized gradient magnitudes accumulated near the deployed epoch and provide a proxy for each modality's contribution to loss reduction under gradient flow. Values are reported as mean $\pm$ standard deviation across the deployed epoch ± 1 snapshot used for reliance estimation.

Several trends emerge. On heterogeneous wearable datasets such as WESAD and DaliaHAR, reliance distributions are highly non-uniform and strongly backbone-dependent, with certain modalities dominating optimization dynamics while others contribute minimally. In contrast, DSADS exhibits substantially more balanced reliance across body locations, reflecting its more homogeneous sensing structure. These observations support two key design choices in `CARAT`: (i) reliance should be estimated in a model-specific manner rather than assumed uniform across modalities, and (ii) routing decisions during test-time adaptation should preserve partially informative modalities rather than treating all corrupted streams equivalently.

Table 9: DaliaHAR – modality reliance scores ρ_j.

Modality	ShaSpec	FlexMoE	MAESTRO
wrist_ACC	0.80 ± 0.00	0.70 ± 0.18	0.65 ± 0.05
wrist_BVP	0.47 ± 0.29	0.19 ± 0.01	0.45 ± 0.05
wrist_EDA	0.71 ± 0.16	0.22 ± 0.01	0.13 ± 0.01
wrist_TEMP	0.17 ± 0.01	0.22 ± 0.02	0.17 ± 0.02
chest_ACC	0.80 ± 0.00	0.58 ± 0.22	0.60 ± 0.03

Table 10: DSADS — modality reliance weights ρ_j.

Modality	ShaSpec	FlexMoE	MAESTRO
torso	0.42 ± 0.18	0.42 ± 0.03	0.40 ± 0.05
right_arm	0.37 ± 0.09	0.40 ± 0.01	0.39 ± 0.06
left_arm	0.45 ± 0.11	0.42 ± 0.01	0.41 ± 0.07
right_leg	0.39 ± 0.05	0.39 ± 0.02	0.37 ± 0.03
left_leg	0.37 ± 0.16	0.37 ± 0.03	0.45 ± 0.10

Table 11: WESAD — modality reliance weights ρ_j.

Modality	ShaSpec	FlexMoE	MAESTRO
chest_ACC	0.80 ± 0.00	0.70 ± 0.14	0.78 ± 0.02
chest_ECG	0.80 ± 0.00	0.58 ± 0.08	0.00 ± 0.00
chest_EMG	0.01 ± 0.00	0.18 ± 0.00	0.00 ± 0.00
chest_RESP	0.70 ± 0.17	0.30 ± 0.01	0.00 ± 0.00
chest_EDA	0.73 ± 0.12	0.28 ± 0.02	0.49 ± 0.05
chest_TEMP	0.65 ± 0.14	0.23 ± 0.02	0.02 ± 0.03
wrist_ACC	0.80 ± 0.00	0.73 ± 0.10	0.75 ± 0.08
wrist_BVP	0.56 ± 0.23	0.33 ± 0.02	0.76 ± 0.08
wrist_EDA	0.72 ± 0.15	0.33 ± 0.03	0.06 ± 0.02
wrist_TEMP	0.55 ± 0.23	0.27 ± 0.00	0.01 ± 0.02

G.2 Sensitivity study

Figures 10 and 11 evaluate the sensitivity of `CARAT` to two adaptation hyperparameters using the same DaliaHAR + ShaSpec setup as the main ablations in Section 4.2. Across both the pooled diverse setting (`Div.`) and representative high-/low-reliance modalities (`w_acc`, `w_temp`), performance

remains stable across a broad range of episodic TTA batch sizes and routing thresholds. In particular, varying the routing threshold ν across $\{0.5, 0.7, 0.8, 0.9\}$ changes macro-F1 by less than 2% relative, suggesting that the routing mechanism depends more on the coarse separation between high- and low-reliance corruption regimes than on precise threshold tuning. We fix $\nu = 0.8$ across all datasets, backbones, and corruption types in the main results. More importantly, ν exposes a useful system-level knob to the deployer: smaller ν shifts more suspect modalities into the modality-composition path (drop and replace with s_0), which avoids the per-modality attenuator forward pass and reduces test-time compute, at the cost of being more aggressive only toward genuinely low-reliance modalities; larger ν shifts more suspect modalities into the latent-attenuation path, preserving partially informative streams at modestly higher cost. This makes ν less of a hyperparameter to tune for accuracy and more of a deployment-time lever to trade efficiency against conservatism, with $\nu = 0.8$ providing a balanced default.

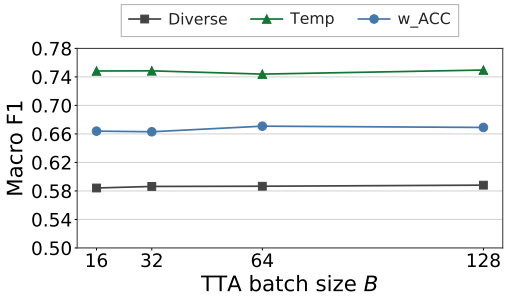


Figure 10: **Batch-size sensitivity.** Macro-F1 under varying episodic TTA batch sizes B using the same DaliaHAR + ShaSpec setup as the ablation analyses. Performance remains stable across batch sizes on `Div.`, `w_temp`, and `w_acc`.

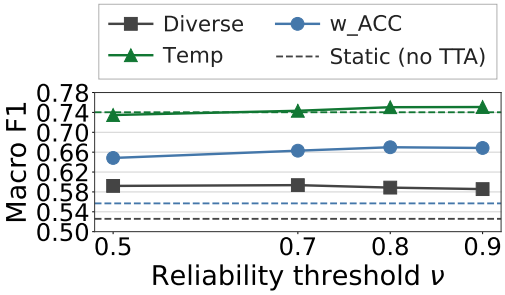


Figure 11: **Routing-threshold sensitivity.** Macro-F1 under varying routing thresholds ν using the same DaliaHAR + ShaSpec setup as the ablation analyses. Performance varies by less than 2% relative macro-F1 across $\nu \in \{0.5, 0.7, 0.8, 0.9\}$.

H Complete Results on DSADS dataset

For completeness, we provide the full DSADS results behind the aggregate numbers in the main text. Tables 12, 13, and 14 show the per-condition F1 scores for MAESTRO, ShaSpec, and Flex-MoE across corrupted modality subsets and noise types.

I Complete Results on DaliaHAR dataset

For completeness, we provide the full DaliaHAR results behind the aggregate numbers in the main text. Tables 15, 16, and 17 show the per-condition F1 scores for MAESTRO, ShaSpec, and Flex-MoE across corrupted modality subsets and noise types.

J Complete Results on PAMAP2 dataset

For completeness, we provide the full PAMAP2 results behind the aggregate numbers in the main text. Tables 18, 19, and 20 show the per-condition F1 scores for MAESTRO, ShaSpec, and Flex-MoE across corrupted modality subsets and noise types.

K Complete Results on WESAD dataset

For completeness, we provide the full WESAD results behind the aggregate numbers in the main text. Tables 21, 22, and 23 show the per-condition F1 scores for MAESTRO, ShaSpec, and Flex-MoE across corrupted modality subsets and noise types.

Table 12: Full per-modality-subset, per-noise F1-score (↑) on **DSADS** with the **MAESTRO** backbone. Each cell is the F1 for one specific (corrupted modality subset, noise type) combination. Noise abbreviations: *miss*=discrete missingness, *wave*=waveform anomaly, *nonlin*=nonlinear degradation, *freq*=frequency attenuation, *nonstat*=nonstationary noise. Clean baseline: F1 = 0.856.

Card.	**Subset**	**Noise**	Static	TENT	EATA	READ	SuMi	SoTTA	AeO	PTA	PSDNorm	**Ours**
1-mod	torso	miss	0.803	0.799	0.788	0.796	0.779	0.792	0.786	0.805	0.741	0.794
		wave	0.763	0.772	0.782	0.757	0.761	0.787	0.759	0.767	0.734	0.799
		nonlin	0.850	0.854	0.851	0.847	0.849	0.853	0.849	0.856	0.822	0.840
		freq	0.850	0.854	0.862	0.862	0.843	0.865	0.855	0.859	0.819	0.856
		nonstat	0.809	0.783	0.790	0.793	0.783	0.787	0.783	0.789	0.745	0.804
	right_arm	miss	0.823	0.826	0.835	0.827	0.820	0.807	0.832	0.829	0.750	0.824
		wave	0.798	0.797	0.802	0.795	0.796	0.810	0.796	0.808	0.755	0.821
		nonlin	0.851	0.853	0.862	0.846	0.843	0.855	0.851	0.863	0.795	0.856
		freq	0.842	0.852	0.858	0.843	0.848	0.853	0.850	0.851	0.800	0.861
		nonstat	0.836	0.828	0.835	0.834	0.823	0.825	0.834	0.840	0.754	0.840
	left_arm	miss	0.794	0.802	0.792	0.792	0.784	0.789	0.798	0.794	0.714	0.793
		wave	0.759	0.762	0.755	0.758	0.762	0.772	0.757	0.783	0.714	0.774
		nonlin	0.857	0.851	0.855	0.850	0.851	0.859	0.855	0.869	0.772	0.865
		freq	0.830	0.844	0.833	0.834	0.821	0.843	0.837	0.848	0.777	0.837
		nonstat	0.803	0.798	0.802	0.795	0.802	0.793	0.784	0.810	0.718	0.795
	right_leg	miss	0.764	0.774	0.763	0.748	0.748	0.766	0.763	0.778	0.728	0.838
		wave	0.705	0.723	0.730	0.680	0.711	0.728	0.722	0.761	0.689	0.822
		nonlin	0.850	0.861	0.857	0.853	0.856	0.853	0.855	0.862	0.780	0.835
		freq	0.829	0.828	0.830	0.825	0.842	0.838	0.853	0.834	0.787	0.811
		nonstat	0.817	0.811	0.822	0.803	0.801	0.816	0.817	0.814	0.740	0.834
	left_leg	miss	0.808	0.797	0.802	0.792	0.799	0.806	0.801	0.808	0.721	0.786
		wave	0.713	0.729	0.749	0.730	0.716	0.721	0.714	0.758	0.685	0.731
		nonlin	0.837	0.851	0.856	0.853	0.845	0.841	0.844	0.862	0.767	0.851
		freq	0.838	0.848	0.835	0.836	0.841	0.850	0.847	0.847	0.785	0.848
		nonstat	0.812	0.826	0.809	0.801	0.807	0.814	0.798	0.822	0.744	0.811
2-mod	arms	miss	0.720	0.727	0.722	0.700	0.699	0.714	0.704	0.713	0.610	0.730
		wave	0.640	0.640	0.643	0.631	0.633	0.642	0.638	0.668	0.589	0.670
		nonlin	0.855	0.853	0.855	0.848	0.852	0.858	0.860	0.867	0.727	0.850
		freq	0.812	0.825	0.834	0.822	0.817	0.835	0.822	0.835	0.750	0.822
		nonstat	0.705	0.713	0.706	0.709	0.679	0.695	0.687	0.712	0.630	0.739
	legs	miss	0.602	0.592	0.603	0.605	0.603	0.616	0.618	0.635	0.412	0.687
		wave	0.344	0.338	0.337	0.357	0.335	0.349	0.373	0.387	0.345	0.527
		nonlin	0.839	0.844	0.852	0.845	0.853	0.852	0.843	0.855	0.679	0.839
		freq	0.767	0.788	0.808	0.779	0.780	0.799	0.787	0.806	0.665	0.781
		nonstat	0.680	0.668	0.668	0.661	0.657	0.667	0.648	0.684	0.547	0.674
3-mod	torso+arms	miss	0.541	0.531	0.529	0.518	0.515	0.520	0.544	0.531	0.410	0.622
		wave	0.377	0.378	0.373	0.359	0.379	0.384	0.386	0.422	0.331	0.532
		nonlin	0.841	0.856	0.856	0.831	0.845	0.843	0.851	0.849	0.704	0.841
		freq	0.757	0.796	0.785	0.788	0.782	0.787	0.783	0.796	0.711	0.774
		nonstat	0.475	0.460	0.458	0.475	0.455	0.469	0.445	0.499	0.494	0.561
	torso+legs	miss	0.466	0.432	0.426	0.445	0.423	0.469	0.462	0.497	0.168	0.581
		wave	0.112	0.089	0.092	0.095	0.089	0.115	0.112	0.121	0.113	0.267
		nonlin	0.834	0.849	0.840	0.837	0.819	0.831	0.831	0.844	0.550	0.821
		freq	0.701	0.742	0.732	0.722	0.722	0.743	0.746	0.747	0.577	0.731
		nonstat	0.432	0.428	0.432	0.417	0.411	0.428	0.432	0.446	0.338	0.524
5-mod	all	miss	0.170	0.162	0.152	0.174	0.157	0.164	0.169	0.172	0.026	0.184
		wave	0.012	0.009	0.009	0.011	0.016	0.017	0.011	0.016	0.016	0.018
		nonlin	0.822	0.824	0.827	0.827	0.824	0.831	0.828	0.834	0.317	0.812
		freq	0.594	0.595	0.591	0.610	0.612	0.599	0.610	0.622	0.404	0.560
		nonstat	0.100	0.094	0.100	0.103	0.097	0.102	0.099	0.106	0.032	0.122

Table 13: Full per-modality-subset, per-noise F1-score (↑) on **DSADS** with the **ShaSpec** backbone. Each cell is the F1 for one specific (corrupted modality subset, noise type) combination. Noise abbreviations: *miss*=discrete missingness, *wave*=waveform anomaly, *nonlin*=nonlinear degradation, *freq*=frequency attenuation, *nonstat*=nonstationary noise. Clean baseline: F1 = 0.765.

Card.	**Subset**	**Noise**	Static	TENT	EATA	READ	SuMi	SoTTA	AeO	PTA	PSDNorm	**Ours**
1-mod	torso	miss	0.531	0.520	0.521	0.506	0.516	0.522	0.529	0.519	0.541	0.708
		wave	0.491	0.468	0.475	0.462	0.463	0.487	0.476	0.493	0.585	0.689
		nonlin	0.617	0.602	0.632	0.615	0.610	0.625	0.618	0.611	0.686	0.720
		freq	0.748	0.726	0.742	0.709	0.740	0.735	0.715	0.751	0.706	0.721
		nonstat	0.719	0.676	0.699	0.658	0.693	0.688	0.670	0.696	0.653	0.727
	right_arm	miss	0.582	0.587	0.567	0.463	0.560	0.555	0.543	0.580	0.639	0.680
		wave	0.497	0.496	0.478	0.451	0.469	0.461	0.474	0.490	0.593	0.647
		nonlin	0.658	0.673	0.659	0.658	0.656	0.657	0.638	0.662	0.680	0.707
		freq	0.775	0.763	0.754	0.694	0.764	0.763	0.753	0.773	0.723	0.738
		nonstat	0.705	0.700	0.693	0.662	0.700	0.703	0.689	0.707	0.667	0.711
	left_arm	miss	0.496	0.496	0.512	0.472	0.497	0.490	0.488	0.527	0.636	0.667
		wave	0.552	0.554	0.552	0.542	0.546	0.559	0.515	0.559	0.600	0.679
		nonlin	0.677	0.648	0.668	0.603	0.663	0.678	0.622	0.668	0.681	0.725
		freq	0.764	0.754	0.755	0.686	0.755	0.728	0.751	0.752	0.723	0.743
		nonstat	0.727	0.722	0.701	0.647	0.703	0.705	0.698	0.723	0.680	0.716
	right_leg	miss	0.246	0.192	0.228	0.191	0.255	0.253	0.338	0.232	0.467	0.553
		wave	0.178	0.118	0.148	0.124	0.185	0.167	0.257	0.145	0.224	0.508
		nonlin	0.520	0.522	0.515	0.412	0.538	0.531	0.507	0.561	0.586	0.644
		freq	0.758	0.736	0.736	0.671	0.725	0.720	0.727	0.739	0.638	0.733
		nonstat	0.528	0.559	0.550	0.513	0.527	0.564	0.500	0.579	0.524	0.628
	left_leg	miss	0.229	0.106	0.112	0.127	0.224	0.191	0.391	0.210	0.541	0.585
		wave	0.154	0.122	0.137	0.144	0.171	0.149	0.210	0.138	0.228	0.479
		nonlin	0.549	0.526	0.541	0.512	0.546	0.552	0.548	0.569	0.618	0.650
		freq	0.761	0.747	0.740	0.694	0.744	0.718	0.743	0.748	0.671	0.734
		nonstat	0.593	0.608	0.592	0.573	0.571	0.608	0.525	0.625	0.544	0.666
2-mod	arms	miss	0.259	0.240	0.240	0.242	0.254	0.270	0.277	0.248	0.450	0.546
		wave	0.299	0.275	0.299	0.294	0.275	0.310	0.306	0.276	0.416	0.554
		nonlin	0.569	0.533	0.545	0.502	0.537	0.550	0.553	0.573	0.591	0.676
		freq	0.760	0.758	0.738	0.676	0.743	0.755	0.727	0.747	0.718	0.724
		nonstat	0.652	0.650	0.654	0.591	0.629	0.625	0.612	0.656	0.590	0.675
	legs	miss	0.038	0.017	0.012	0.011	0.039	0.027	0.060	0.025	0.116	0.067
		wave	0.080	0.056	0.055	0.046	0.072	0.064	0.096	0.061	0.066	0.123
		nonlin	0.243	0.173	0.176	0.173	0.235	0.241	0.319	0.248	0.246	0.360
		freq	0.685	0.668	0.660	0.653	0.679	0.674	0.665	0.679	0.429	0.689
		nonstat	0.109	0.042	0.054	0.034	0.083	0.101	0.176	0.090	0.126	0.195
3-mod	torso+arms	miss	0.212	0.207	0.212	0.109	0.212	0.208	0.218	0.223	0.343	0.436
		wave	0.168	0.157	0.168	0.138	0.177	0.173	0.189	0.157	0.277	0.469
		nonlin	0.485	0.424	0.414	0.361	0.478	0.455	0.478	0.470	0.490	0.609
		freq	0.728	0.723	0.727	0.658	0.696	0.699	0.706	0.722	0.668	0.697
		nonstat	0.525	0.499	0.511	0.448	0.508	0.493	0.466	0.533	0.475	0.610
	torso+legs	miss	0.028	0.024	0.028	0.042	0.032	0.033	0.046	0.027	0.062	0.041
		wave	0.044	0.042	0.044	0.050	0.043	0.042	0.052	0.043	0.034	0.068
		nonlin	0.263	0.215	0.214	0.204	0.267	0.250	0.303	0.254	0.174	0.278
		freq	0.684	0.681	0.675	0.624	0.670	0.660	0.656	0.675	0.366	0.670
		nonstat	0.076	0.052	0.058	0.048	0.079	0.088	0.086	0.057	0.071	0.112
5-mod	all	miss	0.018	0.016	0.018	0.020	0.022	0.018	0.025	0.017	0.012	0.017
		wave	0.024	0.024	0.024	0.022	0.031	0.030	0.028	0.021	0.016	0.024
		nonlin	0.153	0.083	0.123	0.092	0.155	0.148	0.215	0.096	0.097	0.140
		freq	0.653	0.645	0.632	0.589	0.644	0.645	0.620	0.635	0.247	0.631
		nonstat	0.021	0.014	0.020	0.019	0.032	0.027	0.034	0.015	0.030	0.022

Table 14: Full per-modality-subset, per-noise F1-score (↑) on **DSADS** with the **Flex-MoE** backbone. Each cell is the F1 for one specific (corrupted modality subset, noise type) combination. Noise abbreviations: *miss*=discrete missingness, *wave*=waveform anomaly, *nonlin*=nonlinear degradation, *freq*=frequency attenuation, *nonstat*=nonstationary noise. Clean baseline: F1 = 0.740.

Card.	**Subset**	**Noise**	Static	TENT	EATA	READ	SuMi	SoTTA	AeO	PTA	PSDNorm	**Ours**
1-mod	torso	miss	0.680	0.691	0.695	0.671	0.680	0.701	0.679	0.682	0.644	0.719
		wave	0.670	0.697	0.706	0.696	0.700	0.704	0.698	0.710	0.613	0.719
		nonlin	0.693	0.693	0.682	0.679	0.683	0.698	0.684	0.687	0.637	0.709
		freq	0.703	0.733	0.733	0.737	0.725	0.738	0.715	0.739	0.660	0.718
		nonstat	0.670	0.713	0.705	0.690	0.685	0.709	0.694	0.703	0.626	0.685
	right_arm	miss	0.683	0.697	0.686	0.696	0.688	0.694	0.675	0.676	0.646	0.699
		wave	0.655	0.689	0.687	0.685	0.673	0.692	0.674	0.680	0.610	0.698
		nonlin	0.687	0.716	0.704	0.711	0.715	0.698	0.703	0.716	0.659	0.692
		freq	0.733	0.747	0.767	0.765	0.762	0.748	0.733	0.752	0.681	0.753
		nonstat	0.683	0.707	0.700	0.698	0.717	0.693	0.684	0.704	0.643	0.694
	left_arm	miss	0.685	0.667	0.675	0.698	0.669	0.674	0.665	0.690	0.619	0.706
		wave	0.703	0.718	0.718	0.712	0.706	0.699	0.705	0.715	0.647	0.712
		nonlin	0.700	0.707	0.700	0.706	0.712	0.702	0.699	0.702	0.640	0.742
		freq	0.746	0.759	0.760	0.737	0.752	0.738	0.754	0.763	0.685	0.746
		nonstat	0.720	0.730	0.727	0.704	0.713	0.714	0.721	0.724	0.651	0.712
	right_leg	miss	0.627	0.668	0.659	0.663	0.637	0.663	0.633	0.665	0.534	0.705
		wave	0.640	0.645	0.624	0.655	0.652	0.659	0.642	0.658	0.535	0.704
		nonlin	0.736	0.771	0.763	0.744	0.732	0.744	0.734	0.754	0.597	0.725
		freq	0.671	0.681	0.697	0.715	0.693	0.682	0.688	0.705	0.567	0.672
		nonstat	0.532	0.575	0.569	0.646	0.559	0.555	0.563	0.604	0.469	0.663
	left_leg	miss	0.660	0.667	0.668	0.660	0.662	0.666	0.626	0.677	0.579	0.678
		wave	0.631	0.615	0.614	0.617	0.613	0.614	0.608	0.624	0.559	0.674
		nonlin	0.698	0.680	0.679	0.661	0.682	0.679	0.658	0.696	0.610	0.712
		freq	0.696	0.721	0.709	0.731	0.720	0.714	0.708	0.713	0.608	0.708
		nonstat	0.571	0.604	0.610	0.614	0.588	0.608	0.551	0.618	0.556	0.675
2-mod	arms	miss	0.566	0.546	0.563	0.574	0.557	0.579	0.540	0.566	0.533	0.558
		wave	0.555	0.586	0.582	0.583	0.568	0.569	0.557	0.588	0.524	0.578
		nonlin	0.627	0.626	0.644	0.634	0.638	0.635	0.641	0.632	0.581	0.626
		freq	0.732	0.757	0.752	0.746	0.750	0.745	0.763	0.761	0.669	0.744
		nonstat	0.634	0.658	0.660	0.637	0.647	0.640	0.646	0.650	0.608	0.662
	legs	miss	0.317	0.313	0.302	0.355	0.337	0.351	0.340	0.341	0.236	0.336
		wave	0.281	0.267	0.279	0.271	0.285	0.260	0.283	0.278	0.222	0.271
		nonlin	0.556	0.519	0.529	0.517	0.528	0.514	0.518	0.540	0.302	0.519
		freq	0.556	0.584	0.566	0.639	0.605	0.578	0.589	0.586	0.325	0.594
		nonstat	0.179	0.184	0.184	0.174	0.189	0.204	0.219	0.196	0.182	0.194
3-mod	torso+arms	miss	0.293	0.361	0.340	0.352	0.347	0.353	0.363	0.370	0.396	0.370
		wave	0.354	0.387	0.394	0.405	0.392	0.403	0.394	0.404	0.354	0.402
		nonlin	0.552	0.539	0.534	0.552	0.533	0.545	0.536	0.541	0.477	0.528
		freq	0.709	0.728	0.735	0.724	0.729	0.756	0.721	0.732	0.640	0.733
		nonstat	0.533	0.560	0.550	0.538	0.517	0.546	0.489	0.570	0.474	0.567
	torso+legs	miss	0.174	0.166	0.163	0.196	0.185	0.177	0.200	0.194	0.121	0.190
		wave	0.144	0.142	0.136	0.132	0.141	0.147	0.152	0.133	0.110	0.140
		nonlin	0.372	0.354	0.352	0.381	0.354	0.331	0.345	0.360	0.205	0.359
		freq	0.553	0.562	0.567	0.613	0.581	0.551	0.558	0.554	0.309	0.558
		nonstat	0.095	0.090	0.084	0.091	0.092	0.096	0.096	0.092	0.088	0.100
5-mod	all	miss	0.009	0.012	0.012	0.009	0.011	0.011	0.013	0.012	0.024	0.011
		wave	0.028	0.017	0.024	0.026	0.024	0.024	0.027	0.024	0.017	0.025
		nonlin	0.115	0.085	0.087	0.123	0.101	0.094	0.141	0.115	0.074	0.106
		freq	0.555	0.560	0.584	0.606	0.571	0.573	0.582	0.574	0.222	0.568
		nonstat	0.033	0.039	0.036	0.033	0.035	0.034	0.028	0.040	0.024	0.035

Table 15: Full per-modality-subset, per-noise F1-score (↑) on **DaliaHAR** with the **MAESTRO** backbone. Each cell is the F1 for one specific (corrupted modality subset, noise type) combination. Noise abbreviations: *miss*=discrete missingness, *wave*=waveform anomaly, *nonlin*=nonlinear degradation, *freq*=frequency attenuation, *nonstat*=nonstationary noise. Clean baseline: F1 = 0.788.

Card.	Subset	Noise	Static	TENT	EATA	READ	SuMi	SoTTA	AeO	PTA	PSDNorm	**Ours**
1-mod	wrist_ACC	miss	0.751	0.765	0.767	0.748	0.750	0.771	0.733	0.757	0.699	0.768
		wave	0.756	0.785	0.781	0.760	0.756	0.781	0.735	0.755	0.714	0.779
		nonlin	0.756	0.775	0.777	0.762	0.757	0.774	0.736	0.765	0.722	0.776
		freq	0.749	0.773	0.773	0.754	0.748	0.760	0.738	0.739	0.715	0.774
		nonstat	0.750	0.771	0.772	0.749	0.748	0.766	0.727	0.751	0.704	0.771
	wrist_BVP	miss	0.787	0.803	0.800	0.787	0.785	0.796	0.768	0.788	0.757	0.797
		wave	0.785	0.802	0.802	0.783	0.784	0.796	0.768	0.787	0.758	0.800
		nonlin	0.752	0.761	0.759	0.758	0.748	0.764	0.735	0.767	0.722	0.757
		freq	0.787	0.801	0.800	0.782	0.785	0.794	0.775	0.785	0.752	0.801
		nonstat	0.781	0.802	0.794	0.784	0.780	0.799	0.762	0.788	0.750	0.794
	chest_ACC	miss	0.728	0.753	0.758	0.729	0.728	0.753	0.709	0.734	0.564	0.750
		wave	0.606	0.600	0.604	0.590	0.590	0.607	0.591	0.608	0.619	0.599
		nonlin	0.742	0.770	0.771	0.745	0.742	0.765	0.719	0.739	0.616	0.758
		freq	0.735	0.754	0.752	0.729	0.733	0.740	0.717	0.732	0.651	0.745
		nonstat	0.717	0.733	0.742	0.714	0.711	0.737	0.693	0.722	0.617	0.734
	wrist_EDA	miss	0.774	0.788	0.799	0.774	0.770	0.787	0.751	0.773	0.762	0.794
		wave	0.761	0.772	0.773	0.759	0.755	0.773	0.742	0.762	0.732	0.776
		nonlin	0.791	0.811	0.813	0.795	0.790	0.809	0.773	0.789	0.758	0.805
		freq	0.789	0.805	0.807	0.791	0.785	0.802	0.768	0.793	0.758	0.805
		nonstat	0.784	0.805	0.804	0.781	0.780	0.803	0.766	0.778	0.758	0.803
	wrist_TEMP	miss	0.815	0.840	0.839	0.816	0.817	0.824	0.800	0.809	0.786	0.837
		wave	0.821	0.847	0.844	0.820	0.821	0.834	0.799	0.818	0.793	0.841
		nonlin	0.798	0.815	0.816	0.795	0.794	0.813	0.772	0.798	0.769	0.816
		freq	0.790	0.806	0.806	0.787	0.790	0.799	0.766	0.794	0.756	0.804
		nonstat	0.819	0.838	0.835	0.818	0.810	0.834	0.798	0.817	0.786	0.836
2-mod	W_ACC+ W_BVP	miss	0.749	0.765	0.767	0.744	0.752	0.767	0.729	0.755	0.692	0.762
		wave	0.761	0.784	0.788	0.757	0.759	0.776	0.735	0.751	0.713	0.787
		nonlin	0.715	0.735	0.724	0.724	0.725	0.731	0.697	0.730	0.689	0.728
		freq	0.744	0.767	0.767	0.747	0.744	0.761	0.732	0.740	0.707	0.769
		nonstat	0.745	0.752	0.764	0.747	0.744	0.757	0.720	0.752	0.697	0.761
	C_ACC+ W_ACC	miss	0.660	0.681	0.686	0.655	0.654	0.684	0.632	0.656	0.442	0.679
		wave	0.522	0.540	0.532	0.524	0.518	0.535	0.510	0.517	0.501	0.530
		nonlin	0.685	0.715	0.709	0.676	0.677	0.694	0.647	0.664	0.533	0.706
		freq	0.699	0.730	0.725	0.695	0.697	0.704	0.687	0.681	0.588	0.719
		nonstat	0.652	0.676	0.683	0.656	0.648	0.682	0.627	0.662	0.492	0.666
	W_EDA+ W_TEMP	miss	0.802	0.827	0.831	0.797	0.804	0.819	0.783	0.801	0.790	0.822
		wave	0.793	0.809	0.807	0.798	0.795	0.804	0.770	0.806	0.766	0.813
		nonlin	0.800	0.819	0.817	0.800	0.799	0.812	0.782	0.795	0.769	0.819
		freq	0.788	0.810	0.807	0.789	0.787	0.803	0.766	0.791	0.760	0.802
		nonstat	0.808	0.829	0.832	0.803	0.803	0.821	0.782	0.795	0.786	0.832
3-mod	W_ACC+ W_BVP+ C_ACC	miss	0.649	0.672	0.670	0.648	0.646	0.675	0.630	0.657	0.439	0.664
		wave	0.517	0.525	0.537	0.518	0.521	0.526	0.507	0.524	0.491	0.530
		nonlin	0.636	0.652	0.653	0.628	0.631	0.634	0.619	0.624	0.474	0.663
		freq	0.698	0.726	0.718	0.693	0.691	0.705	0.686	0.680	0.589	0.719
		nonstat	0.636	0.655	0.646	0.627	0.633	0.650	0.608	0.647	0.476	0.644
4-mod	all_wrist	miss	0.763	0.780	0.785	0.769	0.768	0.780	0.743	0.756	0.725	0.781
		wave	0.758	0.783	0.781	0.768	0.763	0.784	0.738	0.764	0.727	0.786
		nonlin	0.732	0.750	0.742	0.732	0.732	0.749	0.710	0.741	0.702	0.759
		freq	0.749	0.769	0.773	0.745	0.748	0.765	0.729	0.741	0.707	0.771
		nonstat	0.762	0.780	0.780	0.760	0.755	0.774	0.728	0.742	0.724	0.775

Table 16: Full per-modality-subset, per-noise F1-score (↑) on **DaliaHAR** with the **ShaSpec** backbone. Each cell is the F1 for one specific (corrupted modality subset, noise type) combination. Noise abbreviations: *miss*=discrete missingness, *wave*=waveform anomaly, *nonlin*=nonlinear degradation, *freq*=frequency attenuation, *nonstat*=nonstationary noise. Clean baseline: F1 = 0.741.

Card.	Subset	Noise	Static	TENT	EATA	READ	SuMi	SoTTA	AeO	PTA	PSDNorm	**Ours**
1-mod	wrist_ACC	miss	0.720	0.738	0.733	0.724	0.719	0.726	0.690	0.726	0.637	0.731
		wave	0.557	0.575	0.573	0.562	0.548	0.561	0.520	0.547	0.505	0.662
		nonlin	0.580	0.597	0.588	0.576	0.571	0.573	0.528	0.595	0.531	0.599
		freq	0.586	0.597	0.593	0.585	0.588	0.586	0.566	0.593	0.566	0.592
		nonstat	0.603	0.627	0.623	0.624	0.611	0.618	0.587	0.617	0.585	0.652
	wrist_BVP	miss	0.740	0.749	0.747	0.740	0.736	0.740	0.714	0.746	0.711	0.749
		wave	0.739	0.745	0.745	0.738	0.739	0.735	0.717	0.747	0.707	0.748
		nonlin	0.657	0.669	0.668	0.650	0.643	0.645	0.595	0.663	0.645	0.712
		freq	0.742	0.750	0.750	0.737	0.737	0.742	0.716	0.744	0.711	0.750
		nonstat	0.741	0.753	0.748	0.737	0.735	0.741	0.716	0.744	0.710	0.750
	chest_ACC	miss	0.501	0.500	0.494	0.491	0.493	0.501	0.488	0.499	0.578	0.594
		wave	0.327	0.332	0.323	0.315	0.327	0.325	0.332	0.323	0.362	0.465
		nonlin	0.524	0.529	0.529	0.512	0.512	0.513	0.481	0.518	0.469	0.589
		freq	0.700	0.706	0.705	0.706	0.694	0.700	0.669	0.699	0.643	0.729
		nonstat	0.569	0.557	0.556	0.557	0.556	0.565	0.543	0.561	0.565	0.646
	wrist_EDA	miss	0.712	0.716	0.714	0.713	0.698	0.702	0.672	0.716	0.703	0.734
		wave	0.708	0.713	0.708	0.715	0.695	0.694	0.669	0.715	0.686	0.732
		nonlin	0.740	0.751	0.745	0.741	0.734	0.739	0.710	0.747	0.710	0.748
		freq	0.741	0.752	0.750	0.736	0.738	0.741	0.717	0.747	0.713	0.747
		nonstat	0.715	0.719	0.720	0.717	0.706	0.708	0.689	0.721	0.702	0.740
	wrist_TEMP	miss	0.740	0.753	0.750	0.740	0.737	0.743	0.713	0.746	0.711	0.748
		wave	0.740	0.750	0.752	0.736	0.734	0.743	0.716	0.747	0.711	0.751
		nonlin	0.741	0.751	0.747	0.736	0.736	0.739	0.715	0.744	0.711	0.749
		freq	0.741	0.753	0.748	0.735	0.735	0.742	0.710	0.746	0.711	0.753
		nonstat	0.740	0.754	0.749	0.737	0.737	0.745	0.713	0.746	0.711	0.751
2-mod	W_ACC+ W_BVP	miss	0.719	0.735	0.730	0.729	0.717	0.720	0.686	0.724	0.634	0.730
		wave	0.558	0.577	0.577	0.562	0.551	0.566	0.520	0.547	0.499	0.666
		nonlin	0.436	0.437	0.447	0.429	0.420	0.427	0.414	0.427	0.449	0.518
		freq	0.584	0.599	0.590	0.592	0.591	0.586	0.568	0.595	0.563	0.596
		nonstat	0.600	0.620	0.616	0.615	0.606	0.614	0.584	0.610	0.583	0.650
	C_ACC+ W_ACC	miss	0.481	0.481	0.478	0.473	0.474	0.479	0.462	0.479	0.514	0.561
		wave	0.209	0.208	0.209	0.209	0.209	0.207	0.211	0.210	0.173	0.229
		nonlin	0.298	0.298	0.302	0.302	0.299	0.300	0.286	0.304	0.247	0.306
		freq	0.522	0.525	0.523	0.521	0.519	0.519	0.503	0.514	0.442	0.503
		nonstat	0.497	0.514	0.508	0.499	0.494	0.496	0.456	0.503	0.438	0.527
	W_EDA+ W_TEMP	miss	0.712	0.715	0.714	0.715	0.699	0.697	0.665	0.716	0.701	0.736
		wave	0.707	0.714	0.708	0.712	0.691	0.696	0.666	0.709	0.685	0.732
		nonlin	0.740	0.747	0.745	0.741	0.734	0.741	0.713	0.742	0.709	0.749
		freq	0.741	0.753	0.752	0.739	0.735	0.739	0.716	0.748	0.712	0.752
		nonstat	0.715	0.721	0.721	0.716	0.708	0.708	0.688	0.720	0.700	0.739
3-mod	W_ACC+ W_BVP+ C_ACC	miss	0.478	0.485	0.478	0.474	0.477	0.481	0.463	0.476	0.510	0.548
		wave	0.211	0.211	0.210	0.216	0.207	0.212	0.210	0.209	0.160	0.226
		nonlin	0.261	0.259	0.267	0.259	0.256	0.256	0.248	0.264	0.238	0.270
		freq	0.520	0.525	0.519	0.519	0.517	0.518	0.502	0.520	0.442	0.498
		nonstat	0.491	0.504	0.504	0.493	0.486	0.487	0.450	0.498	0.433	0.523
4-mod	all_wrist	miss	0.683	0.692	0.690	0.685	0.673	0.675	0.647	0.688	0.620	0.709
		wave	0.502	0.508	0.510	0.497	0.487	0.493	0.467	0.497	0.460	0.632
		nonlin	0.428	0.420	0.435	0.417	0.416	0.423	0.411	0.426	0.446	0.510
		freq	0.585	0.602	0.589	0.595	0.586	0.588	0.566	0.591	0.564	0.595
		nonstat	0.563	0.584	0.577	0.581	0.570	0.571	0.544	0.573	0.566	0.627

Table 17: Full per-modality-subset, per-noise F1-score (↑) on **DaliaHAR** with the **Flex-MoE** backbone. Each cell is the F1 for one specific (corrupted modality subset, noise type) combination. Noise abbreviations: *miss*=discrete missingness, *wave*=waveform anomaly, *nonlin*=nonlinear degradation, *freq*=frequency attenuation, *nonstat*=nonstationary noise. Clean baseline: F1 = 0.768.

Card.	Subset	Noise	Static	TENT	EATA	READ	SuMi	SoTTA	AeO	PTA	PSDNorm	**Ours**
1-mod	wrist_ACC	miss	0.743	0.787	0.784	0.748	0.758	0.780	0.726	0.711	0.693	0.788
		wave	0.622	0.653	0.656	0.608	0.637	0.680	0.590	0.552	0.561	0.654
		nonlin	0.624	0.644	0.650	0.608	0.659	0.698	0.627	0.595	0.608	0.650
		freq	0.675	0.736	0.738	0.692	0.676	0.710	0.631	0.613	0.612	0.729
		nonstat	0.586	0.631	0.628	0.604	0.611	0.639	0.573	0.554	0.571	0.625
	wrist_BVP	miss	0.767	0.801	0.809	0.786	0.787	0.801	0.752	0.738	0.745	0.808
		wave	0.767	0.801	0.806	0.782	0.780	0.795	0.759	0.747	0.743	0.801
		nonlin	0.699	0.750	0.753	0.726	0.678	0.697	0.651	0.634	0.687	0.757
		freq	0.770	0.807	0.811	0.788	0.774	0.788	0.755	0.744	0.744	0.800
		nonstat	0.770	0.807	0.807	0.781	0.782	0.799	0.760	0.742	0.744	0.798
	chest_ACC	miss	0.644	0.698	0.703	0.670	0.633	0.638	0.599	0.625	0.638	0.701
		wave	0.510	0.501	0.508	0.498	0.473	0.468	0.461	0.462	0.495	0.508
		nonlin	0.549	0.571	0.572	0.537	0.560	0.592	0.519	0.510	0.482	0.571
		freq	0.731	0.769	0.774	0.755	0.733	0.755	0.706	0.701	0.687	0.769
		nonstat	0.665	0.708	0.718	0.686	0.636	0.638	0.607	0.633	0.636	0.705
	wrist_EDA	miss	0.707	0.760	0.760	0.746	0.713	0.736	0.696	0.702	0.742	0.753
		wave	0.716	0.772	0.769	0.754	0.735	0.763	0.715	0.693	0.717	0.764
		nonlin	0.759	0.804	0.797	0.775	0.778	0.803	0.744	0.730	0.735	0.804
		freq	0.768	0.812	0.807	0.793	0.787	0.812	0.756	0.738	0.743	0.803
		nonstat	0.731	0.788	0.784	0.748	0.762	0.779	0.725	0.705	0.734	0.772
	wrist_TEMP	miss	0.766	0.792	0.794	0.789	0.774	0.793	0.734	0.720	0.737	0.792
		wave	0.763	0.790	0.795	0.778	0.783	0.807	0.758	0.737	0.749	0.798
		nonlin	0.766	0.810	0.801	0.775	0.783	0.802	0.762	0.732	0.742	0.808
		freq	0.767	0.805	0.805	0.793	0.783	0.802	0.756	0.737	0.744	0.803
		nonstat	0.753	0.792	0.799	0.797	0.768	0.784	0.720	0.713	0.733	0.798
2-mod	W_ACC+ W_BVP	miss	0.743	0.783	0.788	0.766	0.753	0.776	0.722	0.709	0.692	0.784
		wave	0.619	0.646	0.649	0.614	0.610	0.639	0.574	0.545	0.560	0.653
		nonlin	0.496	0.518	0.520	0.529	0.422	0.431	0.428	0.394	0.504	0.519
		freq	0.673	0.730	0.735	0.695	0.676	0.705	0.640	0.621	0.609	0.732
		nonstat	0.579	0.626	0.619	0.599	0.605	0.617	0.568	0.568	0.565	0.611
	C_ACC+ W_ACC	miss	0.599	0.644	0.656	0.635	0.575	0.574	0.553	0.560	0.554	0.651
		wave	0.300	0.296	0.293	0.269	0.279	0.284	0.289	0.265	0.268	0.298
		nonlin	0.277	0.270	0.272	0.261	0.255	0.267	0.260	0.238	0.280	0.273
		freq	0.633	0.699	0.695	0.663	0.632	0.655	0.583	0.585	0.499	0.699
		nonstat	0.468	0.515	0.518	0.471	0.443	0.450	0.413	0.390	0.421	0.518
	W_EDA+ W_TEMP	miss	0.688	0.736	0.743	0.707	0.712	0.730	0.693	0.688	0.735	0.741
		wave	0.695	0.754	0.747	0.732	0.701	0.732	0.681	0.678	0.710	0.751
		nonlin	0.757	0.793	0.798	0.757	0.771	0.788	0.743	0.724	0.728	0.797
		freq	0.768	0.801	0.803	0.779	0.783	0.806	0.757	0.746	0.744	0.799
		nonstat	0.710	0.774	0.775	0.745	0.716	0.749	0.687	0.656	0.726	0.762
3-mod	W_ACC+ W_BVP+ C_ACC	miss	0.592	0.644	0.638	0.622	0.589	0.604	0.560	0.562	0.547	0.642
		wave	0.305	0.298	0.296	0.274	0.261	0.251	0.264	0.250	0.261	0.295
		nonlin	0.224	0.204	0.203	0.206	0.185	0.183	0.200	0.199	0.250	0.214
		freq	0.632	0.699	0.698	0.646	0.635	0.666	0.582	0.587	0.498	0.692
		nonstat	0.464	0.515	0.511	0.468	0.453	0.457	0.418	0.414	0.417	0.514
4-mod	all_wrist	miss	0.686	0.749	0.748	0.730	0.739	0.756	0.696	0.691	0.676	0.747
		wave	0.612	0.644	0.636	0.609	0.576	0.597	0.549	0.533	0.577	0.650
		nonlin	0.463	0.486	0.481	0.493	0.407	0.414	0.414	0.385	0.457	0.483
		freq	0.671	0.743	0.747	0.689	0.688	0.713	0.642	0.628	0.608	0.731
		nonstat	0.545	0.600	0.596	0.572	0.519	0.554	0.476	0.479	0.540	0.589

Table 18: Full per-modality-subset, per-noise F1-score (↑) on **PAMAP2** with the **MAESTRO** backbone. Each cell is the F1 for one specific (corrupted modality subset, noise type) combination. Noise abbreviations: *miss*=discrete missingness, *wave*=waveform anomaly, *nonlin*=nonlinear degradation, *freq*=frequency attenuation, *nonstat*=nonstationary noise. Clean baseline: F1 = 0.951.

Card.	**Subset**	**Noise**	Static	TENT	EATA	READ	SuMi	SoTTA	AeO	PTA	PSDNorm	**Ours**
1-mod	IMU_ankle	miss	0.940	0.938	0.946	0.943	0.935	0.940	0.937	0.941	0.901	0.935
		wave	0.926	0.931	0.932	0.926	0.927	0.929	0.924	0.931	0.897	0.925
		nonlin	0.940	0.933	0.938	0.932	0.941	0.935	0.938	0.943	0.903	0.939
		freq	0.941	0.942	0.942	0.936	0.938	0.938	0.939	0.939	0.918	0.943
		nonstat	0.943	0.938	0.943	0.936	0.932	0.937	0.937	0.942	0.892	0.932
	IMU_hand	miss	0.944	0.941	0.945	0.933	0.937	0.944	0.939	0.945	0.910	0.946
		wave	0.945	0.948	0.946	0.939	0.939	0.943	0.944	0.943	0.926	0.948
		nonlin	0.949	0.943	0.948	0.943	0.949	0.947	0.949	0.947	0.914	0.947
		freq	0.937	0.936	0.939	0.941	0.943	0.941	0.944	0.934	0.929	0.937
		nonstat	0.943	0.937	0.944	0.937	0.933	0.941	0.937	0.938	0.923	0.935
	IMU_chest	miss	0.943	0.940	0.943	0.933	0.941	0.940	0.938	0.942	0.906	0.941
		wave	0.915	0.902	0.907	0.906	0.907	0.901	0.906	0.909	0.895	0.921
		nonlin	0.946	0.947	0.941	0.939	0.945	0.942	0.945	0.936	0.927	0.945
		freq	0.944	0.944	0.944	0.945	0.945	0.943	0.944	0.951	0.935	0.943
		nonstat	0.935	0.936	0.934	0.927	0.926	0.935	0.929	0.931	0.913	0.926
	heart_rate	miss	0.950	0.948	0.950	0.947	0.949	0.953	0.947	0.947	0.947	0.945
		wave	0.952	0.948	0.950	0.948	0.948	0.949	0.946	0.950	0.947	0.943
		nonlin	0.951	0.951	0.951	0.948	0.948	0.947	0.948	0.947	0.947	0.948
		freq	0.953	0.946	0.949	0.949	0.946	0.946	0.949	0.950	0.946	0.949
		nonstat	0.950	0.947	0.949	0.944	0.946	0.946	0.948	0.945	0.944	0.949
2-mod	ankle+ hand	miss	0.933	0.933	0.932	0.920	0.927	0.926	0.923	0.932	0.795	0.930
		wave	0.916	0.915	0.920	0.908	0.912	0.917	0.907	0.916	0.827	0.927
		nonlin	0.928	0.921	0.926	0.925	0.924	0.926	0.921	0.932	0.805	0.927
		freq	0.889	0.911	0.928	0.902	0.905	0.905	0.891	0.918	0.860	0.923
		nonstat	0.922	0.915	0.922	0.911	0.903	0.920	0.911	0.921	0.807	0.923
	ankle+ HR	miss	0.941	0.944	0.944	0.935	0.939	0.934	0.940	0.940	0.901	0.936
		wave	0.931	0.933	0.935	0.934	0.930	0.931	0.926	0.931	0.901	0.931
		nonlin	0.943	0.943	0.943	0.932	0.938	0.943	0.938	0.938	0.898	0.939
		freq	0.937	0.945	0.944	0.940	0.936	0.941	0.938	0.942	0.918	0.941
		nonstat	0.938	0.940	0.940	0.934	0.938	0.936	0.936	0.939	0.901	0.936
3-mod	ankle+ hand+ chest	miss	0.896	0.889	0.893	0.883	0.884	0.889	0.881	0.895	0.698	0.880
		wave	0.803	0.805	0.801	0.800	0.791	0.797	0.774	0.807	0.690	0.791
		nonlin	0.901	0.889	0.904	0.885	0.885	0.896	0.887	0.899	0.671	0.894
		freq	0.832	0.855	0.858	0.838	0.835	0.857	0.829	0.850	0.793	0.855
		nonstat	0.848	0.834	0.839	0.848	0.837	0.847	0.842	0.850	0.695	0.841
	ankle+ hand+ HR	miss	0.931	0.927	0.931	0.919	0.921	0.932	0.924	0.931	0.791	0.936
		wave	0.915	0.918	0.924	0.911	0.918	0.917	0.914	0.923	0.834	0.919
		nonlin	0.933	0.925	0.924	0.922	0.916	0.923	0.918	0.929	0.799	0.932
		freq	0.892	0.921	0.911	0.903	0.901	0.906	0.901	0.914	0.857	0.925
		nonstat	0.925	0.924	0.922	0.914	0.917	0.919	0.909	0.925	0.813	0.920
4-mod	all	miss	0.885	0.884	0.883	0.883	0.878	0.888	0.882	0.882	0.696	0.890
		wave	0.798	0.803	0.803	0.795	0.806	0.797	0.776	0.803	0.692	0.788
		nonlin	0.892	0.887	0.895	0.880	0.887	0.889	0.886	0.906	0.675	0.890
		freq	0.830	0.858	0.852	0.845	0.833	0.853	0.829	0.864	0.788	0.847
		nonstat	0.845	0.837	0.833	0.838	0.839	0.835	0.843	0.848	0.704	0.840

Table 19: Full per-modality-subset, per-noise F1-score (↑) on **PAMAP2** with the **ShaSpec** backbone. Each cell is the F1 for one specific (corrupted modality subset, noise type) combination. Noise abbreviations: *miss*=discrete missingness, *wave*=waveform anomaly, *nonlin*=nonlinear degradation, *freq*=frequency attenuation, *nonstat*=nonstationary noise. Clean baseline: F1 = 0.862.

Card.	**Subset**	**Noise**	Static	TENT	EATA	READ	SuMi	SoTTA	AeO	PTA	PSDNorm	**Ours**
1-mod	IMU_ankle	miss	0.819	0.812	0.808	0.763	0.802	0.811	0.803	0.803	0.804	0.824
		wave	0.779	0.764	0.764	0.725	0.770	0.759	0.757	0.767	0.784	0.798
		nonlin	0.270	0.223	0.239	0.214	0.267	0.242	0.289	0.265	0.292	0.511
		freq	0.860	0.850	0.845	0.807	0.851	0.843	0.833	0.843	0.849	0.847
		nonstat	0.850	0.834	0.838	0.794	0.839	0.841	0.828	0.843	0.833	0.822
	IMU_hand	miss	0.839	0.829	0.823	0.816	0.833	0.826	0.825	0.830	0.837	0.845
		wave	0.813	0.798	0.802	0.781	0.806	0.791	0.794	0.801	0.800	0.827
		nonlin	0.724	0.712	0.705	0.691	0.709	0.714	0.699	0.725	0.772	0.798
		freq	0.860	0.851	0.843	0.816	0.857	0.847	0.843	0.847	0.851	0.845
		nonstat	0.857	0.843	0.848	0.817	0.846	0.842	0.837	0.836	0.841	0.848
	IMU_chest	miss	0.784	0.788	0.786	0.759	0.781	0.780	0.766	0.790	0.798	0.794
		wave	0.700	0.693	0.688	0.701	0.695	0.705	0.691	0.701	0.690	0.746
		nonlin	0.432	0.414	0.429	0.434	0.429	0.443	0.436	0.437	0.484	0.568
		freq	0.860	0.854	0.849	0.820	0.852	0.849	0.847	0.847	0.851	0.829
		nonstat	0.800	0.796	0.798	0.763	0.792	0.796	0.788	0.802	0.828	0.783
	heart_rate	miss	0.857	0.845	0.847	0.819	0.849	0.850	0.850	0.856	0.850	0.852
		wave	0.861	0.849	0.849	0.817	0.849	0.839	0.848	0.849	0.852	0.851
		nonlin	0.863	0.851	0.850	0.820	0.853	0.851	0.852	0.848	0.856	0.859
		freq	0.864	0.853	0.846	0.826	0.854	0.847	0.843	0.856	0.859	0.857
		nonstat	0.856	0.857	0.849	0.814	0.850	0.846	0.845	0.854	0.853	0.853
2-mod	ankle+ hand	miss	0.810	0.807	0.816	0.743	0.803	0.800	0.791	0.816	0.771	0.799
		wave	0.743	0.726	0.708	0.673	0.724	0.717	0.717	0.740	0.725	0.748
		nonlin	0.186	0.159	0.145	0.138	0.174	0.164	0.215	0.163	0.224	0.321
		freq	0.852	0.845	0.841	0.813	0.840	0.839	0.844	0.835	0.834	0.839
		nonstat	0.846	0.826	0.834	0.772	0.836	0.832	0.826	0.841	0.821	0.819
	ankle+ HR	miss	0.807	0.800	0.810	0.749	0.788	0.783	0.782	0.807	0.803	0.822
		wave	0.781	0.772	0.774	0.713	0.771	0.776	0.752	0.770	0.783	0.801
		nonlin	0.264	0.220	0.230	0.209	0.259	0.236	0.274	0.282	0.294	0.517
		freq	0.859	0.849	0.846	0.807	0.850	0.846	0.838	0.844	0.852	0.849
		nonstat	0.846	0.839	0.841	0.790	0.849	0.846	0.831	0.833	0.842	0.825
3-mod	ankle+ hand+ chest	miss	0.686	0.678	0.671	0.634	0.662	0.681	0.674	0.678	0.706	0.663
		wave	0.565	0.554	0.553	0.546	0.559	0.569	0.555	0.567	0.577	0.540
		nonlin	0.060	0.056	0.055	0.051	0.064	0.064	0.069	0.062	0.121	0.063
		freq	0.838	0.829	0.827	0.791	0.839	0.832	0.838	0.838	0.819	0.834
		nonstat	0.741	0.735	0.739	0.683	0.716	0.758	0.715	0.751	0.770	0.724
	ankle+ hand+ HR	miss	0.804	0.798	0.789	0.728	0.788	0.790	0.766	0.807	0.768	0.809
		wave	0.710	0.700	0.689	0.662	0.698	0.701	0.693	0.704	0.721	0.746
		nonlin	0.183	0.145	0.147	0.132	0.178	0.160	0.208	0.162	0.223	0.308
		freq	0.853	0.844	0.847	0.808	0.842	0.846	0.838	0.839	0.836	0.847
		nonstat	0.848	0.839	0.837	0.773	0.832	0.825	0.824	0.840	0.830	0.815
4-mod	all	miss	0.682	0.667	0.683	0.627	0.667	0.666	0.652	0.682	0.690	0.669
		wave	0.539	0.530	0.523	0.510	0.532	0.529	0.531	0.532	0.563	0.529
		nonlin	0.059	0.053	0.065	0.048	0.063	0.062	0.070	0.060	0.119	0.056
		freq	0.838	0.834	0.835	0.801	0.834	0.834	0.844	0.839	0.821	0.846
		nonstat	0.744	0.733	0.734	0.700	0.727	0.752	0.730	0.756	0.775	0.728

Table 20: Full per-modality-subset, per-noise F1-score (↑) on **PAMAP2** with the **Flex-MoE** backbone. Each cell is the F1 for one specific (corrupted modality subset, noise type) combination. Noise abbreviations: *miss*=discrete missingness, *wave*=waveform anomaly, *nonlin*=nonlinear degradation, *freq*=frequency attenuation, *nonstat*=nonstationary noise. Clean baseline: F1 = 0.893.

Card.	**Subset**	**Noise**	Static	TENT	EATA	READ	SuMi	SoTTA	AeO	PTA	PSDNorm	**Ours**
1-mod	IMU_ankle	miss	0.855	0.865	0.862	0.834	0.861	0.858	0.863	0.864	0.824	0.862
		wave	0.844	0.864	0.863	0.827	0.862	0.856	0.866	0.859	0.793	0.867
		nonlin	0.709	0.685	0.689	0.684	0.693	0.680	0.678	0.683	0.608	0.685
		freq	0.862	0.874	0.880	0.849	0.866	0.882	0.870	0.885	0.819	0.877
		nonstat	0.861	0.870	0.871	0.830	0.874	0.873	0.870	0.872	0.811	0.875
	IMU_hand	miss	0.883	0.887	0.890	0.864	0.888	0.885	0.888	0.891	0.870	0.892
		wave	0.868	0.872	0.880	0.860	0.882	0.875	0.884	0.877	0.869	0.884
		nonlin	0.819	0.854	0.841	0.813	0.843	0.838	0.849	0.857	0.817	0.852
		freq	0.883	0.896	0.895	0.865	0.892	0.893	0.891	0.901	0.876	0.895
		nonstat	0.866	0.868	0.873	0.849	0.868	0.876	0.873	0.868	0.872	0.876
	IMU_chest	miss	0.805	0.829	0.818	0.798	0.817	0.815	0.818	0.814	0.851	0.818
		wave	0.775	0.776	0.772	0.751	0.769	0.772	0.764	0.770	0.749	0.769
		nonlin	0.754	0.763	0.754	0.734	0.761	0.761	0.757	0.766	0.772	0.763
		freq	0.878	0.895	0.892	0.870	0.887	0.891	0.889	0.886	0.861	0.886
		nonstat	0.809	0.811	0.812	0.809	0.805	0.793	0.812	0.815	0.825	0.802
	heart_rate	miss	0.886	0.895	0.893	0.873	0.891	0.897	0.888	0.895	0.855	0.895
		wave	0.885	0.894	0.900	0.869	0.891	0.892	0.890	0.901	0.836	0.896
		nonlin	0.892	0.899	0.896	0.868	0.898	0.898	0.890	0.895	0.890	0.903
		freq	0.891	0.903	0.896	0.872	0.900	0.894	0.897	0.892	0.883	0.901
		nonstat	0.829	0.858	0.850	0.826	0.855	0.868	0.850	0.860	0.837	0.891
2-mod	ankle+ hand	miss	0.830	0.843	0.844	0.824	0.835	0.836	0.831	0.830	0.788	0.849
		wave	0.806	0.819	0.826	0.797	0.816	0.825	0.821	0.815	0.752	0.825
		nonlin	0.494	0.462	0.450	0.468	0.458	0.453	0.459	0.459	0.423	0.463
		freq	0.857	0.868	0.868	0.847	0.864	0.864	0.867	0.870	0.780	0.872
		nonstat	0.829	0.840	0.843	0.814	0.835	0.840	0.832	0.845	0.765	0.840
	ankle+ HR	miss	0.840	0.858	0.854	0.840	0.851	0.857	0.856	0.857	0.760	0.866
		wave	0.838	0.850	0.848	0.831	0.838	0.846	0.850	0.847	0.713	0.856
		nonlin	0.710	0.686	0.683	0.677	0.689	0.673	0.680	0.693	0.617	0.699
		freq	0.860	0.875	0.876	0.861	0.878	0.874	0.877	0.875	0.813	0.870
		nonstat	0.761	0.807	0.817	0.798	0.808	0.811	0.810	0.802	0.720	0.849
3-mod	ankle+ hand+ chest	miss	0.642	0.675	0.677	0.661	0.665	0.670	0.675	0.676	0.713	0.677
		wave	0.546	0.572	0.561	0.551	0.567	0.566	0.549	0.563	0.515	0.573
		nonlin	0.146	0.142	0.145	0.159	0.146	0.145	0.141	0.139	0.196	0.140
		freq	0.848	0.856	0.857	0.830	0.852	0.855	0.859	0.851	0.750	0.851
		nonstat	0.674	0.695	0.685	0.686	0.687	0.671	0.683	0.685	0.629	0.682
	ankle+ hand+ HR	miss	0.814	0.835	0.837	0.808	0.834	0.831	0.820	0.829	0.728	0.841
		wave	0.789	0.807	0.813	0.781	0.810	0.800	0.798	0.813	0.660	0.823
		nonlin	0.486	0.468	0.458	0.463	0.460	0.463	0.458	0.454	0.434	0.457
		freq	0.856	0.874	0.868	0.845	0.865	0.873	0.869	0.866	0.771	0.867
		nonstat	0.755	0.796	0.797	0.766	0.786	0.798	0.791	0.787	0.680	0.828
4-mod	all	miss	0.635	0.669	0.665	0.650	0.669	0.662	0.657	0.664	0.655	0.677
		wave	0.514	0.534	0.542	0.533	0.536	0.530	0.522	0.535	0.413	0.545
		nonlin	0.142	0.142	0.140	0.157	0.144	0.131	0.142	0.145	0.203	0.140
		freq	0.840	0.853	0.856	0.828	0.851	0.851	0.855	0.853	0.743	0.852
		nonstat	0.605	0.664	0.644	0.647	0.646	0.641	0.644	0.654	0.580	0.675

Table 21: Full per-modality-subset, per-noise F1-score (↑) on **WESAD** with the **MAESTRO** backbone. Each cell is the F1 for one specific (corrupted modality subset, noise type) combination. Noise abbreviations: *miss*=discrete missingness, *wave*=waveform anomaly, *nonlin*=nonlinear degradation, *freq*=frequency attenuation, *nonstat*=nonstationary noise. Clean baseline: F1 = 0.666.

Card.	Subset	Noise	Static	TENT	EATA	READ	SuMi	SoTTA	AeO	PTA	PSDNorm	**Ours**
1-mod	chest_EDA	miss	0.610	0.615	0.618	0.615	0.617	0.606	0.611	0.605	0.640	0.627
		wave	0.634	0.629	0.627	0.626	0.635	0.627	0.641	0.630	0.638	0.628
		nonlin	0.660	0.656	0.656	0.662	0.653	0.659	0.666	0.644	0.652	0.632
		freq	0.657	0.658	0.655	0.654	0.657	0.661	0.662	0.645	0.652	0.624
		nonstat	0.604	0.602	0.594	0.590	0.599	0.602	0.596	0.594	0.637	0.627
	chest_ACC	miss	0.658	0.656	0.656	0.655	0.655	0.653	0.663	0.646	0.665	0.637
		wave	0.666	0.656	0.655	0.664	0.660	0.663	0.671	0.643	0.664	0.645
		nonlin	0.660	0.657	0.658	0.663	0.662	0.651	0.668	0.643	0.665	0.642
		freq	0.655	0.664	0.658	0.658	0.656	0.661	0.666	0.648	0.665	0.649
		nonstat	0.664	0.663	0.661	0.657	0.655	0.660	0.669	0.640	0.665	0.637
	wrist_ACC	miss	0.638	0.637	0.636	0.636	0.637	0.648	0.632	0.626	0.601	0.632
		wave	0.644	0.645	0.642	0.645	0.638	0.629	0.647	0.628	0.637	0.629
		nonlin	0.691	0.688	0.687	0.700	0.702	0.689	0.694	0.700	0.689	0.680
		freq	0.654	0.653	0.653	0.643	0.650	0.649	0.649	0.645	0.625	0.639
		nonstat	0.644	0.631	0.635	0.633	0.636	0.639	0.631	0.638	0.602	0.638
	wrist_EDA	miss	0.630	0.640	0.646	0.630	0.629	0.639	0.625	0.616	0.654	0.635
		wave	0.638	0.636	0.645	0.634	0.634	0.637	0.626	0.620	0.643	0.630
		nonlin	0.663	0.662	0.664	0.660	0.659	0.666	0.658	0.644	0.652	0.638
		freq	0.664	0.660	0.659	0.655	0.665	0.663	0.667	0.644	0.663	0.630
		nonstat	0.635	0.640	0.643	0.626	0.634	0.646	0.630	0.621	0.650	0.633
	chest_ECG	miss	0.658	0.662	0.657	0.657	0.660	0.661	0.657	0.647	0.659	0.643
		wave	0.659	0.659	0.665	0.663	0.659	0.659	0.670	0.645	0.658	0.642
		nonlin	0.663	0.653	0.664	0.653	0.653	0.659	0.661	0.644	0.663	0.640
		freq	0.663	0.656	0.661	0.664	0.660	0.654	0.664	0.644	0.657	0.643
		nonstat	0.661	0.663	0.660	0.658	0.659	0.658	0.661	0.646	0.660	0.640
	chest_EMG	miss	0.664	0.659	0.664	0.657	0.657	0.655	0.662	0.642	0.666	0.647
		wave	0.662	0.658	0.650	0.663	0.663	0.663	0.663	0.650	0.660	0.639
		nonlin	0.667	0.660	0.660	0.659	0.661	0.662	0.660	0.642	0.659	0.644
		freq	0.658	0.667	0.659	0.661	0.653	0.665	0.658	0.646	0.663	0.639
		nonstat	0.661	0.659	0.658	0.662	0.667	0.663	0.659	0.648	0.660	0.636
	chest_RESP	miss	0.660	0.653	0.662	0.655	0.661	0.655	0.654	0.643	0.661	0.640
		wave	0.653	0.656	0.662	0.652	0.654	0.655	0.662	0.639	0.661	0.641
		nonlin	0.633	0.641	0.635	0.646	0.630	0.628	0.624	0.631	0.633	0.620
		freq	0.662	0.663	0.655	0.664	0.666	0.653	0.663	0.646	0.661	0.628
		nonstat	0.656	0.652	0.656	0.650	0.653	0.652	0.655	0.639	0.660	0.627
	wrist_TEMP	miss	0.660	0.662	0.662	0.654	0.659	0.657	0.663	0.645	0.658	0.639
		wave	0.660	0.652	0.661	0.658	0.663	0.664	0.667	0.646	0.660	0.639
		nonlin	0.664	0.659	0.663	0.657	0.667	0.666	0.661	0.647	0.663	0.637
		freq	0.657	0.662	0.668	0.653	0.664	0.660	0.671	0.648	0.664	0.647
		nonstat	0.664	0.659	0.662	0.657	0.661	0.660	0.657	0.641	0.660	0.646
	wrist_BVP	miss	0.658	0.661	0.656	0.655	0.661	0.666	0.671	0.654	0.661	0.656
		wave	0.665	0.662	0.666	0.665	0.662	0.666	0.661	0.650	0.660	0.660
		nonlin	0.669	0.675	0.662	0.674	0.664	0.662	0.669	0.661	0.646	0.665
		freq	0.660	0.664	0.666	0.660	0.666	0.666	0.666	0.649	0.663	0.641
		nonstat	0.661	0.657	0.661	0.661	0.659	0.660	0.671	0.646	0.658	0.655
2-mod	C_EDA+ C_ACC	miss	0.621	0.616	0.616	0.612	0.617	0.615	0.609	0.611	0.644	0.618
		wave	0.633	0.621	0.626	0.624	0.633	0.640	0.630	0.619	0.645	0.630
		nonlin	0.657	0.663	0.665	0.660	0.658	0.655	0.656	0.644	0.651	0.634
		freq	0.654	0.653	0.647	0.652	0.658	0.661	0.657	0.649	0.650	0.632
		nonstat	0.603	0.598	0.602	0.591	0.598	0.605	0.588	0.594	0.636	0.614
	W_EDA+ W_ACC	miss	0.612	0.620	0.620	0.609	0.609	0.627	0.592	0.610	0.592	0.641
		wave	0.601	0.608	0.618	0.599	0.602	0.598	0.582	0.585	0.609	0.613
		nonlin	0.689	0.683	0.680	0.700	0.687	0.694	0.691	0.706	0.680	0.680
		freq	0.651	0.647	0.651	0.655	0.651	0.656	0.647	0.642	0.621	0.647
		nonstat	0.623	0.634	0.633	0.617	0.618	0.632	0.616	0.613	0.591	0.637
	C_ACC+ W_ACC	miss	0.637	0.642	0.644	0.643	0.651	0.639	0.630	0.638	0.596	0.647
		wave	0.638	0.641	0.648	0.634	0.636	0.648	0.649	0.619	0.633	0.634
		nonlin	0.696	0.681	0.681	0.693	0.694	0.691	0.685	0.696	0.686	0.689
		freq	0.655	0.650	0.653	0.651	0.656	0.641	0.647	0.635	0.614	0.641
		nonstat	0.640	0.636	0.631	0.635	0.638	0.637	0.641	0.635	0.598	0.653
	C_ECG+ C_EMG	miss	0.661	0.655	0.659	0.661	0.661	0.657	0.665	0.647	0.657	0.643
		wave	0.658	0.658	0.661	0.660	0.661	0.659	0.665	0.649	0.663	0.644
		nonlin	0.662	0.654	0.659	0.659	0.660	0.667	0.660	0.643	0.663	0.640
		freq	0.661	0.656	0.665	0.654	0.657	0.661	0.665	0.649	0.662	0.643
		nonstat	0.661	0.660	0.659	0.665	0.656	0.658	0.665	0.643	0.661	0.642

Card.	Subset	Noise	Static	TENT	EATA	READ	SuMi	SoTTA	AeO	PTA	PSDNorm	**Ours**
3-mod	C_ACC+ W_ACC+ W_EDA	miss	0.609	0.621	0.622	0.611	0.598	0.625	0.593	0.603	0.589	0.617
		wave	0.604	0.616	0.617	0.605	0.606	0.598	0.585	0.595	0.613	0.592
		nonlin	0.695	0.681	0.687	0.690	0.694	0.686	0.692	0.701	0.676	0.687
		freq	0.653	0.652	0.652	0.651	0.651	0.643	0.642	0.642	0.620	0.653
		nonstat	0.615	0.636	0.628	0.617	0.623	0.628	0.600	0.616	0.591	0.635
	C_ECG+ C_EMG+ C_RESP	miss	0.658	0.657	0.659	0.658	0.660	0.652	0.647	0.653	0.664	0.624
		wave	0.657	0.660	0.659	0.651	0.663	0.653	0.650	0.643	0.657	0.618
		nonlin	0.634	0.636	0.636	0.636	0.635	0.636	0.618	0.626	0.625	0.609
		freq	0.664	0.659	0.659	0.660	0.658	0.661	0.659	0.644	0.658	0.607
		nonstat	0.657	0.655	0.655	0.655	0.653	0.664	0.647	0.642	0.662	0.636
4-mod	C_EDA+ C_ACC+ W_EDA+ W_ACC	miss	0.482	0.462	0.467	0.481	0.479	0.474	0.490	0.483	0.561	0.424
		wave	0.509	0.487	0.491	0.483	0.491	0.491	0.510	0.496	0.583	0.491
		nonlin	0.692	0.683	0.682	0.684	0.692	0.696	0.670	0.692	0.677	0.677
		freq	0.650	0.656	0.651	0.644	0.657	0.655	0.655	0.637	0.607	0.657
		nonstat	0.494	0.467	0.478	0.489	0.475	0.475	0.496	0.486	0.561	0.445
	W_ACC+ W_EDA+ W_TEMP+ W_BVP	miss	0.612	0.623	0.626	0.614	0.619	0.632	0.585	0.607	0.586	0.634
		wave	0.600	0.600	0.615	0.602	0.606	0.619	0.589	0.596	0.617	0.612
		nonlin	0.704	0.707	0.706	0.699	0.697	0.698	0.681	0.685	0.683	0.718
		freq	0.656	0.660	0.656	0.646	0.650	0.658	0.647	0.641	0.625	0.644
		nonstat	0.625	0.637	0.637	0.618	0.620	0.632	0.599	0.613	0.584	0.638
	C_ECG+ C_EMG+ C_RESP+ W_TEMP	miss	0.661	0.658	0.658	0.652	0.653	0.663	0.665	0.644	0.655	0.623
		wave	0.660	0.658	0.658	0.657	0.652	0.659	0.655	0.644	0.657	0.622
		nonlin	0.628	0.642	0.633	0.644	0.635	0.631	0.619	0.618	0.625	0.610
		freq	0.664	0.659	0.659	0.663	0.661	0.666	0.660	0.646	0.662	0.602
		nonstat	0.657	0.658	0.655	0.662	0.656	0.658	0.655	0.640	0.659	0.632
5-mod	all	miss	0.660	0.662	0.661	0.653	0.658	0.659	0.659	0.651	0.660	0.640
		wave	0.663	0.655	0.656	0.662	0.659	0.657	0.654	0.654	0.651	0.625
		nonlin	0.634	0.652	0.645	0.640	0.622	0.647	0.622	0.635	0.608	0.622
		freq	0.658	0.665	0.664	0.659	0.653	0.651	0.675	0.647	0.663	0.606
		nonstat	0.653	0.655	0.652	0.652	0.651	0.655	0.654	0.644	0.661	0.633

Table 22: Full per-modality-subset, per-noise F1-score (↑) on **WESAD** with the **ShaSpec** backbone. Each cell is the F1 for one specific (corrupted modality subset, noise type) combination. Noise abbreviations: *miss*=discrete missingness, *wave*=waveform anomaly, *nonlin*=nonlinear degradation, *freq*=frequency attenuation, *nonstat*=nonstationary noise. Clean baseline: F1 = 0.554.

Card.	Subset	Noise	Static	TENT	EATA	READ	SuMi	SoTTA	AeO	PTA	PSDNorm	**Ours**
1-mod	chest_EDA	miss	0.506	0.505	0.511	0.505	0.507	0.514	0.497	0.496	0.454	0.523
		wave	0.493	0.497	0.496	0.497	0.493	0.501	0.484	0.492	0.465	0.520
		nonlin	0.523	0.521	0.527	0.514	0.520	0.524	0.515	0.506	0.463	0.545
		freq	0.532	0.530	0.539	0.526	0.527	0.531	0.515	0.521	0.478	0.548
		nonstat	0.511	0.506	0.512	0.508	0.511	0.513	0.498	0.501	0.464	0.531
	chest_ACC	miss	0.547	0.547	0.550	0.555	0.542	0.547	0.542	0.551	0.493	0.570
		wave	0.558	0.559	0.557	0.564	0.555	0.552	0.552	0.561	0.507	0.573
		nonlin	0.543	0.543	0.548	0.537	0.539	0.545	0.537	0.524	0.498	0.571
		freq	0.553	0.558	0.555	0.539	0.550	0.557	0.542	0.532	0.506	0.581
		nonstat	0.560	0.559	0.564	0.566	0.558	0.564	0.550	0.554	0.511	0.578
	wrist_ACC	miss	0.549	0.552	0.556	0.546	0.545	0.556	0.533	0.538	0.452	0.558
		wave	0.540	0.540	0.544	0.533	0.535	0.548	0.527	0.535	0.484	0.547
		nonlin	0.559	0.554	0.566	0.552	0.555	0.558	0.546	0.543	0.494	0.568
		freq	0.536	0.541	0.546	0.530	0.531	0.540	0.524	0.520	0.476	0.548
		nonstat	0.511	0.520	0.519	0.519	0.512	0.516	0.501	0.510	0.453	0.532
	wrist_EDA	miss	0.542	0.534	0.544	0.526	0.538	0.541	0.525	0.514	0.502	0.559
		wave	0.539	0.538	0.539	0.532	0.532	0.541	0.519	0.521	0.499	0.557
		nonlin	0.545	0.543	0.548	0.539	0.542	0.551	0.530	0.524	0.506	0.565
		freq	0.553	0.550	0.555	0.543	0.550	0.553	0.544	0.536	0.507	0.579
		nonstat	0.539	0.538	0.541	0.531	0.534	0.544	0.525	0.517	0.500	0.560
	chest_ECG	miss	0.553	0.556	0.559	0.541	0.552	0.558	0.539	0.526	0.494	0.569
		wave	0.500	0.485	0.498	0.497	0.493	0.501	0.495	0.480	0.509	0.516
		nonlin	0.470	0.490	0.478	0.490	0.484	0.485	0.482	0.485	0.471	0.506
		freq	0.433	0.440	0.439	0.446	0.444	0.443	0.457	0.449	0.474	0.476
		nonstat	0.567	0.576	0.579	0.566	0.573	0.584	0.564	0.562	0.517	0.582
	chest_EMG	miss	0.554	0.544	0.557	0.543	0.545	0.552	0.543	0.531	0.504	0.577
		wave	0.550	0.548	0.555	0.544	0.548	0.562	0.533	0.537	0.503	0.575
		nonlin	0.554	0.552	0.556	0.542	0.552	0.549	0.542	0.536	0.507	0.570
		freq	0.554	0.548	0.559	0.545	0.549	0.555	0.543	0.536	0.504	0.581
		nonstat	0.554	0.549	0.558	0.536	0.549	0.556	0.537	0.532	0.504	0.572
	chest_RESP	miss	0.555	0.553	0.559	0.547	0.556	0.559	0.543	0.539	0.509	0.579
		wave	0.553	0.550	0.558	0.540	0.543	0.550	0.537	0.539	0.509	0.571
		nonlin	0.519	0.519	0.520	0.512	0.519	0.529	0.506	0.508	0.474	0.523
		freq	0.553	0.550	0.554	0.544	0.554	0.548	0.535	0.535	0.504	0.578
		nonstat	0.555	0.552	0.560	0.547	0.548	0.554	0.542	0.534	0.505	0.574
	wrist_TEMP	miss	0.555	0.554	0.561	0.548	0.548	0.559	0.538	0.537	0.509	0.577
		wave	0.556	0.555	0.562	0.551	0.548	0.550	0.543	0.538	0.510	0.582
		nonlin	0.558	0.557	0.561	0.550	0.551	0.559	0.538	0.536	0.505	0.573
		freq	0.554	0.555	0.558	0.545	0.550	0.553	0.535	0.537	0.505	0.574
		nonstat	0.559	0.556	0.561	0.551	0.551	0.553	0.539	0.534	0.507	0.575
	wrist_BVP	miss	0.557	0.558	0.558	0.542	0.552	0.551	0.540	0.538	0.506	0.569
		wave	0.557	0.553	0.559	0.545	0.551	0.561	0.535	0.541	0.504	0.566
		nonlin	0.522	0.519	0.522	0.509	0.511	0.527	0.507	0.503	0.500	0.545
		freq	0.553	0.550	0.557	0.544	0.547	0.552	0.539	0.534	0.503	0.570
		nonstat	0.553	0.551	0.556	0.548	0.550	0.555	0.536	0.533	0.507	0.567
2-mod	C_EDA+ C_ACC	miss	0.514	0.509	0.518	0.507	0.511	0.519	0.513	0.506	0.440	0.532
		wave	0.539	0.544	0.542	0.535	0.538	0.535	0.526	0.537	0.479	0.553
		nonlin	0.514	0.512	0.516	0.506	0.505	0.523	0.503	0.497	0.454	0.534
		freq	0.537	0.531	0.542	0.532	0.531	0.534	0.531	0.512	0.480	0.554
		nonstat	0.517	0.517	0.518	0.520	0.513	0.519	0.509	0.508	0.462	0.535
	W_EDA+ W_ACC	miss	0.542	0.541	0.546	0.536	0.537	0.543	0.522	0.527	0.454	0.552
		wave	0.538	0.542	0.540	0.535	0.536	0.546	0.517	0.518	0.495	0.547
		nonlin	0.554	0.554	0.556	0.540	0.547	0.556	0.532	0.528	0.495	0.559
		freq	0.535	0.537	0.541	0.530	0.532	0.539	0.524	0.520	0.475	0.539
		nonstat	0.516	0.522	0.523	0.519	0.513	0.529	0.504	0.509	0.455	0.534
	C_ACC+ W_ACC	miss	0.529	0.541	0.533	0.552	0.537	0.537	0.528	0.537	0.438	0.547
		wave	0.508	0.521	0.512	0.537	0.513	0.509	0.497	0.513	0.474	0.534
		nonlin	0.551	0.548	0.555	0.539	0.547	0.556	0.539	0.537	0.489	0.567
		freq	0.537	0.536	0.545	0.533	0.533	0.540	0.523	0.521	0.473	0.548
		nonstat	0.499	0.504	0.504	0.506	0.500	0.498	0.491	0.507	0.444	0.522
	C_ECG+ C_EMG	miss	0.549	0.550	0.554	0.541	0.547	0.552	0.530	0.527	0.499	0.567
		wave	0.491	0.481	0.493	0.491	0.487	0.500	0.498	0.475	0.507	0.519
		nonlin	0.470	0.493	0.478	0.486	0.483	0.488	0.477	0.484	0.472	0.505
		freq	0.433	0.441	0.437	0.449	0.447	0.446	0.461	0.453	0.474	0.482
		nonstat	0.567	0.576	0.576	0.572	0.566	0.576	0.560	0.560	0.507	0.581

Card.	Subset	Noise	Static	TENT	EATA	READ	SuMi	SoTTA	AeO	PTA	PSDNorm	**Ours**
3-mod	C_ACC+ W_ACC+ W_EDA	miss	0.533	0.534	0.538	0.526	0.537	0.542	0.523	0.526	0.435	0.553
		wave	0.513	0.519	0.518	0.524	0.524	0.523	0.516	0.528	0.477	0.540
		nonlin	0.544	0.537	0.544	0.539	0.541	0.544	0.527	0.526	0.488	0.563
		freq	0.537	0.538	0.545	0.531	0.534	0.541	0.517	0.522	0.472	0.546
		nonstat	0.505	0.511	0.510	0.514	0.505	0.509	0.494	0.508	0.445	0.530
	C_ECG+ C_EMG+ C_RESP	miss	0.555	0.547	0.560	0.544	0.553	0.557	0.551	0.536	0.486	0.570
		wave	0.504	0.493	0.504	0.491	0.495	0.504	0.493	0.479	0.514	0.518
		nonlin	0.518	0.542	0.525	0.537	0.527	0.530	0.496	0.528	0.502	0.535
		freq	0.432	0.443	0.438	0.452	0.449	0.440	0.461	0.449	0.474	0.469
		nonstat	0.568	0.574	0.580	0.575	0.575	0.583	0.561	0.568	0.503	0.593
4-mod	C_EDA+ C_ACC+ W_EDA+ W_ACC	miss	0.502	0.499	0.509	0.498	0.505	0.511	0.492	0.499	0.399	0.514
		wave	0.523	0.525	0.525	0.509	0.522	0.530	0.501	0.517	0.457	0.526
		nonlin	0.506	0.503	0.506	0.507	0.504	0.511	0.490	0.500	0.447	0.520
		freq	0.515	0.518	0.520	0.513	0.516	0.519	0.504	0.503	0.445	0.530
		nonstat	0.483	0.491	0.487	0.482	0.480	0.487	0.461	0.480	0.413	0.511
	W_ACC+ W_EDA+ W_TEMP+ W_BVP	miss	0.538	0.544	0.548	0.536	0.536	0.546	0.524	0.521	0.452	0.547
		wave	0.542	0.542	0.543	0.535	0.540	0.544	0.515	0.530	0.490	0.539
		nonlin	0.511	0.491	0.501	0.492	0.490	0.497	0.485	0.489	0.483	0.522
		freq	0.535	0.541	0.540	0.533	0.533	0.544	0.523	0.525	0.474	0.548
		nonstat	0.515	0.523	0.523	0.519	0.515	0.525	0.500	0.508	0.450	0.532
	C_ECG+ C_EMG+ C_RESP+ W_TEMP	miss	0.553	0.555	0.559	0.554	0.552	0.559	0.551	0.542	0.499	0.582
		wave	0.501	0.503	0.499	0.503	0.500	0.512	0.500	0.482	0.501	0.528
		nonlin	0.512	0.537	0.520	0.536	0.518	0.529	0.486	0.532	0.500	0.533
		freq	0.433	0.440	0.438	0.452	0.449	0.443	0.460	0.452	0.473	0.469
		nonstat	0.563	0.569	0.570	0.563	0.569	0.571	0.553	0.570	0.503	0.575
5-mod	all	miss	0.563	0.554	0.563	0.555	0.556	0.564	0.548	0.547	0.509	0.578
		wave	0.506	0.494	0.504	0.490	0.495	0.507	0.501	0.480	0.498	0.516
		nonlin	0.529	0.541	0.538	0.536	0.525	0.536	0.496	0.530	0.522	0.533
		freq	0.434	0.442	0.440	0.452	0.446	0.442	0.457	0.451	0.472	0.465
		nonstat	0.570	0.578	0.573	0.570	0.571	0.581	0.560	0.571	0.501	0.586

Table 23: Full per-modality-subset, per-noise F1-score (↑) on **WESAD** with the **Flex-MoE** backbone. Each cell is the F1 for one specific (corrupted modality subset, noise type) combination. Noise abbreviations: *miss*=discrete missingness, *wave*=waveform anomaly, *nonlin*=nonlinear degradation, *freq*=frequency attenuation, *nonstat*=nonstationary noise. Clean baseline: F1 = 0.554.

Card.	Subset	Noise	Static	TENT	EATA	READ	SuMi	SoTTA	AeO	PTA	PSDNorm	**Ours**
1-mod	chest_EDA	miss	0.519	0.536	0.535	0.515	0.534	0.556	0.516	0.509	0.492	0.520
		wave	0.515	0.534	0.540	0.511	0.530	0.549	0.501	0.512	0.506	0.522
		nonlin	0.505	0.530	0.524	0.544	0.504	0.517	0.500	0.489	0.484	0.519
		freq	0.532	0.539	0.554	0.561	0.525	0.553	0.492	0.502	0.505	0.543
		nonstat	0.515	0.538	0.538	0.526	0.523	0.538	0.482	0.509	0.495	0.524
	chest_ACC	miss	0.526	0.543	0.555	0.550	0.501	0.499	0.506	0.525	0.540	0.550
		wave	0.545	0.559	0.547	0.549	0.524	0.522	0.521	0.530	0.539	0.558
		nonlin	0.554	0.578	0.567	0.543	0.517	0.524	0.500	0.514	0.544	0.557
		freq	0.544	0.570	0.569	0.573	0.540	0.538	0.529	0.529	0.537	0.559
		nonstat	0.535	0.552	0.552	0.539	0.519	0.521	0.508	0.504	0.535	0.556
	wrist_ACC	miss	0.534	0.544	0.537	0.552	0.502	0.511	0.488	0.515	0.443	0.531
		wave	0.529	0.535	0.536	0.534	0.510	0.527	0.483	0.509	0.514	0.534
		nonlin	0.528	0.540	0.547	0.544	0.554	0.570	0.532	0.519	0.519	0.540
		freq	0.531	0.543	0.542	0.538	0.482	0.499	0.482	0.496	0.508	0.544
		nonstat	0.462	0.464	0.474	0.474	0.447	0.449	0.456	0.437	0.441	0.469
	wrist_EDA	miss	0.557	0.575	0.577	0.600	0.516	0.515	0.486	0.516	0.532	0.571
		wave	0.543	0.550	0.555	0.585	0.474	0.498	0.462	0.491	0.541	0.558
		nonlin	0.544	0.541	0.542	0.561	0.499	0.516	0.469	0.500	0.526	0.544
		freq	0.554	0.568	0.566	0.555	0.524	0.534	0.504	0.519	0.537	0.567
		nonstat	0.555	0.566	0.570	0.581	0.493	0.498	0.457	0.491	0.535	0.568
	chest_ECG	miss	0.552	0.582	0.576	0.543	0.545	0.559	0.523	0.526	0.529	0.561
		wave	0.543	0.559	0.569	0.565	0.504	0.521	0.480	0.493	0.548	0.558
		nonlin	0.528	0.531	0.530	0.524	0.542	0.541	0.538	0.510	0.501	0.521
		freq	0.476	0.470	0.457	0.479	0.516	0.498	0.499	0.512	0.519	0.469
		nonstat	0.571	0.602	0.598	0.563	0.578	0.607	0.546	0.550	0.529	0.584
	chest_EMG	miss	0.554	0.576	0.570	0.552	0.525	0.519	0.508	0.526	0.539	0.558
		wave	0.533	0.548	0.554	0.534	0.504	0.530	0.489	0.495	0.508	0.548
		nonlin	0.550	0.567	0.574	0.565	0.515	0.525	0.495	0.517	0.533	0.561
		freq	0.546	0.564	0.561	0.570	0.512	0.509	0.485	0.519	0.536	0.551
		nonstat	0.551	0.578	0.566	0.560	0.520	0.517	0.512	0.517	0.536	0.563
	chest_RESP	miss	0.554	0.576	0.582	0.547	0.522	0.532	0.500	0.514	0.535	0.568
		wave	0.554	0.569	0.565	0.554	0.526	0.530	0.499	0.511	0.533	0.560
		nonlin	0.539	0.541	0.550	0.534	0.564	0.575	0.535	0.528	0.519	0.532
		freq	0.552	0.572	0.560	0.557	0.519	0.536	0.496	0.514	0.534	0.559
		nonstat	0.550	0.574	0.570	0.563	0.527	0.526	0.491	0.518	0.533	0.562
	wrist_TEMP	miss	0.551	0.580	0.588	0.537	0.534	0.543	0.488	0.524	0.528	0.564
		wave	0.554	0.573	0.591	0.545	0.531	0.544	0.475	0.520	0.529	0.572
		nonlin	0.550	0.571	0.578	0.554	0.523	0.539	0.492	0.523	0.535	0.567
		freq	0.554	0.573	0.571	0.543	0.524	0.543	0.506	0.519	0.537	0.562
		nonstat	0.553	0.582	0.567	0.559	0.521	0.536	0.502	0.514	0.533	0.569
	wrist_BVP	miss	0.554	0.569	0.566	0.527	0.528	0.524	0.512	0.525	0.530	0.562
		wave	0.557	0.585	0.573	0.581	0.524	0.531	0.491	0.516	0.531	0.566
		nonlin	0.477	0.494	0.487	0.503	0.503	0.527	0.472	0.488	0.498	0.488
		freq	0.551	0.563	0.573	0.555	0.523	0.547	0.494	0.519	0.535	0.559
		nonstat	0.549	0.565	0.561	0.550	0.515	0.533	0.504	0.507	0.528	0.554
2-mod	C_EDA+ C_ACC	miss	0.519	0.519	0.520	0.545	0.531	0.549	0.507	0.529	0.508	0.522
		wave	0.524	0.549	0.545	0.520	0.512	0.509	0.492	0.483	0.513	0.538
		nonlin	0.516	0.547	0.550	0.529	0.518	0.533	0.503	0.498	0.507	0.541
		freq	0.542	0.561	0.563	0.568	0.552	0.553	0.546	0.523	0.518	0.561
		nonstat	0.496	0.535	0.534	0.494	0.486	0.500	0.459	0.483	0.510	0.526
	W_EDA+ W_ACC	miss	0.537	0.532	0.533	0.601	0.504	0.499	0.472	0.511	0.446	0.535
		wave	0.531	0.529	0.531	0.565	0.484	0.500	0.450	0.491	0.516	0.525
		nonlin	0.531	0.561	0.565	0.558	0.566	0.572	0.546	0.536	0.510	0.548
		freq	0.529	0.549	0.545	0.552	0.488	0.500	0.491	0.493	0.505	0.542
		nonstat	0.476	0.485	0.485	0.464	0.448	0.441	0.437	0.445	0.449	0.484
	C_ACC+ W_ACC	miss	0.501	0.482	0.478	0.508	0.485	0.468	0.494	0.516	0.455	0.489
		wave	0.507	0.504	0.505	0.504	0.515	0.519	0.498	0.506	0.494	0.512
		nonlin	0.531	0.554	0.563	0.525	0.567	0.582	0.535	0.538	0.518	0.547
		freq	0.526	0.536	0.536	0.539	0.547	0.563	0.538	0.502	0.512	0.525
		nonstat	0.443	0.455	0.461	0.461	0.442	0.439	0.444	0.432	0.446	0.448
	C_ECG+ C_EMG	miss	0.558	0.589	0.586	0.552	0.531	0.547	0.515	0.518	0.537	0.562
		wave	0.515	0.522	0.520	0.481	0.509	0.530	0.506	0.472	0.486	0.524
		nonlin	0.534	0.538	0.539	0.536	0.549	0.546	0.546	0.521	0.504	0.539
		freq	0.479	0.475	0.472	0.492	0.506	0.503	0.493	0.487	0.516	0.476
		nonstat	0.580	0.600	0.603	0.589	0.573	0.601	0.525	0.545	0.551	0.591

Card.	Subset	Noise	Static	TENT	EATA	READ	SuMi	SoTTA	AeO	PTA	PSDNorm	**Ours**
3-mod	C_ACC+ W_ACC+ W_EDA	miss	0.510	0.502	0.493	0.503	0.489	0.474	0.493	0.516	0.455	0.501
		wave	0.522	0.516	0.522	0.508	0.516	0.504	0.530	0.520	0.511	0.522
		nonlin	0.539	0.564	0.553	0.518	0.490	0.501	0.465	0.489	0.504	0.552
		freq	0.526	0.530	0.527	0.549	0.547	0.568	0.537	0.516	0.510	0.526
		nonstat	0.464	0.473	0.477	0.455	0.448	0.448	0.441	0.427	0.445	0.476
	C_ECG+ C_EMG+ C_RESP	miss	0.548	0.574	0.589	0.555	0.523	0.544	0.480	0.508	0.527	0.571
		wave	0.514	0.525	0.539	0.509	0.489	0.511	0.474	0.464	0.485	0.522
		nonlin	0.525	0.526	0.527	0.525	0.524	0.524	0.546	0.513	0.489	0.526
		freq	0.477	0.468	0.473	0.479	0.511	0.501	0.494	0.490	0.518	0.481
		nonstat	0.566	0.599	0.597	0.573	0.575	0.587	0.533	0.541	0.528	0.577
4-mod	C_EDA+ C_ACC+ W_EDA+ W_ACC	miss	0.483	0.493	0.486	0.477	0.410	0.376	0.413	0.458	0.410	0.483
		wave	0.499	0.500	0.495	0.489	0.416	0.394	0.428	0.452	0.500	0.498
		nonlin	0.500	0.530	0.526	0.498	0.521	0.515	0.473	0.507	0.478	0.523
		freq	0.519	0.537	0.534	0.535	0.527	0.532	0.520	0.507	0.492	0.538
		nonstat	0.423	0.448	0.438	0.457	0.377	0.356	0.397	0.383	0.410	0.444
	W_ACC+ W_EDA+ W_TEMP+ W_BVP	miss	0.546	0.571	0.568	0.593	0.504	0.540	0.453	0.476	0.447	0.569
		wave	0.537	0.564	0.558	0.560	0.482	0.527	0.458	0.468	0.518	0.557
		nonlin	0.454	0.465	0.459	0.439	0.472	0.469	0.454	0.472	0.483	0.459
		freq	0.534	0.536	0.537	0.548	0.490	0.483	0.473	0.499	0.500	0.532
		nonstat	0.477	0.487	0.490	0.471	0.448	0.451	0.457	0.438	0.441	0.483
	C_ECG+ C_EMG+ C_RESP+ W_TEMP	miss	0.552	0.583	0.584	0.518	0.521	0.538	0.467	0.508	0.530	0.573
		wave	0.508	0.522	0.538	0.499	0.537	0.550	0.504	0.497	0.464	0.518
		nonlin	0.514	0.516	0.517	0.505	0.525	0.523	0.533	0.489	0.484	0.520
		freq	0.478	0.470	0.474	0.481	0.508	0.507	0.492	0.493	0.520	0.480
		nonstat	0.575	0.603	0.603	0.586	0.573	0.582	0.534	0.532	0.544	0.588
5-mod	all	miss	0.543	0.582	0.581	0.497	0.517	0.537	0.473	0.505	0.527	0.565
		wave	0.503	0.516	0.530	0.483	0.508	0.523	0.471	0.470	0.449	0.518
		nonlin	0.543	0.541	0.545	0.519	0.470	0.455	0.496	0.480	0.501	0.539
		freq	0.480	0.474	0.469	0.474	0.507	0.497	0.493	0.493	0.518	0.478
		nonstat	0.566	0.572	0.579	0.572	0.546	0.580	0.511	0.528	0.522	0.569

L Complete Efficiency Metrics

We report the complete efficiency measurements in Table 24 and Table 25. All measurements are conducted on corrupted test batches with batch size 64. Latency is the average wall-clock time per batch after warm-up, including test-time adaptation and inference. For detector-based methods, the detector is executed as part of the method, so its overhead is included. Energy is measured as the average CPU energy per batch in Joules. GFLOPs are measured using the PyTorch profiler and averaged over test batches. Trainable parameters denote the number of parameters updated during test-time adaptation; non-parametric methods have zero trainable parameters.

For DSADS, the efficiency benchmark corrupts the torso, right-arm, and left-arm modalities with `waveform_anomaly` noise. For DaliaHAR, the benchmark corrupts all wrist modalities. For PAMAP2, the benchmark corrupts the ankle IMU, hand IMU, and heart-rate modalities.

L.1 Detector Training and Deployment Footprint

Table 26 reports the detector footprint for DaliaHAR–MAESTRO. This separate profiling study uses an NVIDIA L40. Detector inference is measured with batch size 64, reporting median latency over 50 synchronized runs after 10 warm-up iterations. Detector training is performed offline; deployment retains the trained detector weights, feature statistics, and frozen reliance profile. These detector-only measurements complement the end-to-end profiles above and should not be directly combined with measurements from different hardware.

Memory and stored artifacts. The reported peak CUDA memory is 4.52 MiB per modality when detectors are evaluated sequentially with buffer reuse. Stored artifacts comprise 164.8 KB of detector weights, 16.3 KB of feature statistics, and a 40-byte reliance profile; storage units follow those reported in the rebuttal. The memory figure describes the detector computation rather than the complete adaptation pipeline. The following subsection evaluates an implementation optimization separately.

Table 24: Efficiency metrics on DSADS.

Backbone	Method	Latency (ms)	Energy (J)	GFLOPs	Trainable Params
MAESTRO					
	Static	435.0	381.1	5.56	0
	TENT	1200.2	1254.5	53.63	3584
	EATA	1038.6	1108.4	53.63	3584
	READ	554.8	528.6	12.59	49920
	SuMi	1238.4	1319.8	53.36	3584
	SoTTA	1349.9	1465.7	45.55	3584
	AeO	4878.5	5651.6	299.49	3584
	PTA	5190.5	5991.9	293.93	3584
	PSDNorm	1073.1	915.1	5.99	0
	CARAT	1713.2	1692.7	48.54	1539.0
ShaSpec					
	Static	464.1	414.0	52.03	0
	TENT	1799.2	2098.0	826.82	8960
	EATA	1124.4	1373.1	826.82	8960
	READ	494.4	478.1	144.18	119192
	SuMi	764.0	820.2	454.49	8960
	SoTTA	565.5	552.1	147.48	8960
	AeO	5987.8	7936.3	4672.61	8960
	PTA	5312.6	7142.5	4600.54	8960
	PSDNorm	1026.3	863.8	52.47	0
	CARAT	1382.9	1434.9	563.22	4355.4
Flex-MoE					
	Static	396.5	336.6	27.68	0
	TENT	1312.3	1348.7	431.79	6912
	EATA	1357.5	1386.2	431.79	6912
	READ	565.8	517.7	77.10	49152
	SuMi	1080.2	1065.5	299.81	6912
	SoTTA	2164.2	2122.8	225.17	6912
	AeO	8049.2	8497.9	2440.03	6912
	PTA	9764.2	10194.9	2402.35	6912
	PSDNorm	1081.8	904.5	28.12	0
	CARAT	2191.3	2147.9	365.35	4867.4

L.2 Detector Implementation Optimization

We assess detector implementation overhead in a separate controlled DSADS–MAESTRO rerun. Reusing the full-channel FITS reconstruction within each batch and batching spectral computations reduces detector latency from 467 to 16 ms and end-to-end latency from 1,713 to 1,262 ms, a reduction of approximately 26%. Both implementations produce identical detector features, suspect flags, and predictions. Table 27 reports this comparison separately from the five reference profiles in Tables 24 and 25.

M Additional Results

N Statistical and Condition-Level Analyses

We complement the setting-level averages in Table 1 with paired comparisons, a repeated-corruption experiment, and condition-level rank and dispersion analyses. The benchmark contains 720 matched conditions: 50 per backbone on DSADS and DaliaHAR, 45 on PAMAP2, and 95 on WESAD. Each condition specifies a dataset, backbone, corrupted-modality subset, and corruption family. Methods are paired on the same corrupted data and source checkpoint within each setting. Unless stated otherwise, Δ denotes CARAT minus the named comparator. Effect sizes and interval bounds below are expressed in *macro-F1 points* (100 times the difference on the 0–1 scale), preserving the original reported precision while displaying two decimals.

Table 25: Efficiency metrics on DaliaHAR and PAMAP2 with MAESTRO.

Dataset	Method	Latency (ms)	Energy (J)	GFLOPs	Trainable Params
DaliaHAR / MAESTRO					
	Static	141.2	147.8	1.44	0
	TENT	508.3	597.9	13.84	1536
	EATA	480.7	554.4	13.84	1536
	READ	177.4	184.8	3.20	6336
	SuMi	521.4	560.2	12.22	1536
	SoTTA	662.0	685.3	6.79	1536
	AeO	2392.8	2911.9	77.26	1536
	PTA	3061.1	3627.8	75.81	1536
	PSDNorm	264.4	243.0	1.63	0
	CARAT	693.3	761.6	13.11	971.8
PAMAP2 / MAESTRO					
	Static	915.1	856.7	19.39	0
	TENT	3161.8	3447.7	185.67	3072
	EATA	2262.1	2521.3	185.67	3072
	READ	1223.6	1228.7	43.92	49920
	SuMi	3016.0	3257.3	147.53	3072
	SoTTA	3895.2	4343.8	242.04	3072
	AeO	9600.4	11220.5	870.11	3072
	PTA	7524.0	9341.5	850.69	3072
	PSDNorm	1640.1	1443.1	21.08	0
	CARAT	3116.1	3250.2	177.50	1603.1

Table 26: **Detector footprint on DaliaHAR–MAESTRO.** Offline training cost and detector-only inference cost on an NVIDIA L40. Inference measurements are per batch of 64.

		Offline training		Detector inference	
Modality	Parameters	Time (s)	GFLOPs	Latency (ms)	GFLOPs
Wrist acceleration	8,512	35.3	97.7	33.9	0.0032
Wrist BVP	8,384	42.2	32.6	12.2	0.0011
Wrist EDA	8,384	42.1	32.6	12.2	0.0011
Wrist temperature	8,384	42.3	32.6	12.1	0.0011
Chest acceleration	8,512	42.6	97.7	33.9	0.0032
Reported total	**42,176**	**204.6**	**293.0**	**104.5**	**0.0095**

Totals retain the reported aggregate measurements; sums of rounded modality-level entries may differ slightly.

N.1 Paired comparisons over matched corruption conditions

Table 29 reproduces the paired Wilcoxon signed-rank summaries against the baseline with the highest mean F1 in each setting. This setting-specific reference may select different methods in different rows; it complements the fixed-method comparisons in Table 1. The W/T/L column instead counts the reported outcomes against *all nine alternatives* (Static and eight TTA baselines), summing to nine in each row. Its T category denotes no declared difference in the original analysis, not statistical equivalence or a practical tie band. Across the 108 comparisons, the reported counts are 68 wins, 28 non-rejections, and 12 losses.

Table 27: **DSADS–MAESTRO detector optimization.** Latency in milliseconds per batch from a controlled rerun. End-to-end latency includes detection, adaptation, and inference.

Implementation	Detector	End-to-end
Reference	467	1,713
Optimized	16	1,262

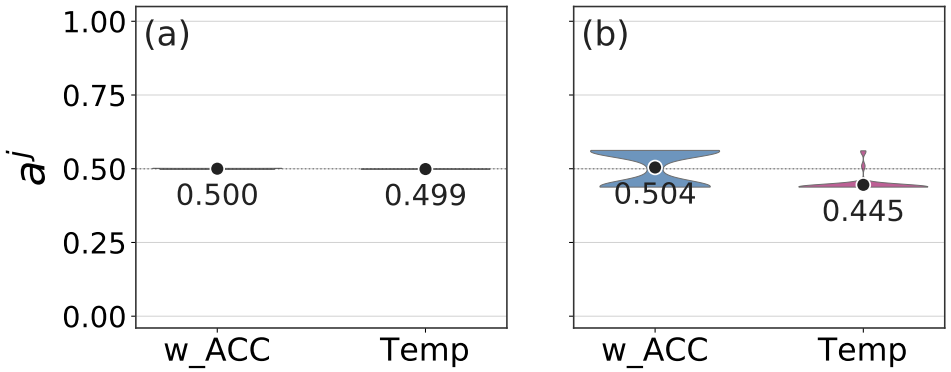


Figure 12: Gate lr sensitivity on DaliaHAR, $\mathcal{S}_i = \{\text{wrist_ACC}, \text{wrist_Temp}\}$. (a) $\eta_\alpha = 10^{-3}$ (F1 = 0.65): gates pinned at ~ 0.5. (b) $\eta_\alpha = 5 \times 10^{-2}$ (F1 = 0.67): gates move within the episodic budget.

Table 28: **Unconditional drop vs. CARAT on MAESTRO.** F1-score (↑) averaged across 10 corrupted-modality scenarios × 5 noise types per dataset. Mean$_{\pm\text{std}}$ across all (subset, noise) combinations. **Bold**: higher of {Unconditional drop, CARAT} per row. We restrict this analysis to MAESTRO, the most stable missingness-resilient backbone in our benchmark, where the routing-vs-dropping mechanism is evaluated without confounding from backbone-specific fragility under modality dropout.

Dataset	Static	Unconditional drop	CARAT
DSADS	0.69 ± 0.22	0.67 ± 0.22	**0.72** ± 0.20
DaliaHAR	0.74 ± 0.07	0.67 ± 0.11	**0.76** ± 0.07
PAMAP2	0.92 ± 0.04	0.68 ± 0.36	**0.92** ± 0.04
WESAD	0.65 ± 0.04	0.62 ± 0.03	**0.63** ± 0.04

Because corruption conditions share examples, checkpoints, and overlapping modality subsets, these condition-level tests are exploratory summaries of the evaluated suites, rather than 720 independent replications. They do not quantify training-seed uncertainty or establish superiority on a population of unseen datasets. The intervals and p-values below retain the values reported in the original analysis.

Table 29: **Paired condition-level comparisons.** n is the number of matched corruption conditions. The reported paired effect and 95% interval compare CARAT with the named setting-specific comparator. Estimates, intervals, and p-values are retained from the original paired analysis. W/T/L counts the reported test outcomes across nine alternatives, not wins across the n conditions. Effect sizes and intervals are in macro-F1 points; positive values favor CARAT.

Setting	n	Comparator	Reported Δ	Reported 95% CI	Reported p	W/T/L
DSADS / MAESTRO	50	PTA	+0.20	[−0.60, +1.90]	1.00	5/4/0
DSADS / ShaSpec	50	PSDNorm	+7.00	[+5.30, +9.10]	6×10^{-8}	9/0/0
DSADS / Flex-MoE	50	PTA	+0.00	[−0.40, +0.70]	1.00	2/7/0
DaliaHAR / MAESTRO	50	EATA	−0.20	[−0.40, −0.10]	0.008	7/0/2
DaliaHAR / ShaSpec	50	TENT	+1.80	[+0.90, +3.70]	6×10^{-4}	9/0/0
DaliaHAR / Flex-MoE	50	EATA	−0.20	[−0.40, +0.00]	0.19	7/2/0
PAMAP2 / MAESTRO	45	EATA	−0.20	[−0.40, −0.10]	0.14	4/5/0
PAMAP2 / ShaSpec	45	PSDNorm	+0.60	[−0.50, +1.70]	0.90	2/7/0
PAMAP2 / Flex-MoE	45	TENT	+0.20	[−0.00, +0.50]	0.30	8/1/0
WESAD / MAESTRO	95	Static	−1.40	[−1.80, −1.10]	6×10^{-9}	0/1/8
WESAD / ShaSpec	95	SoTTA	+1.50	[+1.30, +1.70]	2×10^{-15}	9/0/0
WESAD / Flex-MoE	95	EATA	−0.60	[−0.70, −0.40]	5×10^{-7}	6/1/2

The reported positive paired effects are largest on DSADS–ShaSpec ($+7.00$ points), DaliaHAR–ShaSpec ($+1.80$), and WESAD–ShaSpec ($+1.50$). Negative effects remain visible on DaliaHAR–MAESTRO, WESAD–MAESTRO, and WESAD–Flex-MoE. This breakdown is consistent with the concentration of gains by backbone in the main results.

N.2 Experiment R3.1: repeated corruption realizations

We repeated DSADS–MAESTRO over five trials with redrawn corruption realizations for each of the 50 matched conditions, keeping the source checkpoint fixed. The reported mean macro-F1 is 73.80 for CARAT on the 0–100 scale. Table 30 lists the eight reported comparator summaries; PSDNorm was not included in the reported repeated-trial table. All eight reported paired intervals exclude zero in favor of CARAT. The reported average per-condition standard deviation across trials is 0.60 macro-F1 points. These results are a separate repeated-corruption evaluation and do not replace the single-evaluation means in Table 1. The intervals concern the reported paired comparison, not uncertainty from five independent source-training runs.

Table 30: **Experiment R3.1: DSADS–MAESTRO under five corruption redraws.** Mean F1, paired differences and interval bounds use the 0–100 scale. CARAT achieves 73.80 mean F1. W/T/L sums to 50 matched conditions; the estimates, intervals and p-values retain the original repeated-evaluation reporting.

Comparator	Mean F1	W/T/L	Median Δ	Reported 95% CI	Reported p
Static	69.30	36/0/14	$+2.70$	$[+1.30, +4.80]$	6×10^{-5}
TENT	69.30	36/1/13	$+2.40$	$[+1.20, +4.40]$	9×10^{-5}
EATA	69.30	34/4/12	$+2.40$	$[+1.10, +4.50]$	9×10^{-5}
READ	69.00	38/0/12	$+3.10$	$[+1.60, +5.60]$	5×10^{-5}
SuMi	69.10	37/1/12	$+3.00$	$[+1.60, +5.30]$	6×10^{-5}
SoTTA	69.10	35/1/14	$+2.70$	$[+1.30, +4.90]$	7×10^{-5}
AeO	69.10	36/1/13	$+3.00$	$[+1.50, +5.60]$	6×10^{-5}
PTA	69.30	35/0/15	$+2.80$	$[+1.40, +5.10]$	7×10^{-5}

N.3 Condition-level ranks across corruptions, backbones, and datasets

For this analysis, the ten methods are ranked *within each matched corruption condition*, and those ranks are averaged over the relevant group. Average ranks are assigned to tied condition scores; smaller ranks are better. This differs from Table 1, which first averages F1 within each dataset–backbone setting and then ranks the resulting means. The pooled condition-level average also weights settings by their number of conditions, so WESAD contributes more observations than the other datasets. These rank summaries are descriptive.

Table 31: **Condition-level rank breakdowns.** Left: CARAT across the five corruption families. Right: CARAT and the lowest-rank alternative within each backbone. The reported means pool the conditions in the indicated group.

Corruption	Rank ↓
Waveform anomaly	3.19
Discrete missingness	3.25
Nonstationary noise	3.49
Nonlinear degradation	4.03
Frequency attenuation	4.62

Backbone	Rank ↓	Best alternative
ShaSpec	2.18	Static (4.59)
Flex-MoE	3.59	TENT (3.42)
MAESTRO	5.38	EATA (3.85)

CARAT has its lowest average rank under waveform anomalies and its highest under frequency attenuation. The backbone breakdown is more uneven: CARAT ranks 2.18 on ShaSpec, 3.59 on Flex-MoE, and 5.38 on MAESTRO. Thus, the condition-level analysis also exposes the backbone dependence visible in the setting-level averages.

Table 32 retains the dataset-level and pooled ranks from the same analysis. CARAT has the lowest average rank on DSADS, DaliaHAR, and PAMAP2; EATA has the lowest on WESAD. The difference between PTA's DSADS and WESAD ranks illustrates variation across sensing setups. Because the datasets also differ in task, sensor count, class distribution, and corruption difficulty, this comparison does not identify modality heterogeneity as the sole cause.

Table 32: **Mean condition-level ranks by dataset.** Smaller is better. Pooled averages all 720 condition ranks; Worst is the largest of the four dataset-level ranks. These are not the ranks of dataset–backbone means in Table 1. Bold and underline mark the lowest and second-lowest displayed ranks per column.

Method	DSADS	DaliaHAR	PAMAP2	WESAD	Pooled	Worst
CARAT	**3.56**	**2.39**	**4.09**	4.32	**3.72**	**4.32**
EATA	5.30	2.68	4.47	**3.70**	<u>3.97</u>	<u>5.30</u>
TENT	5.53	<u>2.46</u>	4.65	<u>4.25</u>	4.22	5.53
SoTTA	5.00	4.57	5.41	4.49	4.79	5.41
Static	5.51	5.96	4.60	4.89	5.19	5.96
SuMi	6.02	6.92	5.56	5.71	6.00	6.92
PTA	<u>3.63</u>	6.70	<u>4.32</u>	7.97	6.12	7.97
READ	6.66	5.74	8.21	5.29	6.22	8.21
AeO	5.61	8.77	6.30	7.12	7.00	8.77
PSDNorm	8.17	8.80	7.40	7.25	7.79	8.80

N.4 Dispersion across corruption families

To separate variation across corruption families from variation across corrupted-modality subsets, we compute the sample standard deviation across the five corruption-family scores within each fixed dataset–backbone–subset scenario, then average these standard deviations. Across the 144 scenarios, the reported average is 6.20 macro-F1 points for CARAT, compared with 7.20–7.70 points for the seven parameter-updating baselines (TENT, EATA, READ, SuMi, SoTTA, AeO, and PTA). The paired Wilcoxon comparisons of these scenario-level deviations reported p-values between 3×10^{-11} and 3×10^{-4}. The dependence qualifications of Appendix N.1 also apply to these scenario-level comparisons. PSDNorm has a smaller reported deviation, 5.60 points, alongside a poorer pooled mean rank of 7.79; low dispersion alone therefore does not establish better predictive performance. Table 33 lists the four settings highlighted in the analysis. This dispersion describes differences in corruption difficulty and sensitivity, not uncertainty across training seeds.

Table 33: **Within-subset dispersion across corruption families.** Each entry averages the sample standard deviation over the five corruption scores within a fixed modality-subset scenario. Values are in macro-F1 points. The baseline column gives the reported range across the seven parameter-updating baselines.

Setting	Baseline range	CARAT
DSADS / ShaSpec	18.80–21.60	12.30
PAMAP2 / ShaSpec	20.20–21.90	16.20
DSADS / MAESTRO	12.80–13.50	10.20
DSADS / Flex-MoE	9.90–10.40	9.20

O Source-Training Replacements and Additional Design Analyses

O.1 Alternative source-training regimes

We replace source training for DaliaHAR–ShaSpec with OGM, OPM, or MIAM and evaluate missingness robustness (Table 34) and subsequent TTA (Table 35). Uniform/asymmetric source results reproduce the F1 entries of Table 8, whose full dataset–backbone comparison remains in Appendix E. Asymmetric training achieves the highest F1 at 40% missingness (0.61); uniform training has the highest complete-input F1 (0.77).

Table 34: **Source-training replacement on DaliaHAR–ShaSpec.** Macro-F1 (↑), with the reported dispersions where available. Uniform/asymmetric entries use the original three-seed source-training evaluation. MIAM reports means only at 20%/40% missingness; – denotes an unreported result. **Bold**/underline: best/second distinct displayed mean per column.

Source training	0% missing	20% missing	40% missing
Uniform (Original)	**0.77**$_{\pm 0.02}$	0.57$_{\pm 0.02}$	0.39$_{\pm 0.01}$
OGM	0.73$_{\pm 0.00}$	0.50$_{\pm 0.01}$	0.35$_{\pm 0.02}$
OPM	0.76$_{\pm 0.00}$	0.57$_{\pm 0.02}$	0.45$_{\pm 0.01}$
MIAM	–	**0.68**	0.59
Asymmetric (ours)	0.74$_{\pm 0.00}$	**0.68**$_{\pm 0.02}$	**0.61**$_{\pm 0.02}$

CARAT leads all OGM/OPM columns in Table 35, and the MIAM all_wrist/Div. columns; on MIAM w_acc, TENT/EATA lead (0.64 vs. 0.63). MIAM varies masking rates by batch, so we estimate its prior as the expectation of the checkpoint-specific masking distribution. Our curriculum supplies the terminal dropout profile directly. Across the three curricula, CARAT achieves the highest Div. F1, exceeding the strongest TTA baseline by 1, 2, and 2 points for OGM, OPM, and MIAM, respectively.

Table 35: **Complete TTA comparison on replacement checkpoints.** DaliaHAR–ShaSpec macro-F1 (↑), reported mean$_{\pm SD}$ across the evaluated corruption conditions. all_w: all_wrist; Div.: pooled diverse setting. The spreads describe variation across conditions, not confidence intervals over independent training runs. Styling is applied within each column to displayed means; rounded ties are not statistical equivalence.

	OGM source training			OPM source training			MIAM source training		
Method	w_acc	all_w	Div.	w_acc	all_w	Div.	w_acc	all_w	Div.
Static	0.55$_{\pm 0.10}$	0.53$_{\pm 0.13}$	0.57$_{\pm 0.08}$	0.58$_{\pm 0.10}$	0.55$_{\pm 0.15}$	0.59$_{\pm 0.09}$	0.60$_{\pm 0.08}$	0.55$_{\pm 0.13}$	0.58$_{\pm 0.08}$
TENT	0.60$_{\pm 0.15}$	0.57$_{\pm 0.18}$	0.62$_{\pm 0.10}$	0.63$_{\pm 0.16}$	0.57$_{\pm 0.21}$	0.63$_{\pm 0.12}$	**0.64**$_{\pm 0.15}$	0.57$_{\pm 0.21}$	0.61$_{\pm 0.11}$
EATA	0.61$_{\pm 0.14}$	0.57$_{\pm 0.18}$	0.62$_{\pm 0.10}$	0.63$_{\pm 0.13}$	0.58$_{\pm 0.19}$	0.63$_{\pm 0.10}$	**0.64**$_{\pm 0.11}$	0.57$_{\pm 0.17}$	0.61$_{\pm 0.09}$
READ	0.60$_{\pm 0.16}$	0.58$_{\pm 0.17}$	0.62$_{\pm 0.10}$	0.60$_{\pm 0.18}$	0.57$_{\pm 0.18}$	0.62$_{\pm 0.11}$	0.63$_{\pm 0.12}$	0.55$_{\pm 0.19}$	0.60$_{\pm 0.10}$
SuMi	0.56$_{\pm 0.11}$	0.54$_{\pm 0.14}$	0.58$_{\pm 0.08}$	0.58$_{\pm 0.13}$	0.54$_{\pm 0.16}$	0.59$_{\pm 0.10}$	0.59$_{\pm 0.11}$	0.53$_{\pm 0.16}$	0.57$_{\pm 0.09}$
SoTTA	0.60$_{\pm 0.13}$	0.56$_{\pm 0.15}$	0.61$_{\pm 0.09}$	0.60$_{\pm 0.13}$	0.55$_{\pm 0.17}$	0.61$_{\pm 0.10}$	0.60$_{\pm 0.12}$	0.54$_{\pm 0.17}$	0.58$_{\pm 0.09}$
AeO	0.45$_{\pm 0.04}$	0.42$_{\pm 0.06}$	0.45$_{\pm 0.04}$	0.43$_{\pm 0.05}$	0.41$_{\pm 0.05}$	0.44$_{\pm 0.04}$	0.43$_{\pm 0.04}$	0.41$_{\pm 0.04}$	0.42$_{\pm 0.04}$
PTA	0.54$_{\pm 0.11}$	0.54$_{\pm 0.14}$	0.58$_{\pm 0.08}$	0.59$_{\pm 0.13}$	0.55$_{\pm 0.17}$	0.60$_{\pm 0.10}$	0.61$_{\pm 0.10}$	0.56$_{\pm 0.17}$	0.59$_{\pm 0.09}$
PSDNorm	0.49$_{\pm 0.06}$	0.48$_{\pm 0.07}$	0.53$_{\pm 0.05}$	0.53$_{\pm 0.07}$	0.51$_{\pm 0.09}$	0.56$_{\pm 0.06}$	0.56$_{\pm 0.06}$	0.53$_{\pm 0.08}$	0.56$_{\pm 0.06}$
CARAT	**0.62**$_{\pm 0.16}$	**0.60**$_{\pm 0.18}$	**0.63**$_{\pm 0.09}$	**0.66**$_{\pm 0.15}$	**0.63**$_{\pm 0.19}$	**0.65**$_{\pm 0.10}$	0.63$_{\pm 0.15}$	**0.62**$_{\pm 0.16}$	**0.63**$_{\pm 0.09}$

O.2 Additional component ablations

In Table 2, R8 retains detection and routing, fixes $a^j = \rho^j$ for suspect modalities $j \in \mathcal{S}_i$, and disables the optimizer. Comparison with full CARAT measures the joint benefit of normalization and attenuation updates. R9 applies CARAT unchanged to a uniform-trained ShaSpec checkpoint with a uniform prior; source weights and the prior both change. It measures their combined contribution, while Table 4 provides the complementary intervention on the routing prior itself. These ablations report point estimates.

O.3 UNINFORMATIVE-MODALITY CONTROL

On DaliaHAR–MAESTRO, we replace one, two, or three input modalities with variance-matched white noise before SAX tokenization and keep the replacement fixed across training epochs. This tests source training in the presence of persistently uninformative streams. Table 36 reports the terminal dropout profile ρ^j used by CARAT. Replaced streams receive values of 0.00–0.01, while the remaining wrist/chest accelerometers retain values of 0.80. Classification F1 is 80.40–82.50 points in the replacement settings, compared with 80.20 ± 1.40 for the three-seed control. The available means show no observed F1 reduction, but replacement-run dispersions were not reported, so this comparison does not establish statistical equivalence. It also does not test whether a prior learned before a sensor failure remains appropriate after a persistent deployment shift.

Table 36: **Variance-matched white-noise control on DaliaHAR–MAESTRO.** Bold profile entries identify the replaced streams, not the best values. ρ^j is the terminal dropout probability, rather than a profile normalized to sum to one. F1 is displayed on a 0–100 scale. Only the unmodified control includes a reported three-seed spread.

	Terminal dropout profile ρ^j					
White-noise modalities	w_acc	w_bvp	w_eda	w_temp	c_acc	F1 points
None	0.75	0.80	0.66	0.40	0.80	$80.20_{\pm 1.40}$
w_temp	0.80	0.80	0.44	**0.01**	0.80	81.40
w_temp, w_eda	0.80	0.80	**0.00**	**0.00**	0.80	82.50
w_temp, w_eda, w_bvp	0.80	**0.00**	**0.00**	**0.00**	0.80	80.40

O.4 SENSITIVITY TO INPUT-PROJECTION WIDTH

We separately widen one modality's input projection eightfold and retrain DaliaHAR–MAESTRO with the remaining setup fixed: chest acceleration from hidden width 96 to 768, or wrist temperature from 32 to 256. Table 37 reports the normalized gradient-based reliance ratios $r^j_{e^\star}$, which precede the dropout transformation and should not be confused with the terminal dropout probabilities ρ^j in Table 36. The widened modality's ratio increases by 2.90 percentage points for chest acceleration and 4.80 points for wrist temperature. Mean classification F1 changes from 80.30 to 80.20 and 79.50 points, respectively. These limited retraining controls do not show reliance increasing in proportion to the eightfold width change, but several other modalities' ratios also change; they therefore do not establish parameterization invariance. CARAT uses the reliance profile with the checkpoint from which it was estimated.

Table 37: **Input-projection widening on DaliaHAR–MAESTRO.** Reported mean$_{\pm \mathrm{SD}}$; bold identifies the widened modality. Reliance ratios are percentages; F1 uses a 0–100 scale. These are separate retraining controls from the white-noise experiment.

	Normalized reliance $100\, r^j_{e^\star}$ (%)					
Input projection	w_acc	w_bvp	w_eda	w_temp	c_acc	F1 points
Original widths	$29.50_{\pm 4.40}$	$17.80_{\pm 1.30}$	$19.30_{\pm 3.10}$	$9.80_{\pm 4.10}$	$23.70_{\pm 13.00}$	$80.30_{\pm 0.70}$
c_acc ×8	$22.20_{\pm 2.90}$	$17.80_{\pm 1.70}$	$12.10_{\pm 9.00}$	$21.30_{\pm 6.20}$	$\mathbf{26.60}_{\pm 11.60}$	$80.20_{\pm 0.70}$
w_temp ×8	$19.30_{\pm 4.30}$	$28.20_{\pm 7.80}$	$14.60_{\pm 5.70}$	$\mathbf{14.60}_{\pm 11.50}$	$23.40_{\pm 3.80}$	$79.50_{\pm 2.10}$