ZARA: Training-Free Motion Time-Series Reasoning via Evidence-Grounded LLM Agents

Zechen Li[1] **Baiyu Chen**[1] **Hao Xue**[1,2,3] **Flora D. Salim**[1]

[1]University of New South Wales, Sydney
[2]Hong Kong University of Science and Technology (Guangzhou)
[3]Hong Kong University of Science and Technology

{zechen.li, breeze.chen, flora.salim}@unsw.edu.au haoxue@hkust-gz.edu.cn

Abstract

Motion sensor time-series are central to Human Activity Recognition (HAR), yet conventional approaches are constrained to fixed activity sets and typically require costly parameter retraining to adapt to new behaviors. While Large Language Models (LLMs) offer promising open-set reasoning capabilities, applying them directly to numerical time-series often leads to hallucinations and weak grounding. To address this challenge, we propose **ZARA** (Zero-training Activity Reasoning Agents), a knowledge- and retrieval-augmented agentic framework for motion time-series reasoning in a training-free inference setting. Rather than relying on black-box projections, ZARA distills reference data into a statistically grounded textual knowledge base that transforms implicit signal patterns into verifiable natural-language priors. Guided by retrieved evidence, ZARA iteratively selects discriminative cues and performs grounded reasoning over candidate activities. Extensive experiments on eight benchmarks show that ZARA generalizes robustly to unseen subjects and across datasets, demonstrating strong transferability across heterogeneous sensor domains. These results mark a step toward trustworthy, plug-and-play motion understanding beyond dataset-specific artifacts. Our code is available at https://github.com/zechenli03/ZARA.

1 Introduction

Human activity recognition (HAR) from on-body motion sensors underpins applications from digital health to adaptive interfaces. However, the dominant paradigm in HAR remains heavily supervised. Most existing systems typically rely on task-specific deep neural networks (Ordóñez and Roggen, 2016; Abedin et al., 2021; Vaswani et al., 2017) optimized for fixed sensor setups and classes.

Consequently, existing HAR methods (see Figure 1) face three critical barriers to scalable deployment. **Poor Generalization.** Adapting to new users (cross-subject) or hardware setups (cross-domain) typically necessitates costly model parameter optimization. **Limited Training-Free Adaptation.** Time-series (TS) foundation models like Moment (Goswami et al., 2024) and Mantis (Feofanov et al., 2025) offer transferable representations but still require task-specific classification heads. Contrastive models such as UniMTS (Zhang et al., 2024) eliminate the classifier, yet still struggle to distinguish fine-grained activities in parameter-frozen settings due to limited semantic grounding. **Lack of Interpretability.** Most approaches yield only categorical predictions without transparent reasoning, limiting trust in safety-critical scenarios.

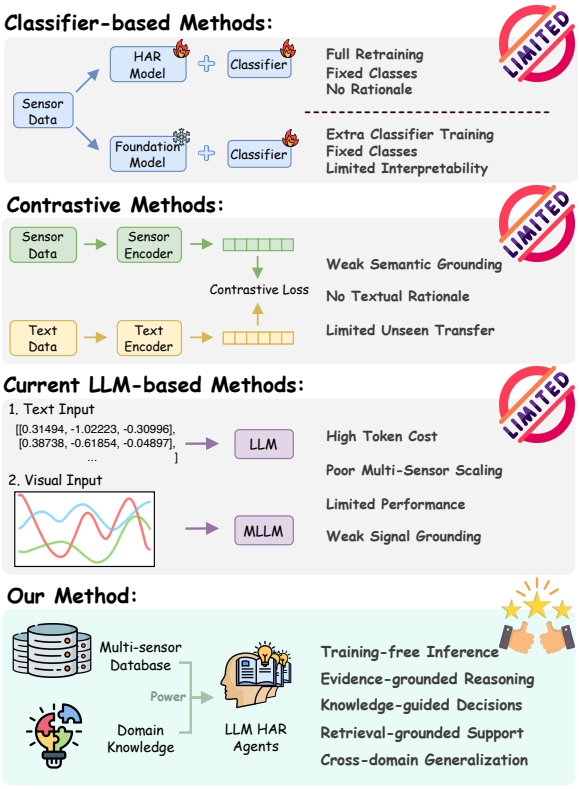


Figure 1: Representative method families for human activity recognition.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 14986–15008
July 2-7, 2026

Meanwhile, LLMs enhanced with retrieval-augmented generation (RAG) have shown strong reasoning capabilities in vision and NLP (Baek et al., 2023; Shen et al., 2022; Xie et al., 2023; Zhang et al., 2025a). However, sensor-based HAR has largely failed to capitalize on this paradigm. Early attempts to apply LLMs to HAR have focused on converting multi-channel signals into token sequences or images. These modality-projection approaches suffer from **excessive token usage**, **significant information loss** during discretization, and **mediocre accuracy**, as LLMs struggle to intuit physical dynamics directly from raw numerical streams.

We argue that the missing link lies in translating implicit signal statistics into explicit, structured language. Just as RAG in NLP relies on a high-quality document corpus, RAG in HAR requires a domain-specific knowledge base that articulates *how* physical movements manifest in sensor data. When equipped with (i) statistically grounded textual priors (e.g., running exhibits higher vertical acceleration variance than walking) and (ii) a retrieval mechanism for relevant signal evidence, LLMs can perform robust reasoning on HAR tasks. This enables evidence-grounded, training-free inference, effectively replacing task-specific classifier training with in-context conditioning on retrieved priors.

Motivated by this insight, we introduce ZARA, a novel agentic framework for HAR in a training-free inference setting via knowledge- and retrieval-augmented reasoning. ZARA bridges the signal-to-language gap via three synergistic components. First, **Offline Statistical Profiling** automatically distills a general-purpose knowledge base from every activity pair by extracting discriminative feature profiles. This pairwise formulation translates implicit signal characteristics into verifiable linguistic priors, enabling the system to accommodate new activities by simply registering their profiles without parameter updates. Second, **Class-Wise Multi-Sensor Retrieval** fetches top-k evidence conditionally per class from a labeled support set to ensure balanced recall across long-tail classes, and then aggregates the heterogeneous sensor-specific rankings using Reciprocal Rank Fusion (Cormack et al., 2009). Finally, **Hierarchical Multi-Agent Reasoning** orchestrates specialized LLM agents to iteratively filter features and prune candidates, progressively narrowing the hypothesis space to produce predictions supported by human-readable explanations.

We benchmark ZARA against 10 established baselines across 8 diverse HAR datasets. Our comprehensive evaluation spans both *Cross-Subject* scenarios (addressing new user adaptation) and *Cross-Dataset* scenarios (testing domain generalization). By fusing structured sensor knowledge with LLM-based reasoning, ZARA offers a plug-and-play alternative to training-intensive pipelines. Concretely, this work contributes:

- **Signal-to-Text Knowledge Grounding.** We propose an automated method to distill motion TS into a pairwise textual knowledge base, enabling LLMs to perform verifiable reasoning in a parameter-frozen setting.
- **Agentic Framework for Interpretable HAR.** ZARA is the first knowledge- and retrieval-driven agentic system for multi-sensor TS classification that also generates concise, evidence-backed rationales, enhancing trust in automated decision-making.
- **Strong Training-Free Generalization.** Extensive experiments confirm ZARA's state-of-the-art performance in parameter-frozen settings, where robust generalization across unseen subjects and heterogeneous domains demonstrates the transferability of the injected motion priors.

2 Related Work

Recent HAR work has shifted from training task-specific networks to using pre-trained foundation models for general time-series (TS) representations. Chronos (Ansari et al., 2024) tokenizes TS via scaling and quantization to enable text-style encoder–decoder training; Moment (Goswami et al., 2024) pre-trains transformers with masked-value prediction; and Mantis (Feofanov et al., 2025) uses a contrastively pre-trained Vision Transformer (Dosovitskiy et al., 2020) tailored for time-series classification. Despite strong representations, these backbones still typically require training downstream classifiers for specific tasks.

To enable parameter-frozen HAR, prior work explores cross-modal alignment, mapping motion signals into shared embedding spaces with text or images. ImageBind (Girdhar et al., 2023) and IMU2CLIP (Moon et al., 2023) align IMU data with VLM spaces (Radford et al., 2021). UniMTS (Zhang et al., 2024) aligns synthetic skeleton motions with text for classifier-free recogni-

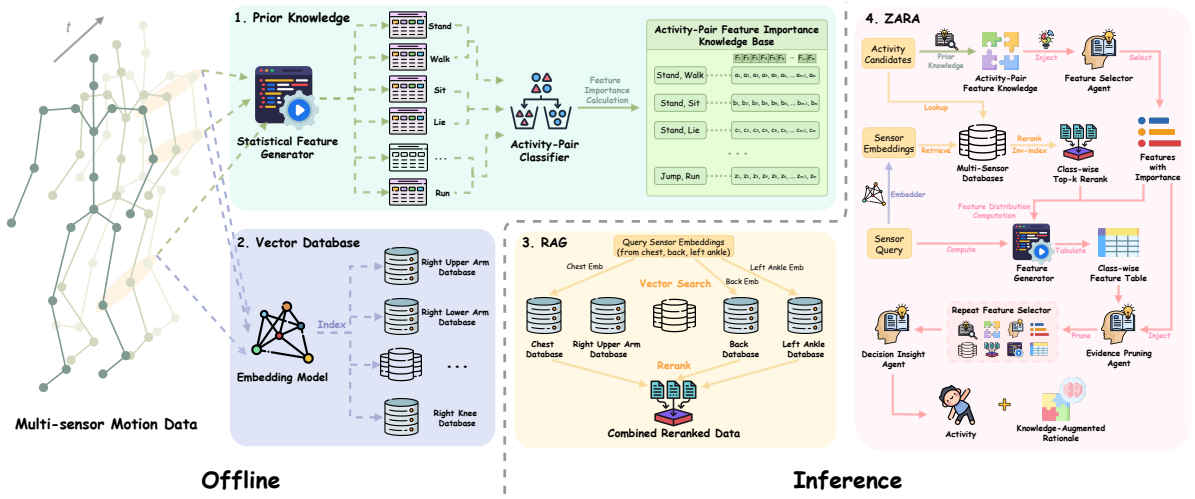


Figure 2: Overall architecture of ZARA, an evidence-grounded agentic framework augmented with knowledge and retrieval for motion time-series reasoning.

tion, but often lacks semantic granularity for complex activities. COMODO (Chen et al., 2025) distills semantics from paired video–IMU data via cross-modal self-supervision, but still relies on task-specific training.

In parallel, recent work has explored LLM agents that leverage long-context memory and external context at test time (Zhang et al., 2025b, 2026), while LLM-based reasoning methods apply similar ideas to HAR. HARGPT (Ji et al., 2024) and Yoon et al. (2024) prompt on raw signals via chain-of-thought or visual transformations, incurring high token cost and information loss. SensorLLM (Li et al., 2025) generates human-readable captions but relies on fine-tuning. ZeroHAR (Chowdhury et al., 2025) and SensorLM (Zhang et al., 2025c) add spatial metadata or hierarchical captions, but still lack verifiable, statistically grounded motion knowledge for robust training-free generalization.

3 Methodology

Figure 2 illustrates the overall framework of ZARA. We first motivate the decoupling of universal knowledge and local evidence, and then describe the construction of the knowledge base, retrieval backbone, and hierarchical agentic workflow.

Decoupling Knowledge and Evidence. Standard RAG systems typically retrieve raw samples directly. However, raw sensor signals lack explicit semantic structure, making it difficult for LLMs to reason about fine-grained physical differences. ZARA addresses this by decoupling information into two sources: *Universal Knowledge ($\mathcal{K}$):* A static reference registry storing *pairwise feature-importance profiles*. Rather than embedding sensor-grounded priors into model weights, it acts as a lookup table identifying which physical properties are most discriminative for separating specific activities (e.g., instructing the agent that "vertical acceleration variance" is the critical metric to distinguish Running from Walking). *Local Evidence ($\mathcal{D}$):* A vector database of raw-signal embeddings that serves as external memory. This provides *local distributional grounding*, allowing the model to adapt to specific sensor placements or users via in-context retrieval rather than weight adaptation.

Offline Statistical Profiling (Global Priors). To equip the LLM with structured, sensor-specific priors, we automatically construct a pairwise Activity Feature Importance Knowledge Base $\mathcal{K}$ through offline statistical analysis. Each wearable unit provides raw sensor channels (typically a 3-axis accelerometer and/or gyroscope). For every labelled window $x_a \in \mathbb{R}^{T\times C}$ of activity a (T time steps, C channels), we derive a feature pool $\mathcal{F}$ comprising low-cost, human-interpretable statistics (see Appendix A.6): *time-domain* measures (mean, variance, RMS, etc.), *frequency-domain* descriptors (spectral entropy, dominant frequency, etc.), and *cross-channel* indicators (correlations, tilt angles). For each ordered activity pair (a_i, a_j), we estimate an importance score $s[f, (a_i, a_j)]$ for every $f \in \mathcal{F}$ using permutation-based feature ranking via AutoGluon (Erickson et al., 2020). Cross-validation with fold-weighted averaging yields ro-

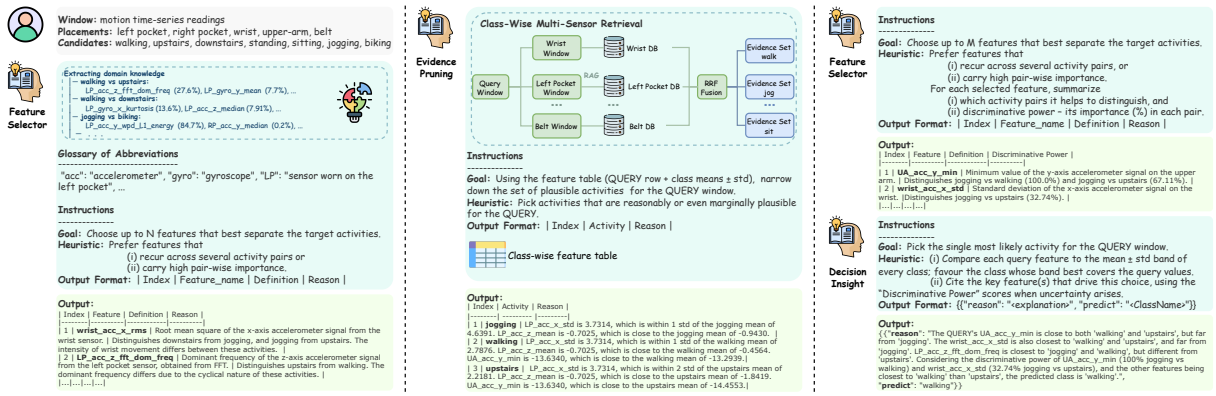


Figure 3: ZARA's multi-agent workflow with placement-specific, class-wise evidence retrieval and rank fusion.

bust estimates that generalize across folds. All feature–score tuples are stored as $\mathcal{K}[(a_i, a_j)] = [(f_1, s_1), \ldots, (f_P, s_P)]$. Because $\mathcal{K}$ is organized pairwise, it translates implicit signal characteristics into verifiable linguistic priors that can be dynamically instantiated for any candidate subset at inference time. Crucially, adding a new activity requires only registering its statistical profile against existing classes, thereby eliminating the need for task-specific retraining or manual rule curation.

Placement-Specific Retrieval (Local Evidence). To ensure retrieved evidence aligns with the query's physical context, we maintain a set of placement-specific vector stores $\{\mathcal{D}^{\text{loc}}\}$, where *loc* denotes the sensor position (e.g., wrist, ankle). Each database acts as a distributional anchor, indexing historical motion windows embedded by a frozen TS foundation encoder $g(\cdot)$ (Mantis (Feofanov et al., 2025) by default), stored alongside their statistical features, labels, and sensor metadata. The resulting embedding vectors are L2-normalized and indexed using FAISS IndexFlatIP (Douze et al., 2025). For a query embedding $u = g(x)$ and a stored vector v, similarity is computed as $\cos(u, v) = u^\top v$. This configuration restricts retrieval to distributionally aligned evidence within each body-location shard, thereby enabling robust local grounding without parameter updates.

Class-Wise Multi-Sensor Retrieval. Given a query window x with sensor placement tag *loc* and candidate activities $\mathcal{A} = \{a_1, \ldots, a_M\}$, we first compute its normalized embedding $u = g(x)$. We then perform *class-conditional retrieval* by scoring u against all historical vectors $v \in \mathcal{D}^{\text{loc}}$ labelled as a_m, producing a similarity-sorted list $\mathcal{L}_m^{\text{loc}}$. In multi-location sensor scenarios (e.g., wrist and ankle), we perform retrieval independently for each placement and fuse the rankings via Reciprocal Rank Fusion (RRF) (Cormack et al., 2009):

$$\text{RRF}(d) = \sum_{\text{loc}} \frac{1}{k_{\text{rrf}} + r_{\text{loc}}(d)}, \quad k_{\text{rrf}} = 60$$

where $r_{\text{loc}}(d)$ is the rank of document index d. Since indices are time-synchronized across sensors, this summation naturally aligns and jointly reranks multi-sensor evidence, promoting time windows that are consistently salient across modalities. Furthermore, as retrieval is performed conditionally per class, the agent receives the best available evidence for every hypothesis, ensuring balanced recall even for long-tail activities often overshadowed in global retrieval.

Hierarchical Multi-Agent Reasoning. ZARA orchestrates a hierarchical reasoning pipeline comprising three specialized agent roles executed in four stages (Figure 3). Initially, a *Feature Selector* agent queries the pairwise knowledge base $\mathcal{K}$ relative to the global candidate set $\mathcal{A}$ to identify n coarse-grained discriminative features. Subsequently, an *Evidence Pruning* agent synthesizes the retrieved class-wise evidence lists $\{\mathcal{N}_a(x)\}$ into a structured statistical comparison table (contrasting query values against class moments) based on these features. It then filters out distributionally mismatched activities, yielding a refined set $\mathcal{A}'$. The *Feature Selector* is then re-engaged on $\mathcal{A}'$ to retrieve m fine-grained features, enabling the system to resolve subtle ambiguities among the remaining candidates. Finally, a *Decision Insight* agent analyzes the updated statistics to derive the final label a', producing a transparent natural-language rationale grounded in the selected statistical features and retrieved evidence. Appendix A.7 provides the full prompts for each agent.

Dataset	Easy						Medium						Hard				Average	
	Opportunity		UCI-HAR		Shoaib		PAMAP2		USC-HAD		MHealth		WISDM		DSADS			
	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 $_{\text{Text}}$	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 $_{\text{Text}}$	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 $_{\text{Table}}$	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 $_{\text{Plot}}$	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 $_{\text{Plot}}$	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 $_{\text{Qwen-30B}}$	84.0	84.2	80.0	79.7	91.9	91.7	71.3	71.3	42.1	41.4	69.6	69.0	53.5	50.9	75.3	73.5	71.0	70.2
ZARA $_{\text{Qwen-80B}}$	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 $_{\text{GPT}}$	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 $_{\text{Gemini}}$	**92.5**	**92.5**	**90.0**	**90.0**	**97.1**	**97.1**	**76.7**	**76.9**	**60.0**	**60.1**	**86.3**	**86.1**	**65.6**	**64.1**	**84.2**	**84.4**	**81.6**	**81.4**

Table 1: Cross-Subject Evaluation: ZARA vs. 10 baselines from three method families. Best scores are shown in **bold**; second-best are underlined.

4 Experiments

We adopt a two-tier evaluation protocol to assess ZARA's generalization: (1) *Cross-Subject Generalization*, which tests robustness to individual variation within a domain; and (2) *Cross-Dataset Generalization* (Section 4.3), which evaluates transfer across heterogeneous sensor hardware and environments.

4.1 Datasets

We benchmark ZARA on 8 open-source and anonymized HAR datasets, ensuring ethical compliance and reproducibility. We group them into three levels: *Easy* (Opportunity (Roggen et al., 2010), UCI-HAR (Anguita et al., 2013), Shoaib (Shoaib et al., 2014)), *Medium* (PAMAP2 (Reiss and Stricker, 2012), USC-HAD (Zhang and Sawchuk, 2012), MHealth (Baños et al., 2014)), and *Hard* (WISDM (Weiss, 2019), DSADS (Altun et al., 2010)).

4.2 Cross-Subject Generalization

Setup. We evaluate ZARA's robustness using both open-source and proprietary LLMs: Qwen-3 (30B [1] and 80B [2]) (Yang et al., 2025), GPT-4.1-mini (OpenAI et al., 2024), and Gemini-2.0-Flash (DeepMind, 2025), with all agents set to temperature 0 for deterministic reproducibility. We employ a rigorous Subject-Hold-Out protocol (Figure 4), where the knowledge base and retrieval index are distilled exclusively from *Seen Subjects*, while inference is performed on a class-balanced split of *Held-out Subjects*. This simulates a realistic deployment where the system must adapt to new users without calibration or fine-tuning. Additionally, to showcase scalability in large-scale settings, we replace static candidate lists with dynamic retrieval for the larger WISDM and DSADS benchmarks. For each query, we calculate the cosine similarity against the vector database to dynamically select the top-10 most relevant classes, thereby decoupling inference cost from the registered activity library size while preserving high recall. Appendix A.3.1 provides sensors, activities, statistics, and preprocessing details for each benchmark.

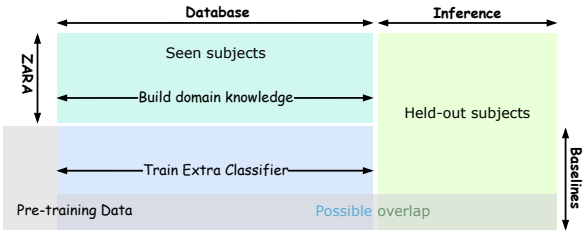
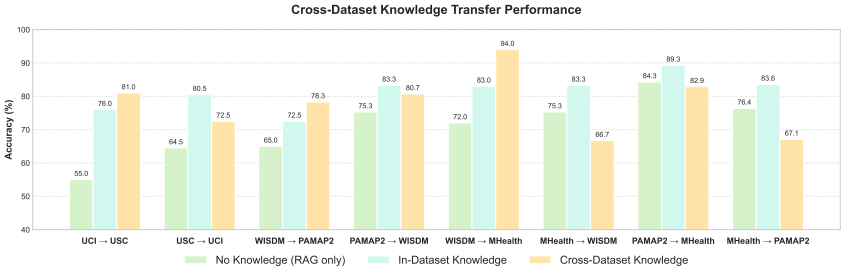


Figure 4: Subject split and data flow. ZARA builds its vector database and domain knowledge from seen subjects and tests on held-out subjects. Baselines may load pretrained weights or train a classifier head on seen subjects before testing on the same held-out subjects.

Baselines. To ensure a rigorous and fair comparison in training-free settings, we benchmark ZARA against 10 representative baselines that align with the data flow in Figure 4. These models rely

[1] qwen3-30b-a3b-instruct-2507
[2] qwen3-next-80b-a3b-instruct



Cross-Dataset Pair	# Act.	Sensor Position Mapping
UCI ↔ USC	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

Figure 5: Cross-Dataset Evaluation. **Left:** Accuracy under No-Knowledge, In-Dataset Knowledge, and Cross-Dataset Knowledge settings. **Right:** dataset compatibility in terms of shared activities and sensor position mapping.

on either pre-training on external datasets or implicit LLM knowledge to perform inference on held-out subjects. We categorize them into three families: (i) *Text-based LLMs*: $\text{HARGPT}_{\text{Text}}$ (Ji et al., 2024), $\text{Gemini}_{\text{Text}}$ and $\text{Gemini}_{\text{Table}}$. The latter follows the structured input protocol described in (Fang et al., 2024; Wang et al., 2025), where time-series data is encoded as a Markdown table; (ii) *Multimodal LLMs*: $\text{HARGPT}_{\text{Plot}}$ and $\text{Gemini}_{\text{Plot}}$, which leverage plotted sensor signals as visual prompts for activity prediction; (iii) *Pretrained HAR Models*: ImageBind (Girdhar et al., 2023), IMU2CLIP (Moon et al., 2023), NormWear (Luo et al., 2024), and UniMTS (Zhang et al., 2024), which learn modality-aligned embeddings for training-free transfer. We also include IMUGPT (Leng et al., 2023), which uniquely differs by pretraining on task-specific virtual motion data for downstream deployment. Further details are provided in Appendix A.1.

Results. Table 1 reports cross-subject performance in parameter-frozen settings. ZARA consistently outperforms all baselines across all difficulty levels. Our best variant, $\text{ZARA}_{\text{Gemini}}$, achieves an average accuracy of 81.6% and a macro F1 of 81.4%, substantially surpassing the strongest baseline, UniMTS. This advantage holds across backbone scales: ZARA variants powered by Qwen3-30B/80B and GPT-4.1-mini all outperform every baseline, indicating that the gains arise from the robustness of ZARA's knowledge- and retrieval-augmented agentic framework rather than backbone size. Performance differences across backbones are mainly attributable to their numerical reasoning ability, particularly when interpreting statistical distribution tables for inference.

In contrast, existing methods exhibit fundamental limitations. IMUGPT fails to transfer from virtual pre-training to real-world benchmarks, while contrastive approaches (ImageBind, IMU2CLIP) are restricted to single-sensor inputs. Even multi-sensor models such as UniMTS and NormWear degrade sharply on activities outside their pre-training distributions. A common failure mode across baselines is a large gap between accuracy and macro F1, revealing a strong bias toward majority classes under distribution shift. By contrast, ZARA maintains close alignment between accuracy and F1, demonstrating robust recognition of long-tail activities via class-balanced retrieval. Moreover, direct prompting methods (HARGPT, $\text{Gemini}_{\text{Text/Table/Plot}}$) fail catastrophically, highlighting that without explicit reference grounding, even capable LLMs cannot reason over numerical sensor streams. Overall, ZARA provides an interpretable, evidence-grounded solution that substantially improves the reliability of training-free inference for HAR. Detailed token-usage statistics for each agent stage are provided in Appendix A.4.

4.3 Cross-Dataset Generalization

Setup. Moving beyond subject variations, this tier evaluates robustness against sensor heterogeneity through transfer experiments across distinct dataset pairs. As illustrated in Figure 5 (Right), we establish a *Common Evaluation Protocol* restricted to the intersection of sensor placements and activity labels. All experiments in this section utilize Gemini-2.0-Flash as the backbone. Appendix A.3.2 provides sensors, activities, and preprocessing details for each cross-dataset pair.

Baselines. To strictly isolate the contribution of transferable prior knowledge, we compare three internal settings. Crucially, all settings share the same retrieval backbone and agent workflow, varying only in the source of knowledge used to guide feature selection: (1) *No Knowledge*, a baseline where the agent is provided with the full list of feature names and definitions, forcing the LLM to rely

Dataset	Easy						Medium						Hard				Time (s)
	Opportunity		UCI-HAR		Shoaib		PAMAP2		USC-HAD		MHealth		WISDM		DSADS		
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Avg
Frozen Embedder + Supervised Head																	
Moment-S	66.0	64.8	77.5	77.5	86.2	85.9	71.7	71.8	54.2	52.8	67.5	66.8	66.3	66.3	72.6	72.3	–
Moment-L	63.5	62.7	78.8	78.6	91.0	90.8	73.3	73.4	47.1	45.2	72.1	72.2	65.3	65.6	74.2	73.7	–
Mantis	90.0	89.9	**91.3**	**91.2**	93.3	92.9	**84.6**	**85.1**	53.8	53.7	86.7	86.0	**71.5**	**71.2**	**90.5**	**90.2**	–
Classifier-Free, Knowledge-Augmented Reasoning																	
ZARA$_{\text{DTW}}$	90.5	90.5	90.4	90.3	96.7	96.7	71.7	71.6	55.4	56.4	86.3	86.1	59.4	57.3	82.6	82.6	0.3826
ZARA$_{\text{Moment-S}}$	88.5	88.4	87.9	87.7	**97.6**	**97.6**	73.3	73.4	53.3	53.0	86.3	86.2	62.2	62.1	86.3	86.0	**0.0438**
ZARA$_{\text{Moment-L}}$	91.0	91.0	87.9	87.8	**97.6**	**97.6**	75.8	76.1	55.8	56.7	**88.3**	**88.0**	65.3	64.2	84.7	83.9	0.1003
ZARA$_{\text{Mantis}}$	**92.5**	**92.5**	90.0	90.0	97.1	97.1	76.7	76.9	**60.0**	**60.1**	86.3	86.1	65.6	64.1	84.2	84.4	0.1826

Table 2: Ablation of Retrieval Backbones. Comparison between ZARA and supervised baselines across diverse representations (DTW, Mantis, Moment). The rightmost column indicates the average retrieval time per query.

solely on its internal parametric knowledge to select discriminative features; (2) *In-Dataset Knowledge*, an upper-bound setting utilizing a knowledge base constructed directly from the target dataset, representing the ideal scenario with perfect domain adaptation; and (3) *Cross-Dataset Knowledge*, the proposed setting where the agent is guided by a knowledge base derived from a distinct source dataset, explicitly testing the transferability of motion priors.

Results. Figure 5 (Left) reports transfer performance across 8 scenarios. To mitigate domain shifts arising from differences in sensor hardware across datasets, we restrict transfer to the *motion priors* derived from the source. This design mirrors practical deployments, where wearable devices often differ across users, brands, and generations, making it impractical to reuse raw-signal evidence collected under a different sensor configuration. Conversely, the retrieval database is constructed from the target domain, ensuring that the *statistical feature evidence* analyzed by the LLM remains distribution-aligned with the query. This setup reveals a clear relationship between knowledge quality and transfer performance. While in-dataset knowledge performs best in 5 cases, cross-dataset knowledge unexpectedly outperforms it in 3 transfers: UCI→USC, WISDM→PAMAP2, and WISDM→MHealth.

This asymmetry is driven by two key factors. First, *User Diversity* strongly influences transferability: knowledge transferred from highly diverse sources such as WISDM (39 subjects) to lower-diversity targets like PAMAP2 (7 subjects) and MHealth (8 subjects) yields substantial gains, suggesting that knowledge derived from diverse populations captures more transferable motion priors that generalize better than local knowledge overfitted to a small cohort. Second, *Data Density* plays a critical role. Transfers from data-rich datasets such as PAMAP2 (∼4.3k samples) to data-scarce targets like MHealth (∼1.7k samples) retain higher performance than the reverse direction, indicating that dense data distributions are essential for learning fine-grained motion priors. Notably, when source knowledge quality is low (e.g., MHealth as the source), cross-dataset performance matches or falls below the No-Knowledge baseline. This behavior suggests that the agent functions as a grounded reasoning engine: it leverages injected knowledge when informative, but does not hallucinate gains beyond the quality of the provided knowledge.

5 Ablation Studies

In this section, we ablate key components of ZARA to quantify their individual impact. We fix Gemini-2.0-Flash as the backbone. Appendix A.5 provides detailed results for each ablation.

Impact of Retrieval Embedder. To evaluate the contribution of our retrieval strategies, we compare ZARA with non-LLM retrieval and supervised baselines built on different representations under four backbone settings (see details in Appendix A.2): classical DTW (Müller, 2007) and three pre-trained foundation models, Moment-Small/Large (Goswami et al., 2024) and Mantis (Feofanov et al., 2025). Moment is pre-trained via masked TS prediction, while Mantis is optimized for TS classification and explicitly incorporates HAR datasets during pre-training. We adhere to strict evaluation protocols: baselines train supervised classifiers on frozen embeddings, whereas ZARA utilizes these embeddings strictly as anchors for parameter-frozen retrieval. As shown in Table 2, ZARA exhibits strong robustness across embedders, maintaining high accuracy

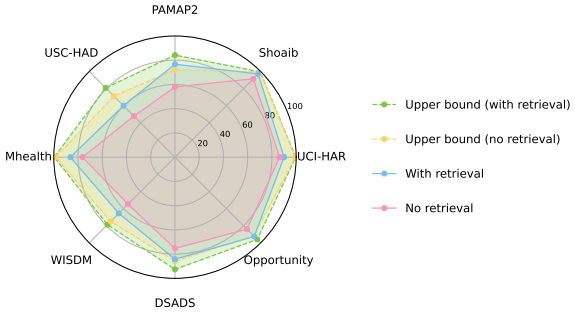


Figure 6: Impact of retrieval. The *upper bound* denotes the proportion of queries for which the pruned candidate set retains the ground-truth label under each setting.

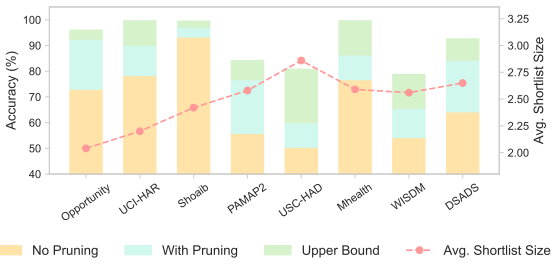


Figure 7: Accuracy with and without the *Evidence Pruning* Agent, along with upper bounds for each setting. The dashed line indicates the average length of the pruned candidate set.

in parameter-frozen settings regardless of backbone choice. Notably, ZARA frequently outperforms its supervised counterparts despite lacking task-specific parameter optimization. With Moment-Small, ZARA surpasses the baseline on 7/8 datasets in F1; with Moment-Large, it wins 7/8 datasets on both metrics; and with Mantis, it improves F1 on 4/8 datasets. Overall, ZARA ranks first on 4 datasets and second on 7. These results confirm that ZARA's knowledge- and retrieval-augmented agentic pipeline delivers substantial reasoning gains beyond raw embedding similarity, enabling classifier-free generalization without the training overhead required by supervised baselines.

Latency profiling (Table 2) reports the average time required to process a single query on an Apple M2 Max CPU with 64GB memory. Moment-Small is the fastest, whereas DTW is substantially slower due to costly pairwise sequence alignment. Although Mantis is lightweight, it incurs higher latency than Moment because it concatenates channel-wise embeddings rather than averaging them; in return, it retrieves more informative samples, reflecting a speed–quality trade-off. While absolute latency varies with query data, database size, and hardware, the relative rankings consistently capture method-level efficiency.

Removing Retrieval Reduces Performance. To assess the necessity of local evidence anchoring, we ablate the retrieval module by replacing top-k retrieval with global class-wise feature distributions computed over the entire database. This forces the LLM to reason solely on global priors without instance-level grounding. As shown in Figure 6, this degradation significantly hampers ZARA's training-free reasoning: average accuracy falls from 81.6% to 71.8%, and the upper bound (the retention rate of the ground-truth label) drops from 91.4% to 86.7%. The decline is particularly pronounced in datasets where individual instance statistics diverge from global averages. These results confirm that parameter-frozen inference requires more than just abstract knowledge; retrieval is essential to surface query-relevant evidence that bridges the gap between global statistics and local signal dynamics.

Skipping Evidence Pruning Hurts. To quantify the role of the Evidence Pruning Agent, we ablate it across all eight benchmarks. Removing pruning causes a sharp drop in average accuracy in parameter-frozen settings, from 81.6% to 68.2%. Figure 7 shows that our pruning agent typically narrows each query to 2–3 candidates per benchmark, with the easy-level datasets yielding even smaller shortlists. Moreover, this aggressive reduction is achieved with minimal information loss, maintaining a 91.4% upper-bound accuracy (ground-truth retention rate). By filtering out clearly mismatched activities early, the system allows the LLM to focus its reasoning capacity on distinguishing the remaining hard negatives using finer-grained evidence. In contrast, omitting pruning forces the LLM to reason over the full candidate pool, degrading focus and performance. Notably, even in the ablated setting, ZARA's 68.2% still outperforms every baseline.

No Prior Knowledge Fails. To quantify the value of injected priors, we disable the pairwise feature-importance knowledge registry, forcing the Feature Selector to rely solely on the LLM's intrinsic world knowledge. As shown in Figure 8, this ablation causes a sharp decline in average accuracy in parameter-frozen settings, from 81.6% to 63.4%. The average upper-bound (ground-truth

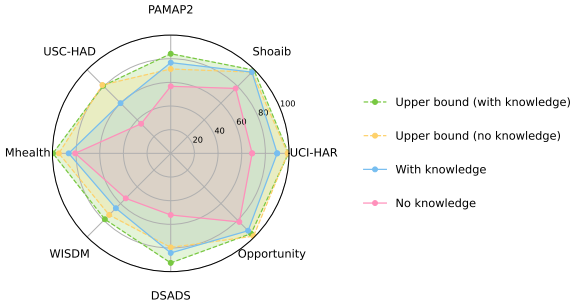


Figure 8: Impact of the prior knowledge base. The *upper bound* denotes the proportion of queries for which the pruned candidate set retains the ground-truth label under each setting.

retention) also drops from 91.4% to 87.0%, indicating that without statistical grounding, the agent struggles to identify discriminative motion properties. Benchmarks with fewer classes (e.g., Opportunity, UCI-HAR, Shoaib) are less affected in the first narrowing stage, which mainly removes clearly irrelevant classes. In the second stage, however, the Feature Selector often chooses suboptimal features due to the absence of effective criteria, leading to overly coarse feature selection that cannot recover the lost accuracy. This confirms that general-purpose LLMs cannot reliably infer discriminative motion priors from text alone; they require statistically grounded references for robust training-free inference.

6 Conclusions

We present ZARA, an agentic framework that combines statistical profiling, retrieval augmentation, and hierarchical reasoning to translate implicit sensor dynamics into explicit linguistic priors. Rather than treating labeled data solely as training targets, ZARA reinterprets them as a reference registry for grounded reasoning. This design anchors LLM decision-making in real-world evidence, enabling training-free inference that is verifiable rather than purely generative, while allowing off-the-shelf LLMs to be deployed without task-specific adaptation.

Extensive experiments across eight HAR benchmarks demonstrate that ZARA achieves state-of-the-art performance in both cross-subject and cross-dataset settings. The results show that ZARA captures transferable motion priors that generalize across diverse user populations and heterogeneous sensor domains. By producing transparent predictions grounded in retrieved evidence, ZARA offers a scalable and trustworthy path toward adaptive HAR in the wild.

7 Limitations

Direct architectural comparisons remain limited, as ZARA occupies a relatively distinct design space at the intersection of domain knowledge, retrieval augmentation, and agentic reasoning for HAR. To ensure a rigorous evaluation, we compare against strong foundation-model baselines (e.g., UniMTS, ImageBind, Mantis, and Moment), whose extensive *domain pre-training* or *task-specific classifiers* serve as practical alternatives to ZARA's retrieval-grounded inference pipeline. We do not include recent methods such as SensorLM (Zhang et al., 2025c), LLaSA (Imran et al., 2025), and RelCon (Xu et al., 2025) because of fundamental input-modality mismatches. While ZARA is designed for synchronized, multi-location motion-sensor data, these methods are typically restricted to single-location inputs, unimodal accelerometer streams, or non-motion modalities (e.g., temperature), making direct comparison infeasible on our standardized multi-view motion benchmarks.

In addition, although ZARA is parameter-frozen and avoids task-specific optimization at inference time, it still requires a labeled retrieval database to ground reasoning in representative evidence. This support set is not used to update model weights, but functions as a reference registry that links implicit sensor patterns to explicit, verifiable comparisons. Finally, the agentic workflow introduces higher computational cost than simpler unimodal pipelines. Future work will therefore explore more efficient agentic designs and domain-adapted LLMs that reduce dependence on external references while preserving transparency and grounded reasoning.

8 Acknowledgements

This work was supported by the ARC Centre of Excellence for Automated Decision-Making and Society (CE200100005).

References

Alireza Abedin, Mahsa Ehsanpour, Qinfeng Shi, Hamid Rezatofighi, and Damith C. Ranasinghe. 2021. Attend and discriminate: Beyond the state-of-the-art for human activity recognition using wearable sensors. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 5(1).

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

D. Anguita, Alessandro Ghio, L. Oneto, Xavier Parra, and Jorge Luis Reyes-Ortiz. 2013. A public domain dataset for human activity recognition using smartphones. In *The European Symposium on Artificial Neural Networks*.

Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, Jasper Zschiegner, Danielle C. Maddix, Hao Wang, Michael W. Mahoney, Kari Torkkola, Andrew Gordon Wilson, Michael Bohlke-Schneider, and Yuyang Wang. 2024. Chronos: Learning the language of time series. *Transactions on Machine Learning Research*.

Jinheon Baek, Alham Fikri Aji, and Amir Saffari. 2023. Knowledge-augmented language model prompting for zero-shot knowledge graph question answering. In *Proceedings of the First Workshop on Matching From Unstructured and Structured Data (MATCHING 2023)*, pages 70–98, Toronto, ON, Canada. Association for Computational Linguistics.

Oresti Baños, Rafael García, Juan Antonio Holgado Terriza, Miguel Damas, Héctor Pomares, Ignacio Rojas, Alejandro Saez, and Claudia Villalonga. 2014. mhealthdroid: A novel framework for agile development of mobile health applications. In *International Workshop on Ambient Assisted Living and Home Care*.

Baiyu Chen, Wilson Wongso, Zechen Li, Yonchanok Khaokaew, Hao Xue, and Flora Salim. 2025. Comodo: Cross-modal video-to-imu distillation for efficient egocentric human activity recognition. *Preprint*, arXiv:2503.07259.

Ranak Roy Chowdhury, Ritvik Kapila, Ameya Panse, Xiyuan Zhang, Diyan Teng, Rashmi Kulkarni, Dezhi Hong, Rajesh K. Gupta, and Jingbo Shang. 2025. Zerohar: Sensor context augments zero-shot wearable action recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(15):16046–16054.

Gordon V. Cormack, Charles L A Clarke, and Stefan Buettcher. 2009. 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*, SIGIR '09, page 758–759, New York, NY, USA. Association for Computing Machinery.

Google DeepMind. 2025. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, and 1 others. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2025. The faiss library. *Preprint*, arXiv:2401.08281.

Nick Erickson, Jonas Mueller, Alexander Shirkov, Hang Zhang, Pedro Larroy, Mu Li, and Alexander Smola. 2020. Autogluon-tabular: Robust and accurate automl for structured data. *arXiv preprint arXiv:2003.06505*.

Xi Fang, Weijie Xu, Fiona Anting Tan, Ziqing Hu, Jiani Zhang, Yanjun Qi, Srinivasan H. Sengamedu, and Christos Faloutsos. 2024. Large language models (LLMs) on tabular data: Prediction, generation, and understanding - a survey. *Transactions on Machine Learning Research*.

Vasilii Feofanov, Songkang Wen, Marius Alonso, Romain Ilbert, Hongbo Guo, Malik Tiomoko, Lujia Pan, Jianfeng Zhang, and Ievgen Redko. 2025. Mantis: Lightweight calibrated foundation model for user-friendly time series classification. *Preprint*, arXiv:2502.15637.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. 2023. Imagebind: One embedding space to bind them all. In *CVPR*.

Mononito Goswami, Konrad Szafer, Arjun Choudhry, Yifu Cai, Shuo Li, and Artur Dubrawski. 2024. Moment: A family of open time-series foundation models. In *International Conference on Machine Learning*.

Sheikh Asif Imran, Mohammad Nur Hossain Khan, Subrata Biswas, and Bashima Islam. 2025. Llasa: A sensor-aware llm for natural language reasoning of human activity from imu data. *Preprint*, arXiv:2406.14498.

Sijie Ji, Xinzhe Zheng, and Chenshu Wu. 2024. Hargpt: Are llms zero-shot human activity recognizers? *Preprint*, arXiv:2403.02727.

Zikang Leng, Hyeokhyen Kwon, and Thomas Ploetz. 2023. 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*, ISWC '23.

Zechen Li, Shohreh Deldari, Linyao Chen, Hao Xue, and Flora D. Salim. 2025. 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.

Yunfei Luo, Yuliang Chen, Asif Salekin, and Tauhidur Rahman. 2024. Toward foundation model for multivariate wearable sensing of physiological signals. *Preprint*, arXiv:2412.09758.

Wannes Meert, Kilian Hendrickx, Toon Van Craenendonck, Pieter Robberechts, Hendrik Blockeel, and Jesse Davis. 2021. DTAIDistance (version v2).

Seungwhan Moon, Andrea Madotto, Zhaojiang Lin, Aparajita Saraf, Amy Bearman, and Babak Damavandi. 2023. IMU2CLIP: Language-grounded motion sensor translation with multimodal contrastive learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13246–13253, Singapore. Association for Computational Linguistics.

Meinard Müller. 2007. Dynamic time warping. *Information Retrieval for Music and Motion*, 2:69–84.

OpenAI. 2024. Gpt-4o mini: advancing cost-efficient intelligence.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Francisco Javier Ordóñez and Daniel Roggen. 2016. Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition. *Sensors*, 16(1).

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Attila Reiss and Didier Stricker. 2012. Introducing a new benchmarked dataset for activity monitoring. In *2012 16th International Symposium on Wearable Computers*, pages 108–109.

Daniel Roggen, Alberto Calatroni, Mirco Rossi, Thomas Holleczek, Kilian Förster, Gerhard Tröster, Paul Lukowicz, David Bannach, Gerald Pirkl, Alois Ferscha, Jakob Doppler, Clemens Holzmann, Marc Kurz, Gerald Holl, Ricardo Chavarriaga, Hesam Sagha, Hamidreza Bayati, Marco Creatura, and José del R. Millán. 2010. Collecting complex activity datasets in highly rich networked sensor environments. *2010 Seventh International Conference on Networked Sensing Systems (INSS)*, pages 233–240.

Sheng Shen, Chunyuan Li, Xiaowei Hu, Yujia Xie, Jianwei Yang, Pengchuan Zhang, Zhe Gan, Lijuan Wang, Lu Yuan, Ce Liu, Kurt Keutzer, Trevor Darrell, Anna Rohrbach, and Jianfeng Gao. 2022. K-lite: Learning transferable visual models with external knowledge. In *Advances in Neural Information Processing Systems*, volume 35, pages 15558–15573. Curran Associates, Inc.

Muhammad Shoaib, Stephan Bosch, Ozlem Durmaz Incel, Hans Scholten, and Paul J. M. Havinga. 2014. Fusion of smartphone motion sensors for physical activity recognition. *Sensors*, 14(6):10146–10176.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Jiahao Wang, Mingyue Cheng, Qingyang Mao, Yitong Zhou, Daoyu Wang, Qi Liu, Feiyang Xu, and Xin Li. 2025. Tabletime: Reformulating time series classification as training-free table understanding with large language models. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, CIKM '25, page 3009–3019, New York, NY, USA. Association for Computing Machinery.

Gary Weiss. 2019. WISDM Smartphone and Smartwatch Activity and Biometrics Dataset . UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5HK59.

Chen-Wei Xie, Siyang Sun, Xiong Xiong, Yun Zheng, Deli Zhao, and Jingren Zhou. 2023. Ra-clip: Retrieval augmented contrastive language-image pre-training. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19265–19274.

Maxwell A Xu, Jaya Narain, Gregory Darnell, Haraldur T Hallgrimsson, Hyewon Jeong, Darren Forde, Richard Andres Fineman, Karthik Jayaraman Raghuram, James Matthew Rehg, and Shirley You Ren. 2025. Relcon: Relative contrastive learning for a motion foundation model for wearable data. In *The Thirteenth International Conference on Learning Representations*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Hyungjun Yoon, Biniyam Aschalew Tolera, Taesik Gong, Kimin Lee, and Sung-Ju Lee. 2024. By my eyes: Grounding multimodal large language models with sensor data via visual prompting. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2219–2241, Miami, Florida, USA. Association for Computational Linguistics.

Mi Zhang and Alexander A. Sawchuk. 2012. Usc-had: a daily activity dataset for ubiquitous activity recognition using wearable sensors. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, UbiComp '12, page 1036–1043, New York, NY, USA. Association for Computing Machinery.

Weizhi Zhang, Yangning Li, Yuanchen Bei, Junyu Luo, Guancheng Wan, Liangwei Yang, Chenxuan Xie, Yuyao Yang, Wei-Chieh Huang, Chunyu Miao, Henry Peng Zou, Xiao Luo, Yusheng Zhao, Yankai Chen, Chunkit Chan, Peilin Zhou, Xinyang Zhang, Chenwei Zhang, Jingbo Shang, and 4 others. 2025a. From web search towards agentic deep research: Incentivizing search with reasoning agents. *Preprint*, arXiv:2506.18959.

Weizhi Zhang, Xiaokai Wei, Wei-Chieh Huang, Zheng Hui, Chen Wang, Michelle Gong, and Philip S. Yu. 2026. Memorycd: Benchmarking long-context user memory of llm agents for lifelong cross-domain personalization. *Preprint*, arXiv:2603.25973.

Weizhi Zhang, Xinyang Zhang, Chenwei Zhang, Liangwei Yang, Jingbo Shang, Zhepei Wei, Henry Peng Zou, Zijie Huang, Zhengyang Wang, Yifan Gao, Xiaoman Pan, Lian Xiong, Jingguo Liu, Philip S. Yu, and Xian Li. 2025b. Personaagent: When large language model agents meet personalization at test time. *Preprint*, arXiv:2506.06254.

Xiyuan Zhang, Diyan Teng, Ranak Roy Chowdhury, Shuheng Li, Dezhi Hong, Rajesh K. Gupta, and Jingbo Shang. 2024. Unimts: Unified pre-training for motion time series. In *Advances in Neural Information Processing Systems*, volume 37, pages 107469–107493. Curran Associates, Inc.

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. 2025c. Sensorlm: Learning the language of wearable sensors. *Preprint*, arXiv:2506.09108.

A Appendix

This appendix provides additional implementation and evaluation details supporting the main findings of the paper. We first describe the baseline models used for comparison, followed by dataset and preprocessing details. We then report per-dataset token usage for each agent and per-class performance across all benchmarks, complementing the aggregate results in the main text. Finally, we include the full prompts used at each reasoning stage in ZARA.

A.1 Baselines

We provide implementation details for all baselines used in our study, including how each method was reproduced or adapted for evaluation in our training-free HAR setting.

HARGPT (Ji et al., 2024). This method directly prompts LLMs to classify motion time-series. We follow the original setup by downsampling input signals to 10 Hz and applying the prompt template for raw numerical input. We additionally evaluate a visual-input variant by providing plotted sensor signals to the underlying LLM (GPT-4o-mini (OpenAI, 2024)). For visual inputs, each 6-channel sensor is rendered as an individual subplot, and multiple sensors are concatenated into a single composite figure to reduce visual clutter in multi-sensor datasets.

Gemini (DeepMind, 2025). To assess the improvements brought by ZARA, we use Gemini-2.0-Flash, the same LLM backbone adopted in our framework, as a standalone baseline. Similar to HARGPT, Gemini is evaluated with raw numerical sequences and plotted sensor signals. In addition, we include a third input modality: Markdown-formatted tables, to test whether more structured input alone improves recognition accuracy. These Gemini baselines allow us to isolate the contribution of ZARA's retrieval, knowledge, and agentic reasoning modules beyond the capability of the base model itself.

ImageBind (Girdhar et al., 2023). ImageBind learns a unified embedding space across six modalities: image, text, audio, depth, thermal, and IMU. We use the publicly released `imagebind_huge` checkpoint for evaluation in our training-free setting. Since ImageBind only supports single-sensor input and requires fixed-length windows of size 6×2000, we evaluate each sensor placement separately. To satisfy the input-length requirement, we apply two strategies—repeat padding and linear interpolation to 2000 steps—and report the best result across sensor placements for each dataset.

IMU2CLIP (Moon et al., 2023). IMU2CLIP aligns inertial measurement unit (IMU) data with video and text by projecting them into CLIP's joint embedding space. Similar to ImageBind, it only supports single-sensor input and requires fixed-length windows of size 6×1000. We therefore evaluate each sensor placement separately and use both repeat padding and interpolation to match the required input length, reporting the best result for each dataset.

NormWear (Luo et al., 2024). NormWear is a foundation model designed to extract generalized representations from multivariate wearable signals. It is pre-trained on a diverse corpus of physiological data, including PPG, ECG, EEG, GSR, and IMU, collected from multiple public datasets. For activity recognition in our training-free setting, we follow the official documentation and use the recommended prompt, "What is the activity being performed currently?", together with the candidate activity labels for inference.

IMUGPT (Leng et al., 2023). IMUGPT generates synthetic training data by first prompting GPT-4o-mini to produce diverse textual activity descriptions. These descriptions are converted into 3D motion sequences and then into virtual IMU streams. For evaluation, we adopt DeepConvLSTM, the best-performing backbone reported in the original paper. To ensure a fair comparison in our training-free setting, we exclude the supervised distribution calibration stage, which depends on labeled downstream data.

UniMTS (Zhang et al., 2024). UniMTS proposes a unified pretraining framework for motion time-series that generalizes across diverse device configurations, including sensor position and orientation. It uses contrastive learning to align motion signals with LLM-enriched text descriptions, enabling classifier-free recognition through semantic matching. For our experiments, we follow the official implementation and evaluate UniMTS in its training-free inference setting, using the released checkpoints and the corresponding text-label matching protocol.

Dataset	# Subjects		# Samples	
	Database	Inference	Database	Inference
Opportunity	3	1	6968	200
UCI-HAR	21	9	7352	240
Shoaib	8	2	5040	210
PAMAP2	7	2	7138	240
USC-HAD	12	2	11889	240
MHealth	8	2	2799	240
WISDM	39	8	14287	288
DSADS	6	2	6840	190

Table 3: Cross-subject dataset statistics.

A.2 Retrieval Strategies

All retrieval strategies in ZARA follow a two-stage pipeline. First, candidates are ranked independently within each sensor placement according to their similarity to the query. The ranked lists are then aggregated across placements using Reciprocal Rank Fusion (RRF) to obtain the final retrieval results. We evaluate four retrieval backbones: DTW and three pretrained time-series encoders.

Dynamic Time Warping (DTW) (Müller, 2007). We implement DTW using the multi-dimensional variant from the `dtaidistance` package (Meert et al., 2021). For each query segment and each database candidate, we first apply z-score normalization independently to each of the six sensor channels. We then compute the DTW distance between the query and each candidate segment using the `distance_fast` routine with pruning enabled to accelerate matching. Distances are negated to obtain similarity scores, and the top-k candidates are returned for retrieval.

Moment (Goswami et al., 2024). Moment is a time-series foundation model based on the T5 architecture, pre-trained on a range of tasks including classification, anomaly detection, and forecasting. We evaluate both the `moment-small` and `moment-large` variants, whose embedding dimensions are 512 and 1024, respectively. For multi-channel inputs, Moment produces per-channel embeddings, which we average to obtain a single representation for retrieval in ZARA. In the supervised baseline comparison reported in Table 2, we additionally follow the standard frozen-encoder protocol by training an SVM classifier on top of the database split and selecting hyperparameters through greedy search.

Mantis (Feofanov et al., 2025). Mantis is a foundation model for time-series classification built on the Vision Transformer (ViT) architecture and pre-trained via contrastive learning. It has also been pre-trained on several HAR-related datasets. For input processing, Mantis rescales each time-series to a fixed length of 512, extracts a 256-dimensional embedding from each channel, and concatenates the channel-wise embeddings into a unified representation. We use this representation as the retrieval backbone in ZARA. In the supervised baseline comparison reported in Table 2, we follow the frozen-encoder setting of the original method by training a random forest classifier on the database split.

A.3 Data Preprocessing

A.3.1 Cross-Subject Generalization

This section introduces the data preprocessing pipeline for the Section 4.2 Cross-Subject Generalization study. Due to the cost constraints of API-based inference and the need for detailed ablation analyses, we evaluate each dataset on a randomly sampled inference subset. For each dataset, we ensure balanced coverage by sampling an equal number of non-overlapping windows per activity class and per subject. This design strikes a practical balance between cost efficiency and diversity across datasets, activity types, and subjects. For fair comparison, we keep the sampled subsets identical across all baselines. Each dataset contains multiple activity classes, and the corresponding sensor placements are summarized in Table 4. Table 3 reports the statistics for the database and test sets.

Opportunity (Roggen et al., 2010). The dataset contains recordings from 4 subjects at a sampling rate of 30 Hz. We designate Subject 4 as the inference user and use data from the remaining subjects to build the retrieval database. Motion sensor data are segmented into non-overlapping 2-second windows (60 timesteps each). For Inference, we randomly sample 50 windows per activity class from the inference split, resulting in a balanced set of 200 samples.

UCI-HAR (Anguita et al., 2013). The dataset contains recordings from 30 volunteers, sampled at 50 Hz. The dataset is pre-segmented using fixed-width sliding windows of 2.56 seconds with 50% overlap. Following the original split, we use data from test set (9 subjects) for inference and the remaining for the database. From the inference set, we randomly sample 40 windows per activity, en-

Dataset	# Classes	Classes	Sensor Placements
Opportunity	4	Stand, Walk, Sit, Lie	Back, upper arms, lower arms
UCI-HAR	6	Standing, Sitting, Laying, Walking, Walking downstairs, Walking upstairs	Waist
Shoaib	7	Walking, Standing, Jogging, Sitting, Biking, Downstairs, Upstairs	right pockets, left pockets, belt, Right upper arm, right wrist
PAMAP2	12	Lying, Sitting, Standing, Ironing, Vacuum cleaning, Ascending stairs, Descending stairs, Walking, Nordic walking, Cycling, Running, Rope jumping	Wrist, chest, ankle
USC-HAD	12	Sleeping, Sitting, Elevator down, Elevator up, Standing, Jumping, Walking downstairs, Walking right, Walking forward, Running forward, Walking upstairs, Walking left	Front right hip
MHealth	12	Climbing stairs, Standing still, Sitting and relaxing, Lying down, Walking, Waist bends forward, Frontal elevation of arms, Knees bending (crouching), Jogging, Running, Jump front & back, Cycling	Chest, right wrist, left ankle
WISDM	18	Walking, Jogging, Stairs, Sitting, Standing, Typing, Brushing Teeth, Eating Soup, Eating Chips, Eating Pasta, Eating Sandwich, Kicking Ball, Playing Catch Ball, Drinking, Dribbling Ball, Writing, Clapping, Folding Clothes	Hand
DSADS	19	Sitting, Standing, Lying on back, Lying on right side, Ascending stairs, Descending stairs, Standing in elevator, Moving around in elevator, Walking slowly, Rowing, Jumping, Walking on a treadmill in flat positions, Walking on a treadmill in inclined positions, Running on a treadmill fast, Exercising on a stepper, Exercising on a cross trainer, Playing basketball, Cycling on an exercise bike in horizontal positions, Cycling on an exercise bike in vertical positions	Torso, right arm, left arm, right leg, left leg

Table 4: Dataset classes and sensor placements.

suring user-balanced representation within each class, resulting in a total of 240 samples.

Shoaib (Shoaib et al., 2014). The dataset contains recordings from 10 subjects, sampled at 50 Hz. We use data from subjects 1 and 9 for inference and the remaining subjects for the database. The recordings are segmented into non-overlapping windows of 2 seconds (100 timesteps). For inference, we randomly sample 30 windows per activity from the inference users (15 from each) yielding a class-balanced test set of 210 samples.

PAMAP2 (Reiss and Stricker, 2012). This dataset contains recordings from 9 subjects at a sampling rate of 100 Hz. We designate subjects 5 and 6 for inference, and use the rest for the database. Recordings are segmented into non-overlapping 2-second windows (200 time steps). For inference, we randomly sample 20 windows per activity from the inference users (10 from each) except for rope jumping, which has limited data. For this activity, we include 18 samples from subject 5 and 2 from subject 6, resulting in a total of 240 samples.

USC-HAD (Zhang and Sawchuk, 2012). This dataset includes motion recordings from 14 subjects at a sampling rate of 100 Hz. We designate subjects 13 and 14 for inference, using the remaining subjects to build the database. Data are segmented into non-overlapping 2-second windows (200 time steps). For inference, we randomly sample 20 windows per activity (10 from each subject), resulting in 240 total samples.

MHealth (Baños et al., 2014). The MHealth dataset contains recordings from 10 subjects at a sampling rate of 50 Hz. We use subjects 1 and 6 for inference and the remaining subjects to construct the database. Signals are segmented into non-overlapping 2-second windows (100 time steps). For inference, we randomly sample 20 windows per activity (10 from each subject), yielding a total of 240 evaluation samples.

WISDM (Weiss, 2019). We use the smartwatch-on-hand subset of the WISDM dataset, recorded at 20 Hz. Accelerometer and gyroscope signals are aligned by timestamp, and we select 47 users whose data show no alignment anomalies. Among

them, 8 users are held out for inference and the rest are used to construct the database. Following the dataset's recommendation, we segment the data into non-overlapping 10-second windows (200 time steps). For inference, we randomly sample 16 windows per activity (2 from each subject), resulting in a total of 288 inference samples.

DSADS (Altun et al., 2010). The DSADS dataset contains recordings from 8 users at a sampling rate of 25 Hz. We designate subjects 2 and 4 for inference and use the remaining users to build the database. We adopt the predefined 5-second windows (125 time steps) provided by the dataset. For inference, we randomly sample 10 windows per activity (5 from each subject), yielding a total of 190 inference samples.

A.3.2 Cross-Dataset Generalization

This section provides supplementary details for the cross-dataset generalization experiments in Section 4.3, including dataset-specific preprocessing, activity-label mapping, and sensor-placement alignment used to establish comparable transfer scenarios. Following the common evaluation protocol defined in the main text, we restrict each transfer pair to the intersection of available sensor locations and activity classes, ensuring that all methods are evaluated under identical conditions.

Importantly, in the cross-dataset setting we transfer only the knowledge component, i.e., guidance on discriminative features, rather than retrieving evidence from the source domain. Because datasets differ in hardware, sampling rate, coordinate conventions, and wearing conditions, source-domain windows often carry dataset-specific signatures that do not reliably align with target-domain queries in the embedding space and may introduce spurious nearest-neighbor matches. We therefore construct the retrieval database exclusively from the target dataset to maintain evidence distribution alignment, while using cross-dataset knowledge to isolate and evaluate the transferability of motion priors.

UCI-HAR↔USC-HAD. This transfer pair evaluates cross-dataset generalization between UCI-HAR and USC-HAD over five shared activities: *Walking, Walking upstairs, Walking downstairs, Sitting,* and *Standing*. To satisfy the common evaluation protocol, we align sensor placement by mapping *Waist* (UCI-HAR) to *Front right hip* (USC-HAD). We additionally downsample USC-HAD to 50 Hz to match the sampling rate of UCI-HAR. For cost-efficient inference with balanced coverage, we randomly sample 20 windows per activity from USC-HAD (10 per subject) and 40 windows per activity from UCI-HAR (4–5 per subject), ensuring non-overlapping instances.

PAMAP2↔WISDM. We evaluate transfer between PAMAP2 and WISDM on five overlapping activities: *Walking, Standing, Jogging, Sitting,* and *Stairs*. Sensor placements are aligned by mapping *Wrist* (PAMAP2) to *Hand* (WISDM). To harmonize temporal resolution, PAMAP2 is downsampled to 20 Hz to match WISDM. We then construct balanced inference subsets by sampling 20 windows per activity from PAMAP2 (10 per subject) and 30 windows per activity from WISDM (3–4 per subject), with all samples non-overlapping.

MHealth↔WISDM. This pair studies transfer across MHealth and WISDM using the same five shared activities: *Walking, Standing, Jogging, Sitting,* and *Stairs*. We align sensor positions by mapping *Right wrist* (MHealth) to *Hand* (WISDM). Since WISDM is collected at 20 Hz, MHealth is downsampled accordingly to ensure consistent input resolution. For inference, we randomly sample 20 windows per activity from MHealth (10 per subject) and 30 windows per activity from WISDM (3–4 per subject), maintaining balanced class and subject coverage with non-overlapping windows.

PAMAP2↔MHealth. For a multi-sensor transfer setting, we consider seven shared activities between PAMAP2 and MHealth: *Lying, Sitting, Standing, Walking, Running, Cycling,* and *Ascending stairs*. We align sensor placements across three locations using the following mappings: *Wrist↔Right wrist*, *Chest↔Chest*, and *Ankle↔Left ankle*. To match MHealth, PAMAP2 is downsampled to 50 Hz. We then sample 20 non-overlapping windows per activity from each dataset (10 per subject), yielding balanced inference subsets for both domains.

A.4 Token Usage Breakdown

To quantify the inference cost of ZARA, we report the token usage of each agent stage across all datasets. For every dataset, we measure the input and output tokens consumed by each agent and report the averages in Table 5. This analysis complements the main results by making the cost–accuracy trade-offs explicit and enabling transparent comparison of compute overhead across evalu-

Dataset	Agent 1		Agent 2		Agent 3		Agent 4	
	In	Out	In	Out	In	Out	In	Out
Opportunity	1505	343	1784	480	941	573	1382	781
UCI-HAR	2836	354	2651	319	604	271	802	400
Shoaib	1815	495	3683	304	511	220	711	306
PAMAP2	13704	1409	15664	1033	690	317	913	454
USC-HAD	8852	1152	14190	1547	830	477	1296	497
MHealth	2870	623	7366	537	451	206	691	271
WISDM	8531	1545	15398	1157	610	274	809	411
DSADS	4834	1036	9695	847	606	314	870	389

Table 5: Per-dataset token usage (input/output) for each agent stage. All token statistics are collected from experiments using Gemini-2.0-Flash for all agent stages.

Dataset	With Retrieval		No Retrieval	
	Acc	UB	Acc	UB
Opportunity	92.5	96.0	84.0	92.0
UCI-HAR	90.0	99.6	86.3	99.2
Shoaib	97.1	99.5	91.4	99.5
PAMAP2	76.7	84.2	57.9	72.1
USC-HAD	60.0	80.8	47.9	71.3
MHealth	86.3	99.6	76.3	98.3
WISDM	65.6	78.8	54.9	75.0
DSADS	84.2	92.6	75.3	85.8
Average	**81.6**	**91.4**	**71.8**	**86.7**

Table 6: Impact of Retrieval on *Evidence Pruning* and *Decision and Insight* Agents. We report training-free accuracy (Acc) and upper-bound accuracy (UB) with and without retrieval across all datasets.

ation settings.

As shown in Table 5, token consumption is dominated by the early agent stages responsible for candidate construction and for assembling knowledge and retrieval evidence, whereas the later reasoning stages are comparatively lightweight. Across datasets, larger input-token counts generally correlate with richer sensor knowledge and a larger number of activity classes, both of which increase the amount of numerical evidence and class-conditioned comparisons included in the prompt. In all cases, input tokens consistently exceed output tokens, indicating that inference cost is driven more by contextual grounding than by generation length. Importantly, our prompts remain substantially shorter than approaches that directly linearize raw sensor signals into text, whose token cost scales with sequence length and channel count and thus quickly becomes impractical in API-based inference. Future work could further reduce inference cost by caching static knowledge, compressing numerical evidence, or using smaller candidate sets.

Dataset	Pruning	Upper Bound	No Pruning	Avg. Length
Opportunity	92.5	96.0	73.0	2.04
UCI-HAR	90.0	99.6	78.3	2.20
Shoaib	97.1	99.5	93.3	2.42
PAMAP2	76.7	84.2	55.8	2.58
USC-HAD	60.0	80.8	50.4	2.86
MHealth	86.3	99.6	76.7	2.59
WISDM	65.6	78.8	54.2	2.56
DSADS	84.2	92.6	64.2	2.65
Average	**81.6**	**91.4**	**68.2**	**2.49**

Table 7: Impact of the Evidence Pruning Agent: Accuracy (%) and Upper Bound with and without pruning, along with the average pruned shortlist length per dataset.

Dataset	With knowledge		No knowledge	
	Acc	UB	Acc	UB
Opportunity	92.5	96.0	82.0	99.0
UCI-HAR	90.0	99.6	68.8	98.9
Shoaib	97.1	99.5	77.6	97.1
PAMAP2	76.7	84.2	56.7	71.3
USC-HAD	60.0	80.8	35.4	81.7
Mhealth	86.3	99.6	80.8	94.6
WISDM	65.6	78.8	53.8	73.6
DSADS	84.2	92.6	52.1	79.5
Average	**81.6**	**91.4**	**63.4**	**87.0**

Table 8: Impact of Prior Knowledge Injection on Feature Selector Accuracy and Upper Bound.

A.5 Ablation

This section provides detailed results for the ablation studies presented in Section 5. All results are obtained using Gemini-2.0-Flash.

Removing Retrieval Reduces Performance. Table 6 reports the exact values underlying Figure 6, comparing ZARA with and without the Evidence Retrieval module. We report both the final accuracy in the training-free setting and the pruning-stage upper-bound accuracy for each dataset.

Skipping Evidence Pruning Hurts. Table 7 reports the dataset-level breakdown of ZARA's accuracy in the training-free setting, with and without the Evidence Pruning Agent, together with the corresponding upper-bound accuracy and the average length of the pruned candidate shortlist. These results complement Figure 7 in the main paper and confirm that pruning substantially improves performance while preserving high-quality candidates.

No Prior Knowledge Fails. Table 8 reports the exact values plotted in Figure 8, comparing ZARA

Dataset	# of Channels	Window Size	Database Size	DTW	Moment-s	Moment-l	Mantis
Opportunity	30	60	6968	0.5814	0.0683	0.1508	0.3072
UCI-HAR	6	128	7352	0.1389	0.0169	0.0342	0.0623
Shoaib	30	100	5040	0.4935	0.0686	0.1700	0.3122
PAMAP2	18	200	7138	0.5104	0.0446	0.1053	0.1777
USC-HAD	6	200	11889	0.2584	0.0180	0.0389	0.0699
Mhealth	15	100	2799	0.1504	0.0405	0.0900	0.1528
WISDM	6	200	14287	0.3152	0.0190	0.0369	0.0679
DSADS	30	125	6840	0.6127	0.0741	0.1762	0.3105

Table 9: Per-query retrieval latency (in seconds) and dataset characteristics.

with and without the prior knowledge base across all eight datasets. These results confirm that prior knowledge plays a critical role in both narrowing the candidate set and distinguishing among activity classes.

Retrieval Latency by Dataset. Table 9 reports the average per-query latency (in seconds) of each retrieval method across all eight datasets. Measurements were taken on an Apple M2 Max CPU with 64GB memory. Although latency varies with window length, number of channels, and database size, the relative ranking remains consistent: DTW is the slowest, Moment-Small is the fastest, and Mantis offers a balanced trade-off between speed and retrieval quality.

Per-Class Evaluation. Table 1 in the main paper reports only the overall accuracy and macro F1 for each dataset. To further assess performance consistency across activity types, we provide detailed per-class accuracy in Figure 9. ZARA consistently achieves strong per-class performance, reflecting its ability to distinguish a wide range of behaviors through structured knowledge and retrieved evidence. By contrast, several baselines exhibit substantially lower F1 than accuracy, indicating a bias toward dominant classes. This further highlights ZARA's stronger and more balanced generalization across HAR activities.

A.6 Predefined Features

To support feature-based reasoning and retrieval, we extract a comprehensive set of handcrafted features from each sensor channel (6 axes per sensor plus 2 magnitude channels). These features span both the time and frequency domains, and are designed to capture fine-grained temporal, statistical, and spectral characteristics of motion signals. The full feature set used for each channel is summarized in Table 10.

A.7 Prompt Template

Below we provide the full prompts used by ZARA during inference. These prompts cover all stages of the reasoning pipeline: the First Feature Selector (Figure 10), Evidence Pruning (Figure 11), the Second Feature Selector (Figure 12), and Decision Insight (Figure 13), together with their corresponding inputs and output formats. Although the exact wording may vary slightly across datasets or instances, the overall structure is consistent throughout all experiments.

Category	Feature Description
Time-domain Features	Mean, Standard Deviation (STD), Variance, Maximum, Minimum, Median, Root Mean Square (RMS), Peak Amplitude, Zero-Crossing Rate, Slope, Mean/RMS/STD of First-Order Differences, Range, Sum, Signal Absolute Value, Mean Absolute Value, Interquartile Range, Skewness, Kurtosis, Signal Magnitude Area
Frequency-domain (FFT)	Band Power (Low, Mid, High), Band Power Ratio (Low, Mid, High), Dominant Frequency, Power of Dominant Frequency, Second Peak Frequency and Power, Spectral Centroid, Spectral Entropy, Spectral Skewness, Spectral Kurtosis, Weighted Average Frequency, Spectral Energy, Max Power Index
Frequency-domain (STFT)	STFT Max / Mean / STD in Low, Mid, High bands, STFT Entropy (Mean, Max, STD), STFT Centroid (Mean, Max, STD)
Autocorrelation	First Peak Lag, First Minimum Lag, First Zero-Crossing Lag
Jerk-based Features	Jerk RMS, Peak, Zero-Crossing Rate
Cross-Channel Features	Pearson Correlation Between Channels

Table 10: Predefined features extracted per sensor channel (6 axes + 2 magnitudes per sensor).

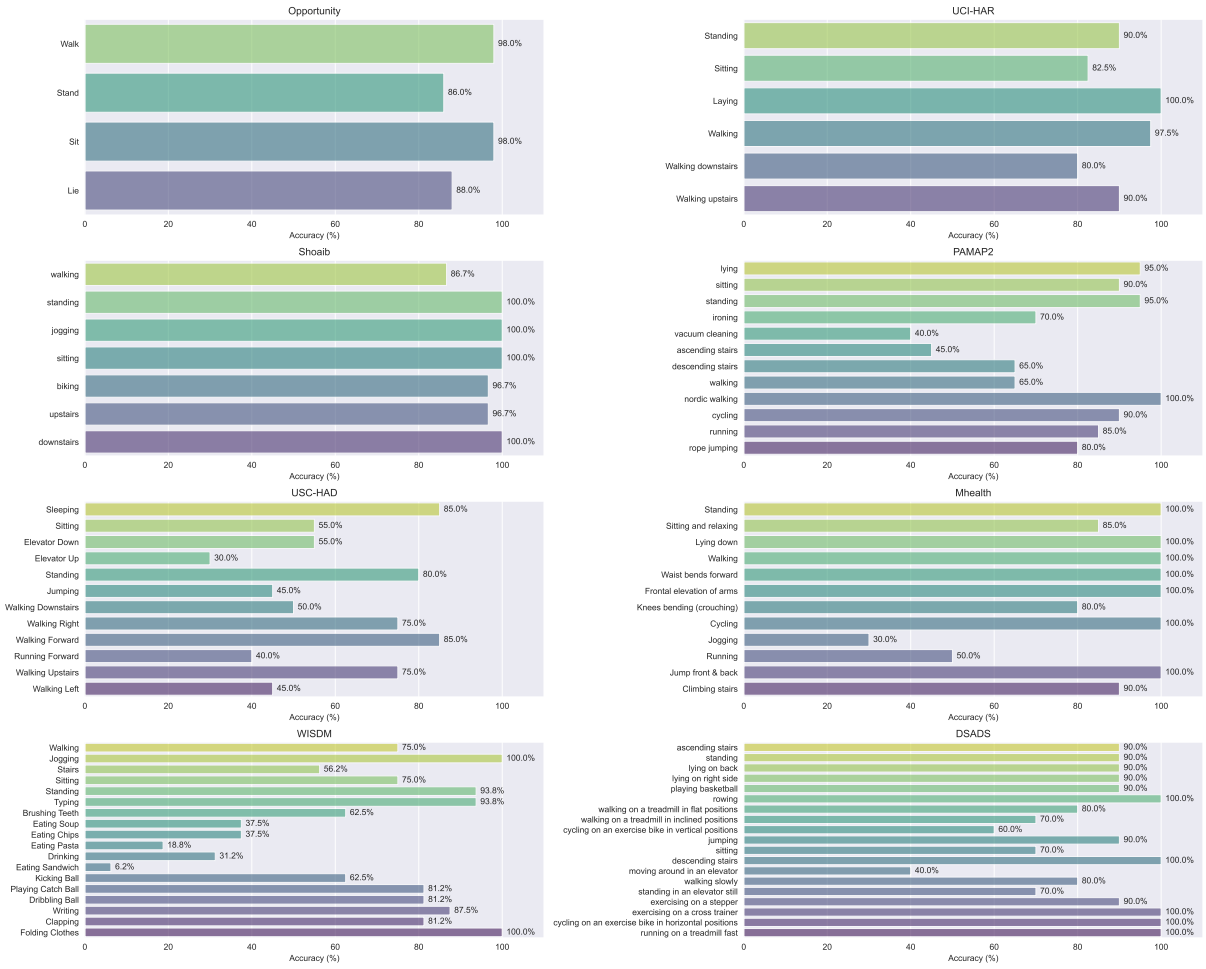


Figure 9: Per-Class Evaluation.

System Instruction

Task
You're an expert in Human Activity Recognition, with a focus on identifying the most effective features for distinguishing between human activities.

Glossary of Abbreviations
`GLOSS_TEXT`

Instructions

- Based on the user-provided "Top Features per Activity Pair", select up to `TOP_N` unique features that best distinguish the specified Target Activities.
- When selecting features, prioritize those that:
  - Appear consistently across multiple activity pairs, or
  - Have relatively high importance scores within specific pairs.

Output Format
Return **only** the following, with no extra text or line breaks:

Index	**Feature Name**

User Prompt

Target Activities
`ACTIVITY_LIST`

Top Features per Activity Pair (Ranked by Importance Score)
(`PAIR_NUM` activity pairs in total)
`PAIR_WISE_KNOWLEDGE`

Figure 10: Prompt template for the first Feature Selector agent.

System Instruction

Task
You're an expert in Human Activity Recognition. Your task is to narrow down the set of plausible activities for the **QUERY** segment, based on the given features and their statistical distributions in the user-supplied activities table.

Instructions

- Each row in the activities table represents an activity class, with cells showing the mean ± std of each feature.
- The QUERY row presents the feature values of the segment to classify.
- You must select at least two activity classes, and ideally all activity classes that are reasonably or even marginally plausible for the QUERY.
- For each selected activity class, provide a brief explanation of why it is a plausible match.

Output Format
Return **only** the following, in this exact order, with no additional text or line breaks:

Index	**Activity**	**Reason**

User Prompt

Activities Table
`ACTIVITIES_FEATURES_TABLE`

Figure 11: Prompt template for Evidence Pruning agent.

System Instruction

Task
You're an expert in Human Activity Recognition, with a focus on identifying the most effective features for distinguishing between human activities.

Glossary of Abbreviations
`GLOSS_TEXT`

Instructions

- Based on the user-provided "Top Features per Activity Pair", select up to `TOP_N` unique features that best distinguish the specified Target Activities.
- When selecting features, prioritize those that:
  - Appear consistently across multiple activity pairs, or
  - Have relatively high importance scores within specific pairs.
- For each selected feature, give:
  - Definition – A concise, clear explanation of the feature.
  - Discriminative Power – Summarize the following:
    * Which activity pairs this feature helps to distinguish.
    * The relative importance rate of this feature within each activity pair, indicating how effectively it differentiates between the two activities in that pair.

Output Format
Return **only** the following, with no extra text or line breaks:

Index	Feature Name	Definition	Discriminative Power

User Prompt

Target Activities
`ACTIVITY_LIST`

Top Features per Activity Pair (Ranked by Importance Score)
(`PAIR_NUM` activity pairs in total)
`PAIR_WISE_KNOWLEDGE`

Figure 12: Prompt template for second Feature Selector agent.

System Instruction

Task
You are an expert in Human Activity Recognition. Your goal is to determine the **most probable activity class** for the **QUERY** segment by comparing its feature values against the statistical distributions in the user-provided activities table.

Sensor Feature Explanation Guide Table
This table describes each feature and indicates which activity classes it helps to distinguish between.
`FEATURES_REFERENCE_TABLE`

Instructions

- Each row in the activities table corresponds to an activity class, with each cell showing the mean ± standard deviation for a feature.
- The QUERY row presents the feature values of the segment to classify.
- Select the single most likely activity class, and base your decision on specific feature(s) in the QUERY row.
- In your explanation:
  - Explicitly compare the Query's feature values to each class's distribution, explaining why the predicted class is a better match than each alternative.
  - When unsure, refer to the Discriminative Power in the guide table to justify how strongly each feature helps distinguish the specific activities.

Output Format
Respond with **exactly one** line in this JSON format (no extra text or line breaks):

```json
{
  "reason": "<your detailed explanation>",
  "predicted_class": "<ClassName>"
}
```

User Prompt

Activities Table
`ACTIVITIES_FEATURES_TABLE`

Figure 13: Prompt template for Decision Insight agent.