MOTION AS COMMON GROUND: ALIGNING HETEROGENEOUS SENSING MODALITIES FOR HUMAN ACTIVITY RECOGNITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Human activity recognition (HAR) using heterogeneous sensors is challenging because modalities such as RGB, depth, LiDAR, mmWave, and WiFi capture the same activity through different physical signals and produce data with distinct structures. Existing methods typically bridge this heterogeneity through modality-invariant learning or LLM-based semantic alignment. We take a different view: *despite different sensing mechanisms, these modalities observe the same underlying human motion*. Since motion is shared across sensors and directly tied to the activity, it provides a natural basis for cross-modal alignment. In this paper, we introduce **UniMotion**, a motion-grounded framework that first establishes a shared temporal–joint motion space using a skeleton-pretrained motion model. UniMotion is then optimized in two stages. In Stage 1, modality-specific Motion Adapters with learnable joint queries map each sensor's native features into joint-indexed tokens in this space, aligning heterogeneous modalities within a common temporal–joint representation. Skeleton sequences are used only during training. In Stage 2, Reliability- and Disagreement-Aware Motion Fusion (RDMF) integrates the aligned modalities for multimodal recognition. RDMF estimates time-varying modality reliability to form a weighted motion consensus and uses cross-modal disagreement to retain complementary evidence beyond the consensus. Evaluated on MM-Fi and XRF55 across seven sensing modalities, UniMotion achieves state-of-the-art average accuracy under random-split, cross-subject, and cross-environment evaluation protocols and remains effective across varying modality combinations. These results support the use of structured human motion as a physically grounded, task-relevant shared representation for multimodal HAR.

1 INTRODUCTION

Human Activity Recognition (HAR) increasingly relies on heterogeneous sensing modalities (Sun et al., 2022; Yang et al., 2023), such as RGB cameras, depth sensors, LiDAR, millimeter-wave (mmWave) radar, and WiFi. Although these modalities provide complementary cues, they observe the same activity through distinct physical processes, yielding representations with substantially different spatial structures, temporal dynamics, and signal statistics (Weng et al., 2024; Chen & Yang, 2025). This heterogeneity makes it difficult to learn a common representation that can bridge sensing modalities while preserving the information needed for activity recognition (Yang et al., 2024; Wu et al., 2024; Bian et al., 2026).

Existing methods mainly bridge sensor heterogeneity through modality-invariant learning or language-grounded alignment. Modality-invariant approaches, such as X-Fi (Chen & Yang, 2025), learn a unified latent representation that can accommodate different sensing modalities and their combinations. More recently, language-grounded approaches project heterogeneous sensor features into pretrained vision-language or LLM semantic spaces (Moon et al., 2023; Li et al., 2025b). HoloLLM (Zhou & Yang, 2026), for example, couples CLIP-aligned sensor representations with an LLM for unified human sensing. Although effective, these approaches primarily define cross-modal commonality in learned latent or semantic spaces, without explicitly grounding the shared representation in the physical process behind the sensor observations. For HAR, that shared pro-

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

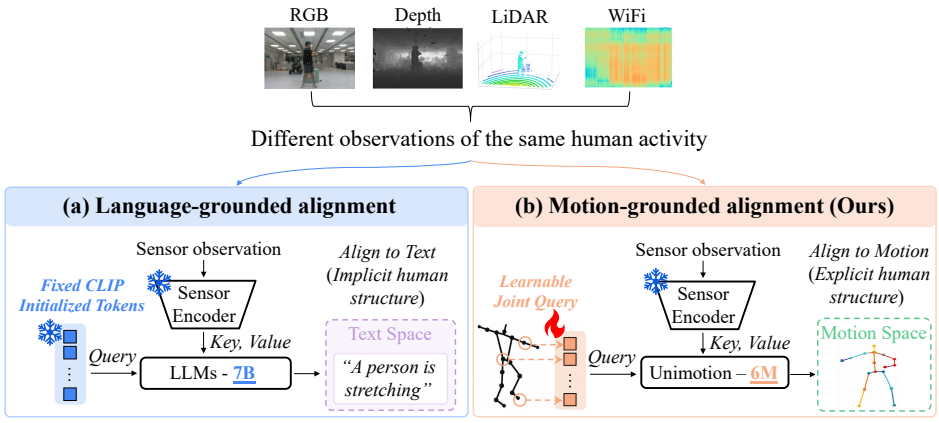

Figure 1: **Motivation for motion-grounded alignment.** **(a)** Language-grounded alignment uses CLIP-initialized queries to aggregate heterogeneous sensor features and map them into an LLM semantic space, where human motion is represented only *implicitly*. **(b)** UniMotion instead uses learnable joint queries to map sensor features into a shared skeleton-derived temporal–joint motion space, making human motion an *explicit* cross-modal reference. By grounding heterogeneous observations in the same underlying human motion, UniMotion provides a physically grounded and task-relevant common representation for HAR.

cess is **human motion**: different sensors capture different manifestations of the same evolving body dynamics. This makes human motion a natural, physically grounded, and task-relevant basis for cross-modal alignment.

This observation motivates a motion-centric perspective on heterogeneous HAR. As illustrated in Fig. 1, compared with language-grounded semantic alignment, our motion-grounded view uses structured human motion as the shared reference across sensors. This provides a physically meaningful common space in which heterogeneous sensor observations can be related through the body dynamics they jointly capture. However, turning this idea into a practical HAR framework raises two key questions: **(1)** how can heterogeneous sensor features be mapped into a common motion space, and **(2)** how can the resulting motion-aligned representations be combined when their reliability and complementary information vary over time?

To address these questions, we introduce **UniMotion**, a motion-grounded framework for heterogeneous-sensor HAR. To align heterogeneous sensor features, UniMotion establishes a shared temporal–joint motion space using a motion model learned from skeleton sequences through motion-aware pretraining and activity supervision. We then design modality-specific Motion Adapters with learnable joint queries to map each sensor's native features into joint-indexed motion tokens in this space, providing a common temporal–joint structure across modalities. Skeleton sequences are used only during training to define the motion anchor.

To integrate the aligned modalities, we introduce Reliability- and Disagreement-Aware Motion Fusion (RDMF). Since different sensors may vary in reliability and provide complementary information over time, RDMF estimates patch-level modality reliability to form a weighted motion consensus and uses cross-modal disagreement to retain information beyond the consensus. Because all modalities are aligned to the same temporal–joint structure, reliability and disagreement can be assessed across corresponding motion tokens.

Our contributions are threefold:

- We introduce a motion-grounded perspective on heterogeneous-sensor HAR, where structured human motion serves as the shared physical reference across sensing modalities.
- We propose **UniMotion**, which defines a shared temporal-joint motion space and aligns heterogeneous sensor features to it via modality-specific motion adapters with learnable joint queries. We further introduce **RDMF**, which combines the aligned representations by modeling time-varying reliability and cross-modal disagreement.

- Extensive experiments on MM-Fi and XRF55, covering seven sensing modalities and three evaluation protocols, show that UniMotion achieves state-of-the-art average accuracy and remains effective across arbitrary combinations of modalities.

2 Related Work

Heterogeneous Sensing for HAR. Recent HAR systems increasingly operate across heterogeneous sensing modalities (Lerch et al., 2025; Liu et al., 2025b), whose differences in spatial structure, temporal resolution, and signal statistics make it challenging to learn a shared representation. Existing methods mainly address this heterogeneity through modality-invariant representation learning, cross-modal interaction, and language-grounded or multimodal foundation models (Chen & Yang, 2025; Zhou & Yang, 2026; Qu et al., 2024; Do & Kim, 2025). For example, X-Fi (Chen & Yang, 2025) learns a modality-invariant latent space for different sensor combinations, while HoloLLM (Zhou & Yang, 2026) aligns multisensory observations through a pretrained semantic space. Related work has also studied representation sharing across structurally different human motion observations. Heterogeneous Skeleton-Based Action Representation Learning (Wang et al., 2025) learns a shared action representation across skeletons with different joint topologies, while recent cross-modal methods align skeleton features with semantic representations for transferable action understanding (Chen et al., 2025b; Zhu et al., 2025). These studies show the value of a common representation across heterogeneous observations, but the shared space is typically either learned from task objectives or inherited from pretrained semantic models. UniMotion instead uses structured human motion itself as the common reference, grounding heterogeneous sensor observations in the physical process they jointly capture.

Motion and Semantic Representation for Action Understanding. Recent research also studies how human activities can be represented beyond sensor-specific features. Vision–language and language-model-based methods map action observations into semantic spaces for knowledge transfer and open-vocabulary reasoning (Qu et al., 2024; Moon et al., 2023; Li et al., 2025b). Recent skeleton–language approaches further connect motion and semantics through semantic prompting (Zhu et al., 2025), evolving skeleton–semantic prototypes (Chen et al., 2025b), and diffusion-based modality alignment (Do & Kim, 2025). A different direction focuses directly on structured human motion. H-MoRe (Huang et al., 2025) learns human-centric motion representations for action analysis, ProtoGCN (Liu et al., 2025a) models fine-grained local motion patterns, and Skeleton Motion Words (Gökay et al., 2025) discovers reusable motion primitives over time. Motion–language models further show that joint structure and temporal dynamics provide a meaningful representation for connecting motion with semantics (Chen et al., 2025a; Hong et al., 2025). Together, these studies show the value of both semantic and motion representations for activity understanding. UniMotion takes a different direction: rather than treating motion as another modality to be aligned with semantics, it uses structured human motion itself as the common temporal–joint space for aligning heterogeneous sensor observations.

3 Problem Formulation

Let $\mathcal{M}$ denote the set of supported sensing modalities. For each training sample, modality $m \in \mathcal{M}$ provides an observation X^m with activity label $y \in \{1, \ldots, C\}$, while a synchronized skeleton sequence X^s is available only during training to define the motion reference. Our goal is to map heterogeneous sensor observations into a shared temporal–joint motion space. Specifically, the skeleton sequence defines motion tokens $\mathcal{Z}^s \in \mathbb{R}^{B\times K\times J\times d}$, where B is the batch size, K the number of temporal patches, J the number of joints, and d the token dimension. For each sensing modality, we learn a corresponding representation

$$\mathcal{Z}^m \in \mathbb{R}^{B\times K\times J\times d}, \qquad m \in \mathcal{M}, \tag{1}$$

such that $\mathcal{Z}^m$ follows the same temporal–joint organization as $\mathcal{Z}^s$ while retaining information needed for activity recognition.

At inference time, skeletons are not required. Given any available modality subset $\mathcal{M}' \subseteq \mathcal{M}$, UniMotion predicts

$$\hat{y} = f(\{\mathcal{Z}^m\}_{m\in\mathcal{M}'}), \tag{2}$$

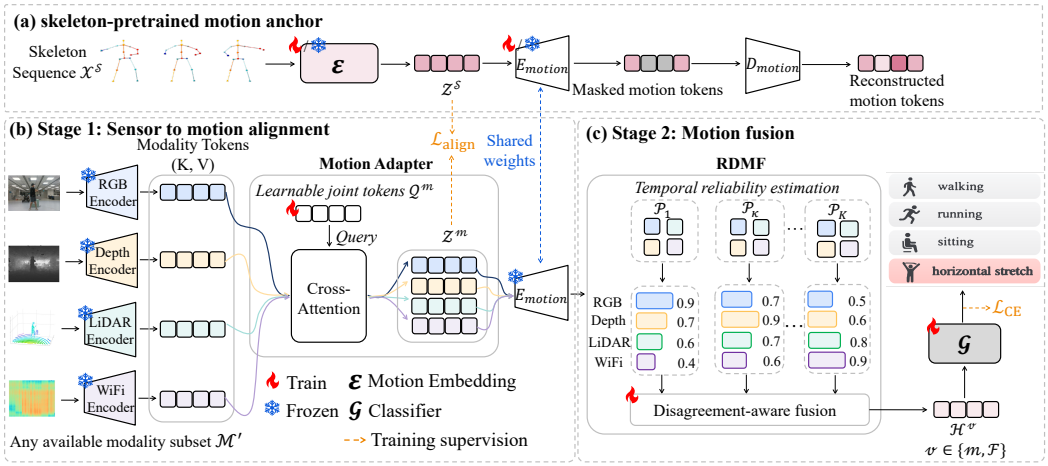

Figure 2: **Overview of UniMotion. (a)** A skeleton motion model is pretrained through masked motion prediction and fine-tuned for HAR to define the shared temporal–joint motion space; the resulting motion model is then frozen. **(b)** In Stage 1, each modality-specific Motion Adapter uses learnable joint queries to map heterogeneous sensor features into this shared space and aligns them with the skeleton-derived motion anchor. **(c)** In Stage 2, the aligned modality branches are frozen, and RDMF integrates their motion representations through temporal reliability estimation and disagreement-aware fusion for activity recognition.

where $\hat{y}$ is the predicted activity label. The same formulation supports both multimodal inputs and the single-modality case $|\mathcal{M}'| = 1$.

4 Method

Figure 2 presents the overall UniMotion framework. After establishing and freezing the skeleton-derived motion anchor, UniMotion is optimized in two stages. In Stage 1, each sensing modality is independently aligned with the shared temporal–joint motion space through a modality-specific Motion Adapter with learnable joint queries. In Stage 2, the aligned modality representations are integrated by RDMF, which models their time-varying reliability and cross-modal disagreement for multimodal activity recognition. At inference time, the skeleton branch is removed, and UniMotion operates on any available subset of sensing modalities.

4.1 Establishing the Motion Anchor from Skeletons

Given a skeleton sequence X^s, the motion embedding $\mathcal{E}_s$ groups every p_t consecutive frames into a structured temporal–joint representation:

$$\mathcal{Z}^s = \mathcal{E}_s(X^s) \in \mathbb{R}^{B\times K\times J\times d}, \tag{3}$$

where each token corresponds to a temporal patch and a human joint. This representation defines the motion space used for subsequent sensor alignment.

We pretrain the motion model following MAMP (Mao et al., 2023), which learns contextual motion representations through masked motion prediction. Given temporal stride δ, the motion target is the signed joint displacement

$$\Delta X_t^s = X_{t+\delta}^s - X_t^s. \tag{4}$$

Following MAMP, temporal–joint regions with larger motion are assigned higher masking probabilities. Let Ω denote a set of masked tokens. The motion encoder E_{motion} processes the visible skeleton tokens, and the decoder reconstructs the corresponding masked motion targets using

$$\mathcal{L}_{\text{motion}} = \frac{1}{|\Omega|}\sum_{i\in\Omega}\left\|\Delta\hat{X}_i^s - \Delta X_i^s\right\|_2^2, \tag{5}$$

where $\Delta\hat{X}_i^s$ denotes the predicted displacement at a masked token i.

After motion-aware pretraining, we fine-tune the model for activity recognition so that the learned representation retains motion patterns relevant to HAR. We then freeze the motion embedding $\mathcal{E}_s$ and motion encoder E_{motion} and discard the pretraining decoder. The resulting $\mathcal{Z}^s$ serves as the skeleton-derived motion anchor for aligning all sensing modalities, while the motion embedding and motion encoder remain frozen.

4.2 Motion Adapter: Aligning Sensors to the Motion Space

Sensor encoders produce features with modality-dependent structures, whereas the motion anchor $\mathcal{Z}^s$ is organized by temporal patches and human joints. We therefore design a modality-specific Motion Adapter that maps each sensor's native features into the same temporal–joint structure using learnable joint queries.

Modality feature preparation. For modality m, a frozen sensor encoder E^m extracts

$$F^m = E^m(X^m) \in \mathbb{R}^{B\times T\times S_m\times C_m}, \tag{6}$$

where S_m and C_m denote the number and dimension of modality-native feature tokens, respectively. We project these features into a common hidden dimension d_h while preserving their temporal order and native token positions:

$$\begin{aligned}\widetilde{F}^m_{t,n} &= f^m_{\text{in}}(F^m_{t,n}) + e^{m,\text{temp}}_{\ell(t)} + e^{m,\text{pos}}_{n},\\ V^m_k &= \text{Flatten}\big(\widetilde{F}^m_{\mathcal{I}_k,:}\big) \in \mathbb{R}^{B\times(p_t S_m)\times d_h},\end{aligned} \tag{7}$$

where $\mathcal{I}_k$ contains the p_t frames of a temporal patch k, $\ell(t)$ denotes the relative frame position within the patch, and n indexes the modality-native feature tokens. Thus, V^m_k collects the temporally and positionally encoded sensor evidence within patch k.

Joint-query adaptation. Although V^m_k retains the native sensor information, it has no explicit correspondence to human joints. We introduce J modality-specific learnable joint queries $Q^m = [q^m_1,\dots,q^m_J]^\top \in \mathbb{R}^{J\times d_h}$, shared across temporal patches to provide consistent joint indexing over time. Each query attends to all sensor tokens within the corresponding patch, producing

$$\mathcal{Z}^m_k = f^m_{\text{out}}\left(Q^m + \text{MHA}\left(Q^m, V^m_k, V^m_k\right)\right) \in \mathbb{R}^{B\times J\times d}. \tag{8}$$

Here, Q^m provides the joint-indexed queries, while V^m_k serves as the keys and values containing the sensor evidence in patch k. Applying the adapter across all K patches yields $\mathcal{Z}^m \in \mathbb{R}^{B\times K\times J\times d}$, which follows the same temporal–joint organization as the skeleton anchor $\mathcal{Z}^s$. This makes token-wise sensor-to-motion alignment well-defined and allows all modalities to be processed by the same frozen motion encoder E_{motion}.

4.3 Reliability- and Disagreement-Aware Motion Fusion

After sensor-to-motion alignment, we freeze the sensor encoders and Motion Adapters, and then learn RDMF on the aligned representations. For each available modality $m \in \mathcal{M}'$, the shared frozen motion encoder produces

$$H^m = E_{\text{motion}}(\mathcal{Z}^m) \in \mathbb{R}^{B\times K\times J\times d}. \tag{9}$$

Because all H^m follow the same temporal–joint indexing, their corresponding tokens describe comparable motion evidence. RDMF first estimates patch-level modality reliability to form a motion consensus, and then uses cross-modal disagreement to condition a residual correction.

Temporal reliability estimation. For each modality, we summarize the joint tokens within temporal patch k and predict its reliability using a shared scoring network q:

$$\begin{aligned} r^m_{b,k} &= \frac{1}{J}\sum_{j=1}^{J} H^m_{b,k,j}, \qquad s^m_{b,k} = q\big(\text{LN}(r^m_{b,k}) + e_m\big),\\ w^m_{b,k} &= (1-\rho)\frac{\exp(s^m_{b,k})}{\sum_{n\in\mathcal{M}'}\exp(s^n_{b,k})} + \frac{\rho}{|\mathcal{M}'|},\end{aligned} \tag{10}$$

where e_m is a learned modality embedding and $\rho \in [0, 1)$ provides a small uniform contribution to every available modality. The weights are normalized across $\mathcal{M}'$ for each temporal patch and shared across its J joint tokens.

Disagreement-aware correction. Using these reliability weights, we compute the motion consensus and its weighted cross-modal deviation as

$$\begin{aligned} \overline{H}_{b,k,j} &= \sum_{m \in \mathcal{M}'} w^m_{b,k} H^m_{b,k,j}, \\ D_{b,k,j} &= \sum_{m \in \mathcal{M}'} w^m_{b,k} \left| H^m_{b,k,j} - \overline{H}_{b,k,j} \right|. \end{aligned} \tag{11}$$

Here, $\overline{H}$ captures motion evidence supported by reliable modalities, while D identifies temporal–joint locations where their observations disagree. Rather than discarding this information during aggregation, RDMF uses it to condition a residual correction:

$$H^F_{b,k,j} = \overline{H}_{b,k,j} + \sigma(a)\, c\big([\mathrm{LN}(\overline{H}_{b,k,j}); \mathrm{LN}(D_{b,k,j})]\big), \tag{12}$$

where $c(\cdot)$ is a lightweight complement network and a is a learnable scalar controlling the correction magnitude. The final H^F therefore combines reliability-weighted shared motion evidence with disagreement-conditioned complementary information for activity recognition.

4.4 Two-Stage Optimization

After fixing the skeleton-derived motion anchor, UniMotion is optimized in two stages: Stage 1 independently aligns each sensing modality with the shared motion space, while Stage 2 freezes the aligned modality branches and learns RDMF for multimodal fusion.

Stage 1: sensor-to-motion alignment. For each modality m, we align the Motion Adapter output $\mathcal{Z}^m$ with the corresponding skeleton representation $\mathcal{Z}^s$ using a token-wise combination of cosine and SmoothL1 losses:

$$\mathcal{L}_{\text{align}} = \mathcal{L}_{\cos} + \mathcal{L}_{\text{SL1}}. \tag{13}$$

The cosine term aligns the direction of corresponding motion tokens, while the SmoothL1 term constrains their coordinate-level discrepancy. The complete definition is provided in Appendix A.4.

Together with activity supervision on $H^m = E_{\text{motion}}(\mathcal{Z}^m)$, the Stage 1 objective is

$$\mathcal{L}^m_{\text{stage1}} = \lambda_{\text{cls}}\, \mathrm{CE}(g^m(H^m), y) + \lambda_{\text{motion}} \mathcal{L}_{\text{align}}(\mathcal{Z}^m, \mathrm{sg}(\mathcal{Z}^s)), \tag{14}$$

where $\mathrm{sg}(\cdot)$ stops gradients through the skeleton anchor. Only the Motion Adapter and classifier g^m are updated, while the sensor encoder and motion model remain frozen.

Stage 2: multimodal fusion. After all modality branches are aligned, we freeze them and optimize RDMF. To directly supervise its reliability weights in the shared motion space, we reuse the patch-level weights to construct an auxiliary fused motion representation:

$$\mathcal{Z}^F_{b,k,j} = \sum_{m \in \mathcal{M}'} w^m_{b,k} \mathcal{Z}^m_{b,k,j}, \qquad \sum_{m \in \mathcal{M}'} w^m_{b,k} = 1. \tag{15}$$

Here, $\mathcal{Z}^F$ supervises the reliability-weighted motion consensus, while activity classification is performed on the disagreement-refined representation H^F. The Stage 2 objective is

$$\mathcal{L}_{\text{stage2}} = \lambda_{\text{cls}}\, \mathrm{CE}\big(g^F(H^F), y\big) + \lambda_{\text{motion}} \mathcal{L}_{\text{align}}\big(\mathcal{Z}^F, \mathrm{sg}(\mathcal{Z}^s)\big). \tag{16}$$

Only RDMF and the fusion classifier g^F are updated in this stage.

At inference time, a single available modality is classified using its Stage 1 classifier g^m, while multimodal inputs are integrated by RDMF and classified using the fusion classifier g^F.

5 Experiment

5.1 Datasets and Implementation Details

We evaluate UniMotion on MM-Fi (Yang et al., 2023) and XRF55 (Sun et al., 2022), following the benchmark settings of X-Fi (Chen & Yang, 2025) and HoloLLM (Zhou & Yang, 2026). MM-Fi provides synchronized RGB, depth, LiDAR, mmWave, WiFi-CSI, and 17-joint 3D skeletons. XRF55

Table 1: Comparison with existing methods under Random, Cross-Subject, and Cross-Environment settings on MM-Fi and XRF55. RGB Video (V), Depth (D), mmWave (M), Infrared (I), RFID (R), LiDAR (L), and WiFi (W) denote the evaluated sensing modalities. Each modality column reports single-modality classification accuracy (%), and Avg. denotes the mean across the five modalities of each dataset. Best results are shown in **bold**, and second-best results are underlined.

Setting	Method	MM-Fi						XRF55					
		V	D	M	L	W	Avg.	V	D	I	R	W	Avg.
Random	TokenPacker (Li et al., 2025a)	15.6	14.5	32.8	18.9	11.5	18.7	9.2	8.2	7.1	2.2	6.5	6.6
	Honeybee (Cha et al., 2024)	9.0	11.3	27.8	12.6	10.3	14.2	2.9	4.3	3.8	1.9	7.4	4.0
	OneLLM (Han et al., 2024)	9.0	9.0	16.2	6.1	9.0	9.9	2.1	2.0	1.5	1.9	2.3	2.0
	ImageBind (Girdhar et al., 2023)	99.5	96.8	63.7	13.2	11.7	57.0	90.8	59.1	93.3	14.6	51.1	61.8
	HoloLLM Zhou & Yang (2026)	99.8	**99.7**	97.3	92.8	83.5	94.6	**98.0**	**96.9**	97.5	39.1	37.7	73.8
	Unimotion (Ours)	**99.8**	99.5	**98.4**	**99.2**	**90.4**	**97.5**	97.6	95.5	**97.5**	**55.5**	**95.5**	**88.3**
Cross-Subject	TokenPacker (Li et al., 2025a)	24.1	14.9	33.9	12.9	**11.1**	19.4	8.6	8.9	6.3	2.8	4.8	6.3
	Honeybee (Cha et al., 2024)	11.6	9.3	27.2	13.9	8.3	14.1	3.0	4.2	4.3	2.4	4.9	3.8
	OneLLM (Han et al., 2024)	8.3	8.3	18.9	6.6	8.3	10.1	2.0	1.8	1.8	2.2	2.6	2.1
	ImageBind (Girdhar et al., 2023)	85.7	70.7	56.9	17.7	11.0	48.4	18.6	14.4	22.0	4.2	5.7	13.0
	HoloLLM (Zhou & Yang, 2026)	**98.4**	**99.0**	93.8	76.6	9.3	75.4	46.4	47.8	**42.2**	4.4	4.5	29.1
	Unimotion (Ours)	95.9	97.3	**95.4**	**93.6**	8.8	**78.2**	**52.0**	**48.3**	41.4	**4.9**	**17.9**	**32.9**
Cross-Environment	TokenPacker (Li et al., 2025a)	15.0	9.4	20.2	11.7	10.7	13.4	9.8	10.1	6.2	2.0	4.5	6.5
	Honeybee (Cha et al., 2024)	13.5	12.9	26.7	12.8	**11.2**	15.4	8.9	4.9	3.1	1.4	6.7	5.0
	OneLLM (Han et al., 2024)	9.5	9.5	5.6	6.8	2.4	6.8	1.5	3.5	1.8	1.9	2.3	2.2
	ImageBind (Girdhar et al., 2023)	45.7	8.0	26.8	8.6	9.5	19.7	6.7	6.8	21.1	2.5	2.5	7.9
	HoloLLM (Zhou & Yang, 2026)	**80.6**	93.2	61.0	53.0	9.5	59.5	**28.9**	12.4	28.3	1.7	2.5	14.8
	Unimotion (Ours)	66.9	**99.9**	**71.6**	**63.9**	10.7	**62.6**	28.5	**13.9**	**35.8**	**37.4**	**41.8**	**31.5**

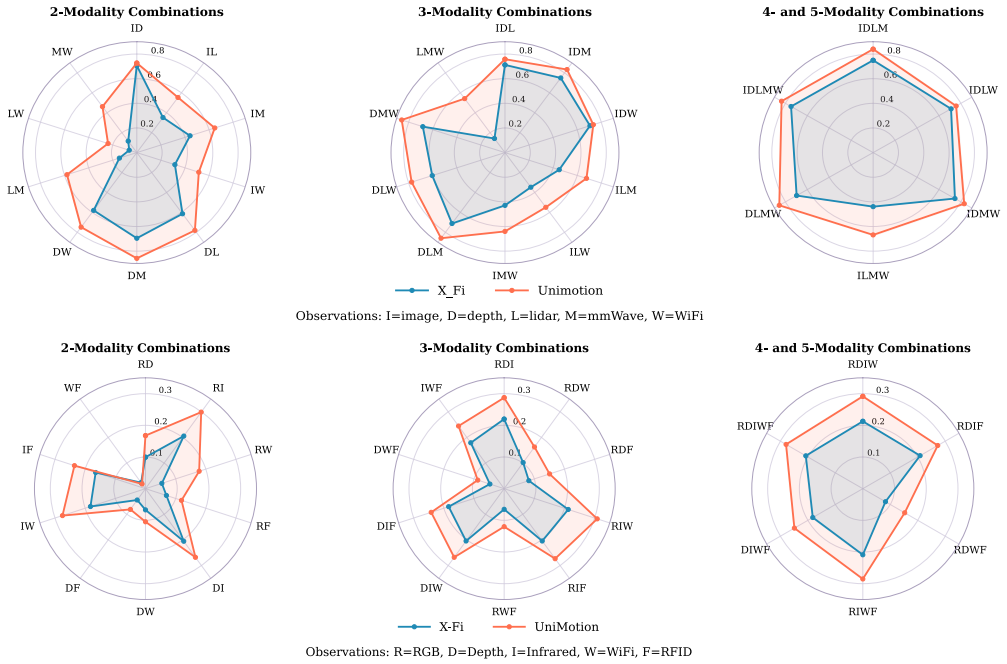

Figure 3: Performance across variable modality combinations on MM-Fi (top) and XRF55 (bottom). The columns report results for two-, three-, and four/five-modality configurations, respectively. UniMotion is compared with X-Fi under identical sensor combinations.

provides RGB, depth, infrared, RFID, and WiFi; we extract 17-joint 2D pose tracks from synchronized RGB videos to construct its motion anchor. Following HoloLLM, we use its 19.8K-sample XRF55 subset, along with the same annotations and evaluation settings. The two benchmarks cover seven distinct sensing modalities in total. Dataset and protocol details are provided in Appendix A.1.

We instantiate the motion anchor with MAMP (Mao et al., 2023) using 17-joint skeleton sequences, with 3D coordinates for MM-Fi and 2D coordinates for XRF55. After motion pretraining and activity fine-tuning, the skeleton embedding and motion encoder are frozen. Each sensing modality uses a frozen HoloLLM encoder and a modality-specific Motion Adapter with 17 learnable joint

Table 2: Ablation of the motion alignment loss $\mathcal{L}_{\text{align}}$ on MM-Fi and XRF55. All other components and training settings are kept unchanged. Best results are shown in **bold**.

Setting	Configuration	MM-Fi						XRF55					
		V	D	M	L	W	Avg.	V	D	I	R	W	Avg.
Random	w/o $\mathcal{L}_{\text{align}}$	98.7	99.3	97.2	98.6	88.7	96.5	97.2	95.1	96.6	54.7	94.2	87.6
	w/ $\mathcal{L}_{\text{align}}$	**99.8**	**99.5**	**98.4**	**99.2**	**90.4**	**97.5**	**97.6**	**95.5**	**97.5**	**55.5**	**95.5**	**88.3**
Cross-Subject	w/o $\mathcal{L}_{\text{align}}$	94.9	96.7	93.8	91.4	7.7	76.9	50.8	47.1	39.5	4.5	17.3	31.8
	w/ $\mathcal{L}_{\text{align}}$	**95.9**	**97.3**	**95.4**	**93.6**	**8.8**	**78.2**	**52.0**	**48.3**	**41.4**	**4.9**	**17.9**	**32.9**
Cross-Environment	w/o $\mathcal{L}_{\text{align}}$	59.6	87.4	67.2	55.8	8.9	55.8	24.8	11.7	33.6	36.1	38.3	28.9
	w/ $\mathcal{L}_{\text{align}}$	**66.9**	**99.9**	**71.6**	**63.9**	**10.7**	**62.6**	**28.5**	**13.9**	**35.8**	**37.4**	**41.8**	**31.5**

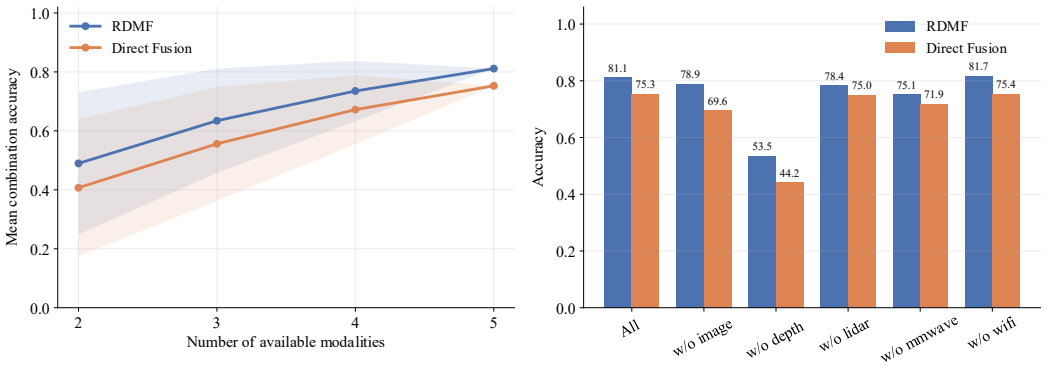

Figure 4: Comparison between RDMF and uniform mean fusion under the cross-environment setting on MM-Fi. Left: mean accuracy over modality combinations with different numbers of available modalities; shaded regions denote the standard deviation across combinations. Right: accuracy using all modalities and under leave-one-modality-out evaluation.

queries. RDMF is then trained over the aligned motion representations while all sensor branches remain frozen, its computational efficiency is evaluated in Tabel 5. Further details on preprocessing, architecture, and optimization are provided in Appendices A.2–A.5.

5.2 Comparison with State-of-the-Art Methods

Overall performance. Table 1 compares UniMotion against existing multimodal and language-grounded methods across the Random, Cross-Subject, and Cross-Environment protocols. UniMotion achieves the highest average accuracy under all three settings on both datasets. On MM-Fi, it obtains 97.5%, 78.2%, and 62.6%, outperforming HoloLLM (Zhou & Yang, 2026) by 2.9, 2.8, and 3.1 percentage points, respectively. On XRF55, the corresponding gains are 14.5, 3.8, and 16.7 points. The improvements are particularly clear for several RF and geometric modalities, supporting motion grounding as an effective common representation across heterogeneous sensors.

Arbitrary modality combinations. Figure 3 further compares UniMotion with X-Fi (Chen & Yang, 2025) under identical sensor combinations. UniMotion performs better across nearly all two-, three-, and four/five-modality configurations on both datasets, with larger gains when only a small number of modalities are available. Its performance remains stable as additional modalities are introduced, indicating that the shared motion space and RDMF can support variable sensor subsets without retraining for each modality combination.

5.3 Ablation Study and Analysis

Effect of motion alignment supervision. Table 2 isolates the contribution of $\mathcal{L}_{\text{align}}$. Adding skeleton-guided motion alignment consistently improves performance across modalities, datasets, and evaluation protocols. The gains are relatively small under the near-saturated random split but become more pronounced under cross-subject and cross-environment shifts. These results suggest that aligning sensor representations with the skeleton-derived motion space provides a useful training

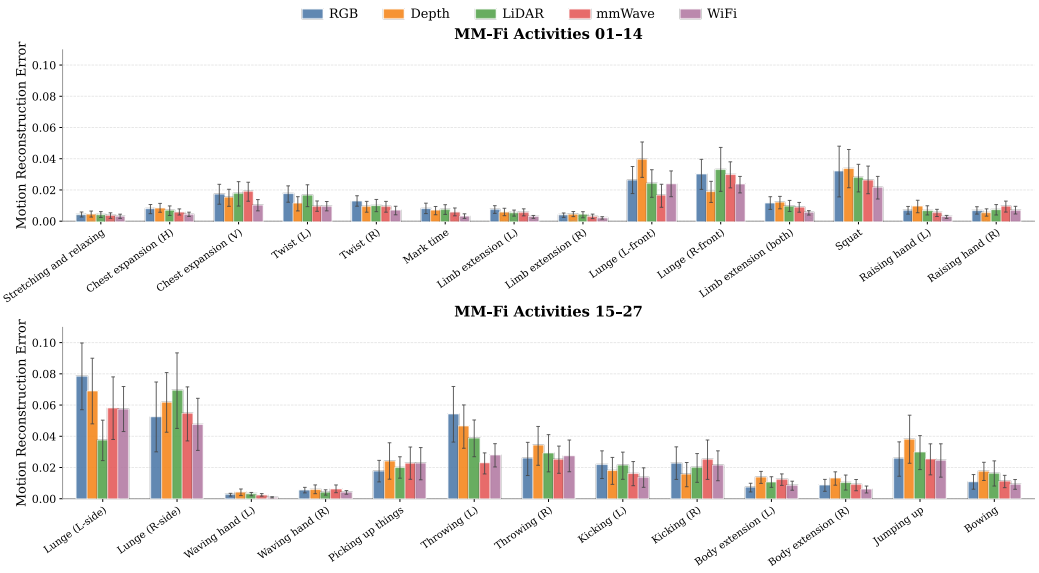

Figure 5: Activity-wise motion reconstruction error across sensing modalities on MM-Fi. Bars report the mean discrepancy between modality-derived and skeleton-derived motion representations, and error bars indicate variation across test samples. Lower values indicate better motion alignment.

constraint, particularly when sensor-specific appearance or signal statistics do not transfer reliably across domains.

Effect of adaptive fusion. We compare RDMF with uniform mean fusion under identical pretrained modality branches, classifier initialization, and training protocol. As shown in Figure 4, RDMF consistently outperforms direct fusion across all modality cardinalities, improving accuracy from 75.3% to 81.1% with all five modalities. The advantage holds under missing modalities, most notably when removing Image (78.9% vs. 69.6%) or Depth (53.5% vs. 44.2%), and persists for LiDAR, mmWave, and WiFi ($+3.4$, $+3.2$, and $+6.3$ points). These results suggest uniform averaging fails to account for modality-dependent reliability, whereas RDMF adaptively preserves informative evidence while suppressing unreliable contributions.

Activity-wise motion alignment. Figure 5 examines the alignment error between sensor-derived and skeleton-derived motion representations across activities on MM-Fi. Spatially constrained or lower-amplitude actions, such as stretching, marking time, hand waving, and limb extension, generally yield lower error rates, whereas activities involving greater whole-body displacement, including side lunges, throwing, squatting, and jumping, exhibit higher errors and greater variance. Despite their different sensing mechanisms, the modalities show broadly consistent activity-wise error patterns. This suggests that the remaining alignment difficulty is strongly related to the underlying motion characteristics rather than a particular sensing modality.

6 Conclusion

We present **UniMotion**, a motion-grounded framework for heterogeneous-sensor HAR. By aligning sensor-specific features to a shared temporal–joint motion space, UniMotion provides a common physical reference across sensing modalities. We further designed **RDMF** to combine the aligned representations by modeling time-varying modality reliability and cross-modal disagreement. Experiments on MM-Fi and XRF55 show consistent gains over the state-of-the-art methods across random, cross-subject, and cross-environment settings, as well as arbitrary modality combinations. These results indicate that structured human motion can serve as an effective common representation for heterogeneous sensing, offering an alternative to latent or semantic alignment in multimodal HAR.

AI USE STATEMENT

Generative AI tools were used to assist with language editing, proofreading, improving readability, and refining the structure and presentation of the manuscript, including selected descriptions of the methodology and experimental results. Generative AI was not used to generate datasets, implement the proposed method, produce experimental results, formulate mathematical claims, or prove theoretical results. All AI-assisted content was reviewed and verified by the authors. The authors take full responsibility for the final content of the paper.

REPRODUCIBILITY STATEMENT

We provide the model architecture, training objectives, and optimization procedure in the main paper. Detailed dataset splits, evaluation protocols, sensor preprocessing, architectural settings, and training schedules are reported in the appendix. We follow the released HoloLLM annotations for MM-Fi and XRF55 to ensure direct comparison with prior work. Code and configuration files will be released upon publication.

REFERENCES

Sizhen Bian, Mengxi Liu, Lala Shakti Swarup Ray, Bo Zhou, Bin Guo, Zhiwen Yu, Thomas Plötz, Paul Lukowicz, Siyu Yuan, and Vitor Fortes Rey. Foundation models defining a new era in sensor-based human activity recognition: A survey and outlook. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 10(2):1–38, 2026.

Junbum Cha, Wooyoung Kang, Jonghwan Mun, and Byungseok Roh. Honeybee: Locality-enhanced projector for multimodal llm. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13817–13827. IEEE, 2024.

Changan Chen, Juze Zhang, Shrinidhi K Lakshmikanth, Yusu Fang, Ruizhi Shao, Gordon Wetzstein, Li Fei-Fei, and Ehsan Adeli. The language of motion: Unifying verbal and non-verbal language of 3d human motion. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6200–6211. IEEE, 2025a.

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

Yang Chen, Jingcai Guo, Song Guo, and Dacheng Tao. Neuron: Learning context-aware evolving representations for zero-shot skeleton action recognition. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8721–8730. IEEE, 2025b.

Jeonghyeok Do and Munchurl Kim. Bridging the skeleton-text modality gap: diffusion-powered modality alignment for zero-shot skeleton-based action recognition. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12757–12768. IEEE, 2025.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind one embedding space to bind them all. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15180–15190. IEEE, 2023.

Uzay Gökay, Federico Spurio, Dominik R Bach, and Juergen Gall. Skeleton motion words for unsupervised skeleton-based temporal action segmentation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12101–12111. IEEE, 2025.

Jiaming Han, Kaixiong Gong, Yiyuan Zhang, Jiaqi Wang, Kaipeng Zhang, Dahua Lin, Yu Qiao, Peng Gao, and Xiangyu Yue. Onellm: One framework to align all modalities with language. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26574–26585. IEEE, 2024.

Seokhyeon Hong, Chaelin Kim, Serin Yoon, Junghyun Nam, Sihun Cha, and Junyong Noh. Salad: Skeleton-aware latent diffusion for text-driven motion generation and editing. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7158–7168. IEEE, 2025.

Zhanbo Huang, Xiaoming Liu, and Yu Kong. H-more: Learning human-centric motion representation for action analysis. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22702–22713. IEEE, 2025.

David Lerch, Bastian Rothenburger, Zeyun Zhong, Manuel Martin, Frederik Diederichs, and Rainer Stiefelhagen. Learning robust aligned representations across multiple visual modalities in human action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2700–2710, 2025.

Wentong Li, Yuqian Yuan, Jian Liu, Dongqi Tang, Song Wang, Jie Qin, Jianke Zhu, and Lei Zhang. Tokenpacker: Efficient visual projector for multimodal llm. *International Journal of Computer Vision*, 133(10):6794–6812, 2025a.

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*, pp. 354–379, 2025b.

Hongda Liu, Yunfan Liu, Min Ren, Hao Wang, Yunlong Wang, and Zhenan Sun. Revealing key details to see differences: A novel prototypical perspective for skeleton-based action recognition. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29248–29257. IEEE, 2025a.

Xiao Liu, Guan Yuan, Yanmei Zhang, Shang Liu, and Qiuyan Yan. Robusthar: Multi-scale spatial-temporal masked self-supervised pre-training for robust human activity recognition. In *IJCAI*, pp. 8563–8571, 2025b.

Yunyao Mao, Jiajun Deng, Wengang Zhou, Yao Fang, Wanli Ouyang, and Houqiang Li. Masked motion predictors are strong 3d action representation learners. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10147–10157. IEEE, 2023.

Seungwhan Moon, Andrea Madotto, Zhaojiang Lin, Aparajita Saraf, Amy Bearman, and Babak Damavandi. Imu2clip: language-grounded motion sensor translation with multimodal contrastive learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 13246–13253, 2023.

Haoxuan Qu, Yujun Cai, and Jun Liu. Llms are good action recognizers. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18395–18406. IEEE, 2024.

Zehua Sun, Qiuhong Ke, Hossein Rahmani, Mohammed Bennamoun, Gang Wang, and Jun Liu. Human action recognition from various data modalities: A review. *IEEE transactions on pattern analysis and machine intelligence*, 45(3):3200–3225, 2022.

Hongsong Wang, Xiaoyan Ma, Jidong Kuang, and Jie Gui. Heterogeneous skeleton-based action representation learning. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19154–19164. IEEE, 2025.

Yuxuan Weng, Guoquan Wu, Tianyue Zheng, Yanbing Yang, and Jun Luo. Large model for small data: Foundation model for cross-modal rf human activity recognition. In *Proceedings of the 22nd ACM Conference on Embedded Networked Sensor Systems*, pp. 436–449, 2024.

Tong Wu, Yanbin Liu, and Sira Yongchareon. Class-aware sample weight learning for cross-modal unsupervised domain adaptation in cross-user wearable human activity recognition. In *ECAI 2024: 27th European Conference on Artificial Intelligence, 19–24 October 2024, Santiago de Compostela, Spain–Including 13th Conference on Prestigious Applications of Intelligent Systems (PAIS 2024)*, pp. 569–576. SAGE Publications Pvt. Ltd 1 Oliver's Yard, 55 City Road, London, EC1Y 1SP, 2024.

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

Xiaoshan Yang, Baochen Xiong, Yi Huang, and Changsheng Xu. Cross-modal federated human activity recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8): 5345–5361, 2024.

Chuhao Zhou and Jianfei Yang. Hololllm: Multisensory foundation model for language-grounded human sensing and reasoning. *Advances in Neural Information Processing Systems*, 38:73543–73568, 2026.

Anqi Zhu, Jingmin Zhu, James Bailey, Mingming Gong, and Qiuhong Ke. Semantic-guided cross-modal prompt learning for skeleton-based zero-shot action recognition. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13876–13885. IEEE, 2025.

A Supplementary Details

A.1 Datasets and Evaluation Protocols

Datasets. MM-Fi (Yang et al., 2023) contains more than 320K synchronized frames from 40 subjects performing 27 activities across four environments. Each frame includes RGB, depth, LiDAR, mmWave, WiFi-CSI, and 17-joint 3D skeleton annotations. XRF55 (Sun et al., 2022) contains 42.9K recordings of 55 activities performed by 39 subjects across four scenes, with synchronized RGB, depth, infrared, RFID, and WiFi observations. Following HoloLLM (Zhou & Yang, 2026), we use its 19.8K-sample XRF55 benchmark subset. For XRF55, we extract 17-joint 2D pose tracks from the synchronized RGB videos and use them only during training to construct the motion anchor.

Evaluation protocols. We use the annotation files released with HoloLLM without regenerating the data splits. The **Random** protocol uses a 3:1 random partition of the annotated samples. **Cross-Subject** uses non-overlapping subjects for training and testing, while **Cross-Environment** evaluates generalization to unseen environments. On MM-Fi, Cross-Environment trains on E01–E03 and evaluates on E04. For XRF55, we retain the scene assignments specified by the released annotations. The resulting sample statistics are reported in Table 3. All recognition results are reported using classification accuracy (%).

Table 3: Dataset statistics under the evaluation protocols adopted from HoloLLM (Zhou & Yang, 2026).

Protocol	MM-Fi		XRF55	
	Train	Test	Train	Test
Random	12,336	4,112	14,850	4,950
Cross-Subject	11,657	4,791	14,300	5,500
Cross-Environment	12,565	3,883	16,500	3,300

The protocol matches HoloLLM at the sample-identity level, but does not reuse its CLIP or language-model preprocessing.

A.2 Temporal Sampling and Sensor Processing

Temporal sampling. Each MM-Fi annotation specifies an inclusive frame interval. We divide this interval into five temporal bins and sample one synchronized frame from each bin, using the same indices for the sensor input and its 3D skeleton target. XRF55 recordings are five seconds long, from which ten frames are sampled following the HoloLLM benchmark and paired with their corresponding 2D pose tracks. The resulting sequence lengths are therefore $T = 5$ for MM-Fi and $T = 10$ for XRF55. The temporal patch size is $p_t = 1$ for both datasets.

Sensor preprocessing. RGB, depth, and infrared frames are resized to 224×224. RGB is normalized using the CLIP image statistics associated with the visual encoder. Kinect depth values are converted from millimetres to metres, while infrared observations are retained as single-channel

inputs. MM-Fi LiDAR and mmWave frames are padded or truncated to 1,536 and 64 points, respectively. WiFi and RFID retain their native structured arrays. For XRF55, missing visual recordings are removed before constructing the data loader, and multimodal evaluation uses the intersection of valid annotation identities across the requested modalities.

Skeleton sequences are represented using 17 joints and centered per sequence. MM-Fi uses the provided 3D coordinates, whereas XRF55 uses the extracted 2D pose coordinates.

A.3 Architecture Details

Motion model. The skeleton motion model follows MAMP (Mao et al., 2023). Its Transformer contains eight blocks and eight attention heads with token dimension $d = 256$. The motion embedding uses 17 joints and a temporal patch size $p_t = 1$. After motion-aware pretraining and supervised activity fine-tuning, the motion embedding $\mathcal{E}_s$ and motion encoder E_{motion} are frozen for all subsequent stages.

Sensor encoders and Motion Adapters. Each sensing modality uses the corresponding protocol-specific HoloLLM (Zhou & Yang, 2026) sensor encoder, which remains frozen. Each Motion Adapter contains 17 learnable joint queries and one eight-head cross-attention layer with a hidden dimension $d_h = 512$. The adapter maps the modality-native features to the $K \times J \times d$ temporal–joint representation described in Section 4.2.

RDMF. The reliability scoring network and disagreement complement network use hidden dimensions 32 and 64, respectively. We set the reliability floor in Eq. 10 to $\rho = 0.1$ and initialize the residual gate a to -2.

A.4 Sensor-to-Motion Alignment Loss

As described in Section 4.4, the sensor-to-motion alignment objective combines token-wise cosine distance and SmoothL1 distance. For a predicted motion representation $\mathcal{Z} \in \mathbb{R}^{B\times K\times J\times d}$ and the corresponding skeleton anchor $\mathcal{Z}^s$, we define

$$\mathcal{L}_{\text{align}}(\mathcal{Z}, \mathcal{Z}^s) = \frac{1}{BKJ} \sum_{b,k,j} \left[1 - \frac{\left\langle \mathcal{Z}_{b,k,j}, \mathcal{Z}^s_{b,k,j} \right\rangle}{\|\mathcal{Z}_{b,k,j}\|_2 \left\| \mathcal{Z}^s_{b,k,j} \right\|_2 + \epsilon} + \ell_{\text{SL1}} \left(\mathcal{Z}_{b,k,j}, \mathcal{Z}^s_{b,k,j} \right) \right], \tag{17}$$

where

$$\ell_{\text{SL1}}(u, v) = \frac{1}{d} \sum_{c=1}^{d} \text{SmoothL1}(u_c - v_c). \tag{18}$$

In Stage 1, $\mathcal{Z} = \mathcal{Z}^m$, so the objective independently aligns each sensor representation with the skeleton-derived motion anchor. In Stage 2, $\mathcal{Z} = \mathcal{Z}^F$, so the same objective directly supervises the reliability-weighted motion consensus used by RDMF.

A.5 Training Details

Motion-anchor training. For each evaluation protocol, we pretrain the motion model for 400 epochs following the MAMP objective, using a masking ratio of 0.9. Pretraining uses AdamW with a learning rate 10^{-3}, weight decay 0.05, batch size 32, and 20 warm-up epochs. The pretrained model is then fine-tuned for activity recognition for 100 epochs with a learning rate 10^{-4} and five warm-up epochs. After fine-tuning, the motion embedding and motion encoder are frozen.

Stage 1: Single-modality adaptation. Each Motion Adapter and its modality-specific classifier are trained for 50 epochs using AdamW. The learning rates are 2×10^{-4} for the Motion Adapter and 10^{-3} for the classifier, with cosine decay to 10^{-6}. We use weight decay 0.05, label smoothing 0.1, and gradient clipping at 1.0. The batch size is 8 for all sensor branches except MM-Fi LiDAR, for which we use 4. The objective weights are $\lambda_{\text{cls}} = \lambda_{\text{motion}} = 1$.

Stage 2: Multimodal fusion. After Stage 1, all sensor encoders, Motion Adapters, and motion encoders are frozen. RDMF and the fusion classifier are trained for 30 epochs using AdamW with learning rate 2×10^{-4}, weight decay 0.01, and batch size 2. All experiments use mixed-precision training and random seed 3407. Results are reported using the prescribed HoloLLM test annotations.

Table 4: Comparison between directly reusing the frozen skeleton-trained classifier and training a fresh sensor-specific classifier on MM-Fi and XRF55. The frozen classifier tests whether adapted sensor representations can be directly recognized by the skeleton-trained motion model, while the fresh classifier is trained separately for each sensing modality. Best results are shown in **bold**.

Setting	Classifier	MM-Fi						XRF55					
		V	D	M	L	W	Avg.	V	D	I	R	W	Avg.
Random	Frozen skeleton classifier	99.6	99.1	96.6	98.2	86.5	96.0	97.1	94.9	96.2	54.4	93.6	87.2
	Fresh sensor classifier	**99.8**	**99.5**	**98.4**	**99.2**	**90.4**	**97.5**	**97.6**	**95.5**	**97.5**	**55.5**	**95.5**	**88.3**
Cross-Subject	Frozen skeleton classifier	94.7	96.5	92.9	89.3	7.4	76.2	50.2	46.5	38.7	4.3	17.2	31.4
	Fresh sensor classifier	**95.9**	**97.3**	**95.4**	**93.6**	**8.8**	**78.2**	**52.0**	**48.3**	**41.4**	**4.9**	**17.9**	**32.9**
Cross-Environment	Frozen skeleton classifier	56.9	82.9	61.8	47.3	7.3	51.2	22.6	10.2	32.4	33.8	37.9	27.4
	Fresh sensor classifier	**66.9**	**99.9**	**71.6**	**63.9**	**10.7**	**62.6**	**28.5**	**13.9**	**35.8**	**37.4**	**41.8**	**31.5**

Table 5: Computational comparison on MM-Fi. "Native E2E" follows each released system's original input setting, while "Controlled E2E" uses the same five-modality front-end for the X-Fi-style fusion model and UniMotion. Total parameters denote all resident model weights, whereas optimized parameters count only parameters updated during the corresponding HAR/fusion stage. "Fusion only" compares the two fusion modules using identical input motion tokens. N/M denotes values not measured under the same hardware protocol.

Method	Scope	Input	Total params (M)	Optimized params (M)	GFLOPs	Latency (ms)
HoloLLM (Zhou & Yang, 2026)	Native E2E	5 modalities	$\geq 7{,}000$	N/M	N/M	N/M
X-Fi (Chen & Yang, 2025)	Native E2E	4 modalities, 1 frame	42.859	15.864	49.895	N/M
UniMotion	Native E2E	5 modalities, 5 frames	165.650	**0.069**	80.061	N/M
X-Fi-style	Controlled E2E	shared 5-modal input	171.390	5.809	81.303	N/M
UniMotion	Controlled E2E	shared 5-modal input	**165.650**	**0.069**	**80.061**	N/M
X-Fi-style	Fusion only	$5\times5\times17\times256$ tokens	5.809	5.809	1.250	11.78
UniMotion	Fusion only	identical tokens	**0.069**	**0.069**	**0.0088**	**1.99**

A.6 TRANSFERABILITY OF THE MOTION SPACE

Table 4 provides a functional test of the learned sensor-to-motion alignment. We directly apply the classifier learned from the skeleton motion model to the adapted sensor representations, without retraining it on sensor features, and compare it with a fresh classifier trained separately for each modality. The frozen skeleton classifier remains close to the fresh classifier across many modalities and settings. Under the Random protocol, for example, the average accuracy decreases only from 97.5% to 96.0% on MM-Fi and from 88.3% to 87.2% on XRF55. This indicates that the Motion Adapters map heterogeneous sensor observations into representations that are directly interpretable by the skeleton-trained motion model, rather than merely matching its tensor structure.

A.7 COMPUTATIONAL EFFICIENCY

Table 5 compares UniMotion with X-Fi (Chen & Yang, 2025) and HoloLLM (Zhou & Yang, 2026). Because the released systems use different input configurations, we report both their native settings and a controlled comparison using the same five-modality front-end. Under the controlled setting, UniMotion optimizes only 0.069M parameters, compared with 5.809M for the X-Fi-style fusion model, while their end-to-end computation is similar. Given identical motion tokens, RDMF requires 0.0088 GFLOPs and 1.99 ms, compared with 1.250 GFLOPs and 11.78 ms for X-Fi-style fusion. Thus, although UniMotion retains frozen sensor encoders as resident parameters, its learned multimodal fusion component is lightweight.