RETHINKING MODALITY FUSION AT TEST TIME: LABEL-FREE SELECTION FOR ROBUST MULTIMODAL SENSING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal sensing benefits from complementary observations, but distribution shift can change the utility of individual modalities, making indiscriminate fusion detrimental. Selecting useful modalities without target labels is challenging: incorrect predictions can remain confident, and cross-modal agreement can be inflated by shared predicted-class preferences. We propose Test-Time Modality Selection (TMS), a label-free framework that selects a shared modality subset for an unlabeled target batch. Using cached predictions from classifiers frozen after source training or few-shot adaptation, TMS combines prediction coverage screening, cross-modal support estimation through empirical-marginal-corrected agreement, and reference-based marginal contribution assessment. Auxiliary calibration supports coalition comparisons, while final predictions use unweighted averaging of the selected modalities' original logits. Selection requires no source calibration data, target labels, parameter updates, or additional classifier forward passes. Experiments on MM-Fi and XRF55 evaluate cross-environment transfer under zero-, one-, and two-shot adaptation and controlled modality corruption. TMS improves mean zero-shot accuracy on MM-Fi from 32.16% to 58.79% and one-shot accuracy under severe mmWave corruption on XRF55 from 20.60% to 35.12%. These results demonstrate the value of reassessing modality participation at test time using unlabeled target predictions.

1 INTRODUCTION

Multimodal sensing combines complementary observations from cameras, depth sensors, radar, LiDAR, and wireless signals to recognize human activities (Vrigkas et al., 2015; Nanzer, 2017). However, the relative reliability of these modalities can change when a system encounters a new target distribution (Yang et al., 2024). As illustrated in Figure 1, a modality that contributes useful information in the source domain may become unreliable in the target domain while remaining available for fusion. Averaging its predictions with those of other modalities can consequently degrade recognition performance. In our zero-shot cross-environment experiments on MM-Fi, full fusion achieves 32.16% accuracy, whereas the proposed modality selector reaches 58.79%. The appropriate combination also changes after few-shot adaptation, as previously unreliable modalities can regain predictive value. These observations motivate deployment-time modality selection that accounts for changes in both the target distribution and classifier state. The central challenge is to determine, without target labels, which modalities remain reliable and whether their inclusion benefits fusion.

Commonly used indicators provide incomplete evidence for this decision. Maximum softmax probability and predictive entropy characterize a classifier's output distribution, but confident predictions can be incorrect, and uncertainty estimates can become unreliable under distribution shift (Guo et al., 2017; Ovadia et al., 2019). Source-calibrated accuracy estimators such as Average Thresholded Confidence provide a more informed alternative, although their validity depends on assumptions about the shift between source and target distributions (Garg et al., 2022). Relationships among classifier predictions offer additional evidence for label-free evaluation under suitable assumptions (Parisi et al., 2014). In our setting, cross-modal agreement compares predictions on the same target instances, but raw agreement can be inflated when different modalities concentrate on the same few classes, motivating correction for agreement expected by chance. Furthermore, identifying a modal-

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.

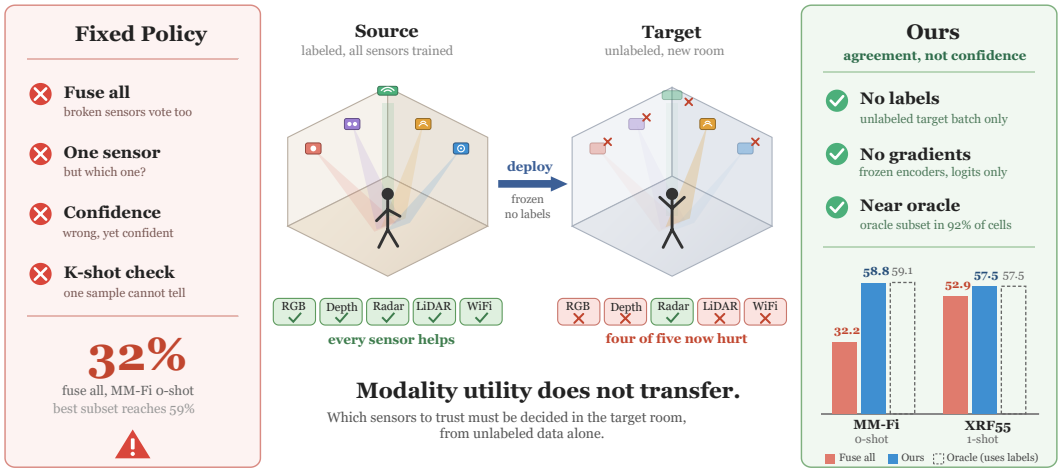

Figure 1: Motivation for Test-Time Modality Selection (TMS). Environmental shift can turn useful source-domain modalities into unreliable contributors to target-domain fusion. Our method uses cross-modal prediction statistics on an unlabeled target batch to select modalities while keeping the classifiers frozen. The subset oracle uses target labels and serves only as an evaluation reference.

ity with strong predictive support does not establish that adding it improves fusion: its information may overlap with other modalities, and its inclusion can change the fused decision unfavorably. Effective selection therefore requires examining both the support for individual modalities and their incremental contributions to a combination, without treating either confidence or agreement as direct evidence of target accuracy.

We propose Test-Time Modality Selection (TMS), a label-free framework that addresses these two aspects through successive screening and contribution assessment. Using cached predictions from classifiers frozen after source training or few-shot adaptation, TMS selects a shared modality subset for an unlabeled target batch. It first examines the effective coverage of each modality's predicted label distribution, then estimates cross-modal support using agreement corrected for empirical prediction marginals. Among the surviving candidates, it assesses marginal contribution by comparing coalition predictions with and without a modality against held-out reference modalities. An auxiliary rank-one agreement model supplies calibration for these comparisons, while final predictions use an unweighted average of the selected modalities' original logits. Selection requires no source calibration data, target labels, classifier updates, or additional classifier forward passes. Experiments on MM-Fi and XRF55 evaluate the same rule across zero-, one-, and two-shot adaptation and controlled modality corruption. Beyond the substantial zero-shot improvement, TMS increases one-shot accuracy from 86.00% to 86.88% on MM-Fi and from 54.85% to 57.48% on XRF55. Under severe mmWave corruption on XRF55, it improves one-shot accuracy from 20.60% to 35.12%, demonstrating the value of reassessing modality participation as deployment conditions change. Our main contributions are as follows:

- We formulate deployment-time modality subset selection from unlabeled target predictions and empirically examine how suitable modality combinations change across environments and adaptation budgets.
- We develop a selector that combines prediction-coverage screening, empirical-marginal-corrected cross-modal support, and reference-based marginal contribution assessment, while retaining frozen classifiers and unweighted final fusion.
- We evaluate the method on two multimodal sensing benchmarks across cross-environment transfer and controlled modality corruption, and use component ablations to assess the roles of the selection mechanisms.

2 Related Work

2.1 Multimodal Human Sensing

Human sensing has been extensively investigated using visual, wearable, and radio frequency sensors (Vrigkas et al., 2015; Nanzer, 2017; Nirmal et al., 2021; Das et al., 2021; Marcard et al., 2016). Visual sensors capture rich appearance and geometric information, but their deployment is affected by illumination, occlusion, and privacy concerns (Nguyen et al., 2016; Elharrouss et al., 2021; Shahroudy et al., 2016; Pavlakos et al., 2018; Goel et al., 2023; Cai et al., 2023). Non-intrusive sensing modalities, including WiFi CSI, mmWave radar, RFID, and LiDAR, offer complementary observations while exhibiting modality-specific limitations in spatial resolution, propagation conditions, environmental sensitivity, and measurement noise (Zhao et al., 2018; Zhou et al., 2023; 2022; D. Gian et al., 2024; An & Ogras, 2022; Wang et al., 2021; Li et al., 2022; Ren et al., 2024; Wang et al., 2024). To exploit this complementarity, multimodal approaches employ feature-level and decision-level fusion, attention mechanisms, and transformer-based representations (Zheng et al., 2022; An et al., 2022; Chen et al., 2023; Man et al., 2023; Xue & Marculescu, 2023; Ma et al., 2022; Chen & Yang, 2025). These approaches primarily investigate how to integrate information across modalities. An additional deployment challenge arises when distribution shift changes the utility of individual streams, so that an available modality can reduce the accuracy of the fused prediction. Our work addresses this challenge through deployment-time modality subset selection, using unlabeled target predictions to assess cross-modal support and marginal contribution while keeping the recognition models frozen.

2.2 Label-Free Reliability Estimation and Modality Selection

Estimating predictive reliability without target labels has been studied through confidence calibration and unsupervised ensemble learning. Calibration improves the correspondence between confidence and accuracy (Guo et al., 2017), but can become unreliable under distribution shift (Ovadia et al., 2019). ATC estimates target accuracy using labeled source validation data and unlabeled target predictions (Garg et al., 2022). Spectral meta-learning instead exploits prediction covariances to rank and combine classifiers under conditional-independence assumptions (Parisi et al., 2014). In multimodal test-time adaptation, READ addresses modality reliability differences by adapting attention-based fusion (Yang et al., 2024). Our method selects modalities with frozen classifiers, combining prediction-coverage screening, chance-corrected agreement, and reference-based marginal contribution assessment. It requires no source calibration data and retains unweighted raw-logit fusion.

3 Method

3.1 Problem Formulation and Overview

We introduce TMS for multimodal classification under distribution shift. Let $\mathcal{M} = \{1, \dots, M\}$ denote the available modalities sharing a label space $\mathcal{Y} = \{1, \dots, C\}$. For an unlabeled target batch $\mathcal{U}_T = \{\mathbf{x}^{(n)}\}_{n=1}^{N}$, each modality-specific classifier produces cached logits $Z_m \in \mathbb{R}^{N \times C}$ and hard predictions $\hat{y}_m^{(n)} = \arg\max_c Z_m[n, c]$. Predictions are aligned across modalities by sample identity. Classifiers may be source-trained or previously adapted using labeled target support data, but remain frozen throughout selection. The selector accesses only the cached target logits, without using target-batch labels or the adaptation support set, and returns a nonempty subset $\widehat{S}$ shared across the batch.

For any nonempty subset S, we define its prediction by unweighted raw-logit averaging:

$$\bar{\mathbf{z}}_S^{(n)} = \frac{1}{|S|} \sum_{m \in S} Z_m[n, :], \qquad \hat{y}_S^{(n)} = \arg\max_c \bar{z}_{S,c}^{(n)}. \tag{1}$$

The ideal subset minimizes target classification risk:

$$S^\star \in \arg\min_{\varnothing \neq S \subseteq \mathcal{M}} R_T(S), \qquad R_T(S) = \mathbb{E}_{(\mathbf{x},y)\sim P_T}\left[\mathbf{1}\{\hat{y}_S(\mathbf{x}) \neq y\}\right]. \tag{2}$$

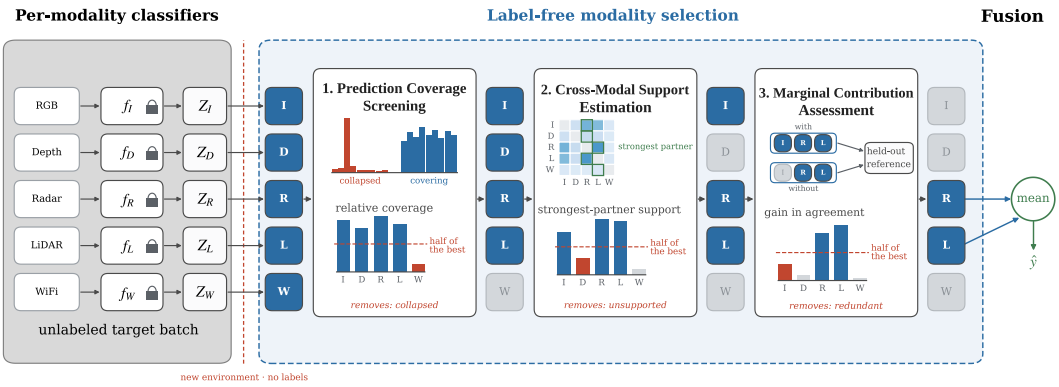

Figure 2: Overview of Test-Time Modality Selection (TMS). Frozen classifiers provide aligned target logits for prediction coverage screening, cross-modal support estimation, and marginal contribution assessment. Blue and gray blocks indicate retained and excluded candidates, respectively; all original predictions remain available for statistical evaluation. Final predictions average the selected modalities' original logits. The illustrated subset is schematic.

Without target labels, this risk cannot be directly evaluated. We instead use observable properties of target predictions to assess whether modalities have sufficient predictive support and whether they contribute additional information to a coalition.

Figure 2 organizes the procedure into three parts. Frozen classifiers on the left produce the cached logits. The central selector successively applies prediction-coverage screening, marginal-corrected agreement screening, and reference-based marginal contribution assessment. The fusion block on the right averages the selected modalities' original logits. The modality blocks between stages represent changing eligibility for final fusion, rather than removal of predictions from subsequent statistical computations. Selection therefore changes the participating modalities without updating classifier parameters or requiring additional classifier forward passes. Numerical details and the complete procedure are provided in Appendix A.

3.2 Prediction-Coverage Screening

Under distribution shift, a modality may confidently predict only a small subset of classes. Stage 1 in Figure 2 examines this batch-level behavior using the empirical predicted-class distribution and effective number of predicted classes:

$$\pi_m(c) = \frac{1}{N}\sum_{n=1}^{N} \mathbf{1}\{\hat{y}_m^{(n)} = c\}, \qquad E_m = \exp\left(-\sum_{c:\pi_m(c)>0} \pi_m(c)\log \pi_m(c)\right). \tag{3}$$

A constant predictor has $E_m = 1$, whereas a predictor uniformly covering k classes has $E_m = k$. Coverage thus measures diversity across predicted labels rather than uncertainty within individual predictions. A classifier can have low predictive entropy on every instance while repeatedly selecting the same class, making these two statistics convey different information.

We retain modalities whose coverage is sufficiently large relative to the widest-coverage modality:

$$\rho_m = \frac{E_m}{\max_{j\in\mathcal{M}} E_j}, \qquad \mathcal{L} = \{m \in \mathcal{M} : \rho_m \geq \tau_{\text{cov}}\}. \tag{4}$$

Relative screening compares modalities on the same target batch without requiring every class to be represented or imposing a fixed absolute coverage threshold. Lower coverage is treated as a screening signal, not an accuracy estimate. This distinction matters because a broad prediction distribution alone does not establish correctness. The next stage therefore evaluates instance-level relationships between modalities. Predictions from modalities excluded from $\mathcal{L}$ remain available for these comparisons, although they are no longer eligible for final fusion. An empty $\mathcal{L}$ falls back to $\mathcal{M}$.

3.3 CROSS-MODAL SUPPORT ESTIMATION

Stage 2 in Figure 2 measures cross-modal support after accounting for predicted-class preferences. We compute observed pairwise agreement and its chance baseline under independent empirical prediction marginals:

$$A_{ij} = \frac{1}{N}\sum_{n=1}^{N} \mathbf{1}\{\hat{y}_i^{(n)} = \hat{y}_j^{(n)}\}, \qquad e_{ij} = \sum_{c=1}^{C} \pi_i(c)\pi_j(c). \tag{5}$$

When modalities favor the same few classes, raw agreement can largely reflect those shared preferences. We normalize the agreement exceeding this baseline using Cohen's κ:

$$\widetilde{\kappa}_{ij} = \frac{A_{ij} - e_{ij}}{1 - e_{ij}}. \tag{6}$$

Degenerate denominators are assigned zero, as specified in Appendix A. Independent prediction variables have zero corrected agreement at the population level when the denominator is nondegenerate. The statistic therefore measures association beyond the marginal-independence baseline rather than classification accuracy.

Each modality is scored by its strongest partner:

$$s_m = \max_{j \in \mathcal{M}\setminus\{m\}} \widetilde{\kappa}_{mj}, \qquad s_{\max} = \max_{j \in \mathcal{L}} s_j. \tag{7}$$

The maximum preserves a modality's strongest available support instead of requiring agreement with every sensor. This choice accommodates heterogeneous modalities whose predictive relationships can differ across partners. It also avoids requiring a global rank-one model to assign the initial selection scores.

Partner search includes all original modalities, while only members of $\mathcal{L}$ are eligible for retention:

$$S_1 = \begin{cases} \{m \in \mathcal{L} : s_m / s_{\max} \geq \tau_{\text{str}}\}, & s_{\max} > \epsilon_s, \\ \mathcal{L}, & \text{otherwise}, \end{cases} \qquad \epsilon_s = 10^{-9}. \tag{8}$$

The fallback avoids ratio tests on zero or negative maxima. We set $\tau_{\text{cov}} = \tau_{\text{str}} = 0.5$. Coverage and strength screening provide complementary checks on prediction distributions and cross-modal support; neither directly evaluates the effect of adding a modality to fusion. The remaining candidates are therefore assessed through coalition comparisons.

3.4 AUXILIARY COALITION CALIBRATION

We construct auxiliary calibrated outputs for the coalition comparisons in Stage 3 of Figure 2. Different classifiers can produce logits with different scales, causing a high-magnitude stream to dominate a coalition prediction. We use rank-one competence estimation and temperature matching to construct the outputs used in these comparisons. This branch supplies neither the initial strength scores nor the weights for final prediction.

The auxiliary fit retains uniformly chance-corrected agreement:

$$K_{ij}^{\text{uni}} = \frac{A_{ij} - 1/C}{1 - 1/C}, \qquad \widehat{\mathbf{q}} \approx \arg\min_{\mathbf{q} \in [0,1]^M} \sum_{i<j} \left(K_{ij}^{\text{uni}} - q_i q_j\right)^2. \tag{9}$$

Under class-conditional independence, class-independent accuracies, and uniformly distributed errors over incorrect classes, $K_{ij}^{\text{uni}} = q_i q_j$ with $q_m = (Cp_m - 1)/(C - 1)$. We use this idealized relationship as an approximate calibration model. Only off-diagonal agreement entries enter the fit, so self-agreement is not treated as evidence of competence. The fitted factors are converted to

$$\widehat{p}_m = \text{clip}_{[1/C,\, 1-10^{-4}]}\left(\frac{1 + (C-1)\widehat{q}_m}{C}\right). \tag{10}$$

These quantities are model-based auxiliary estimates rather than observed target accuracies. The uniform baseline in this branch is distinct from the empirical-marginal baseline used for strength selection.

For each modality, we choose a temperature T_m to approximately match the mean maximum softmax probability to a clipped version of $\widehat{p}_m$. Temperature scaling preserves the modality's hard predictions while changing its contribution to coalition scores. The calibrated outputs and coalition predictions are

$$G_m[n, c] = w_m \log \mathrm{softmax}(Z_m[n, :]/T_m)_c\,, \qquad w_m = \log \frac{(C-1)\check{p}_m}{1-\check{p}_m}, \tag{11}$$

$$\widetilde{y}_B^{(n)} = \arg\max_c \sum_{m \in B} G_m[n, c], \tag{12}$$

where $\check{p}_m$ is a numerically clipped version of $\widehat{p}_m$. Appendix A specifies coordinate updates, temperature search, and clipping bounds. These outputs are used only in the marginal comparisons below; final fusion continues to use the original logits, and the rank-one residual does not trigger abstention.

3.5 Marginal Contribution Assessment

Stage 3 in Figure 2 evaluates whether a supported modality provides additional contribution to a coalition. For modality m and reference $r \neq m$, define $B_r = \mathcal{M} \setminus \{r\}$ and $B_{r,-m} = \mathcal{M} \setminus \{r, m\}$. Both coalitions exclude the reference, preventing its output from directly contributing to the prediction being evaluated, and differ only in whether m is included. Their auxiliary predictions are compared with the reference's original hard predictions. Averaging over references incorporates evidence from multiple modalities instead of treating one sensor as ground truth.

Let $\kappa^{\mathrm{uni}}(\mathbf{a}, \mathbf{b}) = (N^{-1} \sum_n \mathbf{1}\{a_n = b_n\} - 1/C)/(1 - 1/C)$. For $M \geq 3$, define

$$u_m = \frac{1}{M-1} \sum_{r \in \mathcal{M} \setminus \{m\}} \left[\kappa^{\mathrm{uni}}(\widetilde{\mathbf{y}}_{B_r}, \hat{\mathbf{y}}_r) - \kappa^{\mathrm{uni}}\big(\widetilde{\mathbf{y}}_{B_{r,-m}}, \hat{\mathbf{y}}_r\big)\right]. \tag{13}$$

A positive score indicates that including m increases average agreement with the held-out references. The common chance baseline cancels within each difference, making u_m proportional to the mean change in reference agreement. It is a proxy for marginal contribution, not a direct estimate of labeled risk reduction. We evaluate the largest coalition excluding each reference and its single-modality deletion, rather than averaging over all coalition sizes. All coalitions and references use the original set $\mathcal{M}$, including modalities excluded by earlier screening; S_1 only restricts candidates eligible for final retention.

Writing $u_{\max} = \max_{j \in \mathcal{M}} u_j$, we normalize $\eta_m = u_m / u_{\max}$ when $u_{\max} > 10^{-9}$ and identify $m^- \in \arg\min_{m \in S_1} \eta_m$. The final subset is

$$\widehat{S} = \begin{cases} S_1 \setminus \{m^-\}, & |S_1| \geq 3,\ u_{\max} > 10^{-9},\ \eta_{m^-} < \tau_{\mathrm{marg}}, \\ S_1, & \text{otherwise}, \end{cases} \qquad \tau_{\mathrm{marg}} = 0.5. \tag{14}$$

The criterion is relative: a candidate can be removed for contributing substantially less than the strongest modality even if its marginal score is positive. If comparisons are unavailable or no positive normalization maximum exists, S_1 is retained. This update removes at most one modality, without recursive elimination or score recomputation. The rightmost block in Figure 2 then averages the retained modalities' original logits according to Eq. equation 1.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate on MM-Fi (Yang et al., 2023) and XRF55 (Wang et al., 2024). MM-Fi contains 27 activity classes across four environments with five modalities: RGB, depth, mmWave, LiDAR, and WiFi CSI. XRF55 contains 55 activity classes across four scenes; we use mmWave, WiFi, and RFID. We evaluate all 12 directed environment transfers on MM-Fi and transfer from Scene 1 to Scenes 2–4 on XRF55. Dataset descriptions and calibration splits are provided in Appendix B.

Table 1: Complete cross-environment results on MM-Fi and XRF55 without additional corruption. Accuracy is reported in percent. All averages the logits of all available modalities, ATC selects a subset from confidence-based accuracy estimates calibrated on source validation data, and Ours selects a subset from chance-corrected cross-modal agreement.

		0-shot			1-shot			2-shot		
Dataset	Transfer	All	ATC	Ours	All	ATC	Ours	All	ATC	Ours
MM-Fi	E01→E02	26.62	61.95	**64.48**	84.42	85.43	**86.11**	86.14	86.10	**86.77**
	E01→E03	16.02	7.52	**60.13**	83.90	84.25	**85.30**	86.50	86.77	**86.95**
	E01→E04	12.86	26.81	**33.97**	85.03	75.43	**86.20**	86.30	78.36	**87.17**
	E02→E01	50.41	3.72	**76.58**	89.41	82.75	**89.41**	89.74	88.10	**89.75**
	E02→E03	44.54	51.37	**63.62**	86.34	84.58	**86.83**	87.20	85.63	**87.20**
	E02→E04	25.45	3.95	**46.77**	85.86	85.86	**86.67**	86.71	86.63	**87.08**
	E03→E01	47.06	5.16	**63.17**	88.09	88.33	**88.72**	88.93	87.85	**89.57**
	E03→E02	47.11	32.78	**65.19**	85.19	71.33	**85.92**	86.13	84.98	**86.13**
	E03→E04	34.51	5.45	**60.19**	86.05	86.05	**86.41**	87.18	87.18	**87.50**
	E04→E01	26.99	56.25	**56.25**	88.94	88.94	**89.47**	90.08	77.49	**90.28**
	E04→E02	25.91	3.78	**53.85**	83.65	83.65	**85.03**	86.06	78.45	**86.67**
	E04→E03	28.42	3.78	**61.25**	85.12	80.25	**86.45**	87.05	83.57	**87.20**
XRF55	S1→S2	5.01	2.01	**9.70**	51.36	53.16	**54.13**	62.53	64.56	**65.86**
	S1→S3	5.49	2.65	**19.65**	61.69	59.39	**61.69**	74.51	73.55	**74.51**
	S1→S4	5.45	0.96	**21.82**	51.52	56.62	**56.62**	64.65	69.93	**69.93**

Evaluation protocol and baselines. We evaluate source-trained classifiers under target adaptation budgets of $K \in \{0, 1, 2\}$ shots, with $K = 0$ denoting no target adaptation. After adaptation, classifiers are frozen and selectors access only cached target-query logits, without support samples or query labels. We report mean top-1 accuracy over the 12 MM-Fi transfers or three XRF55 target scenes; corruption results are averaged over scenes separately for each condition. **All** averages all available modalities' raw logits. **Conf.** uses mean maximum softmax probability (Hendrycks & Gimpel, 2016), while **Ent.** uses one minus mean predictive entropy normalized by $\log C$; both retain modalities scoring at least half the maximum. **ATC** adapts Average Thresholded Confidence (Garg et al., 2022) to subset selection: each subset is temperature-scaled on labeled source validation data (Guo et al., 2017), and a confidence threshold is calibrated to reproduce its validation accuracy. The fraction of target predictions exceeding this threshold estimates the subset's target accuracy, and the highest-scoring subset is selected. Calibration uses the corresponding checkpoint, including after adaptation. **Ours** applies the selector in Section 3. All these methods use the same unweighted raw-logit fusion. **Oracle** exhaustively evaluates all $2^M - 1$ nonempty subsets using query labels, providing a reference upper bound under this fusion rule.

4.2 Cross-Environment Recognition Performance

Table 1 reports the complete cross-environment results on MM-Fi and XRF55 under zero-, one-, and two-shot adaptation. Additional experimental results are provided in the appendix.

Without target adaptation, our method substantially improves over full fusion on both datasets. On MM-Fi, mean accuracy increases from 32.16% to 58.79%, with improvements on all 12 transfers ranging from 16.11 to 44.11 percentage points. ATC instead achieves 21.88% and falls below full fusion on eight transfers, indicating that confidence estimates calibrated on source data can become unreliable under target shift. On XRF55, our method selects radar alone in all three target scenes and improves mean accuracy from 5.32% to 17.05%, whereas ATC reaches only 1.87%. Few-shot adaptation substantially improves full-fusion performance, reducing the gains available from selection. Nevertheless, our method reaches 86.88% and 87.69% on MM-Fi under one- and two-shot adaptation, respectively, compared with 86.00% and 87.34% for full fusion and 83.07% and 84.26% for ATC. Confidence-based selection therefore remains vulnerable after adaptation: for example, ATC retains only RGB and depth on E03→E02 at one shot, reducing accuracy to 71.33%, compared with 85.19% for full fusion. On XRF55, ATC becomes more competitive after adaptation, reaching 56.39% and 69.35%, while our method achieves 57.48% and 70.10%, exceeding full fusion by 2.62 and 2.87 percentage points. Across all reported transfers and adaptation budgets, our

Table 2: XRF55 accuracy (%) averaged over Scenes 2–4 under eight corruption conditions for each adaptation budget. Conf. and Ent. select a subset from confidence- and entropy-based reliability scores, ATC selects a subset from confidence-based accuracy estimates calibrated on source validation data, and Ours selects a subset from chance-corrected cross-modal agreement. All methods fuse the selected modalities by unweighted logit averaging. Oracle searches all nonempty subsets with target labels and is shown for reference only.

Shots	Condition	All	Conf.	Ent.	ATC	Ours	Oracle
0	Clean	5.32	5.32	5.32	1.87	**17.05**	17.05
	WiFi mild	5.05	5.05	5.05	2.21	**17.05**	17.05
	WiFi moderate	5.59	5.59	5.59	2.13	**17.05**	17.05
	WiFi severe	6.08	6.08	6.08	7.58	**17.05**	17.05
	mmWave mild	2.13	2.13	2.13	1.87	**2.30**	3.12
	mmWave moderate	1.80	1.80	1.80	2.08	**2.30**	2.59
	mmWave severe	1.69	1.69	1.69	2.14	**2.30**	2.47
	Both moderate	2.22	2.22	2.22	2.26	**2.43**	2.72
1	Clean	54.85	54.09	54.85	56.39	**57.48**	57.48
	WiFi mild	52.49	51.24	52.49	53.69	**54.95**	54.95
	WiFi moderate	40.21	37.84	40.21	40.55	**41.66**	42.17
	WiFi severe	31.33	29.10	31.33	26.78	**38.20**	39.79
	mmWave mild	23.67	21.95	23.67	19.44	**35.12**	39.90
	mmWave moderate	20.95	19.33	20.95	18.97	**35.12**	39.90
	mmWave severe	20.60	18.44	20.60	19.05	**35.12**	39.90
	Both moderate	10.37	8.91	10.37	8.18	**12.84**	13.64
2	Clean	67.23	67.23	67.23	69.35	**70.10**	70.10
	WiFi mild	64.31	64.31	64.31	66.41	**66.41**	66.41
	WiFi moderate	51.98	51.98	51.98	54.19	**54.22**	54.93
	WiFi severe	42.56	42.56	42.56	51.67	**52.01**	53.11
	mmWave mild	33.56	33.56	33.56	44.58	**45.40**	52.44
	mmWave moderate	31.38	31.38	31.38	43.40	**45.40**	52.44
	mmWave severe	31.40	31.40	31.40	43.33	**45.40**	52.44
	Both moderate	13.30	13.30	13.30	10.56	**19.12**	19.39

method matches or exceeds both full fusion and ATC. These results show that the same selection rule can recover substantial performance under severe zero-shot shift while retaining smaller, consistent benefits after few-shot adaptation.

4.3 Robustness to Modality Corruption

We evaluate controlled sensor degradation on XRF55. WiFi corruption adds Gaussian noise with standard deviation α times the per-sample standard deviation, where $\alpha \in \{0.5, 1.0, 2.0\}$, followed by z-score normalization. For mmWave, we drop spatial bins using a mask shared across 17 frames and add Gaussian noise before normalization. Mild, moderate, and severe corruption pair drop ratios of $\{0.30, 0.60, 0.90\}$ with noise coefficients of $\{0.5, 1.0, 2.0\}$, respectively. Including the clean condition and simultaneous moderate corruption of both modalities yields eight conditions per adaptation budget. Table 2 reports all results averaged over Scenes 2–4, while Figure 3 visualizes single-modality corruption under one- and two-shot adaptation.

Our method matches or exceeds full fusion, Conf., Ent., and ATC in every reported condition. Figure 3 shows that its advantage over full fusion becomes more pronounced as corruption increases, particularly under mmWave degradation. Across the three mmWave corruption levels, our accuracy remains at 35.12% and 45.40% under one- and two-shot adaptation, exceeding full fusion by 11.45–14.52 and 11.84–14.02 percentage points, respectively. Under severe WiFi corruption, our method achieves 38.20% and 52.01%, compared with 31.33% and 42.56% for full fusion. The comparison with ATC depends on the adaptation budget: its one-shot accuracy falls to 26.78% under severe WiFi corruption, whereas its two-shot result of 51.67% approaches ours. Entropy-based selection consistently reproduces full-fusion performance, and confidence-based selection offers no consistent improvement. Table 2 further shows gains under simultaneous moderate corruption, from 10.37% to 12.84% at one shot and from 13.30% to 19.12% at two shots. Selection nevertheless remains con-

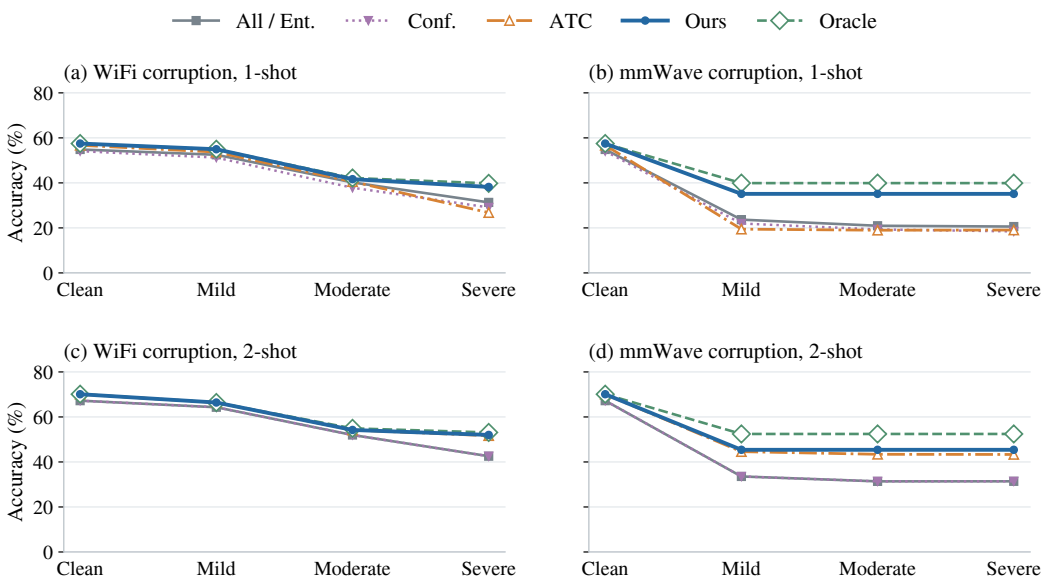

Figure 3: Accuracy under increasing WiFi and mmWave corruption on XRF55, averaged over Scenes 2–4.

strained by the available predictive information: two-shot mmWave results fall below the oracle's 52.44%, while zero-shot accuracy drops from 17.05% under WiFi-only corruption to 2.30–2.43% when mmWave is corrupted.

4.4 Ablation Study

Table 3 shows that removing coverage screening or marginal contribution assessment, or replacing empirical-marginal correction with a uniform chance baseline, reduces accuracy at every adaptation budget. Removing marginal contribution assessment causes the largest zero-shot drop (6.87 percentage points), removing coverage screening causes the largest one-shot drop (2.54 points), and uniform chance correction reduces zero-shot accuracy by 4.69 points. These results support the contribution of each component under the same fixed thresholds across adaptation budgets.

Table 3: Component ablation on MM-Fi. Accuracy (%) is averaged over 12 directed environment transfers for each adaptation budget.

Variant	0-shot	1-shot	2-shot
Full method	**58.79**	**86.88**	**87.69**
Uniform chance correction	54.10	84.96	87.57
w/o marginal utility	51.92	85.96	86.75
w/o coverage screening	52.68	84.34	86.75

5 Conclusion

We presented Test-Time Modality Selection (TMS), a label-free framework for reassessing modality participation under distribution shift. TMS combines prediction coverage screening, cross-modal support estimation, and reference-based marginal contribution assessment to select a batch-level modality subset while keeping classifiers frozen and final fusion unweighted. Experiments on MM-Fi and XRF55 show improvements over full fusion across adaptation budgets and corruption conditions, while ablations support the contribution of each selection component. These findings highlight the importance of evaluating which modalities should participate in fusion at test time, as their predictive value changes with the target environment and classifier state.

AI USE STATEMENT

In this work, we used generative AI tools to provide feedback on research methodology or experiments, assist with translation, and interpret results. We have not used generative AI tools to generate synthetic datasets, help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, design research methodology or experiments, implement methods, clean or reformat datasets, or support qualitative and thematic data analysis. We have reviewed all AI-assisted work. Feedback provided by generative AI on research methodology and experiments was independently evaluated by the authors, all translations were manually reviewed and revised, and AI-assisted interpretations of experimental results were verified against the actual experimental results. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Sizhe An and Umit Y Ogras. Fast and scalable human pose estimation using mmwave point cloud. In *Proceedings of the 59th ACM/IEEE Design Automation Conference*, pp. 889–894, 2022.

Sizhe An, Yin Li, and Umit Ogras. mri: Multi-modal 3d human pose estimation dataset using mmwave, rgb-d, and inertial sensors. *Advances in Neural Information Processing Systems*, 35: 27414–27426, 2022.

Zhongang Cai, Wanqi Yin, Ailing Zeng, Chen Wei, Qingping Sun, Wang Yanjun, Hui En Pang, Haiyi Mei, Mingyuan Zhang, Lei Zhang, et al. Smpler-x: Scaling up expressive human pose and shape estimation. *Advances in Neural Information Processing Systems*, 36:11454–11468, 2023.

Anjun Chen, Xiangyu Wang, Kun Shi, Shaohao Zhu, Bin Fang, Yingfeng Chen, Jiming Chen, Yuchi Huo, and Qi Ye. Immfusion: Robust mmwave-rgb fusion for 3d human body reconstruction in all weather conditions. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2752–2758. IEEE, 2023.

Xinyan Chen and Jianfei Yang. X-fi: A modality-invariant foundation model for multimodal human sensing. In *International Conference on Learning Representations*, volume 2025, pp. 97378–97397, 2025.

Toan D. Gian, Tien Dac Lai, Thien Van Luong, Kok-Seng Wong, and Van-Dinh Nguyen. Hpe-li: Wifi-enabled lightweight dual selective kernel convolution for human pose estimation. In *European Conference on Computer Vision*, pp. 93–111. Springer, 2024.

Avigyan Das, Pritam Sil, Pawan Kumar Singh, Vikrant Bhateja, and Ram Sarkar. Mmhar-ensemnet: A multi-modal human activity recognition model. *IEEE Sensors Journal*, 21(10):11569–11576, 2021. doi: 10.1109/JSEN.2020.3034614.

Omar Elharrouss, Noor Almaadeed, and Somaya Al-Maadeed. A review of video surveillance systems. *Journal of Visual Communication and Image Representation*, 77:103116, 2021.

Saurabh Garg, Sivaraman Balakrishnan, Zachary C Lipton, Behnam Neyshabur, and Hanie Sedghi. Leveraging unlabeled data to predict out-of-distribution performance. *arXiv preprint arXiv:2201.04234*, 2022.

Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4d: Reconstructing and tracking humans with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14783–14794, 2023.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.

Jialian Li, Jingyi Zhang, Zhiyong Wang, Siqi Shen, Chenglu Wen, Yuexin Ma, Lan Xu, Jingyi Yu, and Cheng Wang. Lidarcap: Long-range marker-less 3d human motion capture with lidar point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 20502–20512, 2022.

Mengmeng Ma, Jian Ren, Long Zhao, Davide Testuggine, and Xi Peng. Are multimodal transformers robust to missing modality? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18177–18186, 2022.

Yunze Man, Liang-Yan Gui, and Yu-Xiong Wang. Bev-guided multi-modality fusion for driving perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21960–21969, 2023.

Timo von Marcard, Gerard Pons-Moll, and Bodo Rosenhahn. Human pose estimation from video and imus. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(8):1533–1547, 2016. doi: 10.1109/TPAMI.2016.2522398.

Jeffrey A Nanzer. A review of microwave wireless techniques for human presence detection and classification. *IEEE Transactions on Microwave Theory and Techniques*, 65(5):1780–1794, 2017.

Duc Thanh Nguyen, Wanqing Li, and Philip O Ogunbona. Human detection from images and videos: A survey. *Pattern Recognition*, 51:148–175, 2016.

Isura Nirmal, Abdelwahed Khamis, Mahbub Hassan, Wen Hu, and Xiaoqing Zhu. Deep learning for radio-based human sensing: Recent advances and future directions. *IEEE Communications Surveys & Tutorials*, 23(2):995–1019, 2021.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems*, 32, 2019.

Fabio Parisi, Francesco Strino, Boaz Nadler, and Yuval Kluger. Ranking and combining multiple predictors without labeled data. *Proceedings of the National Academy of Sciences*, 111(4):1253–1258, 2014.

Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3d human pose and shape from a single color image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 459–468, 2018.

Yiming Ren, Xiao Han, Chengfeng Zhao, Jingya Wang, Lan Xu, Jingyi Yu, and Yuexin Ma. Livehps: Lidar-based scene-level human pose and shape estimation in free environment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1281–1291, 2024.

Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1010–1019, 2016.

Michalis Vrigkas, Christophoros Nikou, and Ioannis A Kakadiaris. A review of human activity recognition methods. *Frontiers in Robotics and AI*, 2:28, 2015.

Fei Wang, Yizhe Lv, Mengdie Zhu, Han Ding, and Jinsong Han. Xrf55: A radio frequency dataset for human indoor action analysis. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(1):1–34, 2024.

Yuheng Wang, Haipeng Liu, Kening Cui, Anfu Zhou, Wensheng Li, and Huadong Ma. m-activity: Accurate and real-time human activity recognition via millimeter wave radar. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8298–8302. IEEE, 2021.

Zihui Xue and Radu Marculescu. Dynamic multimodal fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2575–2584, 2023.

Jianfei Yang, He Huang, Yunjiao Zhou, Xinyan Chen, Yuecong Xu, Shenghai Yuan, Han Zou, Chris Xiaoxuan Lu, and Lihua Xie. Mm-fi: Multi-modal non-intrusive 4d human dataset for versatile wireless sensing. *Advances in Neural Information Processing Systems*, 36:18756–18768, 2023.

Mouxing Yang, Yunfan Li, Changqing Zhang, Peng Hu, and Xi Peng. Test-time adaptation against multi-modal reliability bias. In *International Conference on Learning Representations*, volume 2024, pp. 55036–55055, 2024.

Mingmin Zhao, Tianhong Li, Mohammad Abu Alsheikh, Yonglong Tian, Hang Zhao, Antonio Torralba, and Dina Katabi. Through-wall human pose estimation using radio signals. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7356–7365, 2018.

Jingxiao Zheng, Xinwei Shi, Alexander Gorban, Junhua Mao, Yang Song, Charles R Qi, Ting Liu, Visesh Chari, Andre Cornman, Yin Zhou, et al. Multi-modal 3d human pose estimation with 2d weak supervision in autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4478–4487, 2022.

Yue Zhou, Aichun Zhu, Caojie Xu, Fangqiang Hu, and Yifeng Li. Perunet: Deep signal channel attention in unet for wifi-based human pose estimation. *IEEE Sensors Journal*, 22(20):19750–19760, 2022.

Yunjiao Zhou, He Huang, Shenghai Yuan, Han Zou, Lihua Xie, and Jianfei Yang. Metafi++: Wifi-enabled transformer-based human pose estimation for metaverse avatar simulation. *IEEE Internet of Things Journal*, 10(16):14128–14136, 2023.

A Algorithm

This appendix specifies the numerical implementation of the selector in Section 3. All inputs are cached logits aligned by target-sample identity. Selection uses neither target labels nor adaptation support samples.

A.1 Screening and Numerical Conventions

We set $\widetilde{\kappa}_{ij} = 0$ when $e_{ij} \geq 1 - 10^{-12}$ and otherwise apply Eq. equation 6. Coverage screening falls back to $\mathcal{M}$ if $\mathcal{L}$ is empty, and strength screening retains $\mathcal{L}$ when $s_{\max} \leq 10^{-9}$. With the thresholds used here, a positive maximum always satisfies its own relative retention test.

The implementation has two configurable thresholds: a shared retention threshold $\tau_{\text{cov}} = \tau_{\text{str}} = 0.5$ and a marginal threshold $\tau_{\text{marg}} = 0.5$. Their distinct symbols indicate different roles in the procedure. Partner search, auxiliary fitting, reference selection, and coalition construction always use the complete modality set $\mathcal{M}$. Candidate masks determine only which modalities can participate in final fusion.

A.2 Auxiliary Rank-One Fitting

The auxiliary fit follows an idealized classifier model. Suppose modality m has class-independent accuracy p_m, assigns each incorrect class probability $(1 - p_m)/(C - 1)$, and is conditionally independent of other modalities given the true class. Then

$$\Pr(\hat{y}_i = \hat{y}_j \mid y) = p_i p_j + \frac{(1 - p_i)(1 - p_j)}{C - 1}. \tag{15}$$

Since this expression does not depend on y, it also gives the unconditional agreement probability. Substituting $q_m = (Cp_m - 1)/(C - 1)$ yields

$$\frac{A_{ij} - 1/C}{1 - 1/C} = q_i q_j \tag{16}$$

at the population level. Equation equation 9 fits this relationship approximately to empirical off-diagonal agreements.

Define $\mathrm{clip}_{[a,b]}(v) = \min\{b, \max\{a, v\}\}$. We initialize

$$q_i^{(0)} = \sqrt{\mathrm{clip}_{[10^{-6},1]}\left(\frac{1}{M-1}\sum_{j\neq i}\max\{K_{ij}^{\mathrm{uni}}, 0\}\right)} \tag{17}$$

and sequentially update each coordinate using the latest available values:

$$q_i \leftarrow \mathrm{clip}_{[0,1]}\left(\frac{\sum_{j\neq i} K_{ij}^{\mathrm{uni}} q_j}{\max\{\sum_{j\neq i} q_j^2, 10^{-12}\}}\right). \tag{18}$$

We perform at most 500 sweeps and stop when the maximum absolute coordinate change between consecutive sweeps falls below 10^{-10}. The fitted factors are converted to $\widehat{p}_m$ using Eq. equation 10. Neither the fitting residual nor a factor reaching its constraint boundary changes the selection rule.

A.3 Temperature Matching and Coalition Prediction

For each modality, define

$$g_m(T) = \frac{1}{N}\sum_{n=1}^{N}\max_c \mathrm{softmax}(Z_m[n,:]/T)_c\,, \qquad \widetilde{p}_m = \mathrm{clip}_{[1/C+10^{-4},\,1-10^{-4}]}(\widehat{p}_m). \tag{19}$$

Starting from $T_{\mathrm{lo}} = 0.02$ and $T_{\mathrm{hi}} = 500$, we perform 60 bisection steps in log-temperature space. Each step evaluates $T_{\mathrm{mid}} = \sqrt{T_{\mathrm{lo}}T_{\mathrm{hi}}}$. If $g_m(T_{\mathrm{mid}}) > \widetilde{p}_m$, the lower bound becomes T_{mid}; otherwise, the upper bound becomes T_{mid}. The final temperature is $T_m = \sqrt{T_{\mathrm{lo}}T_{\mathrm{hi}}}$. An unattainable target leads to a temperature near the corresponding search boundary.

The competence weight uses

$$\check{p}_m = \mathrm{clip}_{[1/C+10^{-6},\,1-10^{-6}]}(\widehat{p}_m), \qquad w_m = \log\frac{(C-1)\check{p}_m}{1-\check{p}_m}. \tag{20}$$

For a nonempty coalition B, the auxiliary prediction has the equivalent form

$$\widetilde{y}_B^{(n)} = \arg\max_c \sum_{m\in B} G_m[n,c] = \arg\max_c \sum_{m\in B}\frac{w_m}{T_m} Z_m[n,c]. \tag{21}$$

The equality follows because each log-softmax normalization term is constant across classes for a fixed instance. Auxiliary coalition predictions therefore use effective coefficients w_m/T_m, whereas final predictions use equal weights on the original logits.

A.4 Marginal Comparison and Complete Procedure

Writing $\mathrm{Agr}(\mathbf{a},\mathbf{b}) = N^{-1}\sum_n \mathbf{1}\{a_n = b_n\}$, the marginal score can be expressed as

$$u_m = \frac{1}{(M-1)(1-1/C)}\sum_{r\neq m}\left[\mathrm{Agr}(\widetilde{\mathbf{y}}_{B_r}, \hat{\mathbf{y}}_r) - \mathrm{Agr}\big(\widetilde{\mathbf{y}}_{B_{r,-m}}, \hat{\mathbf{y}}_r\big)\right]. \tag{22}$$

Only the common additive chance baseline cancels; correlated prediction errors do not generally cancel. Each pair excludes the reference modality from both coalitions. For five input modalities, this compares a four-modality coalition with its three-modality counterpart, irrespective of how many candidates survive the initial screening.

The normalization maximum is taken over $\mathcal{M}$, while the removal candidate is selected from S_1. If valid comparisons are unavailable, $u_{\max} \leq 10^{-9}$, or $|S_1| \leq 2$, no marginal deletion is performed. Scores are not recomputed after deletion. Algorithm 1 summarizes the procedure; marginal-score computation can be skipped when $|S_1| \leq 2$ because it cannot affect the output.

Algorithm 1 Test-Time Modality Selection (TMS)

Require: Aligned target logits $\{Z_m\}_{m\in\mathcal{M}}$; thresholds $\tau_{\mathrm{cov}} = \tau_{\mathrm{str}} = \tau_{\mathrm{marg}} = 0.5$

Ensure: Selected subset $\widehat{S}$ and predictions $\hat{\mathbf{y}}_{\widehat{S}}$

Compute hard predictions $\hat{\mathbf{y}}_m$, marginals π_m, and effective coverage E_m
$\mathcal{L} \leftarrow \{m : E_m / \max_j E_j \geq \tau_{\mathrm{cov}}\}$
if $\mathcal{L} = \varnothing$ **then**
    $\mathcal{L} \leftarrow \mathcal{M}$
end if
Compute A_{ij} and $\widetilde{\kappa}_{ij}$ for all modality pairs in $\mathcal{M}$
$s_m \leftarrow \max_{j\in\mathcal{M}\setminus\{m\}} \widetilde{\kappa}_{mj}$
$s_{\max} \leftarrow \max_{j\in\mathcal{L}} s_j$; $S_1 \leftarrow \mathcal{L}$
if $s_{\max} > 10^{-9}$ **then**
    $S_1 \leftarrow \{m \in \mathcal{L} : s_m / s_{\max} \geq \tau_{\mathrm{str}}\}$
end if
Fit $\widehat{\mathbf{q}}$ from K^{uni} and compute $\widehat{p}_m$
Find temperatures T_m and construct auxiliary outputs G_m
$\widehat{S} \leftarrow S_1$
if $|S_1| \geq 3$ **then**
    for each $m \in \mathcal{M}$ **do**
        Compare $\mathcal{M} \setminus \{r\}$ and $\mathcal{M} \setminus \{r, m\}$ for every $r \neq m$
        Average the reference-agreement differences to obtain u_m
    end for
    if valid scores are available **then**
        $u_{\max} \leftarrow \max_{j\in\mathcal{M}} u_j$
        if $u_{\max} > 10^{-9}$ **then**
            $m^- \leftarrow \arg\min_{m\in S_1} u_m$
            if $u_{m^-} / u_{\max} < \tau_{\mathrm{marg}}$ **then**
                $\widehat{S} \leftarrow S_1 \setminus \{m^-\}$
            end if
        end if
    end if
end if
$\hat{y}_{\widehat{S}}^{(n)} \leftarrow \arg\max_c |\widehat{S}|^{-1} \sum_{m\in\widehat{S}} Z_m[n, c]$ for all n
return $\widehat{S}, \hat{\mathbf{y}}_{\widehat{S}}$

B Datasets and Evaluation Splits

B.1 MM-Fi

MM-Fi provides over 320,000 synchronized multimodal frames from 40 participants performing 27 daily and rehabilitation activities in four environments. It includes RGB images, depth maps, mmWave radar, LiDAR, and WiFi CSI, denoted by I, D, R, L, and W, respectively. The dataset provides both activity labels and human-pose annotations; our experiments use the activity classification task. Participants are divided into four environment-specific groups, so cross-environment evaluation also involves subject variation. We treat environments E01–E04 as separate domains and evaluate every ordered pair of distinct environments, yielding 12 transfer tasks. Each task uses the source environment for classifier training and the target environment for few-shot adaptation and query evaluation.

B.2 XRF55

XRF55 records 55 indoor activities across four scenes, covering human–object interactions, human–human interactions, fitness activities, body motions, and human–computer interactions. Its RF sensing system includes mmWave radar, WiFi CSI, and RFID, denoted by R, W, and F. The recording setup uses nine WiFi links and 23 RFID tags, together with synchronized Kinect recordings. Our experiments use only the three RF modalities. Scene 1 serves as the source domain, and Scenes 2–4

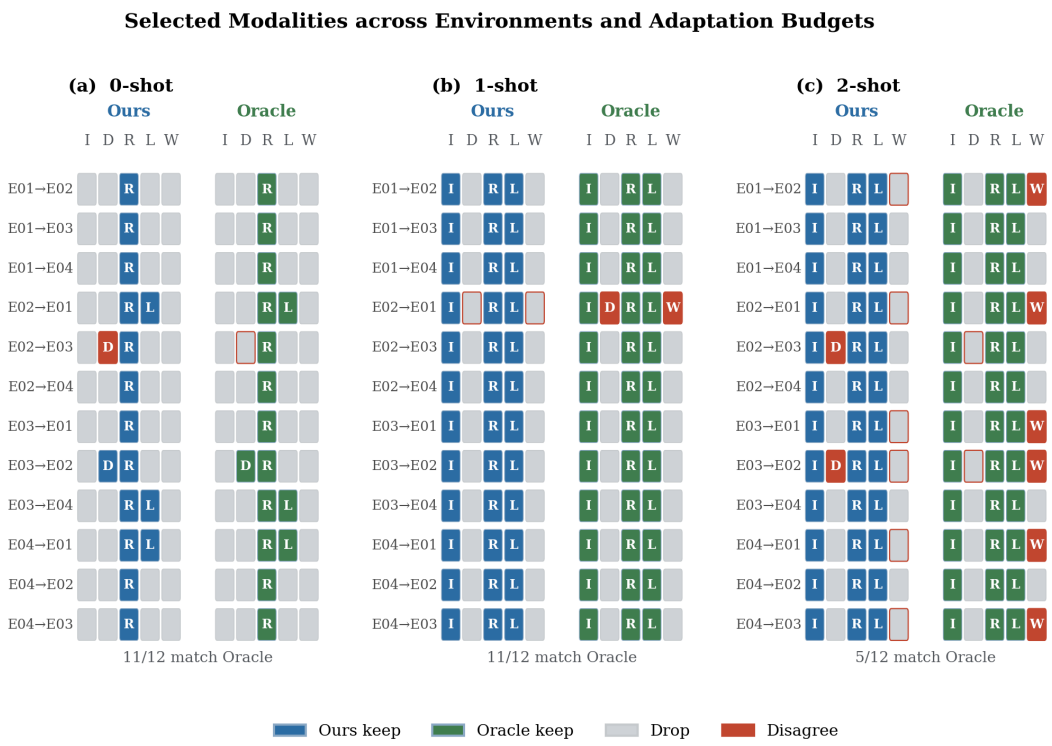

Figure 4: Modality subsets selected by TMS and the label-informed oracle across MM-Fi transfers under zero-, one-, and two-shot adaptation. I, D, R, L, and W denote RGB, depth, mmWave, LiDAR, and WiFi, respectively. Blue and green indicate retained modalities, gray indicates excluded modalities, and red highlights selection disagreements. Counts report exact matches with the displayed oracle subsets.

are evaluated separately as target domains. Besides clean transfer, we evaluate three WiFi corruption levels, three mmWave corruption levels, and simultaneous moderate corruption of both modalities. Including the clean condition, this yields eight conditions per adaptation budget. Corruption results are averaged over the three target scenes within each condition.

B.3 Source Calibration and Target Evaluation

ATC requires labeled source data to calibrate its accuracy estimates. On MM-Fi, the calibration split consists of the last two subjects in the source environment; on XRF55, it is the official Scene 1 test split. For each evaluated checkpoint, including checkpoints obtained after target adaptation, source logits are computed on the same calibration split. Each candidate subset is temperature-scaled, and its confidence threshold is fitted to match its source validation accuracy. These calibration parameters are then applied to the corresponding target logits for subset selection. Our selector requires no source calibration data. For both selectors, target support samples are used only in the preceding adaptation stage, and target-query labels are reserved for accuracy evaluation and oracle construction.

C Additional Results and Modality Selection Analysis

C.1 Modality Selection across Adaptation Budgets

Figure 4 compares the modality subsets selected by TMS with the reported oracle subsets across the 12 directed MM-Fi transfers. Without target adaptation, TMS predominantly retains mmWave, sometimes together with depth or LiDAR, whereas one-shot adaptation leads it to consistently re-

Table 4: Additional cross-environment accuracy (%) on MM-Fi and XRF55 without added corruption. All uses full fusion; Conf. and Ent. use confidence- and entropy-based selection. Oracle uses target labels to identify the best subset under the same fusion rule. Bold denotes the best label-free result among the methods shown.

		0-shot					1-shot					2-shot				
Dataset	Transfer	All	Conf.	Ent.	Ours	Oracle	All	Conf.	Ent.	Ours	Oracle	All	Conf.	Ent.	Ours	Oracle
MM-Fi	E01→E02	26.62	26.62	26.62	**64.48**	64.48	84.42	84.42	84.42	**86.11**	86.11	86.14	86.14	86.14	**86.77**	87.10
	E01→E03	16.02	16.02	16.02	**60.13**	60.13	83.90	83.90	83.90	**85.30**	85.30	86.50	86.50	86.50	**86.95**	86.95
	E01→E04	12.86	12.86	12.86	**33.97**	33.97	85.03	85.03	85.03	**86.20**	86.20	86.30	86.30	86.30	**87.17**	87.17
	E02→E01	50.41	50.41	50.41	**76.58**	76.58	89.41	89.41	89.41	**89.41**	89.41	89.74	89.74	89.74	**89.75**	90.48
	E02→E03	44.54	44.54	44.54	**63.62**	67.80	86.34	86.34	86.34	**86.83**	86.83	87.20	87.20	87.20	**87.20**	87.23
	E02→E04	25.45	25.45	25.45	**46.77**	46.77	85.86	85.86	85.86	**86.67**	86.67	86.71	86.71	86.71	**87.08**	87.08
	E03→E01	47.06	47.06	47.06	**63.17**	63.17	88.09	88.09	88.09	**88.72**	88.72	88.93	88.93	88.93	**89.57**	89.88
	E03→E02	47.11	47.11	47.11	**65.19**	65.19	85.19	85.19	85.19	**85.92**	85.92	86.13	86.13	86.13	**86.13**	86.86
	E03→E04	34.51	34.51	34.51	**60.19**	60.19	86.05	86.05	86.05	**86.41**	86.41	87.18	87.18	87.18	**87.50**	87.50
	E04→E01	26.99	26.99	26.99	**56.25**	56.25	88.94	88.94	88.94	**89.47**	89.47	90.08	90.08	90.08	**90.28**	90.78
	E04→E02	25.91	25.91	25.91	**53.85**	53.85	83.65	83.65	83.65	**85.03**	85.03	86.06	86.06	86.06	**86.67**	86.67
	E04→E03	28.42	28.42	28.42	**61.25**	61.25	85.12	85.12	85.12	**86.45**	86.45	87.05	87.05	87.05	**87.20**	87.20
XRF55	S1→S2	5.01	5.01	5.01	**9.70**	9.70	51.36	51.36	51.36	**54.13**	54.13	62.53	62.53	62.53	**65.86**	65.86
	S1→S3	5.49	5.49	5.49	**19.65**	19.65	61.69	59.39	61.69	**61.69**	61.69	74.51	74.51	74.51	**74.51**	74.51
	S1→S4	5.45	5.45	5.45	**21.82**	21.82	51.52	51.52	51.52	**56.62**	56.62	64.65	64.65	64.65	**69.93**	69.93
MM-Fi	Mean	32.16	32.16	32.16	**58.79**	59.14	86.00	86.00	86.00	**86.88**	86.88	87.34	87.34	87.34	**87.69**	87.91
XRF55	Mean	5.32	5.32	5.32	**17.05**	17.05	54.85	54.09	54.85	**57.48**	57.48	67.23	67.23	67.23	**70.10**	70.10

tain RGB, mmWave, and LiDAR. The selected subsets match the displayed oracle subsets in 11 of 12 transfers under both zero- and one-shot adaptation, and in 5 of 12 transfers under two-shot adaptation. Subset agreement and recognition accuracy nevertheless capture different aspects of selection: for one-shot E02→E01, TMS retains three modalities while the oracle retains all five, yet both achieve 89.41% accuracy. Likewise, despite lower subset agreement at two shots, the reported mean accuracy remains close to the oracle, at 87.69% versus 87.91%. These results illustrate how modality participation changes after adaptation and why exact subset agreement should be interpreted alongside predictive performance.

C.2 Extended Cross-Environment Comparisons

Table 4 supplements the main cross-environment results with confidence-based selection, entropy-based selection, and the subset oracle. Entropy-based selection reproduces full-fusion accuracy in every reported setting, while confidence-based selection differs only on one-shot XRF55 S1→S3, where accuracy decreases from 61.69% to 59.39%. TMS matches or exceeds all three label-free baselines shown in the table across all transfers and adaptation budgets. On MM-Fi, its reported mean accuracy is 0.35 percentage points below the oracle at zero shots, matches it at one shot, and is 0.22 points below it at two shots; on XRF55, TMS matches oracle accuracy in every reported setting. These comparisons show that cross-modal support and contribution assessment can identify useful modality combinations in settings where confidence- and entropy-based selection provide little benefit.