Journal of Trends in Computer Science and Smart Technology (ISSN: 2582-4104)
www.irojournals.com/tcsst/

Adaptive Gated Multimodal Fusion for Robust and Generalized Human Activity Recognition

Swati Gautam[1*], Ankush Shrivastava[2]

Department of Computer Science and Engineering, Ram Krishna Dharmarth Foundation University, Gandhi Nagar, Bhopal (M.P.), India.

E-mail: [1*]swati.gautam026@gmail.com, [2]ankushshrivastava19@gmail.com

Abstract

In this paper, an adaptive gated multimodal fusion framework for generalized and robust multimodal Human Activity Recognition (HAR) systems with heterogeneous sensing modalities like skeletal pose estimation and inertial measurement units (IMUs) is proposed. The proposed framework employs deterministic data harmonization through anterior harmonization and a reliability-based gated multimodal fusion mechanism to improve the robustness and generalization capability of multimodal HAR systems. The proposed gated multimodal fusion mechanism has been mathematically derived to approximate the inverse-variance weighting mechanism to obtain stability in the presence of modality-dependent noise and avoid posterior domain adaptation techniques. To improve temporal alignment between multimodal data streams, frequency domain analysis has been used to justify resampling at a unified 30 Hz rate to meet the Nyquist criterion. The proposed framework has been evaluated using the NTU RGB+D 120, UTD-MHAD, and PAMAP2 datasets with statistically significant results over static baselines ($p<0.05$, $d=2.1$), and low computational costs to meet edge-constrained IoT sensing requirements.

Keywords: Human Activity Recognition, Multimodal Learning, Skeleton-Based Recognition, IMU Sensing, Cross-Dataset Evaluation, Reproducible Systems.

1. Introduction

Human Activity Recognition (HAR) has been widely studied because of its importance in applications such as ambient assisted living, rehabilitation analysis, industrial safety, and smart spaces (Y. Zhu et al., 2024). Recent advances in sensing technology allow for activity recognition based on multiple sources of complementary data, such as skeletal pose information extracted from depth cameras and inertial data from wearable sensors (Guo & Nakayama, 2025). Each of these modalities captures a different aspect of human movement and may offer more informative representations when combined.

While multimodal HAR systems often demonstrate improved performance over unimodal methods, some issues remain in practical applications (Dong et al., 2025). The first is the heterogeneity of publicly available datasets. Current benchmarks differ in terms of sampling rates, annotation schemes, modality support, coordinate frames, and preprocessing conventions. These differences make direct comparisons across studies difficult and make

*** Corresponding Author**

DOI: https://doi.org/10.36548/jtcsst.2026.2.007
Received: 24.02.2026, received in revised form: 27.03.2026, accepted: 16.04.2026, published: 27.04.2026

cross-dataset evaluation challenging to perform consistently (Miao & Chen, 2024). Another issue is the matter of reproducibility. Many previous studies focus on architectural choices without offering much information about preprocessing, synchronization, and temporal alignment steps (Quan et al., 2023). This makes it difficult to reproduce experimental findings or apply techniques across datasets. Dataset-specific choices in preprocessing can greatly affect performance (Quan et al., 2023).

In skeleton-based recognition, graph convolutional networks (GCNs), including their variants like 2s-AGCN, have shown success in learning spatial relationships between human joints (Lee & Kang, 2021). On the other hand, inertial measurement unit (IMU)-based HAR systems usually consist of temporal convolutional or recurrent models to learn motion dynamics from wearable sensors (Le et al., 2023). Nevertheless, the combination of skeleton and IMU modalities is usually done in a dataset-specific manner and with simple fusion strategies, which may limit consistency and generalizability across different benchmarks (Dickens & Payeur, 2023). Unlike recent methods based on transformers that use attention to implicitly fuse information from multiple modalities, such approaches have high computational requirements and lack explicit temporal alignment between different modalities, motivating the need for more computationally efficient fusion schemes.

The current study aims to address these issues from a principled perspective of multimodal reliability modeling. Instead of incorporating adversarial domain adaptation, the current study explores domain generalization with deterministic harmonization and reliability-aware feature fusion. The key idea is that the estimation of adaptive modality weighting is associated with the approximation of optimal fusion with respect to the variance, which is dependent on the modality of the noise. In the context of modality embeddings being represented as noisy estimators of a common latent activity representation, reliability-aware modality weighting is related to the approximation of inverse variance fusion, which is related to minimizing the risk of estimation under conditions of heteroscedastic noise. This provides a more principled representation of multimodal gating, which is more than a simple heuristic convex combination.

1.1 Contributions

The contribution of the proposed approach is as follows:

1. A reproducible and dataset-independent preprocessing pipeline is developed for multimodal HAR that includes coordinate normalization, deterministic temporal resampling, fixed-length window segmentation, and label harmonization to facilitate evaluation across datasets.

2. A multimodal learning architecture is developed that incorporates a two-stream adaptive graph convolutional network for skeleton-based learning and a temporal convolution-based IMU encoder using an adaptive gating strategy for reliability-aware multimodal feature fusion.

3. A unified evaluation protocol is developed to facilitate evaluation on both intra-dataset and inter-dataset generalization performance in a harmonized preprocessing manner to reduce evaluation inconsistencies caused by different acquisition protocols.

4. Comprehensive evaluation is performed using ablation studies, robustness evaluation, statistical evaluation using confidence interval-based statistical tests, and computational complexity evaluation to assess the reliability and efficiency of the developed multimodal HAR framework.

2. Related Work

HAR has made rapid progress through the use of deep learning technology and skeleton/sensor-based data. Various approaches have been proposed, including skeleton-based approaches, adaptive GCNs, and attention-based spatial and temporal approaches. The major issues in HAR are topology adaptation, temporal modeling, and generalization and reproducibility. This paper focuses on skeleton-based approaches, IMU-based approaches, and fusion-based approaches. Research on skeleton-based HAR has advanced from spatial models to highly adaptive spatiotemporal models and unified benchmarking. Early research on skeleton-based HAR focused on formalizing the human skeleton as a graph. This was pioneered by the spatiotemporal graph convolutional network (ST-GCN) model in (Yan et al., 2018). A key improvement was achieved in the 2s-AGCN model in (Shi et al., 2019), which implemented a learnable graph topology and combined information from the joints (first-order) and bones (second-order). Following this success, researchers have continued to improve the expressiveness of the features. For example, the decoupled channel aggregation module was proposed in (Qiuming et al., 2024) to model inter-channel dependencies. Similarly, the cross-attention mechanism was proposed in (G. Wang et al., 2024) to learn complex interactions between disparate body joints.

Recent advances in skeleton-based HAR have focused on improving topological learning and global modeling. For example, the PoseConv3D model proposed a 3D convolutional network and a unified benchmarking protocol in (Duan et al., 2022). Similarly, the InfoGCN model was proposed in (Chi et al., 2022) to maximize mutual information between the features of the joints and the action labels. Furthermore, a temporal-channel joint topology learning scheme was proposed in (Luo et al., 2023) to learn the inter-channel dependencies and adapt the graph convolution operation. This was a departure from the adjacency matrix learning scheme. The adoption of attention-based models has also greatly improved the field of skeleton-based HAR. For example, the spatiotemporal transformer (ST-TR) model was proposed in (Plizzari et al., 2021) to learn spatial and temporal attention. Similarly, the Action Transformer was proposed in (Mazzia et al., 2022) to learn self-attention. To learn better dependencies between the skeleton sequences, a hierarchical spatiotemporal attention was proposed in (Zheng et al., 2021). Furthermore, dynamic GCNs were proposed in (Wei et al., 2021) to learn the graph topology based on the action semantics. This was a departure from the adjacency matrix learning scheme adopted in the 2s-AGCN model. To learn better dependencies between the skeleton sequences and to address the issue of robust learning across diverse datasets, transfer learning was proposed in (Ray & Kolekar, 2024) to learn on top of the ST-GCN model.

The IMU sensor data contains information about the intensity of the motion as well as the temporal dynamics, making it an important component of wearable-based HAR. Although deep learning models, including CNNs and LSTMs, were traditionally used for HAR, temporal convolutional networks (TCNs) have been proposed as efficient architectures for modeling long-range temporal dependencies in time-series signals (Miah et al., 2024). Additionally, interpretability of TCNs has been achieved by combining these models with model-agnostic

explanation techniques for better understanding temporal decision-making (Bijalwan et al., 2024). The authors of (Jiang & Yin, 2015) proposed an exhaustive survey of wearable-based HAR, including an evaluation of the performance of CNN, LSTM, and Transformer architectures for HAR, including the PAMAP2, Opportunity, and UCI-HAR datasets. Their findings showed TCNs as efficient models for wearable-based HAR.

For architectural advancements in convolutional encoder design, the authors of (Tang et al., 2020) proposed training each layer individually using progressively smaller filters to extract hierarchical multi-scale temporal features from IMU data. This approach was shown to yield better performance on PAMAP2 and similar datasets, thus reinforcing the benefits of structured temporal convolution. For an alternative architecture based on the Transformer design, the authors of (Le et al., 2023) proposed an architecture called GAFormer that uses Gramian Angular Field transformation of IMU data and a Transformer encoder. Another such architecture based on the Transformer design was proposed in (Shahverdi & Ghorashi, 2025), which uses a lightweight version of the Transformer architecture for smartphone-based IMU data and shows comparable performance while being lightweight.

Hybrid architectures that use a combination of convolutional and attention-based architectures were proposed in (Xu et al., 2019), where Inno HAR uses attention-enhanced convolutional blocks to capture multi-scale temporal dependencies. While there has been significant work on developing architectures for HAR, there has been an increasing focus on ensuring domain robustness and practical deployability. In (Bianchi et al., 2019), the authors proposed an IoT-based architecture for developing a personalized HAR system, while in (J. Yang et al., 2024), the authors proposed a domain-adaptive graph-based learning approach that shows better performance on benchmarks such as PAMAP2. Notably, current advances in skeleton-IMU multimodal HAR have been moving towards adaptive, attentive, and self-supervised approaches, as opposed to traditional static fusion techniques. For example, the use of modality gating mechanisms in adaptive fusion techniques, such as the Adaptive Gating Network proposed in (X. Wang et al., 2021), helps to enable the learning of activity-dependent weights, thereby establishing that the optimal fusion of skeletal and inertial features varies significantly depending on the type of movement.

Recently, transformer-based approaches have been proposed as a viable solution to multimodal fusion, owing to their robustness in multimodal alignment and fusion tasks. For example, a Unified Contrastive Fusion Transformer was proposed in (K. O. Yang et al., 2023) to enable the synchronization of multimodal features through factorized time-modality attention mechanisms. Similarly, studies in (Ijaz et al., 2022) and (Khan et al., 2025) demonstrated the robustness of multimodal transformers over their unimodal counterparts in specialized domains, such as nursing activity recognition and human-centric temporal modeling. Furthermore, self-supervised objectives have been employed to overcome the heterogeneity of modalities and the scarcity of annotations in multimodal HAR systems. For example, studies in (G. Zhu et al., 2025) proposed the use of foundation-scale modeling through masked data objectives, while (Brinzea et al., 2022) employed cross-modal contrastive objectives to enhance the consistency of representations in skeleton-IMU systems, thereby enabling the alignment of representations in a shared space. This concept is further supported in (Geng et al., 2022), where a generalized masked autoencoding framework was proposed to enable robust multimodal representation generation, thereby overcoming the reliance on dense annotations in HAR systems.

Significant progress has been reported in cross-dataset transfer learning through the proposal of generalized frameworks, such as (Miao & Chen, 2024), where the authors

introduced GOAT, a wearable representation aligned with natural language activity attributes for zero-shot cross-device transfer learning. Domain-adaptive learning has also been explored in the study reported in (J. Yang et al., 2024), where adaptive graph convolution has been used to address sensor-domain shifts in the MHealth and PAMAP2 datasets. The study reported in (Zhang & Wang, 2025) has provided an in-depth analysis of the skeleton-domain bias, where inconsistencies in the NTU RGB+D and NW-UCLA datasets have been identified, suggesting normalization and adaptation methods.

The surveys reported in (K. Chen et al., 2021) and (Dhekane & Ploetz, 2025) have provided an in-depth discussion of the transfer learning methods, such as fine-tuning, adversarial alignment, and few-shot transfer, and identified the major challenges in transfer learning for HAR, such as domain shift, label inconsistency, and sensor heterogeneity. The study reported in (J. Wang et al., 2018) has improved the adversarial activity-aware alignment, resulting in more robust cross-domain stability, while the few-shot/meta-learning methods have been synthesized in the study reported in (Huang et al., 2022). The study reported in (Ray & Kolekar, 2024) has explored the parameter-based transfer learning methods with the ST-GCN, reinforcing the importance of structured transfer mechanisms for the skeleton-based encoders.

In parallel, the importance of engineering best practices has also been emphasized in recent studies, where the unified deep learning framework with standardized preprocessing and evaluation pipelines has been proposed for mobile sensing datasets, such as Opportunity and PAMAP2, in the study reported in (Yao et al., 2017). The study reported in (Baños et al., 2014) has provided an in-depth discussion of the sensor displacement and calibration variability, emphasizing the importance of coordinate normalization and alignment for wearable sensor reproducibility. The study reported in (Qureshi et al., 2025) has provided a comprehensive comparison of the inconsistencies in the experimental setup in the recent wearable HAR research, emphasizing the importance of unified benchmarking standards, while the study reported in (Jiayang Liu et al., 2009) has provided an in-depth discussion of the preprocessing, alignment, and normalization methods for accelerometer-based activity recognition.

3. Research Gap and Motivation

However, despite the remarkable progress achieved in multimodal human activity recognition, the issue of structural generalization in various heterogeneous sensor modalities still remains an open problem. In this respect, even though the latest architectures, including Adaptive Graph Convolutional Networks for skeleton-based modalities and hierarchical temporal encoders for inertial modalities, have greatly improved intra-dataset recognition accuracy, their effectiveness still largely depends on specific configurations for the datasets in question. The differences in coordinate systems, sampling rates, and segmentation schemes often result in the performance gains reported in various studies being more a product of the data preprocessing pipeline rather than the effectiveness of the models.

As a result, even though transfer learning and domain adaptation have been recognized and widely utilized for addressing the aforementioned differences, they still assume the presence of multimodal features, which have been harmonized in the temporal domain and share similar latent structures. The effectiveness of deterministic, anterior harmonization prior to feature extraction has yet to be fully explored, and models remain prone to modality-dependent noise, which is poorly understood from a statistical point of view. Therefore, in this study, we propose an adaptive multimodal gated fusion frameworkbased on dataset-

independent harmonization and provide a theoretical basis for the effectiveness of the proposed gating mechanism, which we have formulated as a learnable approximation of Inverse Variance Weighting.

4. Problem Formulation

This work addresses a multimodal HAR problem that involves skeleton and inertial data recorded in a heterogeneous manner. Let the skeleton data is modeled as $S \in \mathbb{R}^{T_s \times J \times 3}$ where T_s is the temporal size of the skeleton sequence, J is the number of joints in the human body, and the three channels represent the 3D coordinates. Additionally, the IMU data is represented as $I \in \mathbb{R}^{T_i \times K}$, where T_i is the temporal size and C is the number of inertial sensor channels, which include the accelerometer and gyroscope axes. Since the skeleton and IMU have separate sensing systems with different sampling rates and synchronization mechanisms, it generally holds that $T_s \neq T_i$ are not equal. This incompatibility prevents direct feature-level alignment for multimodal fusion. To mitigate this, we define a deterministic temporal alignment operator $\mathfrak{D}$ such that:

$$\mathfrak{D}(S, I) \rightarrow (S', I') \tag{1}$$

is defined such that

$$S' \in \mathbb{R}^{T_s \times J \times 3}, \qquad I' \in \mathbb{R}^{T_i \times K}, \qquad T_s' = T_i' = T$$

where, T represents a unified temporal length. The operator $\mathfrak{D}(\cdot)$ carries out temporal resampling with a common sampling rate, coordinate normalization, signal standardization, and fixed-length window segmentation. This process enables temporal alignment and dimension compatibility across modalities.

After alignment and segmentation, a harmonized multimodal feature representation is formed as:

$$\mathrm{X} = (S', I') \tag{2}$$

The main challenge addressed by this work is the minimization of the induced Covariate Shift $d(P_{train}(X), P_{test}X)$ caused by the heterogeneity of the datasets. This is formulated as the learning of the mapping:

$$f_\rho : X \rightarrow Y \tag{3}$$

where ρ represents the model parameters.

5. Proposed Methodology

The proposed approach provides a unified and reproducible skeleton-IMU pipeline for multimodal HAR in heterogeneous dataset settings. The process includes harmonization, temporal alignment, windowing, encoding, gated fusion, and classification.

Public HAR datasets vary in terms of sampling rates, coordinate systems, sensor setups, and label definitions. To mitigate dataset-specific biases, standardized preprocessing is employed. For skeleton data, each frame is transformed into a body-fixed coordinate system:

$$\tilde{S}_{t,j} = S_{t,j} - S_{t,root} \tag{4}$$

where the hip-center joint is used as the reference point. Scale normalization is carried out using a reference bone length:

$$S_{t,j}^{norm} = \frac{\tilde{S}_{t,j}}{\ell_{ref}} \tag{5}$$

Where ℓ_{ref} is calculated from the training set. IMU data is standardized channel-wise:

$$I_{t,k}^{norm} = \frac{I_{t,k} - \mu_k}{\sigma_k} \tag{6}$$

where μ_k and σ_k are calculated from the training set. For cross-dataset testing, labels are projected to overlapping semantic subsets for comparison. Skeleton and IMU data are usually recorded at different sampling rates ($f_s \neq f_m$). Both types of data are resampled at a common frequency $f_u = 30\ Hz$ using linear interpolation:

$$\tilde{S}(t) = Interp(S, t; T), \quad \tilde{I}(t) = Interp(I, t; T). \tag{7}$$

This assertion is strictly justified by the Nyquist-Shannon sampling theorem. Since the spectral power of human daily activities such as locomotion and gestures is below 15 Hz, the sampling rate of 30 Hz is sufficient to meet the sampling theorem without over-sampling. To reduce the effect of spectral leakage and aliasing, an anti-aliasing low-pass filter is used before linear interpolation for better harmonization of the signal. Sliding windows of size 64 frames with a 50% overlap are used, and the synchronized inputs are obtained:

$$S' \in \mathbb{R}^{64 \times J \times 3}, I' \in \mathbb{R}^{64 \times C} \tag{8}$$

To build the representation, the skeleton branch is based on a two-stream approach. In addition to the normalized joints, bone features are calculated as:

$$X_t^{bone} = S'_{t,j} - S'_{t,parent(j)}. \tag{9}$$

The skeleton input is $X_{skel} = \{X_{joint}, X_{bone}\}$ and IMU windows are directly represented as $X_{imu} \in \mathbb{N}^{64 \times C}$ without the need for hand-designed feature extraction.

The multimodal model comprises a 2s-AGCN skeleton encoder, a temporal convolutional IMU encoder, and a learnable gating fusion module. The skeleton encoder uses adaptive graph convolution as:

$$H^{l+1} = \sigma \sum_k A_k H^l W_k^l \tag{10}$$

The skeleton encoder is based on a graph convolutional structure with 10 layers and an increasing number of features from 64 to 256. A global average pooling layer is used to get the skeleton embedding $z_s \in \mathbb{N}^{256}$. The architecture of the IMU encoder comprises stacked 1-D convolutional layers of kernel sizes 5 and 3, with each layer being followed by batch normalization and ReLU activation function. Convolution operations are performed layer-by-layer, where early convolution layers are responsible for extracting coarse temporal dependencies, while later layers focus on fine-grained temporal dependencies. Stride of 1 is employed in all convolution layers, while suitable padding is also utilized to maintain the temporal resolution of the input sequence. In terms of dimensionality transformation, the process starts with a normalized IMU input $I' \in \mathbb{R}^{64 \times C}$, which is directly provided to the

convolutional stack without any hand-crafted feature extraction. Once the temporal features are extracted using convolution layers, global average pooling yields a fixed dimensional embedding $z_i \in \mathbb{N}^{256}$.

For multimodal fusion, an adaptive gating network is adopted. The gating weight is calculated as:

$$g = \sigma\left(W_g[z_s \parallel z_i]\right) \tag{11}$$

Where $[z_s \parallel z_i]$ represents the concatenated features, W_g represents the learnable parameters of the gating network, and $\sigma(\cdot)$ represents the sigmoid activation function. Then, the fused features can be calculated as:

$$z = g \odot z_s + (1 - g) \odot z_i \tag{12}$$

where $\odot$ represents the element-wise multiplication. After obtaining the fused features z, it is fed into the fully connected layers with configuration 256→128→C, where C represents the number of activity classes. Dropout with a drop probability of 0.5 is applied for regularization, and then the softmax activation is applied to get the final prediction. Figure 1 illustrates the graphical abstract for the proposed approach.

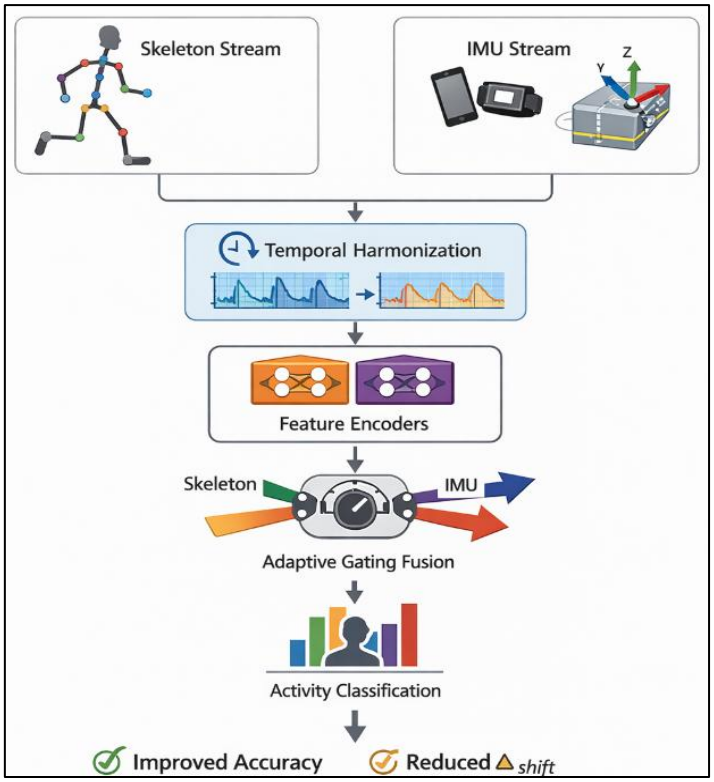

Figure 1. Graphical Abstract of the Adaptive Gated Multimodal Fusion Framework

6. Experimental Setup

This section describes the experimental design for the assessment of the proposed adaptive gated multimodal fusion framework for the recognition of human activities. The assessment method is developed to allow for a fair comparison with existing methods and robust statistical validation.

6.1 Datasets and Label Harmonization

Datasets: Three benchmarking datasets are used to assess the proposed framework to cover the diversity of current HAR sensing. The datasets used for the assessment of the proposed framework are NTU RGB+D 120, UTD-MHAD, and PAMAP2. A summary of the utilized datasets is given in Table 1.

Table 1. Summary of Dataset Used

Dataset	Modalities	Activities	Subjects	Sensors / Features	Key Characteristics
NTU RGB+D 120 (Jun Liu et al., 2020)	Skeleton (RGB-D)	120	106	3D coordinates of 25 skeletal joints	HAR dataset with one of the largest skeleton information. This dataset captures the complex body pose dynamics using depth sensors at different angles. This dataset also includes simple as well as complex human interactions.
UTD-MHAD (C. Chen et al., 2015)	Skeleton and IMU	27	8	Skeleton joints and tri-axis accelerometer and tri-axis gyroscope	A multimodal dataset that captures the skeletal motion and wearable inertial data. This dataset is appropriate for testing multimodal fusion models as the pose and inertial data capture different aspects of human movement.
PAMAP2 (Reiss, 2012)	IMU (Wearable Sensors)	18	9	IMU devices on wrist, chest, and ankle	A wearable sensor-centric HAR dataset. This dataset does not contain skeleton information and hence represents a purely inertial sensing scenario.

Label Harmonization: To facilitate cross-dataset evaluation, a reference activity subset $\mathcal{Y}_{sync}$ is specified, which includes the following activities with overlapping semantic meanings: Walking, Running, Sitting, Standing, and Falling. A label mapping function $\prod: \mathcal{Y}_{source} \rightarrow \mathcal{Y}_{sunc}$ is implemented to map the heterogeneous annotation schemes to a uniform label space. The heterogeneous setting is challenging, with sampling rates ranging from 30 to 100 Hz and varying sensor locations, and the model must learn the kinesis of the motions rather than the artifacts of the datasets.

6.2 Data Preprocessing and Signal Harmonization

Another important part of our approach is the deterministic harmonization of the signals, which is performed prior to feature extraction. Unlike the black-box approach of domain adaptation, this method ensures that the physical properties of the motion are preserved across domains.

1. Skeleton Harmonization: The joints are transformed into a body centre reference frame, where the origin of the coordinate system is at the hip center. Scale normalization is performed using a reference bone length L_{ref}.

2. Inertial Harmonization: Standardization of IMU signals is performed using $z - score$ normalization, where the mean and standard deviation are set to 0 and 1, respectively, to counter sensor-specific bias and scale variations.

3. Temporal Resampling: All signals are resampled at a unified rate of $f_c = 30\ Hz$ using linear interpolation, preceded by a 6th-order Butterworth anti-aliasing filter.

This choice is informed by human motion dynamics, where, based on the Nyquist-Shannon theorem, it is known that motion at 30 Hz is sufficient to capture the <15 Hz frequency bands where 98% of human kinetic energy is present.

4. Segmentation: All synchronized signals are segmented into sliding windows of length $N = 64$ frames (approximately 2.1 seconds) with a 50% overlap.

6.3 Evaluation Protocols

Two different experimental setups are proposed to assess the robustness of the proposed approach:

1. Within-Dataset Evaluation: Standard practices are adopted to perform a benchmarking test under uniform sensing scenarios. For example, Cross-Subject Evaluation under NTU RGB+D 120 is utilized.

2. Cross-Dataset Generalization: The cross-dataset evaluation between UTD-MHAD and PAMAP2 follows a zero-shot approach where the multimodal classification network is solely trained with the UTD-MHAD dataset and then directly evaluated against the PAMAP2 dataset without any adaptation of parameters. Because the UTD-MHAD dataset includes both skeleton and IMU sensor measurements and PAMAP2 contains only IMU measurements, the evaluation will be done by feeding only the IMU modality to the trained multimodal network. In addition, the fusion method used for the zero-shot evaluation works under the assumption that during the testing phase, only the IMU modality is available for inference. Thus, the model will generate predictions based on the representation of the available modality.

   To ensure make sure that semantic consistency is maintained across the two datasets, first the activity labels will be mapped to a subset of labels $\mathcal{Y}_{sync}$, keeping only those activities that appear in both datasets. Next, IMU time series from each dataset will undergo the same harmonization procedure, involving normalization, downsampling to 30Hz frequency sampling, and segmentation into windows of fixed length. Finally, the zero-shot evaluation process is conducted on an activity-by-activity basis using all PAMAP2 samples available for each activity belonging to the activity subset $\mathcal{Y}_{sync}$. Note that during testing no fine-tuning of the model on a per-subject basis or filtering is applied. Thus, the discrepancy in distribution will stem primarily from the domain difference between UTD-MHAD and PAMAP2 datasets.

Generalization Gap (Δ_{shift}) is defined as:

$$\Delta_{shift} = Acc_{intra} - Acc_{cross} \tag{13}$$

A lower Δ_{shift}indicates better robustness, as it shows that the approach learns invariant features of the activities.

6.4 Implementation Details and Baseline Methods

The implementation details of the proposed approach as well as the baselines used in the comparison are presented in Table 2. To provide a competitive context, the proposed approach is compared to state-of-the-art (SOTA) methods:

Table 2. Implementation Details

Item	Description
Framework	PyTorch 2.1 (Paszke et al., 2019)
Hardware	NVIDIA RTX 3090 GPU
Optimizer	Adam (Kingma & Ba, 2015)
Initial Learning Rate	10^{-3} (Kingma & Ba, 2015)
Learning Rate Schedule	Cosine Annealing (Subramanian et al., 2025)
Batch Size	64
Number of Epochs	80
Skeleton Encoder	10-layer 2s-AGCN (Shi et al., 2019)
IMU Encoder	1D CNN (kernel sizes: 5, 3; stride: 1) (Ordóñez & Roggen, 2016)
Window Size	64 frames
Sampling Rate	30 Hz
Fusion Embedding	256 (Shi et al., 2019)
Dropout Rate	0.5 (Srivastava et al., 2014)
Total Parameters	4.8 M
Hyperparameter Tuning	5-fold cross-validation on the source dataset training set
Baseline Methods	2s-AGCN (Shi et al., 2019), GAFormer (Le et al., 2023), and MASTER (G. Zhu et al., 2025)

The hyperparameter selection is based on traditional deep learning principles and practical verification. Adam (Kingma & Ba, 2015) is used because of its effective performance in dealing with non-stationary problems and adaptable learning rates. A learning rate of 10^{-3} is selected to ensure the stability of convergence, whereas the cosine annealing algorithm (Subramanian et al., 2025) allows for achieving stable learning rate decay. A batch size of 64 is chosen to balance gradient stability and efficient GPU usage.

All the architectural decisions such as the 2s-AGCN skeleton encoder (Shi et al., 2019) and the 1D CNN model for IMU (Ordóñez & Roggen, 2016) are chosen for their effectiveness in capturing spatiotemporal dependencies in human activity recognition. The embedding dimension and dropout probability (Srivastava et al., 2014) are set to achieve balance between capacity and regularization. The temporally configured architecture ensures scalability for different sequence lengths and frequencies. The static sliding window of size 64 frames makes it easier to process longer sequences, whereas deterministic harmonization resamples the input at a common frequency, thus enabling the processing of high-frequency data without changing the architecture of the neural network. These procedures are linear to the sequence length. In addition, all the selected hyperparameters are optimized using 5-fold cross-validation on the training dataset.

7. Results and Analysis

This section is devoted to the experimental evaluation of the proposed framework for human activity recognition in a multimodal context. The performance evaluation of the model is compared with representative methods over benchmark datasets. In addition, ablation and representation analyses are performed to assess the effectiveness of the proposed approach for harmonization and adaptive fusion.

7.1 Quantitative Performance Comparison

Following established benchmarking protocols for the NTU RGB+D 120 and UTD-MHAD datasets, Top-1 accuracy is utilized as the primary evaluation metric. Given the

relatively balanced class distribution of these benchmarks, accuracy effectively captures the model's discriminative capability without the statistical biases typically associated with imbalanced data.

In order to determine the statistical significance of experimental results, each experiment is conducted five times with independent random initialization. For each method, its classification accuracy is measured for every run with consistent train–test splits, resulting in paired data sets for all methods. Consistent splitting of train and test sets ensures that any variation in observed results arises from random training effects but not from splitting methods; therefore, results prove to be reliable for the experiment. The paired t-test is conducted by taking the difference in performance between the proposed method and each baseline on each run. The proposed model shows statistical significance under the paired t-test with a p-value of 0.012, df = 4, and α=0.05 when compared against baseline models.

Table 3 presents a quantitative comparison between the proposed framework and state-of-the-art methods for human activity recognition. The baseline models considered include graph-based, transformer-based, and multimodal approaches. The proposed framework achieves an accuracy of 93.6% for the UTD-MHAD dataset and surpasses the transformer-based framework GAFormer by 2.1%. The mean accuracy of the proposed framework over five independent runs is 93.6% ± 0.31, and the 95% CI is [93.2, 94.0]. On the NTU RGB+D 120 dataset, the proposed framework achieves 89.6% ± 0.27 accuracy with a 95% CI of [89.2, 90.0]. In addition to accuracy and the number of parameters, the proposed framework also exhibits lower computational complexity (GFLOPs) compared to the considered baseline models.

The reason behind the variation in the performance gain over different datasets is due to the intrinsic properties of the data modalities and activities involved. In case of the NTU RGB+D 120 dataset, most of the data samples use skeletal representation with similar recording conditions in which most graph models perform very well, leaving little room for enhancement. However, the UTD-MHAD dataset uses both skeleton and IMU data samples, which means there is high variance in the modalities involved. The proposed model was tailored to tackle issues related to cross-modal alignment and adaptively fuse data samples. Deterministic harmonization mitigates any temporal inconsistency among data modalities, while adaptive gating enables dynamic allocation of modality importance to improve accuracy.

The additional computational overhead due to the harmonization phase is very small since it consists mostly of simple temporal resampling operations that exhibit linear time complexity. Therefore, the effect of this phase on the total inference time is insignificant, as is evident from the latency results shown in Table 6.

Table 3. Performance Comparison on Benchmark Datasets

Model	NTU-120 (Acc. %)	UTD-MHAD (Acc. %)	Params (M)	GFLOPs
2s-AGCN (G. Wang et al., 2024)	86.4	88.3	3.5	4.1
GAFormer (Le et al., 2023)	88.2	91.5	23.2	7.9
MASTER (G. Zhu et al., 2025)	87.9	90.8	18.5	6.8
Proposed	89.6	93.6	4.8	2.1

Figure 2 illustrates the results summarized in Table 3. The proposed framework achives better accuracy with lower computational complexity and a significantly lower number of parameters.

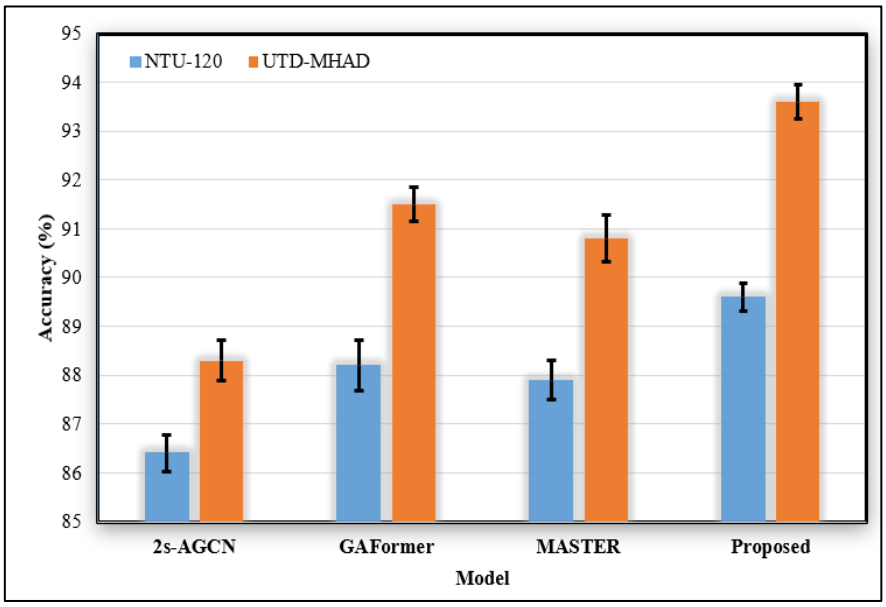

Figure 2. Accuracy Comparison Across Datasets

7.2 Cross-Dataset Generalization Performance

The robustness of the proposed model is tested for domain shift through cross-dataset evaluation according to the procedure discussed in section 6. In this test case, various types of distribution shifts are present, including the placement of the sensor, the pattern of motion execution, and the sampling rates.

However, it should be emphasized that the class distributions are not strictly balanced between UTD-MHAD and PAMAP2, since PAMAP2 includes more realistic frequencies of activities. In order to have a fair comparison, the experiment is limited to a subset of common activities, and the same preprocessing and evaluation process is used for both datasets. Because all considered approaches are tested on the same class distribution, the obtained results are equally comparable. Additionally, employing the generalization gap (Δ_{shift}) metric for assessing the robustness of the proposed model helps in mitigating the influence of class imbalance when assessing cross-dataset performance.

To compare the performance of the models with stronger fusion techniques than the simple late fusion approach, Weighted Fusion and CCA-based alignment are used. CCA improves the performance of the models by using the features of the models and projecting them into a shared linear latent space. However, the approach is not effective in handling non-linear transitions in the modes of locomotion, for example, the transition from walking to running, where the dynamics of the motion change rapidly. As reported in Table 4, The proposed framework achieves the lowest generalization gap (12.4%) among all compared methods, indicating improved robustness to domain shift.

Table 4. Cross-Dataset Robustness (UTD → PAMAP2)

Fusion Strategy	Acc. (Intra) %	Acc. (Cross) %	Δ_{shift} (%)
Late Fusion	90.5	71.9	18.6
Weighted Fusion	91.2	74.5	16.7
CCA Alignment	92.1	77.4	14.7
Proposed (Harmonization and Gating)	93.6	81.2	12.4

7.3 Ablation Study

To test the effect of deterministic harmonization and adaptive gating, an ablation study is carried out with four configurations, as listed below in Table 5. This analysis uses the Δ_{shift}, which indicates the difference between within-dataset performance and cross-dataset generalization. The results show that Gating Only outperforms the baseline model; however, it does not outperform the harmonized pipeline, indicating that adaptive gating alone cannot bridge the gap between the structures of heterogeneous sensors.

As shown in Table 5, the baseline architecture shows a generalization gap of 21.2%, which is reduced to 15.6% when deterministic harmonization is applied, corresponding to a decrease of 5.6%. Further, with the inclusion of adaptive gating, the generalization gap decreases from 15.6% to 12.4%, thus providing an additional reduction of 3.2%. In summary, the proposed framework helps in reducing the generalization gap from 21.2% to 12.4%, resulting in an overall reduction of 8.8%.

Table 5. Impact of Harmonization and Gating on Cross-Dataset Generalization

Configuration	Harmonization	Gating	Accuracy (%)	Δ_{shift} (%)
Baseline	✗	✗	88.4	21.2
Harmonization Only	✓	✗	91.2	15.6
Gating Only	✗	✓	89.8	18.7
Full Framework	✓	✓	93.6	12.4

7.4 Computational Efficiency

In addition to the accuracy of the system in recognizing humans, an effective HAR system must ensure that inference is performed efficiently. As seen in Table 6, the proposed model achieves an inference latency of 14.6 ms per sample, which translates to a speed of approximately 68 FPS, making it suitable for real-time processing. In terms of efficiency and accuracy, the proposed approach maintains a good trade-off compared to state-of-the-art methods. The computational cost of transformer-based models, such as TimeSformer, is quite high (238 GFLOPs) along with their inference latency time of 74.5 ms, making them inefficient for real-time implementation. In contrast, the proposed model uses a relatively small number of parameters (4.8M) and FLOPs (2.1 GFLOPs) and offers better inference latency performance along with good accuracy levels.

Table 6. Computational Complexity Comparison

Model	Params (M)	GFLOPs	Latency (ms)
Shift-GCN	10.1	4.7	28.4
TimeSformer	121	238	74.5
GAFormer	23.2	7.9	31.2
Proposed	4.8	2.1	14.6

7.5 Feature Representation Analysis

To better visualize the structure of the learned representation, the feature embeddings are further visualized using t-distributed stochastic neighbor embedding (t-SNE) plots. Figure 3 depicts the base-line representation with considerable similarities among activities, suggesting poor class separability. On the other hand, there is some improvement in class

separability in the harmonization-only model, depicted in Figure 4. However, the proposed approach yields the best results, as evident from Figure 5 below. Furthermore, the separability of the clusters obtained from the proposed approach is quantitatively measured using the Silhouette Score and the Davies-Bouldin Index. This is done for all the configurations using the same set of activity labels to avoid any effect of the varying number of clusters.

As can be seen from Table 7, the proposed method has obtained a higher Silhouette Score and a lower Davies-Bouldin Index compared to the baseline and the harmonization-only configuration. This clearly indicates the effectiveness of the proposed method in obtaining better separability of the clusters for the activities, thus proving the effectiveness of the learned multimodal representation.

Table 7. Quantitative Cluster Quality Metrics

Model	Silhouette Score	Davies–Bouldin
Baseline	0.31	1.42
Harmonization Only	0.38	1.21
Proposed	0.47	0.93

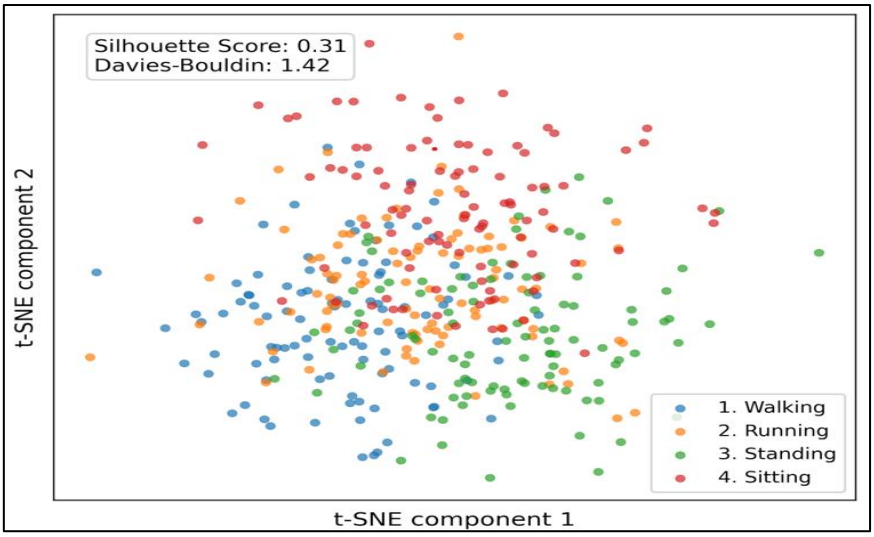

Figure 3. Visualization of t-SNE Analysis of the Baseline Feature Embeddings

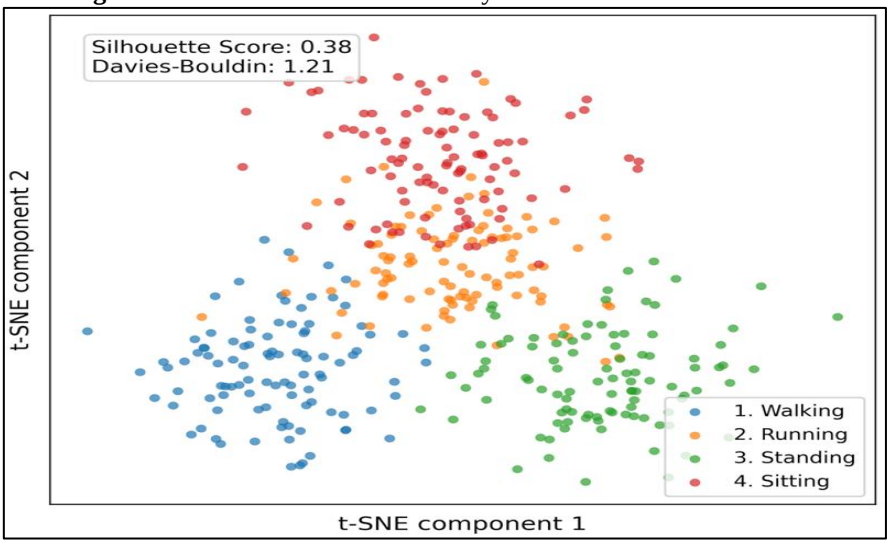

Figure 4. Visualization of t-SNE Analysis of the Harmonization-Only Model

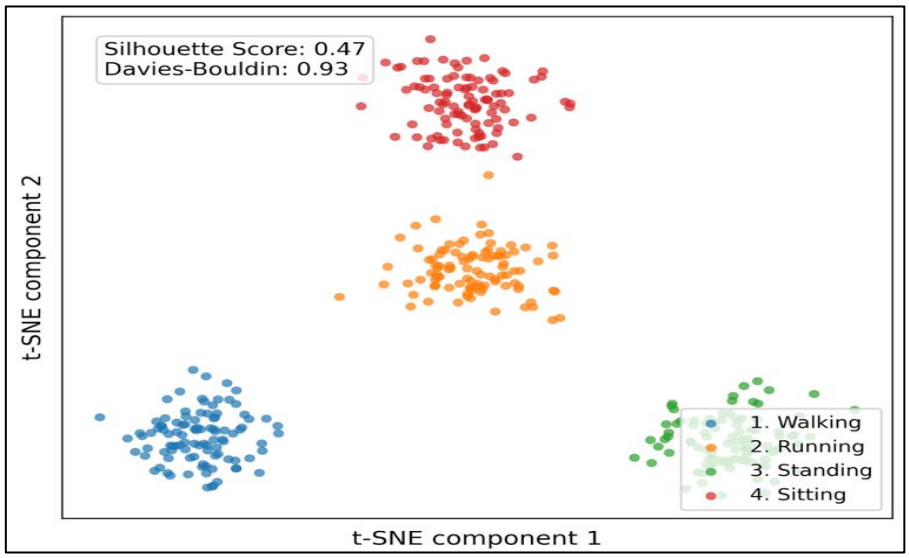

Figure 5. Visualization of t-SNE Analysis of the Proposed Approach

7.6 Gating Weight Distribution and Temporal Dynamics

To gain a deeper understanding of how the adaptive fusion strategy behaves, the distribution of gating coefficients is studied with respect to different levels of noise. Under standard evaluation conditions, the mean value of gating is about 0.51, meaning that there is balanced use of both modalities. However, under the influence of Gaussian noise ($\sigma = 0.1$) on the IMU signal, the value of the gating coefficient rises to 0.84. This change implies that the model adaptively decreases the weight of the degraded modality and depends more on the reliable modality.

Moreover, by studying the time behavior of gating weights, it can be noted that during dynamic activities like falling, the fusion scheme becomes aware of the context and utilizes more skeletal information. The above observations clearly indicates that the proposed adaptive weighting approach is robust against modality impairment, since it adjusts the contribution of each modality according to its performance, resulting in consistent performance even with significant degradation of modality inputs.

7.7 Failure Case and Error Analysis

A confusion matrix analysis is performed for the cross-dataset evaluation scenario. The analysis indicates that the most common misclassification occurs between walking and running, as they have similar periodic motion features. In the late fusion baseline, Figure 6, the confusion between walking and running is 14.7%. However, the proposed framework, Figure 7, reduces this confusion to 8.1%.

Static actions such as standing and sitting have almost perfect classification results. This reveals that the model has correctly learned the basic motion patterns. The results further show that the majority of the errors occur in the classification of fine-grained locomotion actions, which remain challenging due to differences in speed and style of movement. Based on these findings, cross-dataset generalization appears to be especially challenging when the activities have similar characteristics, there are differences in sensor positioning, and there are variations in motion dynamics. This creates confusion for the learned features and makes it difficult to differentiate between similar tasks.

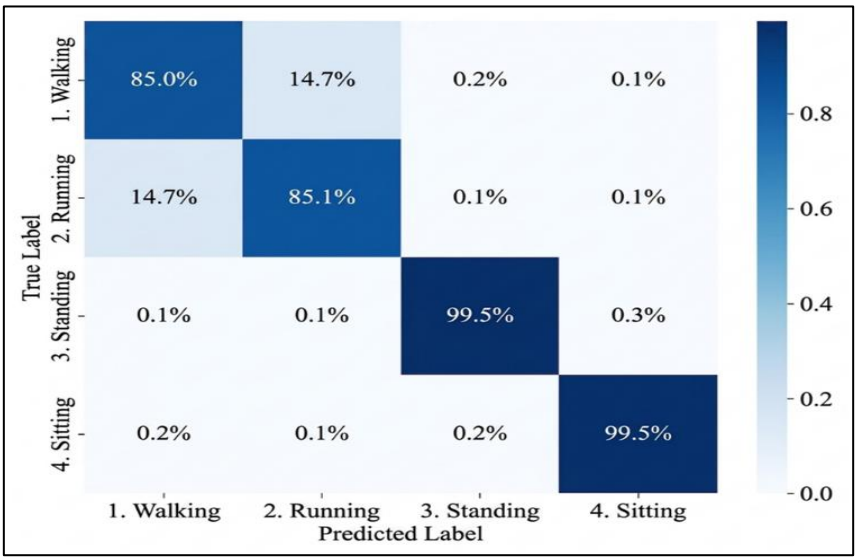

Figure 6. Confusion Matrix for Late Fusion Baseline Approach

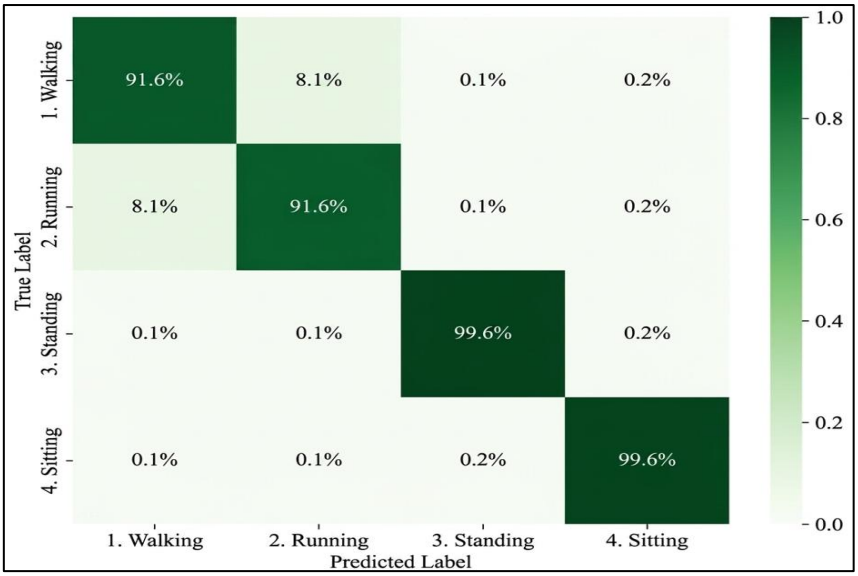

Figure 7. Confusion Matrix for the Proposed Framework

7.8 Cross-Modality Contribution Analysis

Lastly, the contribution of each modality is determined by conducting modality dropout experiments. It was observed that the absence of the skeleton modality caused a drop in the performance rate of 9.4%, while the removal of the IMU modality caused a decline of 5.7%. This clearly shows the effectiveness of skeletal data in providing structural information, whereas IMU data helps in learning dynamic motion. Another piece of evidence for the adaptive fusion process is provided by the variability of gating scores per activity.

8. Limitations

Although the proposed multimodal harmonization and adaptive gating technique achieves better recognition accuracy and minimizes the cross-dataset generalization gap, there are some limitations to this proposed technique. Firstly, the proposed technique is evaluated over datasets with limited locomotion and posture activities. Although these datasets are

commonly used in various HAR scenarios, they may not represent all the activities encountered in the real world. Therefore, future work should involve the evaluation of the proposed technique over larger activity taxonomies and various environmental conditions.

Secondly, the proposed harmonization technique assumes the availability of a deterministic temporal alignment technique that requires resampling heterogeneous sensor streams into a single frequency domain. Although this technique effectively handles sampling rate inconsistencies among different sensors, there is still a need to explore the incorporation of other temporal alignment models that could further enhance the robustness of the proposed technique in highly asynchronous sensing scenarios. Thirdly, the proposed adaptive gating technique is applied at the feature fusion level and estimates the modality reliability levels indirectly. Although the temporal weight trace analysis demonstrates the effectiveness of the proposed adaptive gating technique in response to changes in modality reliability levels, there is still a need to explore the incorporation of other probabilistic feature fusion models and Bayesian-based gating models that could provide more interpretable estimates of modality reliability levels. Finally, the proposed technique is limited to the fusion of skeleton and inertial sensing modalities.

However, in various ambient intelligence scenarios, other sensing modalities such as RGB video and depth sensors could provide more information about the context. Therefore, future work should involve extending the proposed technique to allow the integration of more than two sensing modalities. The proposed technique has demonstrated that deterministic-based multimodal harmonization and adaptive gating is an effective approach for improving cross-dataset robustness while maintaining computational efficiency.

9. Conclusion and Future Work

This paper proposes a multimodal fusion framework for HAR that addresses heterogeneous sensor alignment and multimodal fusion in the presence of domain shifts. The framework integrates deterministic temporal harmonization for cross-sensor alignment and an adaptive gating strategy for multimodal fusion. The proposed framework is evaluated using experiments on the NTU RGB+D 120, UTD-MHAD, and PAMAP2 datasets. The results show that the framework achieves 93.6% accuracy on the UTD-MHAD dataset and 89.6% accuracy on the NTU RGB+D 120 dataset, with a compact architecture comprising only 4.8M parameters and 2.1 GFLOPs. The performance gains are also consistent over multiple runs, as confirmed through statistical tests.

The framework is also evaluated using cross-dataset experiments, where it is trained on the UTD-MHAD dataset and evaluated on the PAMAP2 dataset. The results show that the framework reduces the generalization gap between the two datasets to 12.4%. The framework also outperforms other fusion strategies, such as late fusion, weighted fusion, and CCA-based fusion. Overall, the results demonstrated that the proposed framework can improve the accuracy of human activity recognition, enhance cross-dataset robustness, and increase computational efficiency, making it applicable in real-world scenarios. Future work will involve exploring other sensing modalities, learnable temporal alignment for asynchronous sensors, and uncertainty-aware multimodal fusion strategies.

References

[1] Baños Legrán, Oresti, Mate Attila Toth, Miguel Damas Hermoso, Héctor Emilio Pomares Cintas, and Ignacio Rojas Ruiz. "Dealing with the Effects of Sensor Displacement in Wearable Activity Recognition." (2014). 9995–10023.

[2] Bianchi, Valentina, Marco Bassoli, Gianfranco Lombardo, Paolo Fornacciari, Monica Mordonini, and Ilaria De Munari. "IoT Wearable Sensor and Deep Learning: An Integrated Approach for Personalized Human Activity Recognition in a Smart Home Environment." IEEE Internet of Things Journal 6, no. 5 (2019): 8553-8562.

[3] Bijalwan, Vishwanath, Abdul Manan Khan, Hangyeol Baek, Sangmin Jeon, and Youngshik Kim. "Interpretable Human Activity Recognition with Temporal Convolutional Networks and Model-Agnostic Explanations." IEEE Sensors Journal 24, no. 17 (2024): 27607-27617.

[4] Brinzea, Razvan, Bulat Khaertdinov, and Stylianos Asteriadis. "Contrastive Learning with Cross-Modal Knowledge Mining for Multimodal Human Activity Recognition." In 2022 International Joint Conference on Neural Networks (IJCNN), IEEE, 2022, 01-08.

[5] Chen, Chen, Roozbeh Jafari, and Nasser Kehtarnavaz. "UTD-MHAD: A Multimodal Dataset for Human Action Recognition Utilizing a Depth Camera and a Wearable Inertial Sensor." In 2015 IEEE International conference on image processing (ICIP), IEEE, 2015, 168-172.

[6] Chen, Kaixuan, Dalin Zhang, Lina Yao, Bin Guo, Zhiwen Yu, and Yunhao Liu. "Deep Learning for Sensor-Based Human Activity Recognition: Overview, Challenges, and Opportunities." ACM Computing Surveys (CSUR) 54, no. 4 (2021): 1-40.

[7] Chi, Hyung-gun, Myoung Hoon Ha, Seunggeun Chi, Sang Wan Lee, Qixing Huang, and Karthik Ramani. "Infogcn: Representation Learning for Human Skeleton-Based Action Recognition." In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, 20186-20196.

[8] Dhekane, Sourish Gunesh, and Thomas Ploetz. "Transfer Learning in Sensor-Based Human Activity Recognition: A Survey." ACM Computing Surveys 57, no. 8 (2025): 1-39.

[9] Dickens, James, and Pierre Payeur. "Multi-Modal Human Action Segmentation Using Skeletal Video Ensembles." Engineering Proceedings 58, no. 1 (2023): 30.

[10] Dong, Hao, Moru Liu, Kaiyang Zhou, Eleni Chatzi, Juho Kannala, Cyrill Stachniss, and Olga Fink. "Advances in Multimodal Adaptation and Generalization: From Traditional Approaches to Foundation Models." IEEE Transactions on Pattern Analysis and Machine Intelligence (2026). 5672-5691.

[11] Duan, H., Zhao, Y., Chen, K., Lin, D., & Dai, B. (2022). Revisiting Skeleton-based Action Recognition. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2959–2968. https://doi.org/10.1109/cvpr52688.2022.00298

[12] Geng, Xinyang, Hao Liu, Lisa Lee, Dale Schuurmans, Sergey Levine, and Pieter Abbeel. "Multimodal Masked Autoencoders Learn Transferable Representations." arXiv preprint arXiv:2205.14204 (2022).

[13] Guo, Pengyu, and Masaya Nakayama. "Towards User-Generalizable Wearable-Sensor-Based Human Activity Recognition: A Multi-Task Contrastive Learning Approach." Sensors 25, no. 22 (2025): 6988.

[14] Huang, Sipeng, Yang Chen, Dingchao Wu, Guangwei Yu, and Yong Zhang. "Few-Shot Learning for Human Activity Recognition Based on CSI." In 2022 Asia Conference on Algorithms, Computing and Machine Learning (CACML), IEEE, 2022, 403-409.

[15] Ijaz, Momal, Renato Diaz, and Chen Chen. "Multimodal Transformer for Nursing Activity Recognition." In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, 2065-2074.

[16] Jiang, Wenchao, and Zhaozheng Yin. "Human Activity Recognition Using Wearable Sensors by Deep Convolutional Neural Networks." In Proceedings of the 23rd ACM international conference on Multimedia, 2015, 1307-1310.

[17] Khan, Samee Ullah, Maryam Sultana, Sufyan Danish, Deepak Gupta, Norah Saleh Alghamdi, Suchang Woo, Dong-Gyu Lee, and Sangtae Ahn. "Multimodal Feature Fusion for Human Activity Recognition Using Human Centric Temporal Transformer." Engineering Applications of Artificial Intelligence 160 (2025): 111844.

[18] Kingma, Diederik P., and Jimmy Ba. "Adam: A Method for Stochastic Optimization." arXiv preprint arXiv:1412.6980 (2014). https://doi.org/10.48550/arxiv.1412.6980

[19] Le, Trung-Hieu, Thai-Khanh Nguyen, Trung-Kien Tran, Thanh-Hai Tran, and Cuong Pham. "Gaformer: Wearable Imu-Based Human Activity Recognition with Gramian Angular Field and Transformer." In 2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), IEEE, 2023, 297-303.

[20] Lee, James, and Suk-ju Kang. "Skeleton Action Recognition Using Two-Stream Adaptive Graph Convolutional Networks." In 2021 36th International Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC), IEEE, 2021, 1-3.

[21] Liu, Jiayang, Lin Zhong, Jehan Wickramasuriya, and Venu Vasudevan. "uWave: Accelerometer-Based Personalized Gesture Recognition and Its Applications." Pervasive and Mobile Computing 5, no. 6 (2009): 657-675.

[22] Liu, Jun, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C. Kot. "Ntu rgb+ d 120: A Large-Scale Benchmark For 3D Human Activity Understanding." IEEE transactions on pattern analysis and machine intelligence 42, no. 10 (2019): 2684-2701.

[23] Luo, Jinzhao, Lu Zhou, Guibo Zhu, Guojing Ge, Beiying Yang, and Jinqiao Wang. "Temporal-Channel Topology Enhanced Network for Skeleton-Based Action Recognition." In Chinese Conference on Pattern Recognition and Computer Vision (PRCV), Singapore: Springer Nature Singapore, 2023, 109-119.

[24] Mazzia, Vittorio, Simone Angarano, Francesco Salvetti, Federico Angelini, and Marcello Chiaberge. "Action Transformer: A Self-Attention Model for Short-Time Pose-Based Human Action Recognition." Pattern Recognition 124 (2022): 108487.

[25] Miah, Abu Saleh Musa, Yong Seok Hwang, and Jungpil Shin. "Sensor-Based Human Activity Recognition Based on Multi-Stream Time-Varying Features with Eca-Net Dimensionality Reduction." IEEE Access 12 (2024): 151649-151668.

[26] Miao, Shenghuan, and Ling Chen. "Goat: A Generalized Cross-Dataset Activity Recognition Framework with Natural Language Supervision." Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 8, no. 4 (2024): 1-28.

[27] Ordóñez, Francisco Javier, and Daniel Roggen. "Deep Convolutional and Lstm Recurrent Neural Networks for Multimodal Wearable Activity Recognition." Sensors 16, no. 1 (2016): 115.

[28] Paszke, Adam, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen et al. "Pytorch: An Imperative Style, High-Performance Deep Learning Library." Advances in neural information processing systems 32 (2019).) 8026–8037.

[29] Plizzari, Chiara, Marco Cannici, and Matteo Matteucci. "Skeleton-Based Action Recognition via Spatial and Temporal Transformer Networks." Computer Vision and Image Understanding 208 (2021): 103219.

[30] Qiuming, Liu, Chen Longping, Wang Da, Xiao He, Zhou Yang, and Wu Dong. "Decoupled 2S-AGCN Human Behavior Recognition Based on New Partition Strategy." In International Conference on Mobile Networks and Management, Cham: Springer Nature Switzerland, 2023, 70-82.

[31] Quan, Zhenzhen, Qingshan Chen, Wei Wang, Moyan Zhang, Xiang Li, Yujun Li, and Zhi Liu. "SMTDKD: A Semantic-Aware Multimodal Transformer Fusion Decoupled Knowledge Distillation Method for Action Recognition." IEEE Sensors Journal 24, no. 2 (2023): 2289-2304.

[32] Qureshi, Tayyab Saeed, Muhammad Haris Shahid, Asma Ahmad Farhan, and Sultan Alamri. "A Systematic Literature Review on Human Activity Recognition Using Smart Devices: Advances, Challenges, And Future Directions." Artificial Intelligence Review 58, no. 9 (2025): 276.

[33] Ray, Abhisek, and Mahesh Kolekar. "Skeleton-Based Action Recognition Using Graph Convolution and Cross-Domain Transfer Learning." In 2024 National Conference on Communications (NCC), IEEE, 2024, 01-06.

[34] Reiss, A. (2012). PAMAP2 Physical Activity Monitoring. UCI Machine Learning Repository, 10, C5NW2H. https://doi.org/10.24432/C5NW2H

[35] Shahverdi, Hossein, and Seyed Ghorashi. "Lightweight Transformer for Robust Human Activity Recognition Using Smartphone IMU Data." In 14th International Conference on Human Interaction and Emerging Technologies: Artificial Intelligence & Future Applications, IHIET-FS 2025, June 10-12, 2025, University of East London, London, United Kingdom., vol. 196, AHFE International, 2025, 238-248.

[36] Shi, Lei, Yifan Zhang, Jian Cheng, and Hanqing Lu. "Two-Stream Adaptive Graph Convolutional Networks for Skeleton-Based Action Recognition." In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, 12026-12035.

[37] Srivastava, Nitish, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. "Dropout: A Simple Way to Prevent Neural Networks from Overfitting." The journal of machine learning research 15, no. 1 (2014): 1929-1958.

[38] Subramanian, Shreyas, Bala Krishnamoorthy, and Pranav Murthy. "Dynamic Learning Rate Scheduling based on Loss Changes Leads to Faster Convergence." arXiv preprint arXiv:2512.14527 (2025).

[39] Tang, Yin, Qi Teng, Lei Zhang, Fuhong Min, and Jun He. "Layer-Wise Training Convolutional Neural Networks with Smaller Filters for Human Activity Recognition Using Wearable Sensors." IEEE Sensors Journal 21, no. 1 (2020): 581-592.

[40] Wang, Guanbo, Jiapeng Guo, Jiazhong Zhang, Xiangting Qi, and Hang Song. "Design of Human Action Recognition Method Based on Cross Attention and 2s-AGCN Model." In 2024 IEEE 6th International Conference on Civil Aviation Safety and Information Technology (ICCASIT), IEEE, 2024, 1341-1345.

[41] Wang, Jindong, Vincent W. Zheng, Yiqiang Chen, and Meiyu Huang. "Deep Transfer Learning for Cross-Domain Activity Recognition." In proceedings of the 3rd International Conference on Crowd Science and Engineering, 2018, 1-8.

[42] Wang, Xiaojuan, Tianqi Lv, Ziliang Gan, Mingshu He, and Lei Jin. "Fusion of Skeleton and Inertial Data for Human Action Recognition Based on Skeleton Motion Maps and Dilated Convolution." IEEE Sensors Journal 21, no. 21 (2021): 24653-24664.

[43] Wei, Jinfeng, Yunxin Wang, Mengli Guo, Pei Lv, Xiaoshan Yang, and Mingliang Xu. "Dynamic Hypergraph Convolutional Networks for Skeleton-Based Action Recognition." arXiv preprint arXiv:2112.10570 (2021).

[44] Xu, Cheng, Duo Chai, Jie He, Xiaotong Zhang, and Shihong Duan. "InnoHAR: A Deep Neural Network for Complex Human Activity Recognition." Ieee Access 7 (2019): 9893-9902.

[45] Yan, Sijie, Yuanjun Xiong, and Dahua Lin. "Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition." In Proceedings of the AAAI conference on artificial intelligence, vol. 32, no. 1. 2018.

[46] Yang, Jing, Tianzheng Liao, Jingjing Zhao, Yan Yan, Yichun Huang, Zhijia Zhao, Jing Xiong, and Changhong Liu. "Domain Adaptation for Sensor-Based Human Activity Recognition with a Graph Convolutional Network." Mathematics 12, no. 4 (2024): 556.

[47] Yang, Kyoung Ok, Junho Koh, and Jun Won Choi. "Unified Contrastive Fusion Transformer for Multimodal Human Action Recognition." arXiv preprint arXiv:2309.05032 (2023).

[48] Yao, Shuochao, Shaohan Hu, Yiran Zhao, Aston Zhang, and Tarek Abdelzaher. "Deepsense: A Unified Deep Learning Framework for Time-Series Mobile Sensing Data Processing." In Proceedings of the 26th international conference on world wide web, 2017, 351-360.

[49] Zhang, Yumin, and Yanyong Wang. "A Comprehensive Survey on RGB-D-Based Human Action Recognition: Algorithms, Datasets, and Popular Applications." EURASIP Journal on Image and Video Processing 2025, no. 1 (2025): 15.

[50] Zheng, Ce, Sijie Zhu, Matias Mendieta, Taojiannan Yang, Chen Chen, and Zhengming Ding. "3d Human Pose Estimation with Spatial and Temporal Transformers." In Proceedings of the IEEE/CVF international conference on computer vision, 2021, 11656-11665.

[51] Zhu, Guanzhou, Dong Zhao, Chunliang Li, Mingyue Zhao, Zhengyuan Zhang, Hefeng Quan, and Huadong Ma. "MASTER: A Multi-Modal Foundation Model for Human Activity Recognition." Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 9, no. 3 (2025): 1-26.

[52] Zhu, Yida, Haiyong Luo, Runze Chen, and Fang Zhao. "DiamondNet: A Neural-Network-Based Heterogeneous Sensor Attentive Fusion for Human Activity Recognition." IEEE Transactions on Neural Networks and Learning Systems 35, no. 11 (2023): 15321-15331.