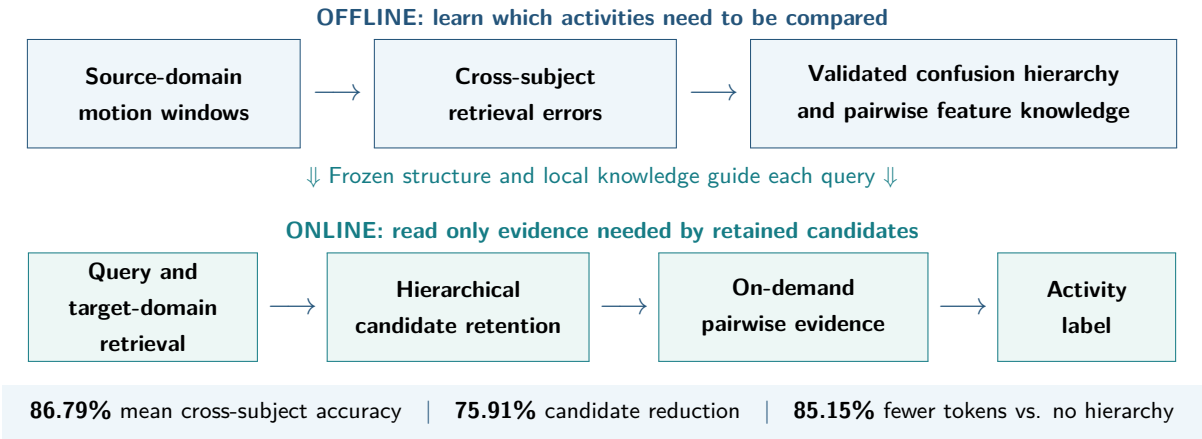


Graphical Abstract

CHEAR: Confusion-Hierarchy Evidence-Aware Reasoning for Cost-Efficient Human Activity Recognition

Xiangqing Lu, Tao Zhu, Mingxing Nie, Liming Chen

CHEAR: Organizing evidence before online activity recognition



Highlights

CHEAR: Confusion-Hierarchy Evidence-Aware Reasoning for Cost-Efficient Human Activity Recognition

Xiangqing Lu, Tao Zhu, Mingxing Nie, Liming Chen

- Confusion-guided organization of motion evidence for activity recognition.
- Offline construction of activity hierarchies and local discriminative knowledge.
- On-demand evidence use within hierarchically retained activity candidates.
- Accurate recognition with fewer candidates and lower language-model token use.

CHEAR: Confusion-Hierarchy Evidence-Aware Reasoning for Cost-Efficient Human Activity Recognition

Xiangqing Lu^a, Tao Zhu^{a,*}, Mingxing Nie^a and Liming Chen^b

^a*School of Computer, University of South China, Hengyang, China*

^b*School of Computer Science and Technology, Dalian University of Technology, Dalian, China*

ARTICLE INFO

Keywords:

human activity recognition
motion time series
retrieval-augmented reasoning
large language models
hierarchical inference
cross-dataset transfer

Abstract

Large language models (LLMs) offer a new approach to wearable-sensor-based human activity recognition. Retrieval-augmented generation (RAG) further introduces similar motion examples and discriminative sensor knowledge, enabling decisions grounded in observations relevant to the current query. More evidence, however, does not necessarily mean better-organized evidence. When online prompts and LLM agents primarily orchestrate inference, decisions about which activities to compare and when to invoke discriminative knowledge remain concentrated at inference time. As the activity vocabulary grows, the model must repeatedly process more candidates and associated knowledge, whereas pruning candidates too early risks excluding the true activity. We propose confusion-hierarchy evidence-aware reasoning (CHEAR), which moves part of this knowledge organization offline. CHEAR constructs and validates an activity hierarchy from recurrent cross-subject retrieval errors and builds corresponding activity-pair knowledge. At inference time, the hierarchy orders candidate comparisons; the LLM retains candidates progressively, and local discriminative knowledge is accessed only for the remaining activities. Across eight datasets, CHEAR achieves 86.79% mean cross-subject accuracy, a 75.91% candidate reduction rate, and 96.44% true-activity retention. Per-sample token use is 85.15% lower on average than in a variant without the hierarchical structure. Mean accuracy across eight cross-dataset transfer directions is 88.81%. These results show that organizing candidates and discriminative knowledge around observed activity confusions reduces online LLM input while maintaining recognition performance.

1. Introduction

Wearable-sensor-based human activity recognition (HAR) maps multichannel motion signals to activity labels for applications such as health monitoring, sports analytics, and mobile interaction. Recent uses of LLMs in HAR have moved from direct interpretation of sensor sequences toward evidence-grounded decisions. HARGPT [1] prompts an LLM with sensor sequences and activity descriptions. RAG-HAR [2] retrieves similar sensor windows and statistical descriptors as examples for the current query. ZARA [3] combines position-specific motion retrieval with sensor knowledge in an LLM-based reasoning agent. Together, these studies show that observed motion patterns and signal knowledge can enter the decision context directly, reducing reliance on an LLM's interpretation of raw numerical sequences alone.

Retrieval augmentation addresses where decision evidence comes from, but adding evidence does not by itself organize it effectively. Similar examples are generally returned by activity, whereas local features describe distinctions between particular pairs of activities. Their usefulness depends on which activities must currently be distinguished and when the corresponding features are consulted. For example, walking versus ascending stairs involves motion rhythm and vertical-axis changes; sitting versus standing depends more on body posture; and different eating activities may share hand movement while differing in local wrist dynamics. In

pipelines orchestrated primarily by online prompts and LLM agents, these comparison relationships, candidate-screening order, and knowledge-access decisions are still largely determined after each query arrives. More activity classes therefore entail more candidates and related knowledge to process; removing candidates prematurely may eliminate the true activity. This reliance on online organization also blurs the division between knowledge processing that could be completed in advance and activity decisions that must remain online, complicating future edge-cloud task allocation.

We therefore study not only how to retrieve similar examples but also how observed activity confusions can organize candidate comparisons in advance and bring discriminative knowledge into inference only when needed. We propose CHEAR (confusion-hierarchy evidence-aware reasoning). Offline, we measure the direction and frequency of errors made by leave-one-subject-out retrieval in the source domain, construct an activity-confusion hierarchy, select groups using an independent source validation set, and build local feature knowledge for validated activity pairs. Online, the LLM uses current retrieval evidence to retain a variable number of candidates at each level of the hierarchy. Pairwise knowledge is accessed only when both relevant activities remain, supporting local comparison and final label selection. Stable activity relationships are thus supplied by an offline structure, while the LLM makes choices from evidence for the current query without having to organize every activity and its discriminative knowledge on its own.

Across eight public datasets, CHEAR achieves 86.79% mean cross-subject accuracy and 86.89% macro-F1. The

*Corresponding author

 xqlu@stu.usc.edu.cn (X. Lu); tzhu@usc.edu.cn (T. Zhu);
niemx@usc.edu.cn (M. Nie); limingchen0922@dlut.edu.cn (L. Chen)

final candidate set contains 2.38 activities on average, corresponding to a 75.91% candidate reduction rate and 96.44% true-activity retention. Relative to the variant without the hierarchical structure, average per-sample token use decreases by 85.15%. Mean accuracy across eight cross-dataset transfer directions is 88.81%. These results evaluate the proposed organization of activity relationships and discriminative knowledge in terms of classification, candidate retention, and model input.

The principal contribution is to turn reproducible source-domain activity confusions from a post-hoc error analysis into an offline candidate structure that determines the order of online comparisons and the scope of local knowledge access. The framework retains the LLM's ability to reason over query-specific evidence while reducing the number of candidates and the amount of knowledge it processes online. The separation between offline knowledge organization and online activity decisions also provides a basis for future work on allocating tasks between edge devices and the cloud.

2. Related Work

Research on wearable-sensor HAR has addressed transferable representation learning from continuous signals, the alignment of motion signals with activity semantics, and the use of external evidence for individual queries. MOMENT [4] pretrains on time series from diverse sources to provide general representations for downstream tasks, including classification. Mantis [5] studies lightweight representations for time-series classification through contrastive pretraining on synthetic data. Scaling Wearable Foundation Models [6] examines the effects of data, model, and compute scale across everyday wearable signals and evaluates pretrained representations for motion and activity recognition. These studies provide encoders for sensor windows, while the relationship between motion signals and activity descriptions requires additional modeling.

Cross-modal studies introduce semantic information into motion representations through images, text, and sensor context. ImageBind [7] uses images as a bridge to map six modalities, including inertial measurements and text, into a shared representation. IMU2CLIP [8] aligns inertial motion sequences with egocentric video and text, enabling cross-modal retrieval and activity recognition. To account for changes in body placement and device orientation, UniMTS [9] generates motion signals from skeletal actions and aligns them with text to learn activity semantics. ZeroHAR [10] incorporates sensor type, coordinate axis, and wearing position into motion representations for recognizing activities unseen during training. IMUGPT [11] follows another cross-modal route: it generates activity descriptions, synthesizes three-dimensional motion, and converts it to virtual sensor sequences for training activity classifiers.

Language models have also been used to interpret sensor inputs directly, rather than only to learn motion representations or generate training data. HARGPT [1] examines whether an LLM can recognize activities from inertial

signals and activity descriptions supplied through prompts. SensorLLM [12] first aligns sensor signals with text describing their trends and then trains for HAR. SensorLM [13] constructs training data through hierarchically generated sensor captions and learns a joint sensor-language representation. LLM4HAR [14] adapts pretrained language models through sensor embeddings and sensor knowledge, using partial training and model compression to support on-device inference. These approaches address prompting, model alignment, and model efficiency, respectively.

When a decision requires examples relevant to the current sample, retrieval-augmented methods bring similar labeled motion windows into inference. RAG-HAR [2] retrieves similar windows using sensor statistical descriptors and passes the results to a language model for classification. RAG-HAR+ [15] selects dataset-specific retrieval features offline, classifies samples with sufficient retrieval evidence by nearest-neighbor voting, and invokes an LLM only for uncertain samples. ZARA [3] combines motion-example retrieval with discriminative sensor knowledge, allowing a language-model agent to select cues and reason progressively over candidate activities. From general representations and cross-modal semantics to retrieved evidence, these studies expand the information available to HAR and motivate treating the organization of candidates and discriminative knowledge as a separate problem.

3. Problem Definition and Design Principles

Let $\mathcal{Y} = \{1, \dots, K\}$ be the activity set. A query window x contains inertial-sensor sequences from one or more body locations. Source-domain data are separated by purpose. Within the construction pool, subjects rotate between query and reference roles in leave-one-subject-out folds: all windows of the held-out subject are queries, and windows of the other construction subjects form the reference set. An independent source validation set is used only to select the hierarchy, confirm activity pairs, and choose feature budgets. Target test labels are not used for hierarchy construction, feature selection, or online decisions; target test windows enter as unlabeled queries only after these components have been frozen.

The system must output one label $\hat{y} \in \mathcal{Y}$. In addition to accuracy and macro-F1, we record the mean size of the final leaf-activity candidate set $S(x)$. For a fixed dataset of N queries, candidate reduction is $1 - \frac{1}{N} \sum_{n=1}^N |S(x_n)|/K$, and true-activity retention is $\frac{1}{N} \sum_{n=1}^N \mathbb{I}[y_n \in S(x_n)]$. Candidate-set metrics measure whether routing preserves an opportunity for subsequent decisions; they do not replace final classification accuracy.

Meaning of the activity-confusion hierarchy. In practical terms, it places activities that retrieval repeatedly confuses into the same candidate group. Its inputs are directed confusion matrices from source construction queries and an independent validation set. A null model first removes the confusion expected from overall class difficulty. Candidate groups are then evaluated by their modularity gain, and

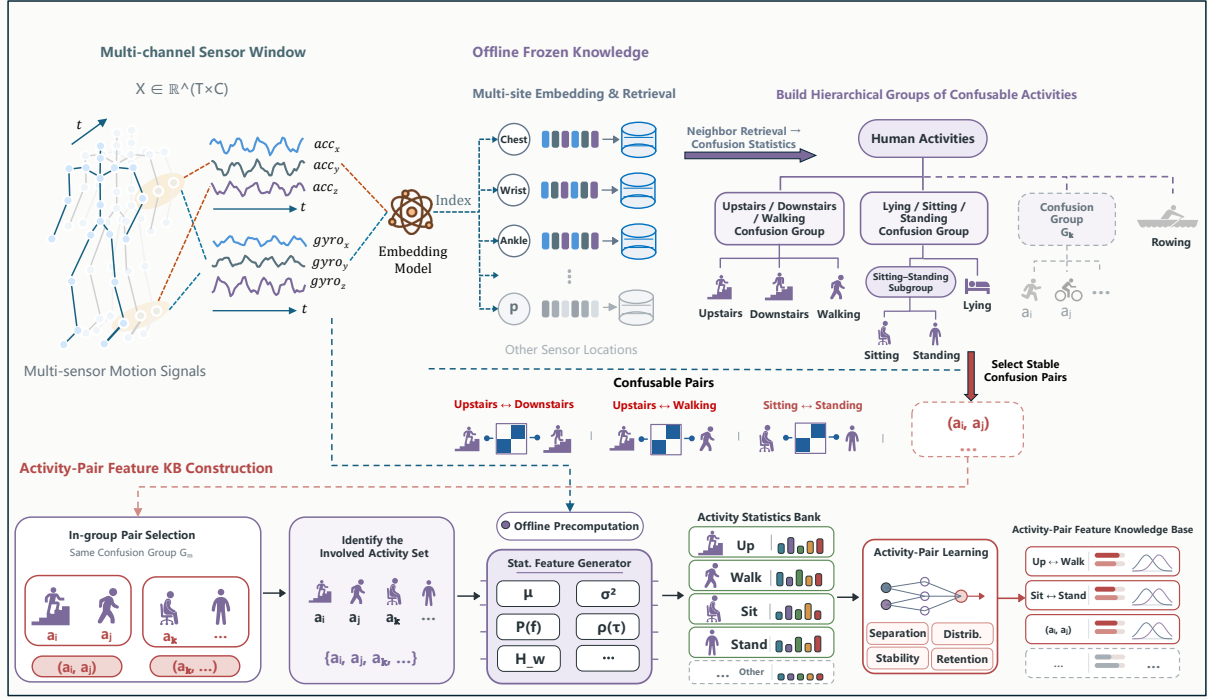


Fig. 1: Offline knowledge construction in CHEAR.

only groups with strictly positive gains in both datasets are retained. Online, the model selects a nonempty subset of current candidate units. Unselected groups and their activities are removed; selected groups expand by one level, while singleton activities remain unchanged. The outputs are root candidate units, their within-group children, and confirmed activity pairs. For example, if walking and ascending stairs exhibit above-expected confusion in both directions in both source evidence sets, they may first appear as one group and expand into two activities only after that group is retained.

The design follows two principles. First, **the hierarchy is derived only from reproducible source-domain evidence**; neither an LLM nor a person modifies it. Second, **local evidence applies only to its registered endpoints**: an activity-pair report compares its two specified activities and cannot become a global vote over other candidates.

4. Method

4.1. Overview

Offline knowledge construction. Fig. 1 summarizes the preparation completed before testing. The input consists of labeled, multi-position inertial-sensor windows from the source domain. A frozen encoder produces retrieval vectors for each sensor position; position-specific reference stores are then used to measure activity confusion under leave-one-subject-out retrieval. Candidate groups are generated from the directed confusion matrix and filtered using an independent source validation set. For difficult activity pairs confirmed by the hierarchy, the system computes predefined

statistical features from the corresponding reference windows, trains activity-pair models, and selects feature budgets using independent source validation. The stage produces three frozen outputs: position-specific retrieval stores, the activity-confusion hierarchy, and local feature knowledge for registered activity pairs.

Online hierarchical inference. Fig. 2 shows inference for one query window. The same frozen encoder processes the query, and a compact retrieval summary is obtained from the target dataset's position-specific reference stores. Inference starts from root candidate units. Given the current candidates and retrieval evidence, the LLM retains a variable-size subset. Each retained confusion group expands by one level, whereas a singleton activity proceeds unchanged. This continues until every remaining candidate is an individual activity. The system then reads local feature reports only for registered pairs whose activities both remain among the candidates; unpaired or unregistered activities continue to participate in the final decision based on retrieval evidence. The output is one activity label from the remaining candidates. Offline preparation determines the available hierarchy and local knowledge, while online inference accesses only the portions relevant to the query and remaining candidates rather than comparing every activity and every pairwise feature at once.

4.2. Frozen Encoding and Position-Specific Retrieval

Input and processing. A query window from each sensor position is linearly resampled to 512 time points and passed to a frozen Mantis-8M encoder. The encoder

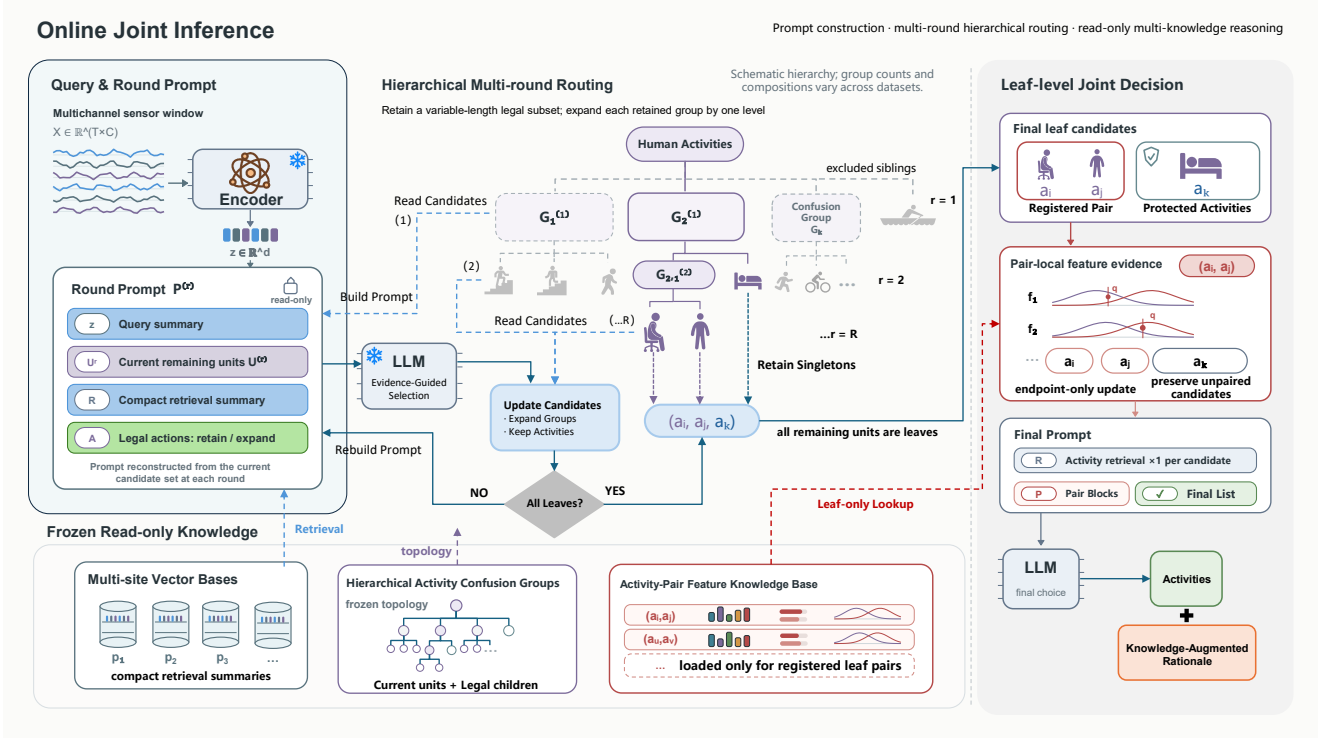


Fig. 2: Online hierarchical inference in CHEAR.

produces a 256-dimensional vector per channel, followed by ℓ_2 normalization. Reference vectors are indexed by sensor position and activity using FAISS IndexFlatIP[16]. For each activity and position, at most the 100 most similar reference items are retained; the globally top 100 results are also retained per position.

Activity scores. The similarity between normalized query vector u and reference vector v is

$$s(u, v) = u^T v = \cos(u, v). \quad (1)$$

An activity score is the sum of raw cosine similarities across positions. It ranks retrieval results and forms part of the compact evidence description; it is neither interpreted as a probability nor used directly as a group confidence score. Reciprocal rank fusion[17] is used only to select reference examples for activity pairs: an example ranked $r \geq 1$ in a list contributes $1/(60+r)$, with at most 100 items from each list.

Retrieval evidence. For each eligible activity, the module summarizes hit counts, similarity quality, and results across sensor positions. Online inference reads these compact summaries to compare candidate units at the current hierarchy level.

4.3. Constructing a Hierarchy from Directed Confusion

4.3.1. Confusion Matrix and Null Model

We perform leave-one-subject-out retrieval among construction subjects and retain only off-diagonal errors. The conditional probability that the retrieval rule incorrectly

predicts j when the true activity is i is

$$A_{ij} = P(\hat{y} = j \mid y = i), \quad A_{ii} = 0. \quad (2)$$

Both $i \rightarrow j$ and $j \rightarrow i$ are retained. Let

$$k_i^{\text{out}} = \sum_j A_{ij}, \quad k_j^{\text{in}} = \sum_i A_{ij}, \quad m = \sum_{i,j} A_{ij}, \quad (3)$$

The excess confusion after accounting for overall activity difficulty is

$$B_{ij} = A_{ij} - \frac{k_i^{\text{out}} k_j^{\text{in}}}{m}. \quad (4)$$

$B_{ij} > 0$ means that the observed $i \rightarrow j$ confusion exceeds the expectation of a null model preserving in- and out-strengths; it does not merely indicate that an error occurred.

4.3.2. Candidate Merging and Independent Validation

Given an activity partition c , directed modularity[18] is

$$Q(c) = \frac{1}{m} \sum_{i,j} B_{ij} \delta(c_i, c_j). \quad (5)$$

Starting from singleton activities, the method repeatedly merges the two groups with the largest strictly positive modularity gain and stops when no positive gain remains.

The construction set generates candidate merge paths. The independent source validation set may select only candidates on these paths whose gain over the all-singleton partition is strictly positive; it cannot introduce a structure absent from construction. The same procedure is then applied

recursively to the induced submatrix of each accepted group. Thus, the method imposes neither a fixed depth nor a binary-tree constraint, and activities without stable confusion are not forced into artificial intermediate categories.

4.3.3. Registering Difficult Activity Pairs

Bidirectional excess confusion between specific activities i and j is

$$E(i, j) = B_{ij} + B_{ji}. \quad (6)$$

An activity pair is registered only if $E(i, j) > 0$ in both the construction and independent source validation sets and both classes have validation examples. A within-group pair is attached to the deepest confusion group containing both endpoints. A cross-group pair is stored as a root-level lateral relationship without changing the main hierarchy.

Meaning of activity-pair feature knowledge. This is a description of sensor differences between two activities that are easily confused. During construction, the method reads source-domain acceleration, angular velocity, and retrieved reference examples for a registered pair; trains candidate tabular models; computes permutation importance; and selects the number of feature families on the independent source validation set. It outputs a pair identifier, available features, source statistics, and frozen reliability. During online inference, the corresponding report is loaded only if both endpoints remain among the current candidates. The report can compare its left and right endpoints but cannot affect a third activity. For example, a walking–ascending-stairs report provides local evidence only for those two activities and cannot be used to remove standing.

4.4. Local Feature Knowledge for Activity Pairs

Inputs. Feature knowledge is built only for registered activity pairs. Candidate feature families include time-domain and frequency-domain statistics, short-time Fourier transforms, autocorrelation, impact rate of change, fourth-order autoregression, recurrence rate, wavelet packets, permutation entropy, gravity direction, and inter-axis correlation.

Models and importance. Leave-one-subject-out evaluation is performed on construction subjects. In each fold, AutoGluon[19] fits candidate tabular models on the training subjects: GBM (gradient boosting machine), XGB (extreme gradient boosting), CAT (CatBoost), RF (random forest), XT (extremely randomized trees), and LR (logistic regression). Macro-F1 and permutation importance over five repetitions are computed on the held-out subject. Negative importance values are clipped to zero, and stable ties are resolved by feature identifier. Positive-class weights are restricted to $[1, 50]$ so that majority-class examples do not dominate minority-class activity-pair training.

Budget selection. Candidate feature-family budgets are $M \in \{4, 8, 12\}$. The independent source validation set selects M by activity-pair macro-F1, choosing the smaller budget in a tie. Feature knowledge is enabled online only if the best budget strictly outperforms a similarity-only activity-pair baseline; all other pairs use retrieval evidence alone.

Online use. The tabular models above are used only offline for feature screening and budget selection; they are not called for online queries. For a query, the selected feature value x is computed from the target window and compared with frozen source statistics μ_c, s_c of candidate activity c :

$$z = \frac{x - \mu_c}{\max(s_c, 10^{-6})}. \quad (7)$$

A smaller $|z|$ indicates that the query value is closer to the activity's source statistics; it is not itself a programmatic activity decision. The system summarizes distances and frozen reliability for both endpoints by feature family and supplies them, together with endpoint-specific retrieval summaries, to the LLM. The model judges whether the left endpoint, right endpoint, or neither can be selected for that pair. If the pair does not meet the feature-enablement criterion or required features are unavailable, no feature report is loaded and the remaining candidates are judged from retrieval evidence.

4.5. Variable-Size Hierarchical Candidate Retention

Online root candidates comprise top-level confusion groups and root-level singleton activities. Let U_t be the candidate units at round t . The LLM reads each unit's identifier and compact retrieval summary and outputs a variable-size set of retained units $R_t \subseteq U_t$. Candidates are updated by

$$U_{t+1} = \bigcup_{u \in R_t} \begin{cases} \text{children}(u), & u \text{ is a confusion group,} \\ \{u\}, & u \text{ is a singleton activity.} \end{cases} \quad (8)$$

Thus, the system neither fixes a top- N candidate count nor requires inference to follow a single path:

1. A retained confusion group expands by one level; an unretained group and its activities leave the candidate set.
2. A retained singleton activity proceeds unchanged to the next round.
3. If hit counts, similarity quality, and conclusions across sensor positions conflict, the model retains the smallest reasonable set covering the conflict rather than being forced onto one path.
4. Registered and available activity pairs are enumerated only after every remaining candidate is a leaf activity.

At the activity-pair stage, the model labels each available report as favoring the left endpoint, favoring the right endpoint, or inconclusive. Each report compares only its two endpoints. An inconclusive result provides no basis for deletion, and candidates without a direct pairwise comparison remain. A candidate is removed only when all its direct comparisons favor the opposing endpoint. The final prompt presents one global retrieval summary per remaining activity and appends local feature evidence for registered pairs. The model then selects one activity label from the remaining leaf activities.

Retrieval-Augmented HAR: Experimental Overview

8 datasets • 1,848 samples • ablations, Token trade-offs, LLM backbones, candidate pruning, and knowledge transfer

Experimental Overview



Fig. 3: Overview of cross-subject accuracy, token cost, and results across foundation models.

5. Experimental Setup

5.1. Datasets and Evaluation Protocol

We evaluate on eight public motion datasets: DSADS[20], Mhealth[21], Opportunity[22], PAMAP2[23], Shoaib[24], UCIHAR[25], USCHAD[26], and WISDM[27]. The retrieval encoder, indexing rules, hierarchy, activity-pair registry, and feature budgets are frozen before testing. Activities, sensor positions, subject splits, and window preprocessing are described in the appendix.

Cross-dataset experiments align only shared activities and comparable body locations. The transferred components are limited to the source-domain hierarchy and activity-pair feature knowledge. Online retrieval always uses a target-domain reference set and never retrieves raw source-domain windows. This separates the transferability of knowledge from direct reuse of source examples in the target domain.

5.2. Comparators and Metrics

Comparators fall into three categories: direct prompting, pretrained recognition, and retrieval-augmented reasoning. Direct prompting includes HARGPT and Gemini reading numerical sequences or sensor plots, as well as Gemini reading Markdown tables. Pretrained methods include NormWear[28], ImageBind, IMU2CLIP, and UniMTS, which recognize activities through pretrained representations or text matching; IMUGPT uses synthetic motion data to train a classifier. The retrieval-augmented comparator, ZARA, combines position-specific retrieval, sensor statistical knowledge, and LLM-agent reasoning.

The main configuration and cross-dataset experiments use Gemini 3 Flash Preview[29] (gemini-3-flash-preview). The multi-model study also uses GPT-4.1 mini[30], DeepSeek-V3.2[31], and Qwen3-Next-80B-A3B-Instruct[32].

The main metrics are accuracy and macro-F1. Structural effects are also assessed by the number of root candidate units, mean final candidate count, candidate reduction, and true-activity retention. Cost is measured in tokens per sample and wall-clock latency. Cross-subject means are arithmetic means over the eight datasets. In the cross-dataset table, accuracy and macro-F1 are equally weighted across eight transfer directions; candidate count, reduction, and retention are aggregated over all query windows.

6. Results

Fig. 3 summarizes cross-subject accuracy, token cost, and results across foundation models; the corresponding tables give the exact values.

6.1. Cross-Subject Main Experiment

Table 1 orders the eight datasets by increasing number of activities. Among prior methods, ZARA with Gemini has the highest mean accuracy (81.6%) and macro-F1 (81.4%). CHEAR with Gemini reaches 86.79% and 86.89%, exceeding those tabulated values by 5.19 and 5.49 percentage points. Its accuracy is higher than that of ZARA with Gemini on seven of the eight datasets. The differences are particularly large on WISDM (65.6% versus 83.68%) and USCHAD (60.0% versus 69.17%); on Opportunity, CHEAR reaches 85.00%, compared with 92.5% for ZARA.

Table 1

Cross-subject comparison with prior methods (accuracy and macro-F1, %). Results for prior methods are taken from ZARA[3]. Means are arithmetic means across eight datasets. For each metric in each dataset and in the mean column, the highest value is bold and the second-highest is underlined; tied values receive the same marking.

Model/Method	Easy						Medium						Hard				Mean	
	Opportunity		UCIHAR		Shoaib		PAMAP2		USCHAD		Mhealth		WISDM		DSADS		Mean	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
# Classes	4		6		7		12		12		12		18		19		–	
# Channels	30		6		30		18		6		15		6		30		–	
HARGPT _{Text}	21.0	19.2	29.6	17.4	27.1	19.2	12.1	6.2	13.8	7.3	12.1	6.0	5.6	1.8	10.5	6.2	16.5	10.4
Gemini _{Text}	26.5	19.8	24.2	13.0	27.1	17.6	15.0	10.2	14.2	5.4	25.4	20.8	11.1	7.6	13.2	8.6	19.6	12.9
Gemini _{Table}	29.0	22.3	21.3	9.8	27.6	18.7	11.7	7.2	17.1	9.7	22.9	18.3	10.1	7.8	16.3	10.3	19.5	13.0
HARGPT _{Plot}	21.5	15.6	28.3	15.7	24.3	14.2	10.0	6.9	14.6	8.7	15.0	11.0	5.9	2.8	7.9	4.5	15.9	9.9
Gemini _{Plot}	23.5	21.3	20.6	31.7	31.4	24.1	10.4	6.9	10.8	5.3	19.2	17.4	9.4	7.2	10.0	4.8	18.3	13.5
NormWear	23.0	23.8	17.9	11.4	15.2	11.7	9.2	2.7	10.0	5.8	8.3	2.2	4.2	1.4	3.7	2.2	11.4	7.7
IMUGPT	38.5	28.7	32.5	21.6	26.7	15.2	12.9	3.8	2.9	1.9	8.3	2.8	5.9	2.1	7.4	3.6	16.9	10.0
ImageBind	35.5	30.0	28.8	19.9	36.7	30.2	18.8	10.2	7.9	1.8	17.9	11.1	8.0	4.7	10.5	5.7	20.5	14.2
IMU2CLIP	36.5	34.4	33.3	22.8	39.5	34.5	15.8	11.6	16.3	10.5	16.3	14.3	10.1	5.9	13.7	9.2	22.7	17.9
UniMTS	33.5	24.8	37.1	23.9	51.9	40.6	32.9	29.2	29.6	24.2	65.4	58.8	30.2	28.5	34.7	27.0	39.4	32.1
ZARA _{Qwen-80B}	80.5	79.8	76.7	75.4	93.8	93.5	67.5	67.0	47.1	49.2	82.1	81.6	57.6	56.4	81.6	81.3	73.4	73.0
ZARA _{GPT}	86.5	86.5	85.0	85.0	93.3	93.2	72.5	72.8	56.7	57.4	82.5	80.6	60.8	59.5	82.6	82.3	77.5	77.2
ZARA _{Gemini}	92.5	92.5	90.0	90.0	<u>97.1</u>	<u>97.1</u>	76.7	76.9	<u>60.0</u>	<u>60.1</u>	86.3	86.1	65.6	64.1	84.2	84.4	81.6	81.4
CHEAR _{GPT}	77.00	76.78	90.42	90.44	93.81	93.90	79.17	80.68	57.92	56.96	87.50	87.41	75.00	75.03	85.79	85.40	80.83	80.83
CHEAR _{DeepSeek}	81.50	81.37	<u>95.00</u>	<u>94.93</u>	99.05	99.05	<u>80.83</u>	<u>82.34</u>	57.08	57.33	84.17	83.85	<u>80.21</u>	<u>79.67</u>	<u>90.00</u>	<u>89.66</u>	83.48	83.53
CHEAR _{Qwen}	81.50	81.36	<u>95.00</u>	<u>94.93</u>	99.05	99.05	79.58	81.28	57.08	57.12	86.25	85.97	79.17	78.46	90.53	90.24	83.52	83.55
CHEAR _{Gemini}	85.00	85.13	96.67	96.64	99.05	99.05	82.50	83.43	69.17	69.58	88.75	88.62	83.68	83.50	89.47	89.18	86.79	86.89

Table 2

Shared activities and sensor-position mappings in cross-dataset experiments.

Dataset pair	Activities	Sensor-position mapping
UCIHAR ↔ USCHAD	5	Waist ↔ front right hip
PAMAP2 ↔ WISDM	5	Wrist ↔ hand
Mhealth ↔ WISDM	5	Right wrist ↔ hand
PAMAP2 ↔ Mhealth	7	Wrist ↔ right wrist
		Chest ↔ chest
		Ankle ↔ left ankle

Mean accuracy across the four CHEAR configurations ranges from 80.83% to 86.79%. The GPT, DeepSeek, and Qwen configurations achieve 80.83%, 83.48%, and 83.52%, respectively. These results indicate that hierarchical candidates and local evidence can be used with different foundation models, although final accuracy still depends on the model. The larger gain on WISDM is consistent with the design of first narrowing candidates and then loading activity-pair knowledge; the ablations below further assess these components. Per-activity accuracies are shown in Appendix Fig. E.1.

6.2. Cross-Dataset Experiments

Table 3 reports eight transfer directions across four dataset pairs. The full method achieves mean accuracy of 88.81% and macro-F1 of 88.74%. Removing transferred

feature knowledge yields 86.60% mean accuracy, a difference of 2.21 percentage points. The full method has higher accuracy in six directions; for PAMAP2→Mhealth, it reaches 97.14%, compared with 90.71% without transferred feature knowledge.

In cross-dataset experiments, the full method has 1.681 final candidates on average, with 69.55% mean candidate reduction and 98.24% mean true-activity retention; retention ranges from 94.00% to 100.00% across directions. Online retrieval still uses target-domain reference examples. The comparison without transferred feature knowledge retains the source-domain hierarchy but removes source-domain activity-pair features, isolating the contribution of that transferred knowledge.

6.3. Ablation Results and Token Cost

We compare the full method and ablated variants in terms of accuracy and tokens per sample. Accuracy tests whether each component improves final classification; token count measures changes in interface input and output cost due to candidate reduction. The variant without the hierarchical structure does not retain candidates level by level. Instead, it sends all activities to the final decision while retaining retrieval evidence and registered activity-pair knowledge.

6.3.1. Accuracy

Table 4 reports accuracy. The full method averages 86.79% across the eight datasets, exceeding the variant without feature knowledge by 3.31 percentage points, the fixed-rule variant without an LLM by 3.84 points, and the

Table 3

Cross-dataset transfer results and accuracy without transferred feature knowledge. Candidate-set metrics are aggregated over query windows.

Transfer direction	Accuracy (%)	Macro-F1 (%)	Accuracy without transferred features (%)	Mean final candidates	Candidate reduction (%)	True-activity retention (%)
UCIHAR→USCHAD	94.00	94.00	92.00	2.270	54.60	100.00
USCHAD→UCIHAR	93.00	92.90	94.00	1.130	77.40	97.00
WISDM→PAMAP2	78.00	78.25	77.00	1.910	61.80	94.00
PAMAP2→WISDM	92.00	91.99	88.67	1.313	73.73	98.67
WISDM→Mhealth	75.00	74.00	71.00	1.780	64.40	100.00
Mhealth→WISDM	91.33	91.38	88.67	1.400	72.00	98.00
PAMAP2→Mhealth	97.14	97.13	90.71	1.850	73.57	100.00
Mhealth→PAMAP2	90.00	90.27	90.71	2.336	66.63	98.57
Arithmetic mean	88.81	88.74	86.60	1.681	69.55	98.24

Table 4

Cross-subject accuracy (%). Highest values are bold; ties are marked equally.

Dataset	Full	No features	No LLM	No hierarchy
DSADS	89.47	85.79	84.74	87.37
Mhealth	88.75	87.08	86.67	88.75
Opportunity	85.00	81.00	80.00	85.00
PAMAP2	82.50	80.83	80.42	81.67
Shoaib	99.05	99.05	99.05	99.05
UCIHAR	96.67	95.00	92.08	96.25
USCHAD	69.17	57.50	62.50	61.67
WISDM	83.68	81.60	78.13	83.33
Arithmetic mean	86.79	83.48	82.95	85.39

variant without the hierarchical structure by 1.40 points. These comparisons assess local feature knowledge, LLM candidate selection, and the hierarchical inference procedure, respectively. The hierarchical procedure yields both an accuracy gain and token savings.

6.3.2. Token Use

Table 5 reports tokens per sample for the full method and the variants without feature knowledge or the hierarchical structure. Relative to the latter, the full method reduces total input and output tokens on all eight datasets. Savings range from 78.75% to 90.68%, with an arithmetic mean of 85.15%. The full method uses more tokens than the variant without features because it reads activity-pair reports after candidate reduction. This additional input accompanies the mean accuracy gain of 3.31 percentage points in Table 4. Token use for the full method ranges from 1,013.4 per sample on Shoaib to 12,143.1 on WISDM, indicating that cost depends not only on the final candidate count but also on the number of registered pairs, available feature reports, and routing rounds.

6.4. Does the Hierarchy Retain the True Activity?

Table 6 reports candidate-set size alongside true-activity retention. Across eight datasets, the final set contains 2.38

Table 5

Tokens per sample and the full method's savings relative to the variant without the hierarchical structure.

Dataset	Full	No features	No hierarchy	Savings (%)
DSADS	4,679.3	935.0	44,674.6	89.53
Mhealth	1,910.9	1,409.3	9,238.3	79.32
Opportunity	3,111.1	1,104.6	14,642.6	78.75
PAMAP2	10,038.2	1,141.6	63,655.0	84.23
Shoaib	1,013.4	787.4	8,484.7	88.06
UCIHAR	1,803.5	814.9	10,666.5	83.09
USCHAD	7,879.8	1,059.8	63,146.4	87.52
WISDM	12,143.1	1,093.9	130,267.7	90.68
Arithmetic mean	–	–	–	85.15

candidates on average, with 75.91% mean reduction and 96.44% true-activity retention. These metrics must be read together: a smaller set is useful only if it still contains the true activity. On DSADS, 19 activities are reduced to 3.09 candidates on average, with 99.47% retention. On USCHAD, the set shrinks to 2.35 candidates, but retention is 90.42%, indicating that some subsequent performance loss originates when upper-level routing removes the true activity. This diagnosis separates candidate-reduction errors from errors in final label selection.

6.5. Local Processing Latency

Table 7 reports mean time spent on retrieval, activity-pair report generation, and query feature extraction in the full method. Each operation takes less than one second per sample on every dataset; retrieval takes 0.005–0.043 seconds per sample. Report generation varies with the number of registered pairs among current candidates, whereas feature extraction varies with sensor positions and available modalities.

7. Conclusion

We proposed CHEAR, which uses stable source-domain activity confusions to determine which activities should be compared first. Variable-size candidate retention then narrows the set level by level, and validated sensor-feature

Table 6

Hierarchical candidate reduction. Root compression is calculated from the number of root activities and root candidate units; final reduction and true-activity retention are reported together.

Dataset	Root activities	Root units	Root compression (%)	Final candidates	Final reduction (%)	True-activity retention (%)
DSADS	19.00	6.00	68.42	3.09	83.71	99.47
Mhealth	12.00	8.00	33.33	3.01	74.93	100.00
Opportunity	4.00	2.00	50.00	1.55	61.38	93.00
PAMAP2	12.00	6.00	50.00	2.65	77.95	92.50
Shoaib	7.00	5.00	28.57	1.77	74.69	99.52
UCIHAR	6.00	2.00	66.67	1.80	70.00	98.33
USCHAD	12.00	2.00	83.33	2.35	80.42	90.42
WISDM	18.00	5.00	72.22	2.85	84.18	98.26
Arithmetic mean	–	–	–	2.38	75.91	96.44

Table 7

Local processing latency of the full method (seconds per sample). Timings for individual components may overlap slightly.

Dataset	Retrieval	Pair reports	Query features
DSADS	0.043	0.369	0.150
Mhealth	0.012	0.102	0.076
Opportunity	0.018	0.510	0.160
PAMAP2	0.016	0.270	0.097
Shoaib	0.019	0.244	0.106
UCIHAR	0.005	0.045	0.036
USCHAD	0.007	0.067	0.041
WISDM	0.010	0.070	0.036

knowledge is loaded only for registered pairs that remain. Across eight datasets, the full method achieves 86.79% mean accuracy, with 2.38 final candidates, 75.91% candidate reduction, and 96.44% true-activity retention. It uses 85.15% fewer tokens on average than the variant without the hierarchical structure. Across eight cross-dataset directions, mean accuracy is 88.81%, compared with 86.60% without transferred feature knowledge. These results support a clear principle for organizing inference: use reproducible confusion evidence to narrow the comparison scope before an LLM reads local knowledge directly relevant to the current candidates.

A. Dataset Descriptions

The experiments use eight public HAR datasets. Table A.1 summarizes their numbers of activities, sensor positions, and activity labels. Retrieval, hierarchy construction, and inference are performed within each dataset's label set.

B. Dataset Preprocessing

UCIHAR uses its publicly provided windows, and DSADS uses its predefined windows. All other signals are divided into nonoverlapping windows.

- **Opportunity:** 30 Hz; subject 4 provides queries and the other subjects provide reference data; 2-second windows; 50 query windows per class.

- **UCIHAR:** 50 Hz; the public training/test split is retained; 2.56-second windows with 50% overlap; 40 windows per class are sampled from test subjects.
- **Shoaib:** 50 Hz; subjects 1 and 9 provide queries; 2-second windows; 30 windows per class.
- **PAMAP2:** 100 Hz; subjects 5 and 6 provide queries; 2-second windows; 20 windows per class.
- **USCHAD:** 100 Hz; subjects 13 and 14 provide queries; 2-second windows; 20 windows per class.
- **Mhealth:** 50 Hz; subjects 1 and 6 provide queries; 2-second windows; 20 windows per class.
- **WISDM:** The 20-Hz smartwatch hand subset is used. Accelerometer and gyroscope readings are aligned by timestamp. Forty-seven subjects without alignment anomalies are retained, eight of them for queries. Ten-second windows are used, with 16 windows per class.
- **DSADS:** 25 Hz; subjects 2 and 4 provide queries; the dataset's 5-second windows are used, with 10 windows per class.

C. Cross-Dataset Preprocessing

Cross-dataset experiments retain shared activities and map sensor positions according to Table 2. Only the hierarchy and activity-pair feature knowledge are transferred from the source domain; queries always retrieve from the target-domain reference set and never access raw source-domain windows. The four bidirectional dataset pairs are processed as follows:

- **UCIHAR and USCHAD:** Walking, ascending stairs, descending stairs, sitting, and standing are aligned. USCHAD is downsampled to 50 Hz; 40 and 20 windows per class are sampled from the respective datasets.
- **PAMAP2 and WISDM:** Walking, standing, jogging, sitting, and stairs are aligned. PAMAP2 is downsampled to 20 Hz; 20 and 30 windows per class are sampled, respectively.
- **Mhealth and WISDM:** Walking, standing, jogging, sitting, and stairs are aligned. Mhealth is downsampled to 20 Hz; 20 and 30 windows per class are sampled, respectively.
- **PAMAP2 and Mhealth:** Lying, sitting, standing, walking, running, cycling, and ascending stairs are aligned. PAMAP2 is downsampled to 50 Hz; 20 windows per class are sampled from each dataset.

Sampling is balanced across activities, and no window is sampled twice within a dataset.

Table A.1

Experimental datasets, sensor positions, and activity labels.

Dataset	Activities	Sensor positions	Activity labels
Opportunity	4	Back, upper arms, forearms	Standing, walking, sitting, lying
UCIHAR	6	Waist	Standing, sitting, lying, walking, descending stairs, ascending stairs
Shoaib	7	Left and right pockets, belt, right upper arm, right wrist	Walking, standing, jogging, sitting, cycling, descending stairs, ascending stairs
PAMAP2	12	Wrist, chest, ankle	Lying, sitting, standing, ironing, vacuum cleaning, ascending stairs, descending stairs, walking, Nordic walking, cycling, running, rope jumping
USCHAD	12	Front right hip	Sleeping, sitting, elevator down, elevator up, standing, jumping, descending stairs, walking right, walking forward, running forward, ascending stairs, walking left
Mhealth	12	Chest, right wrist, left ankle	Ascending stairs, standing still, sitting relaxed, lying, walking, bending forward, lifting arms forward, bending knees, jogging, running, jumping forward and backward, cycling
WISDM	18	Hand	Walking, jogging, stairs, sitting, standing, typing, brushing teeth, eating soup, eating chips, eating pasta, eating a sandwich, kicking a ball, catching a ball, drinking, dribbling, writing, clapping, folding clothes
DSADS	19	Torso, left and right arms, left and right legs	Sitting, standing, lying supine, lying on right side, ascending stairs, descending stairs, standing in an elevator, moving in an elevator, walking slowly, rowing, jumping, treadmill walking on level ground, treadmill walking on an incline, fast treadmill running, stepping, cross-training, basketball, and two exercise-bicycle postures

Table D.1

Predefined features used for activity-pair knowledge.

Feature family	Representative statistics or derived quantities
Time-domain statistics	Mean, standard deviation, variance, maximum, minimum, median, root mean square, peak, range, interquartile range, skewness, kurtosis, and signal magnitude area
Frequency-domain statistics	Low-, mid-, and high-frequency band powers and their ratios; dominant frequency and power; secondary peak frequency; spectral centroid, entropy, skewness, and kurtosis; weighted mean frequency; and spectral energy
Short-time Fourier transform	Maximum, mean, and standard deviation in each frequency band, plus statistics of short-time spectral entropy and spectral centroid
Change and impact	Mean, root mean square, and standard deviation of first differences; zero-crossing rate; impact rate of change; and peaks
Periodicity and dependence	First autocorrelation peak, first minimum, and first zero crossing; fourth-order autoregressive coefficients; and recurrence rate
Multiscale complexity	Wavelet-packet energy and permutation entropy
Posture and inter-axis relations	Gravity-direction statistics and inter-axis Pearson correlation coefficients

D. Predefined Features

Table D.1 groups the candidate features used for activity-pair knowledge by computation type. Features are computed from the corresponding sensor axes and magnitude sequences. Activity-pair models then select the feature families to load using cross-subject permutation importance and independent source validation.

E. Per-Class Metrics

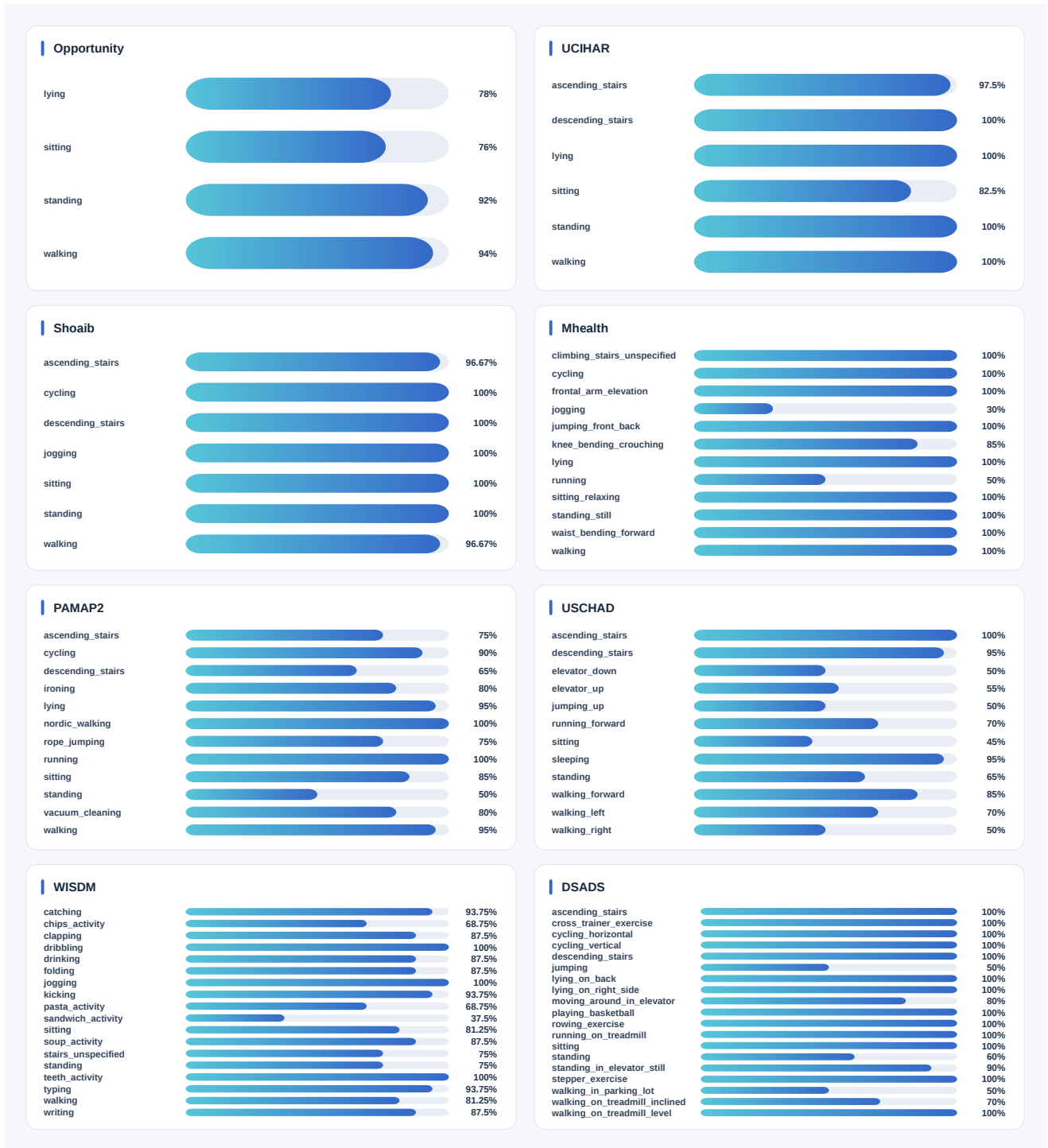


Fig. E.1: Per-activity accuracy of the full method across eight datasets.

CRedit authorship contribution statement

Xiangqing Lu: Conceptualization, Data curation, Investigation, Methodology, Visualization, Writing – original draft.

Tao Zhu: Funding acquisition, Resources, Project administration, Supervision, Validation.

Mingxing Nie: Project administration, Supervision, Validation.

Liming Chen: Resources, Supervision, Validation.

References

- [1] Sijie Ji, Xinzhe Zheng, and Chenshu Wu. HARGPT: Are LLMs Zero-Shot Human Activity Recognizers? In 2024 IEEE International Workshop on Foundation Models for Cyber-Physical Systems & Internet of Things (FMSys), pages 38–43, 2024.
- [2] Nirhoshan Sivaroopan, Hansi Karunaratna, Chamara Madarasingha, Anura Jayasumana, and Kanchana Thilakarathna. RAG-HAR: Retrieval Augmented Generation-based Human Activity Recognition. In IEEE International Conference on Pervasive Computing and Communications (PerCom), pages 1–12, 2026.
- [3] Zechen Li, Baiyu Chen, Hao Xue, and Flora D. Salim. ZARA: Training-Free Motion Time-Series Reasoning via Evidence-Grounded LLM Agents. In Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics, pages 14986–15008, 2026. doi:10.18653/v1/2026.acl-long.684.
- [4] Mononito Goswami, Konrad Szafer, Arjun Choudhry, Yifu Cai, Shuo Li, and Artur Dubrawski. MOMENT: A Family of Open Time-series Foundation Models. In Proceedings of the 41st International Conference on Machine Learning, PMLR 235:16115–16152, 2024.
- [5] Vasilii Feofanov, Songkang Wen, Shifeng Xie, et al. Mantis: Lightweight Foundation Model for Time Series Classification. In Proceedings of the 43rd International Conference on Machine Learning, 2026.
- [6] Girish Narayanswamy, Xin Liu, Kumar Ayush, Yuzhe Yang, Xuhai Xu, Shun Liao, Jake Garrison, Shyam Tailor, Jacob Sunshine, Yun Liu, et al. Scaling Wearable Foundation Models. In International Conference on Learning Representations, volume 2025, pages 59382–59413, 2025.
- [7] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, et al. ImageBind: One Embedding Space To Bind Them All. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 15180–15190, 2023.
- [8] Seungwhan Moon, Andrea Madotto, Zhaojiang Lin, et al. IMU2CLIP: Multimodal Contrastive Learning for IMU Motion Sensors from Egocentric Videos and Text. arXiv:2210.14395, 2022.
- [9] Xiyuan Zhang, Diyan Teng, Ranak Roy Chowdhury, et al. UniMTS: Unified Pre-training for Motion Time Series. Advances in Neural Information Processing Systems, 37:107469–107493, 2024.
- [10] Ranak Roy Chowdhury, Ritvik Kapila, Ameya Panse, et al. ZeroHAR: Sensor Context Augments Zero-Shot Wearable Action Recognition. Proceedings of the AAAI Conference on Artificial Intelligence, 39(15):16046–16054, 2025.
- [11] Zikang Leng, Hyeokhyun Kwon, and Thomas Plötz. Generating Virtual On-body Accelerometer Data from Virtual Textual Descriptions for Human Activity Recognition. In Proceedings of the 2023 ACM International Symposium on Wearable Computers, pages 39–43, 2023.
- [12] Zechen Li, Shohreh Deldari, Linyao Chen, Hao Xue, and Flora D. Salim. SensorLLM: Aligning Large Language Models with Motion Sensors for Human Activity Recognition. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 354–379, 2025. doi:10.18653/v1/2025.emnlp-main.19.
- [13] Yuwei Zhang, Kumar Ayush, Siyuan Qiao, A. Ali Heydari, Girish Narayanswamy, Maxwell A. Xu, Ahmed A. Metwally, Shawn Xu, Jake Garrison, Xuhai Xu, Tim Althoff, Yun Liu, Pushmeet Kohli, Jiening Zhan, Mark Malhotra, Shwetak Patel, Cecilia Mascolo, Xin Liu, Daniel McDuff, and Yuzhe Yang. SensorLLM: Learning the Language of Wearable Sensors. Preprint, arXiv:2506.09108, 2025c.
- [14] Zhiqing Hong, Yiwei Song, Zelong Li, et al. LLM4HAR: Generalizable On-device Human Activity Recognition with Pretrained LLMs. In Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pages 4511–4521, 2025.
- [15] Hansi Karunaratna, Nirhoshan Sivaroopan, Chamara Madarasingha, Anura Jayasumana, and Kanchana Thilakarathna. RAG-HAR+: Towards Cost-Efficient LLM-Based Human Activity Recognition for Edge Deployment. arXiv:2607.26631, 2026.
- [16] Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. The faiss library. IEEE Transactions on Big Data, 2025.
- [17] Gordon V. Cormack, Charles L. A. Clarke, and Stefan Buettcher. Reciprocal Rank Fusion Outperforms Condorcet and Individual Rank Learning Methods. In Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 758–759, 2009.
- [18] Elizabeth A. Leicht and Mark E. J. Newman. Community Structure in Directed Networks. Physical Review Letters, 100(11):118703, 2008.
- [19] Nick Erickson, Jonas Mueller, Alexander Shirkov, et al. AutoGluon-Tabular: Robust and Accurate AutoML for Structured Data. arXiv:2003.06505, 2020.
- [20] 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.
- [21] [dataset] Oresti Baños, Rafael García, and Alejandro Saez. MHEALTH. UCI Machine Learning Repository, 2014. doi:10.24432/C5TW22.
- [22] Daniel Roggen, Alberto Calatroni, Mirco Rossi, et al. Collecting Complex Activity Datasets in Highly Rich Networked Sensor Environments. In Seventh International Conference on Networked Sensing Systems, pages 233–240, 2010.
- [23] Attila Reiss and Didier Stricker. Introducing a New Benchmarked Dataset for Activity Monitoring. In 16th International Symposium on Wearable Computers, pages 108–109, 2012.
- [24] Muhammad Shoaib, Stephan Bosch, Ozlem Durmaz Incel, Hans Scholten, and Paul J. M. Havinga. Fusion of Smartphone Motion Sensors for Physical Activity Recognition. Sensors, 14(6):10146–10176, 2014.
- [25] Davide Anguita, Alessandro Ghio, Luca Oneto, Xavier Parra, and Jorge L. Reyes-Ortiz. A Public Domain Dataset for Human Activity Recognition Using Smartphones. In European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, pages 3–4, 2013.
- [26] Mi Zhang and Alexander A. Sawchuk. USC-HAD: A Daily Activity Dataset for Ubiquitous Activity Recognition Using Wearable Sensors. In Proceedings of the 2012 ACM Conference on Ubiquitous Computing, pages 1036–1043, 2012. doi:10.1145/2370216.2370438.
- [27] [dataset] Gary M. Weiss. WISDM Smartphone and Smartwatch Activity and Biometrics Dataset. UCI Machine Learning Repository, 2019. doi:10.24432/C5HK59.
- [28] Yunfei Luo, Yuliang Chen, Asif Salekin, and Tauhidur Rahman. Toward Foundation Model for Multivariate Wearable Sensing of Physiological Signals. ACM Transactions on Computing for Healthcare, 7(3):1–43, 2026.
- [29] Gemini Team et al. Gemini: A Family of Highly Capable Multimodal Models. arXiv:2312.11805v5, 2025.
- [30] OpenAI et al. GPT-4 Technical Report. arXiv:2303.08774v6, 2024.
- [31] DeepSeek-AI et al. DeepSeek-V3.2: Pushing the Frontier of Open Large Language Models. arXiv:2512.02556, 2025.
- [32] An Yang et al. Qwen3 Technical Report. arXiv:2505.09388, 2025.