frontiers | Frontiers in Computer Science

TYPE Original Research
PUBLISHED 12 August 2025
DOI 10.3389/fcomp.2025.1569205

OPEN ACCESS

EDITED BY
Kristof Van Laerhoven,
University of Siegen, Germany

REVIEWED BY
Laura Belli,
University of Parma, Italy
Md Yusuf Sarwar Uddin,
University of Missouri–Kansas City,
United States

*CORRESPONDENCE
Qingxin Xia
✉ qingxinxia@hkust-gz.edu.cn

RECEIVED 31 January 2025
ACCEPTED 26 June 2025
PUBLISHED 12 August 2025

CITATION
Ray LSS, Xia Q, Rey VF, Wu K and Lukowicz P (2025) Improving IMU based human activity recognition using simulated multimodal representations and a MoE classifier. *Front. Comput. Sci.* 7:1569205. doi: 10.3389/fcomp.2025.1569205

Improving IMU based human activity recognition using simulated multimodal representations and a MoE classifier

Lala Shakti Swarup Ray[1], Qingxin Xia[2]*, Vitor Fortes Rey[1,3], Kaishun Wu[2] and Paul Lukowicz[1,3]

[1]Embedded Intelligence Group, German Research Centre for Artificial Intelligence, Kaiseslautern, Germany, [2]Information Hub, Hong Kong University of Science and Technology, Guangzhou, China, [3]Department of Computer Science, Rhineland-Palatinate University of Technology Kaiserslautern-Landau, Kaiserslautern, Germany

The lack of labeled sensor data for Human Activity Recognition (HAR) has driven researchers to synthesize Inertial Measurement Unit (IMU) data from video, utilizing the rich activity annotations available in video datasets. However, current synthetic IMU data often struggles to capture subtle, fine-grained motions, limiting its effectiveness in real-world HAR applications. To address these limitations, we introduce Multi^{3}Net+, an advanced framework leveraging cross-modal, multitask representations of text, pose, and IMU data. Building on its predecessor, Multi^{3}Net, it uses improved pre-training strategies and a mixture of experts classifier to effectively learn robust joint representations. By leveraging refined contrastive learning across modalities, Multi^{3}Net+ bridges the gap between video and wearable sensor data, enhancing HAR performance for complex, fine-grained activities. Our experiments validate the superiority of Multi^{3}Net+, showing significant improvements in generating high-quality synthetic IMU data and achieving state-of-the-art performance in wearable HAR tasks. These results demonstrate the efficacy of the proposed approach in advancing real-world HAR by combining cross-modal learning with multi-task optimization.

KEYWORDS

HAR, sensor simulation, multi-modal learning, pretraining, MoE classifier

1 Introduction

Human Activity Recognition (HAR) using wearable devices has gained significant attention in various real-world applications, including healthcare (Inoue et al., 2019), manufacturing (Xia et al., 2020), and fitness (Ray et al., 2024; Czekaj et al., 2024). However, compared to fields like computer vision and natural language processing, HAR with wearable sensors has made slower progress in leveraging recent advancements in Deep Learning. This problem is largely due to the lack of large-scale, labeled datasets for sensor-based HAR tasks, which are readily available in fields like computer vision (e.g., ImageNet, COCO).

To address the limitation of lacking large-scale labeled datasets for sensor-based HAR tasks, several studies have investigated the synthesis of IMU data from video, capitalizing on the wealth of labeled activity data available in video formats. While this approach has proven useful for certain HAR tasks (Kwon et al., 2020; Rey et al., 2019; Moon et al., 2023), it faces significant challenges in accurately generating IMU data for fine-grained and subtle movements (Leng et al., 2024, 2023a). These types of movements, which are critical in real-world activities such as detailed tasks in manufacturing or playing a musical instrument, remain difficult to capture effectively using video-based methods. Existing studies about generation of synthetic IMU data from monocular video presents three major challenges:

(1) Inaccurate IMU data generation: current methods for generating synthetic IMU data from video often rely on kinetic detection and motion capture. However, these techniques can introduce errors due to factors like lighting variations, body shape differences, and occlusions, all of which limit the accuracy of IMU data generation, particularly for complex activities. Subtle motions, such as those involving the wrist, are especially challenging to simulate from monocular video due to the relatively small wrist's degrees of movements and representation in pixels.

(2) Unmodeled human attribute variations: monocular video capture is inherently limited by factors such as the shooting angle and occlusion, which prevents the task of accurately modeling consistent human attributes like height, gender, and body proportions in the video. These variations can lead to significant differences in the generated synthetic IMU data, especially when simulating fine-grained movements. Moreover, current methods are unable to capture IMU data of the same activity from individuals with different characteristics (e.g., different heights or body shapes) in a single video, which makes the cost of synthesizing IMU data very high, thus preventing the widespread application of synthetic IMU data technology.

(3) Loss of information from video to IMU Data: despite advancements in synthetic data generation, the synthetic IMU data cannot fully capture the motion details present in the video, leading to a loss of information. As a result, models trained with synthetic IMU data may not always demonstrate superior performance in HAR, particularly for complex, fine-grained activities.

In this paper, we propose Multi^{3}Net+, a novel multi-modal framework designed to improve the IMU based HAR with synthetic IMU data generated by video. To tackle the challenge of Inaccurate IMU Data Generation, we employ the Skinned Multi-Person Linear model (SMPL) (Loper et al., 2015), a highly effective tool for capturing complex human poses with exceptional fidelity. The SMPL model enables precise pose calibration, allowing us to generate high-quality synthetic IMU especially for fine-grained activities.

To address the issue of Unmodeled Human Attribute Variations, we apply data augmentation techniques to the human poses generated by SMPL. First, we introduce variations in key human attributes such as height, weight, and body proportions to ensure that human attributes remain consistent within each video. Next, we modify these attributes to efficiently generate a diverse set of IMU data from the source video, capturing a wide range of real-world variations in human characteristics.

To address the Loss of Information from Video to IMU Data, we preserve intermediate representations during the IMU generation process. In addition to using synthetic data for pretraining the network, we also incorporate latent representations from video descriptions and human poses. Specifically, we apply contrastive learning across video descriptions $\leftrightarrow$ pose, video descriptions $\leftrightarrow$ synthetic IMU, and pose $\leftrightarrow$ synthetic IMU. This contrastive framework helps the model learn joint representations across these diverse modalities, enabling the model to capture complex relationships between visual cues, human poses, and IMU data.

Finally, we fine-tune the pretrained model using a small amount of real IMU data for downstream HAR tasks. This fine-tuning step leverages the learned representations from the multimodal training phase, allowing the model to effectively transfer knowledge from synthetic data to real-world applications.

The key contributions of this paper are as follows:

(1) We introduce a multi-modal, multi-task approach which uses the contrastive learning to integrate video, pose, and synthetic IMU data, enabling joint representation learning for improved HAR performance. (2) We introduce a Mixture of Expert (MoE) downstream classifier built upon the predecessor work to further improve the HAR results. (3) We demonstrate the use of SMPL for generating high-fidelity synthetic IMU data, effectively addressing inaccuracies in current IMU generation methods along with a novel data augmentation strategy that accounts for human attribute variations, ensuring better generalization across diverse real-world scenarios. (4) We show how the proposed approach can effectively adapt to real-world IMU-based HAR tasks by validating it on three publicly available datasets and comparing it to two State of the Art (SoTA) IMU simulation methods that proves even with limited real data we can have superior performance, by leveraging pre-trained multimodal representations.

2 Related work

HAR using wearable sensors offers advantages such as reduced privacy concerns and lower energy consumption (Lyu et al., 2024) compared to video-based HAR, making it widely applicable for daily life activity recognition. However, wearable sensor-based HAR faces significant challenges (Bian et al., 2022), particularly in recognizing complex activities and adapting to professional domains such as nursing care and industrial activity monitoring, which have gained increasing attention in recent years. A primary limitation lies in the scarcity of labeled datasets, driven by the high costs associated with data collection and annotation, posing a significant obstacle to the development of robust and generalizable HAR models.

2.1 IMU simulation

To address the limitations of labeled data, synthetic IMU data generation has become a prominent approach. One technique involves leveraging videos to synthesize IMU data. For example, IMUTube utilizes kinetic 3D pose estimation models to track joint

movements from online videos and employs physical simulation (IMUSim) to convert these 3D poses into IMU data. However, due to challenges in generating high-quality IMU data, this approach has shown promising results primarily for simple and repetitive activities, such as dumbbell exercises, while its effectiveness diminishes for more complex activities. Xiao et al. (2021) and Multi^{3}Net (Fortes Rey et al., 2024) generate SMPL model (Loper et al., 2015) parameters from motion capture data to track body movements in videos, offering the advantage of capturing detailed movements compared to kinetic-based methods. However, these approaches fail to account for variations in human attributes, producing only a single IMU data stream corresponding to each video.

Another emerging approach leverages language-based cross-modality transfer models, such as T2M-GPT (Zhang J. et al., 2023), MotionDiffuse (Zhang et al., 2024), ReMoDiffuse (Zhang M. et al., 2023), and IMUGPT (Leng et al., 2023b, 2024), which generate 3D human movements from textual descriptions. These generated movements are subsequently converted into virtual IMU data streams. However, these motion synthesis models depend on datasets like HumanML3D (Guo et al., 2022), which lack diversity in body morphology. Consequently, their ability to generate complex activities not represented in the motion datasets is significantly limited.

Other recent studies, such as V2IMU (Santhalingam et al., 2023), aim to directly map video inputs to IMU outputs using supervised learning. However, the inherent modality gap and motion ambiguity between video and sensor data lead to decreased accuracy, especially in real-world tasks involving fine-grained or occluded motion.

2.2 Representation learning

In recent years, deep learning models have demonstrated the ability to transfer video data into IMU data directly (Santhalingam et al., 2023). However, the inherent domain gap between video and IMU data often results in poor performance in generating accurate IMU data especially when the activities are complex (Leng et al., 2024). To address this challenge, contrastive learning-based approaches have emerged as promising techniques for learning joint representations from different domains.

For instance, CLIP (Radford et al., 2021) aligns visual and text representations using paired images and text, achieving impressive generalization performance. Similarly, Moon et al. (2023) proposed a multimodal contrastive framework that aligns IMU data with text and video, projecting multimodal data into a unified representation space. Building on this, Yang et al. (2024) enhanced contrastive learning for text and IMU alignment by introducing a hierarchical temporal transformer to better capture important representations.

In wearable sensing, recent works such as IMUCLIP (Moon et al., 2023) and CoHAR (Keyvanpour et al., 2024) have extended these ideas by learning modality-invariant embeddings specifically tuned for activity recognition, demonstrating improved robustness under missing modalities and cross-subject generalization.

Inspired by these advancements, the previous approach Multi^{3}Net (Fortes Rey et al., 2024) leverages representation learning techniques to assist in IMU representation learning. Multi^{3}Net utilizes video representations, text representations, and synthetic IMU representations to enhance IMU-based HAR tasks. While the quality of synthetic IMU data may not yet match that of real IMU data due to information loss during the data generation process, representation learning helps bridge this gap.

In this study, we propose Multi^{3}Net+, which includes a novel data augmentation strategy for generating IMU data that accounts for human attribute variations. Additionally, we introduce a Mixture of Experts (MoE) classifier to further enhance the performance of downstream HAR tasks by automatically selecting important learned features. This architecture builds on previous contrastive learning pipelines by tightly integrating pose, text, and IMU branches in a joint training scheme, achieving state-of-the-art results on multiple benchmark datasets.

3 Background on key frameworks and components

To improve the readability of the paper, firstly we describe the key tools and frameworks involved in the Multi^{3}Net+ system as visualized in Figure 1. The overall pipeline begins with motion capture (MoCap) files, which are converted into 3D body meshes via SMPL and processed in Blender to standardize skeleton geometry. Data augmentation is applied to modify human body features such as height and limb proportions. The enhanced poses are then passed through the Orient3IMU model to generate virtual IMU signals. These multimodal representations (pose, text, IMU) are used to pretrain the Multi^{3}Net+ model, which is fine-tuned using a Mixture of Experts (MoE) classifier.

3.1 SMPL

SMPL (Loper et al., 2015) is a parametric 3D human body model that maps pose and shape vectors to a consistent mesh. It preserves bone-length consistency and anatomical fidelity, enabling accurate and structured motion representation. SMPL is used in the pipeline to extract realistic full-body poses for downstream IMU simulation.

3.2 Blender for skeleton transformation

Blender is a 3D graphics tool used to apply geometric normalization on SMPL skeletons. We align the global origin to the feet and standardize body scale (e.g., height set to 1.7 m) to ensure consistent coordinate systems across MoCap inputs.

3.3 Human attribute augmentation

To model inter-individual variability, we augment pose sequences by modifying SMPL shape parameters (e.g., limb length, shoulder width). This produces diverse human geometries

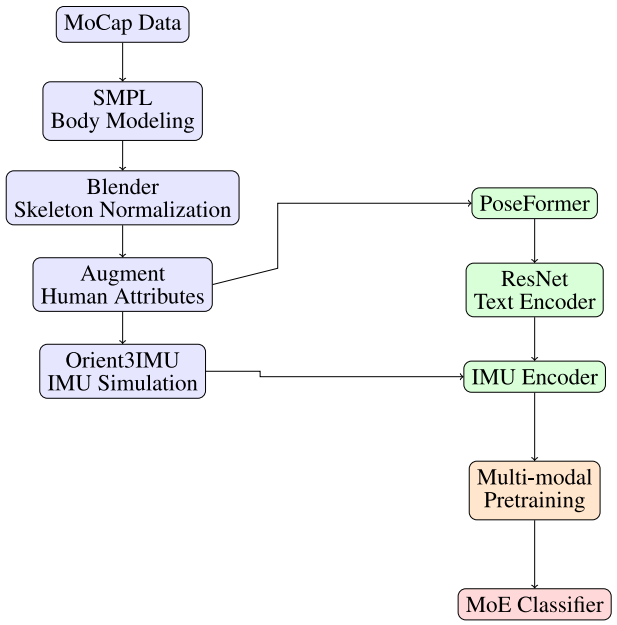

FIGURE 1
Overview of the Multi^{3}Net+ pipeline: starting from MoCap-based pose data, we generate synthetic IMU signals via SMPL modeling, geometric transformation, augmentation, and IMU simulation. The pose, text, and IMU modalities are encoded and pretrained in a multi-task contrastive framework, followed by fine-tuning with a MoE classifier.

while maintaining biomechanical plausibility, thereby improving generalizability in downstream tasks.

3.4 Orient3IMU and IMUSim

Orient3IMU is a module in IMUSim (Young et al., 2011) that computes local-frame acceleration and angular velocity from 3D pose data. We modify it by removing sensor noise and aligning the body frame using rotation matrices, producing high-fidelity IMU signals based on kinematic ground truth.

3.5 PoseFormer (pose encoder)

PoseFormer (Zheng et al., 2021) is a spatial-temporal transformer designed for human motion sequences. It applies attention mechanisms across both joints and frames, capturing fine-grained motion patterns. We use PoseFormer as the pose encoder in the contrastive learning setup.

3.6 ResNet (text encoder backbone)

A lightweight ResNet (He et al., 2016) backbone processes text embeddings to extract hierarchical features. Residual connections help stabilize training and ensure effective learning even in deep architectures.

3.7 Instructor (large) language model

Instructor (Su et al., 2022) is a pretrained language model that generates semantic embeddings from free-form video descriptions. These embeddings are used to align textual descriptions with IMU and pose modalities via contrastive learning.

3.8 MoE classifier

The MoE classifier (Shazeer et al., 2017) consists of multiple expert networks and a soft gating mechanism that dynamically routes samples to relevant experts. This architecture improves adaptability and generalization in downstream HAR tasks with diverse motion profiles.

4 Data generation

4.1 Simulation pipeline

By leveraging MoCap files, we can accurately compute both the linear acceleration and angular velocity of objects in motion. These calculations are made possible by tracking the precise positions and orientations of markers attached to the body over time. This meticulous tracking enables the generation of highly accurate data related to the movement dynamics of the body. Such data is invaluable for various applications that require a detailed understanding of human motion, especially in the context of synthetic IMU data generation.

While synthetic IMU data generated using IMUSim (Young et al., 2011) as employed in prior works such as IMUTube (Kwon et al., 2020) and IMUGPT (Leng et al., 2023c) is already a valuable resource, the proposed method significantly enhances it through multi-modal representation learning. Specifically, we align IMU, pose, and text modalities using contrastive learning to create a robust, semantically meaningful latent space. This is further regularized through Pose2IMU regression and IMU reconstruction tasks. Additionally, we introduce a Mixture of Experts (MoE) classifier that dynamically routes latent representations through specialized expert branches, improving adaptability across varied activity types. This integrated strategy helps close the domain gap between synthetic and real IMU data, resulting in better generalization and downstream performance even under partial modality conditions.

The proposed approach to generating synthetic IMU data draws inspiration from the Orient3IMU model, a component of IMUSim, but incorporates several modifications to improve the quality and consistency of the data. Notably, we exclude noise parameters from the IMUSim framework to ensure a cleaner, noise-free dataset. The process begins with MoCap motion data formatted in SMPL, a widely used human body model. Using Blender, we transform the skeletal data to generate a shape approximation corresponding to a human body with specific measurements, such as an average height of 1.7 m. This ensures uniformity across all MoCap files, which is critical for standardizing the dataset. Additionally, we reposition the skeleton's origin to align the center of the feet with the global origin (0,0,0) in terms of position and

orientation. This adjustment simplifies the downstream tasks of calculating movement and motion dynamics in a consistent and uniform reference frame.

Transforming linear acceleration to local coordinates considering Gravity:

$$\mathbf{a}_{\text{local}}(t) = \mathbf{R}_{\text{local}}(t) \cdot \left(\frac{d^2 \mathbf{r}_{\text{global}}(t)}{dt^2} - \mathbf{g}_{\text{global}} \right) \tag{1}$$

where $\mathbf{a}_{\text{local}}(t) \in \mathbb{R}^3$ is the linear acceleration of the rigid body in the local (body-fixed) coordinate system at time t. $\mathbf{R}_{\text{local}}(t) \in \mathbb{R}^{3\times3}$ is the rotation matrix that transforms vectors from the global coordinate system to the local coordinate system at time t. $\mathbf{r}_{\text{global}}(t) = [x(t), y(t), z(t)]^T \in \mathbb{R}^3$ is the position vector of the rigid body in the global coordinate system. $\frac{d^2 \mathbf{r}_{\text{global}}(t)}{dt^2} \in \mathbb{R}^3$ is the global linear acceleration, i.e., the second derivative of the global position vector with respect to time. $\mathbf{g}_{\text{global}} = [0, -9.8, 0]^T (\text{m/s}^2)$ is the gravitational acceleration vector in the global coordinate system, assuming gravity acts in the negative y-direction.

Similarly, after calculating global angular velocity from orientation, we transform it to local coordinates:

$$\boldsymbol{\omega}_{\text{local}}(t) = \mathbf{R}_{\text{local}}(t)^T \cdot \boldsymbol{\omega}_{\text{global}}(t) \tag{2}$$

where $\boldsymbol{\omega}_{\text{local}}(t)$ is the angular velocity of the rigid body in the local coordinate system.

The motivation of this study is to adopt a custom model instead of relying solely on IMUSim stems from certain limitations inherent in the IMUSim framework. One significant drawback is the absence of IMU calibration signals, which are essential for producing accurate and high-quality IMU data. Without these calibration signals, the generated data can deviate significantly from the expected range, resulting in inaccuracies and inconsistencies. By developing the proposed approach, we ensure greater uniformity and control over the generated dataset as visualized in Figure 2. For instance, we maintain consistent initial positions and orientations across all motion capture files and enforce uniform body dimensions, simplifying the neural network's task of learning meaningful correlations within the dataset.

Moreover, employing SMPL bodies for pose generation offers distinct advantages over traditional kinematic 3D pose estimation techniques. One notable benefit is that SMPL bodies maintain constant bone lengths, ensuring anatomical consistency across the dataset. Additionally, the SMPL model provides 3D joint angles, which are more informative than mere positional data. In contrast, kinematic 3D pose estimations often require the application of inverse kinematics to convert 3D poses into MoCap files. This additional step can introduce inaccuracies or lead to the loss of valuable information, as highlighted in the Vi2IMU paper. By using SMPL bodies, we circumvent these issues, preserving the fidelity and accuracy of the generated data.

In summary, the proposed approach leverages human motion capture data to generate high-quality IMU data that accurately captures the linear acceleration and angular velocity of objects in motion. By addressing the limitations of existing frameworks such as IMUSim or V2IMU and incorporating SMPL-based modeling, we achieve a standardized, consistent, and information-rich dataset that is well-suited for training neural networks and advancing synthetic IMU data generation as shown by the metrics in Table 1.

4.2 Source dataset

The How2Sign dataset (Duarte et al., 2021) contains over 80 h of sign language videos accompanied by corresponding transcripts, providing extensive information on hand and wrist movements. While How2Sign includes text annotations, it lacks ground-truth IMU labels, making it suitable for self-supervised multimodal pretraining. In this approach, pose sequences are extracted from the video and processed using SMPL and Orient3IMU to synthesize high-fidelity IMU signals. The associated text transcripts are not treated as class labels but are instead used as semantic anchors in a contrastive learning setup. This alignment of text, pose, and IMU modalities within a shared latent space effectively creates a pseudo-labeled IMU dataset. By leveraging these semantically enriched video–text pairs, we can train HAR models without requiring manually labeled IMU data. This enables robust joint representation learning across modalities and supports the transferability of the model to downstream IMU-based tasks. The generated synthetic IMU signals based on How2Sign are used for training the IMU encoder in a self-supervised manner. We have showcased How2Sign accompanying label and generated IMU signal in Figure 3.

The GRAB dataset (Taheri et al., 2020) consists of ∼4 h of MoCap data from subjects engaging in the action of grabbing various everyday objects. It includes contributions from 10 subjects interacting with 51 different objects and is also designated solely for pretraining. Unlike How2Sign, GRAB does not provide natural language descriptions, but it includes detailed 3D pose and hand-object interaction sequences and annotation cllasses. We use GRAB similarly like How2Sign for pre-training.

4.3 Simulated IMU signal fidelity

To check the fidelity of the simulated IMU Signals we use OpenPack dataset simulate the IMU signal and compute the magnitude of the acceleration and angular velocity vectors, we eliminate the dependency on the initial orientations of the real and simulated signals. This orientation-invariant method ensures that the comparison focuses solely on the overall behavior of the signals rather than their alignment in a specific coordinate system.

For the acceleration magnitude, $\|a\| = \sqrt{a_x^2 + a_y^2 + a_z^2}$, we capture the total acceleration experienced by the sensor, which accounts for all directional components. Similarly, for the angular velocity magnitude, $\|\omega\| = \sqrt{\omega_x^2 + \omega_y^2 + \omega_z^2}$, we measure the overall rotational speed regardless of axis orientation.

After calculating these magnitudes over all time steps in the signal, we evaluate the fidelity of the simulated signals against the real ones using orientation-invariant metrics such as Mean Squared Error (MSE). The MSE is computed as:

$$MSE = \frac{1}{N} \sum_{i=1}^{N} \left(\|x_{\text{sim},i}\| - \|x_{\text{real},i}\| \right)^2$$

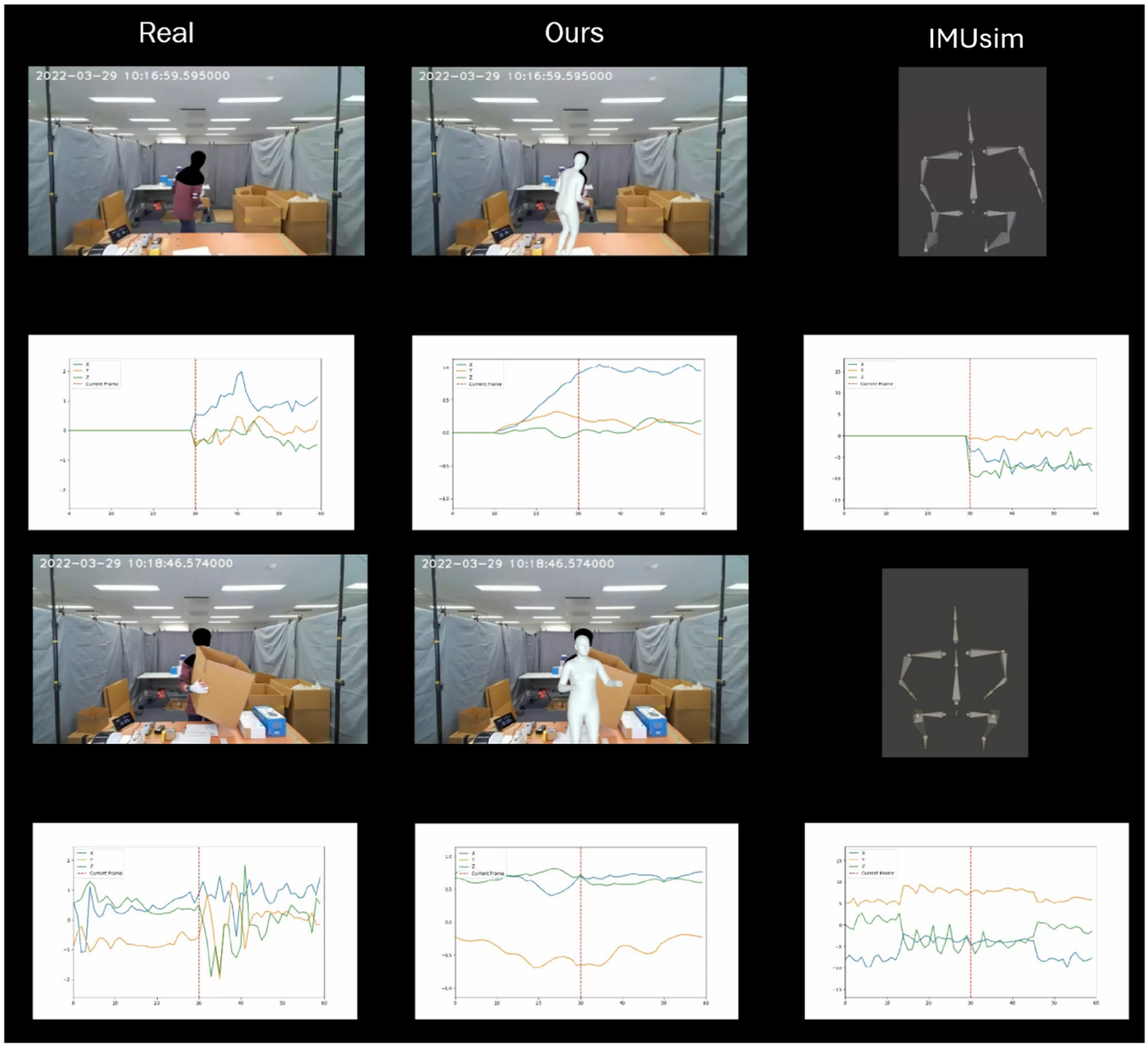

FIGURE 2
Comparison of synthetic IMU signals generated by the proposed method and IMUTube (IMUSim) on the OpenPack dataset. The proposed method produces signals that more closely resemble real sensor measurements, particularly in terms of motion fidelity and temporal dynamics.

TABLE 1 Fidelity of different IMU simulators tested on normalized IMU generated from 3D pose of MM-Fit dataset vs. the real collected IMU data on the wrist.

Simulator	$MSE_{\|a\|}$	$MSE_{\|\omega\|}$
IMUTube(IMUSim)	0.173 ± 0.018	0.244 ± 0.032
V2IMU	0.488 ± 0.021	0.479 ± 0.044
Proposed	$\mathbf{0.149 \pm 0.014}$	$\mathbf{0.312 \pm 0.037}$

Bold value depicts best performing model instance.

Here: - N represents the total number of time steps. - $\|x_{\text{sim},i}\|$ and $\|x_{\text{real},i}\|$ are the magnitudes of the simulated and real signals at the i-th time step.

By relying on the magnitudes and using this MSE-based evaluation, we can objectively assess the quality of the generated IMU signals without being influenced by variations in orientation, making it a robust method for evaluating simulated IMU data.

5 Multi^{3}Net+ architecture

5.1 Pre-training

After getting the pose and IMU data generated from video data, we then pretrain joint representations of text, pose, and IMU data via Multi^{3}Net+, which consists of three tasks (1) multi-modal contrastive learning, (2) Pose2IMU regression, and (3) IMU reconstruction.

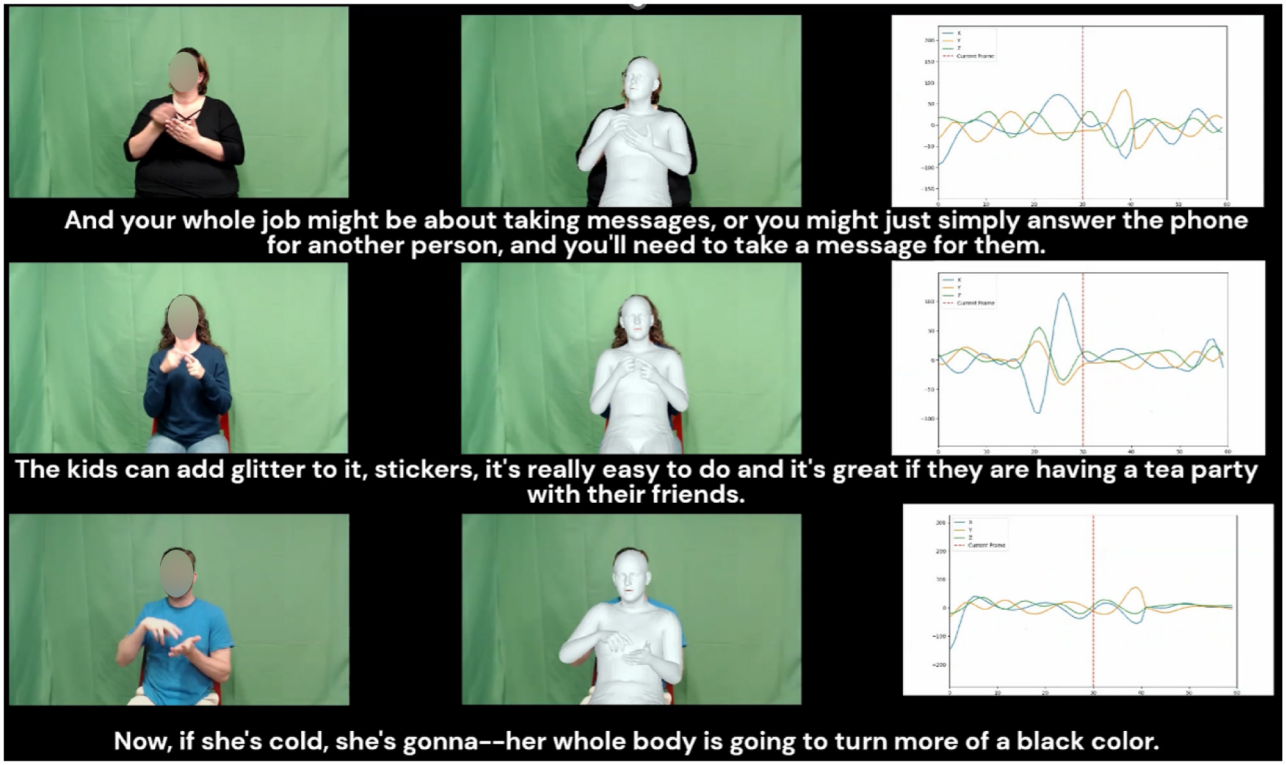

FIGURE 3
Pipeline for generating synthetic IMU signals from How2Sign videos. The process begins with extracting high-quality pose sequences using SMPL-based modeling, followed by simulation of IMU signals via Orient3IMU. The text annotations accompanying the videos are used as semantic anchors during contrastive pretraining, enabling the alignment of pose, text, and IMU representations. This multimodal pipeline supports the creation of pseudo-labeled IMU datasets from raw video-text inputs.

5.1.1 Multi-modal contrastive learning

As illustrated in Figure 4, the pretraining model comprises three encoders, each mapping text, pose, and IMU data to a respective latent space. Regarding the **Text encoder**, the input consists of the embedding of the text description of the corresponding video, derived from the output of the last hidden layer of a large pretrained model Instructor (Large) (Su et al., 2022). The output of the Text encoder is denoted as e_t. The encoder architecture is based on ResNet architecture with three residual blocks each containing a 1D CNN layer, followed by a batch normalization layer, and a Residual layer. In contrast to IMU2CLIP, where the text encoder is frozen to facilitate modality transitivity, in the proposed approach, the text encoder is trainable during pretraining to acquire joint representations for multi-modality data. Similar to the Text encoder, the **Pose encoder** takes the SMPL pose parameters of the body with (22, 3) tensor except for the two hand parameters, and both left and right hand as Mano parameters (30, 3) tensors to generate the output embedding of e_p. The pose encoder is based on the spatial-temporal transformer architecture of PoseFormer (Zheng et al., 2021) where each module is passed to a spatial attention block followed by a temporal attention block to generate intermediate embedding. For the **IMU encoder**, to facilitate adaptable processing across diverse scenarios, we utilize identical multi-headed attention blocks with positional embedding for data collected from both the left and right wrists. The input to the IMU encoder consists of synthetic data segments for each wrist, and the output comprises embedding for the left and right wrists, denoted as e_{sl} and e_{sr}, respectively. Although both encoders share identical architecture the learnable weights are different.

5.1.2 Pose2IMU regression

The Pose2IMU regression block consists of a Pose encoder and a Pose2IMU decoder, which has a CNN architecture with three ConvTranspose and Unpooling layers along with Batch normalization and Dropout blocks. Since activities in real scenarios typically involve fine-grained motions, the Pose2IMU decoder is designed to guarantee that the input pose encoder encompasses the required features by reconstructing the IMU data from the encoder. For the Pose2IMU decoder, the input is e_p, and the output is the predicted result of the corresponding synthetic IMU data, denoted as X_p. The decoder architecture is based on PSN from PresSim (Ray et al., 2023).

5.1.3 IMU reconstruction

Similar to the Pose2IMU regression block, the IMU reconstruction block comprises two IMU encoders and an IMU reconstructor. This IMU reconstructor features an identical CNN architecture to that of Pose2IMU but that takes the

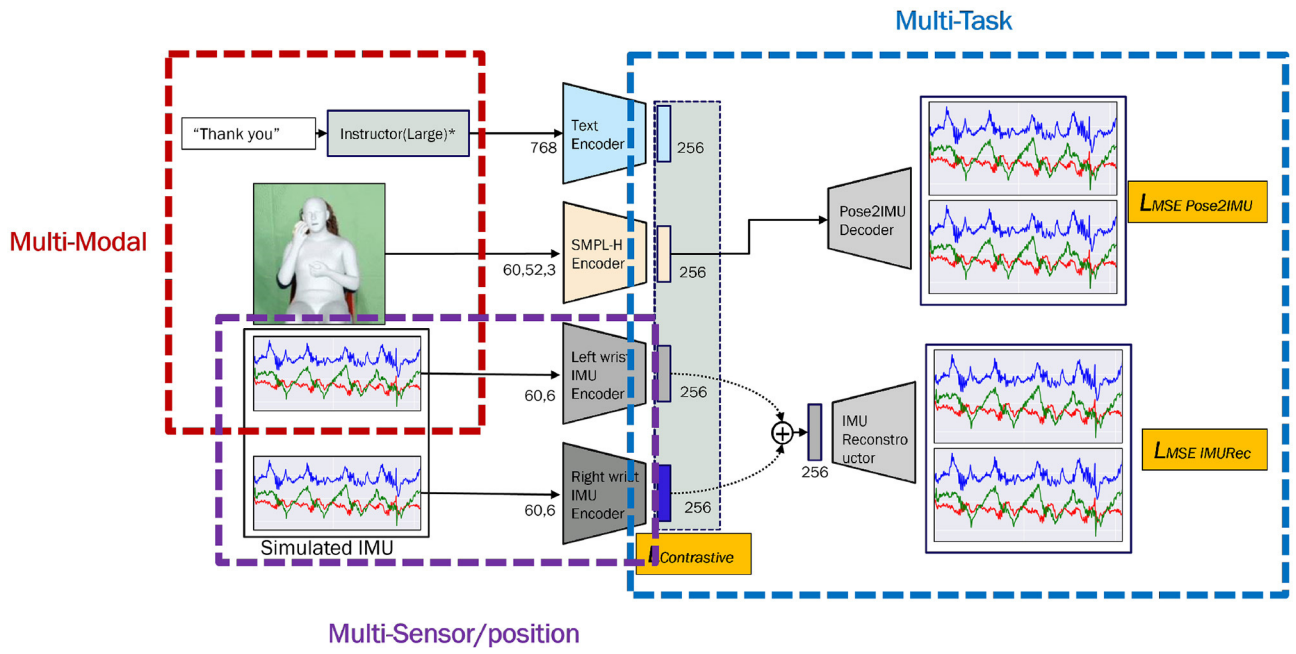

FIGURE 4
Overview of the Multi^{3}Net+ pretraining architecture. The model consists of three input modalities—text, pose, and synthetic IMU—encoded by three modality-specific encoders and supervised with two auxiliary tasks. **Text encoder (top)**: Textual descriptions from the Instructor (Large) model are fed into a lightweight ResNet-based encoder to produce a semantic embedding. Although this embedding captures contextual information from high-level descriptions (e.g., "a person grabs a box"), it is not used in the contrastive or reconstruction branches. This is because free-form natural language descriptions may not correspond to fine-grained sensor-level motion patterns, and enforcing strict alignment could degrade generalization. Instead, text features are used to enrich joint latent space representations. **Pose encoder (middle)**: SMPL-H body pose data is encoded via a PoseFormer-based encoder into a 256-dimensional latent embedding. This embedding is used in two downstream tasks: (1) as input to the **Pose2IMU decoder** to reconstruct IMU signals, and (2) for contrastive alignment with the IMU encodings. **IMU encoders (bottom)**: Separate encoders are used for left and right wrist synthetic IMU data. Their 256-dimensional outputs are used in two ways: (1) concatenated and passed to an **IMU reconstructor** to reconstruct the original IMU input, and (2) used in contrastive loss with the pose embedding.

concatenation of $[e_{sl}, e_{sr}]$ and uses one linear layer to map it back from (256, 2) to 256 vectors to reconstruct one frame input IMU instance. The predicted IMU data is denoted as X_s.

5.1.4 Loss function

To acquire joint representations for text, pose, and IMU from the encoders, we propose using instance discrimination by minimizing the InfoNCE loss (Oord et al., 2018) for each pair of encoders, which encourages similar representations for positive pairs closer and pushes representations of negative pairs apart, leading to meaningful feature representations, the InfoNCE loss is defined as follows:

$$\text{InfoNCE}(q,k) = -\frac{1}{N}\sum_{i=1}^{N}\log\left(\frac{e^{s(q_i,k_i)/\tau}}{e^{s(q_i,k_i)/\tau}+\sum_{j=1,j\neq i}^{N}e^{s(q_i,k_j)/\tau}}\right) \tag{3}$$

where N is the batch size, q_i and k_i are the representations of the i-th data sample under two different augmentations, $s(q_i, k_i)$ is the cosine similarity score between q_i and k_i, normalized by the temperature τ of 0.07. The cosine similarity score $s(q_i, k_i)$ is computed as:

$$s(q_i,k_i) = \frac{q_i \cdot k_i}{\|q_i\| \cdot \|k_i\|} \tag{4}$$

where $\cdot$ denotes the dot product and $\|\cdot\|$ denotes the L2 norm.

Thus, the overall InfoNCE loss of the pairs of encoders is formulated as follows:

$$\begin{aligned}\mathbf{L}_{Contrastive} = &\text{InfoNCE}(e_t,e_p) + \text{InfoNCE}(e_t,e_{sl})\\ &+\text{InfoNCE}(e_t,e_{sr}) + \text{InfoNCE}(e_p,e_{sl})\\ &+\text{InfoNCE}(e_p,e_{sr}) + \text{InfoNCE}(e_{sl},e_{sr})\end{aligned} \tag{5}$$

To ensure that the pose and IMU encoders preserve the necessary features to represent activities, MSE loss is applied for Pose2IMU regression and IMU reconstruction. Taking X_v as the synthetic IMU data, the MSE loss is defined as follows:

$$\begin{aligned}\mathbf{L}_{MSE} &= \mathbf{L}_{MSE\ Pose2IMU} + \mathbf{L}_{MSE\ IMURec}\\ &= \frac{1}{N}\sum_{j=1}^{N}\frac{1}{l}\sum_{i=t}^{t+l}(X_v^i - X_p^i)^2 + (X_v^i - X_s^i)^2\end{aligned} \tag{6}$$

where X_v^i represents the i-th synthetic IMU data point, X_p^i and X_s^i denote the prediction values of the pose2IMU and IMU2IMU decoders, respectively. l represents the segment length.

The overall loss for pertaining is $\mathbf{L}_{Contrastive} + \mathbf{L}_{MSE}$.

5.2 Downstream training

After pretraining the model using video resources containing rich hand movements, we can subsequently finetune the IMU

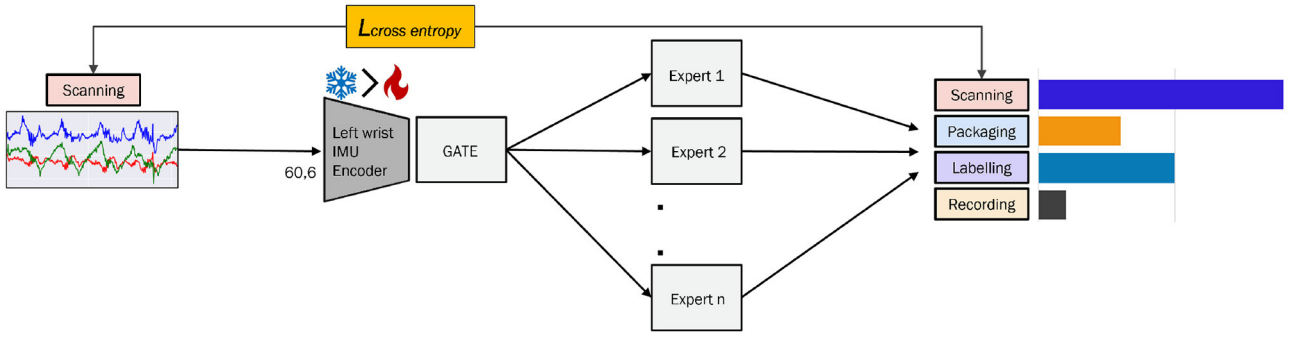

FIGURE 5
Downstream training of the HAR encoder using a MoE classifier on target data. The pretrained multimodal encoder is fine-tuned on specific HAR datasets, with the MoE module dynamically selecting the most relevant expert pathways based on input features. This enables adaptive representation learning and improves classification performance across diverse activity types.

encoder with a small amount of target IMU data. This process enables us to achieve robust HAR performance even with limited data.

The model structure consists of the pretrained IMU encoder and a classifier. To prevent the bottleneck problem, we take the intermediate output $\mathbf{H} \in \mathbb{R}^{6\times256}$ of the pretrained encoder instead of the final 1D feature $\mathbb{R}^{256}$ during pretraining.

We used a hybrid decoder where the classifier processes input feature $\mathbf{H} \in \mathbb{R}^{6\times256}$ through a sequence of neural network layers, starting with a convolutional layer:

$$\mathbf{H}_{\text{conv}} = \text{ReLU}(\text{BN}(\text{Conv1D}(\mathbf{H})))$$

where Conv1D(·) applies a 1D convolution, BN(·) denotes batch normalization, and ReLU(·) is the activation function.

The features are then reshaped and passed through two stages of multi-head attention mechanisms:

$$\mathbf{H}^{(i)}_{\text{attn}} = \text{MHA}(\text{LN}(\mathbf{H}^{(i-1)}_{\text{attn}})) + \mathbf{H}^{(i-1)}_{\text{attn}}, \quad i = 1, 2$$

where MHA(·) represents multi-head self-attention, and LN(·) is layer normalization. Each attention stage is followed by a feedforward transformation:

$$\mathbf{H}^{(i)}_{\text{ffn}} = \text{ReLU}(\text{LN}(\mathbf{W}_i\mathbf{H}^{(i)}_{\text{attn}} + \mathbf{b}_i)), \quad i = 1, 2$$

where $\mathbf{W}_i, \mathbf{b}_i$ are learnable parameters.

To enhance model expressivity and generalization, we introduce a MoE classifier with N experts as visualized in Figure 5. The MoE classifier learns a weighted combination of expert predictions, where the gating network determines the contribution of each expert. Given the final feature representation $\mathbf{H}^{(2)}_{\text{ffn}}$, the expert outputs are:

$$\mathbf{y}_j = f_j(\mathbf{H}^{(2)}_{\text{ffn}}), \quad j = 1, \ldots, N$$

where $f_j(\cdot)$ represents the j-th expert network. The gating network computes a softmax-weighted combination:

$$g_j = \frac{\exp(\mathbf{w}_j^\top \mathbf{H}^{(2)}_{\text{ffn}})}{\sum_{k=1}^{N} \exp(\mathbf{w}_k^\top \mathbf{H}^{(2)}_{\text{ffn}})}$$

where $\mathbf{w}_j$ are learnable parameters of the gating network. The final classification output is given by:

$$\mathbf{y} = \sum_{j=1}^{N} g_j \mathbf{y}_j$$

Finally, the predicted class is obtained by averaging across a specific dimension:

$$\hat{y} = \frac{1}{6}\sum_{i=1}^{6} \mathbf{y}_i$$

This MoE-based classifier allows for dynamic selection of relevant experts, improving robustness in HAR tasks with limited IMU data.

6 Evaluation

6.1 Datasets and evaluation metrics

We utilized two types of datasets: (1) large video datasets rich in hand activity representations for pretraining, which were described at Section 4.2, and (2) target inertial HAR datasets utilizing wrist IMUs, which will describe as follows. To ensure consistency across the datasets, all video data were resampled to 60 frames per second.

To clarify, both How2Sign and GRAB are multimodal datasets that provide pose data and either text annotations (How2Sign) or structured action labels (GRAB). In the proposed pipeline, we ensure that text, pose, and synthetic IMU signals are derived from the same source video within each dataset. That is, for each training sample, all modalities come from a single video in How2Sign or GRAB. We do not mix modalities across datasets for contrastive learning. Instead, we perform staged pretraining using one dataset at a time and optionally evaluate joint training in ablation studies. The goal is to build a generalizable multi-modal embedding space from diverse data sources, without requiring perfect sample-level alignment across datasets.

The MM-Fit dataset (Strömbäck et al., 2020) includes data from 10 subjects performing 10 different gym exercises. IMU data was captured using Mobvoi TicWatch Pro devices, sampled at 100 Hz,

providing detailed movement information from the participants' wrists. Furthermore, RGB video data was captured at 30 Hz to provide visual context for the exercises performed.

The OpenPack dataset (Yoshimura et al., 2024) features acceleration data collected from both the left and right wrists of 5 workers using an Empatica E4 wristband, with a sampling rate of 30 Hz. The data was gathered while the workers performed a packaging task involving 11 distinct activity classes. Additionally, the workers' activities were recorded on video to serve as ground truth.

The ALS-HAR dataset (Ray et al., 2025) has three distinct scenarios. In this experiment, we focus on the IMU data collected in an outdoor environment (scenario 3). The dataset includes data from three subjects, each of whom wore a Samsung Galaxy S20 smartphone on their left wrist, with a sampling rate of 30 Hz. The activities performed consist of six unique upper body fitness exercises, along with three additional hand-focused tasks, each lasting ∼20 min.

To evaluate the performance of the proposed model, we utilized the macro F1 score as the primary evaluation metric. We employed leave-one-user-out experiments for the downstream task. Subjects in every dataset is divided into five subsets, and the model is trained on four of these subsets while being validated on the remaining subset. This process is repeated five times with different random seeds.

6.2 Results

In this section, we present the Macro F1-scores obtained from experiments on three distinct datasets: OpenPack, MM-Fit, and ALS-HAR. Each dataset was evaluated using various model architectures and training strategies, including DCL, Base, IMU Reconstruction, Contrastive Pretraining, Multi^{3}Net, and Multi^{3}Net+.

DCL: This approach employs the DeepConvLSTM (DCL) architecture as the foundation for HAR tasks, which is widely used as the baseline method for IMU-based HAR. We use the content in parentheses to indicate what data is used for pre-training the model structure. "Real" means this model uses only the real IMU data from the target dataset for training. "Real + Synthetic IMUTube" means the model utilizes real and synthetic IMU data for training, with the synthetic data created using IMUTube. The process begins by extracting 2D skeletal poses from videos using AlphaPose, then mapping these 2D poses to 3D using VideoPose3D, and employing IMUSim to generate synthetic IMU data for specific body joints. Finally, the simulated IMU is calibrated using real IMU data to ensure a similar range of variability. "Real + Synthetic IMUGPT" means the model utilizes real and synthetic IMU data for training, with the synthetic data created using IMUGPT. Unlike IMUTube, IMUGPT utilizes ChatGPT to generate synthetic IMU data from activity word descriptions, making it more flexible in generating diverse activity data.

Base: We do the downstream training without any pretraining. The IMU encoder weights are initialized randomly. "Real", "Real + Synthetic IMUTube", and "Real + Synthetic IMUGPT" correspond to the training processes using real IMU data, real and synthetic IMU data created using IMUTube, and real and synthetic IMU data created using IMUGPT, respectively

IMU Reconstruction: In this method, only the IMU reconstruction model is applied for pretraining. The content in parentheses indicates which large video dataset (how2sign or GRAB) is used for pre-training the model structure and whether the pretrained model weights are frozen during fine-tuning of the downstream HAR tasks. "Frozen" means that the learned weights of the IMU encoder remain unchanged during fine-tuning, which focuses on measuring the quality of the features learned from IMU reconstruction. "Not frozen" means that the method also employs the IMU reconstruction model for pre-training, but the IMU encoder's weights are kept frozen during training until the loss stops decreasing and reaches the patience P for the first time; after that, the encoder is unfrozen to allow the classifier to learn.

TABLE 2 Macro F1-score for OpenPack dataset.

Model	Left wrist	Both wrists
DCL (real)	**43.3 ± 0.81**	**43.1 ± 0.50**
DCL (real + synthetic IMUTube)	42.5 ± 1.56	41.3 ± 1.48
DCL (real + synthetic IMUGPT)	38.4 ± 0.88	36.4 ± 1.31
Base (real)	33.8 ± 0.39	42.3 ± 0.25
Base (real + synthetic IMUTube)	35.3 ± 0.74	40.3 ± 0.89
Base (real + synthetic IMUGPT)	35.4 ± 0.79	37.2 ± 1.23
IMU reconstruction (how2sign:frozen)	33.2 ± 0.53	41.2 ± 0.33
IMU reconstruction (how2sign:not frozen)	39.7 ± 0.24	48.7 ± 0.45
Contrastive pretrain (how2sign:frozen)	39.4 ± 0.37	53.7 ± 0.18
Contrastive pretrain (how2sign:not frozen)	45.3 ± 0.18	58.2 ± 0.26
IMU reconstruction (GRAB:frozen)	31.4 ± 0.13	39.1 ± 0.19
IMU reconstruction (GRAB:not frozen)	37.3 ± 0.41	46.2 ± 0.20
Contrastive pretrain (GRAB:frozen)	40.2 ± 0.23	53.8 ± 0.53
Contrastive pretrain (GRAB:not frozen)	44.1 ± 0.26	57.2 ± 0.38
Multi^{3}Net (how2sign:frozen)	40.4 ± 0.17	54.2 ± 0.28
Multi^{3}Net (how2sign:not frozen)	47.3 ± 0.13	59.8 ± 0.27
Multi^{3}Net (GRAB:frozen)	41.2 ± 0.16	55.1 ± 0.34
Multi^{3}Net (GRAB:not frozen)	45.2 ± 0.28	58.3± 0.28
Multi^{3}Net (Both:frozen)	41.9 ± 0.28	56.4 ± 0.16
Multi^{3}Net (Both:not frozen)	48.4 ± 0.18	61.1 ± 0.39
Multi^{3}Net+ (how2sign:frozen)	41.4 ± 0.38	55.3 ± 0.19
Multi^{3}Net+ (how2sign:not frozen)	48.6 ± 0.28	60.6 ± 0.38
Multi^{3}Net+ (GRAB:frozen)	42.6 ± 0.48	56.7 ± 0.62
Multi^{3}Net+ (GRAB:not frozen)	46.3 ± 0.38	59.2 ± 0.67
Multi^{3}Net+ (Both:frozen)	42.6 ± 0.71	57.6 ± 0.58
Multi^{3}Net+ (Both:not frozen)	**49.6 ± 0.31**	**62.8 ± 0.28**

Bold value depicts best performing model instance.

Contrastive pretrain: only the multimodal contrastive model is applied for pre-training. "Frozen" means that all the encoders' learned weights remain frozen. "Not frozen" means that the multimodal contrastive model is applied for pre-training, but all the encoders' weights are kept frozen during training until the loss stops decreasing and reaches the patience P for the first time; after that, the encoder is unfrozen to allow the classifier to learn.

Multi^{3}Net: This method utilizes both IMU reconstruction and contrastive pre-training methods to learn joint representations. The large video datasets used for pre-training the model structure include either how2sign, GRAB, or both datasets. "Frozen" means that all the learned weights of the encoders are kept frozen during fine-tuning for downstream tasks. "Not frozen" means that in the downstream task, all encoder weights remain frozen during training until the loss stops decreasing and reaches the patience P for the first time; after that, the encoder is unfrozen to allow the classifier to learn.

Multi^{3}Net+: This is the proposed model structure that utilizes improved multi-task pretraining methods with a MoE classfier to train a better joint representation for downstream HAR tasks. Similar to the Multi^{3}Net approach, "Frozen" and "Not frozen" refer to whether the encoders' weights are always frozen during training for the downstream HAR tasks, respectively.

Table 2 displays the Macro F1-scores for the OpenPack dataset. The Multi^{3}Net+ model achieved the highest scores across both evaluation conditions, with a score of 49.6 $\pm$ 0.31 for the Left wrist and 62.8 $\pm$ 0.28 for Both wrists when the weights were not frozen. This demonstrates the model's superior capability in learning joint representations. In comparison, the standard Multi^{3}Net model also performed well, with scores of 47.3 $\pm$ 0.13 (Left wrist) and 59.8 $\pm$ 0.27 (Both wrists) that utilized how2sign dataset for pre-training. The improvement in Multi^{3}Net+ can be attributed to its enhanced pre-training strategy and the incorporation of a MoE classifier, which allows for more adaptive decision-making and better feature representation for multi-task model structures. The Contrastive pretrain method, particularly when using the how2sign dataset with weights not frozen, yielded scores of 45.3 $\pm$ 0.18 for the Left wrist and 58.2 $\pm$ 0.26 for Both wrists. This indicates that while

TABLE 3 Macro F1-score for MM-Fit dataset.

Model	Left wrist	Both wrists
DCL (real)	75.5 $\pm$ 2.53	75.8 $\pm$ 2.02
DCL (real + synthetic IMUTube)	75.6 $\pm$ 1.56	76.0 $\pm$ 2.35
DCL (real + synthetic IMUGPT)	78.8 $\pm$ 1.37	80.1 $\pm$ 2.18
Base (real)	85.2 $\pm$ 0.31	88.1 $\pm$ 0.57
Base (real + synthetic IMUTube)	83.4 $\pm$ 0.26	88.9 $\pm$ 0.25
Base (real + synthetic IMUGPT)	**86.4 $\pm$ 1.51**	**89.3 $\pm$ 0.54**
IMU reconstruction (how2sign:frozen)	75.6 $\pm$ 0.18	78.4 $\pm$ 0.41
IMU reconstruction (how2sign:not frozen)	82.7 $\pm$ 0.38	86.6 $\pm$ 0.22
Contrastive pretrain (how2sign:frozen)	80.7 $\pm$ 0.61	84.5 $\pm$ 0.33
Contrastive pretrain (how2sign:not frozen)	89.2 $\pm$ 0.74	93.5 $\pm$ 0.71
IMU reconstruction (GRAB:frozen)	77.2 $\pm$ 0.34	82.1 $\pm$ 0.68
IMU reconstruction (GRAB:not frozen)	83.5 $\pm$ 0.64	87.2 $\pm$ 0.38
Contrastive pretrain (GRAB:frozen)	80.5 $\pm$ 0.49	86.6 $\pm$ 0.53
Contrastive pretrain (GRAB:not frozen)	88.3 $\pm$ 0.61	90.4 $\pm$ 0.18
Multi^{3}Net (how2sign:frozen)	80.6 $\pm$ 0.18	86.4 $\pm$ 0.82
Multi^{3}Net (how2sign:not frozen)	91.0 $\pm$ 0.13	93.8 $\pm$ 0.29
Multi^{3}Net (GRAB:frozen)	82.3 $\pm$ 0.17	87.6 $\pm$ 0.21
Multi^{3}Net (GRAB:not frozen)	89.7 $\pm$ 0.17	92.0 $\pm$ 0.20
Multi^{3}Net (Both:frozen)	81.4 $\pm$ 0.81	86.3 $\pm$ 0.26
Multi^{3}Net (Both:not frozen)	91.2 $\pm$ 0.26	93.4 $\pm$ 0.07
Multi^{3}Net+ (how2sign:frozen)	82.1 $\pm$ 0.38	87.5 $\pm$ 0.62
Multi^{3}Net+ (how2sign:not frozen)	92.5 $\pm$ 0.28	95.1 $\pm$ 0.38
Multi^{3}Net+ (GRAB:frozen)	83.6 $\pm$ 0.48	88.7 $\pm$ 0.52
Multi^{3}Net+ (GRAB:not frozen)	90.7 $\pm$ 0.38	93.3 $\pm$ 0.37
Multi^{3}Net+ (Both:frozen)	82.6 $\pm$ 0.71	87.6 $\pm$ 0.58
Multi^{3}Net+ (Both:not frozen)	**92.9 $\pm$ 0.31**	**95.8 $\pm$ 0.28**

Bold value depicts best performing model instance.

TABLE 4 Macro F1-score for ALS-HAR dataset.

Model	Left wrist
DCL (real)	**72.2 $\pm$ 0.033**
DCL (real + synthetic IMUTube)	68.8 $\pm$ 0.040
DCL (real + synthetic IMUGPT)	63.4 $\pm$ 0.045
Base (real)	58.0 $\pm$ 0.025
Base (real + synthetic IMUTube)	59.8 $\pm$ 0.030
Base (real + synthetic IMUGPT)	57.2 $\pm$ 0.034
IMU reconstruction (how2sign:frozen)	56.3 $\pm$ 0.028
IMU reconstruction (how2sign:not frozen)	60.9 $\pm$ 0.030
Contrastive pretrain (how2sign:frozen)	60.4 $\pm$ 0.025
Contrastive pretrain (how2sign:not frozen)	66.0 $\pm$ 0.020
IMU reconstruction (GRAB:frozen)	53.2 $\pm$ 0.018
IMU reconstruction (GRAB:not frozen)	58.4 $\pm$ 0.022
Contrastive pretrain (GRAB:frozen)	62.2 $\pm$ 0.028
Contrastive pretrain (GRAB:not frozen)	65.5 $\pm$ 0.021
Multi^{3}Net (how2sign:frozen)	63.5 $\pm$ 0.022
Multi^{3}Net (how2sign:not frozen)	69.2 $\pm$ 0.018
Multi^{3}Net (GRAB:frozen)	64.5 $\pm$ 0.023
Multi^{3}Net (GRAB:not frozen)	67.9 $\pm$ 0.027
Multi^{3}Net (Both:frozen)	65.3 $\pm$ 0.026
Multi^{3}Net (Both:not frozen)	70.8 $\pm$ 0.019
Multi^{3}Net+ (how2sign:frozen)	66.8 $\pm$ 0.024
Multi^{3}Net+ (how2sign:not frozen)	72.4 $\pm$ 0.021
Multi^{3}Net+ (GRAB:frozen)	68.2 $\pm$ 0.031
Multi^{3}Net+ (GRAB:not frozen)	71.0 $\pm$ 0.029
Multi^{3}Net+ (Both: frozen)	68.9 $\pm$ 0.030
Multi^{3}Net+ (Both:not frozen)	**74.4 $\pm$ 0.025**

Bold value depicts best performing model instance.

the IMU reconstruction and Contrastive pretrain methods show promise, they do not reach the performance levels of Multi^{3}Net and Multi^{3}Net+.

Table 3 summarizes the results for the MM-Fit dataset. The Multi^{3}Net+ model again achieved the best results with a Macro F1-score of 92.9 $\pm$ 0.31 for the Left wrist and 95.8 $\pm$ 0.28 for Both wrists when weights were not frozen. The standard Multi^{3}Net model also performed admirably, scoring 91.0 $\pm$ 0.13 (Left wrist) and 93.8 $\pm$ 0.29 (Both wrists), but still lagged behind Multi^{3}Net+. The IMU Reconstruction and Contrastive pretrain methods showed a relatively lower scores as well, with 86.6 $\pm$ 0.53 and 93.5 $\pm$ 0.71 for Both wrists when weights were not frozen, respectively.

The results for the ALS-HAR dataset are presented in Table 4. Here, Multi^{3}Net+ achieved the highest Macro F1-score of 74.4 $\pm$ 0.025 when weights were not frozen, indicating its robustness across different datasets. The standard Multi^{3}Net model achieved a score of 69.2 $\pm$ 0.018, highlighting the advantages brought by the enhancements in Multi^{3}Net+. The IMU Reconstruction method, while exhibiting potential with a score of 60.9 $\pm$ 0.030 (not frozen), and Contrastive pre-training achieving 66.0 $\pm$ 0.020, did not perform as well as the Multi^{3}Net models, further emphasizing the effectiveness of the Multi^{3}Net+ architecture.

Across all three datasets, the Multi^{3}Net and Multi^{3}Net+ architectures consistently outperformed traditional models like DCL and Base, particularly when employing contrastive pretraining methods and unfrozen weights. The proposed Multi^{3}Net+ model, with its improved pretraining strategy and MoE classifier, demonstrated superior performance compared to the standard Multi^{3}Net when pre-trained with different large video datasets, indicating that leveraging advanced training strategies can significantly enhance model robustness and effectiveness in HAR tasks.

6.3 Analysis

6.3.1 Impact of proportion of pre-training datasets

In this section, we increase the amount of pre-training data across different large video datasets to measure its impact on downstream HAR performance. Figure 6 shows the macro F1-scores of OpenPack, MM-Fit, and ALS-HAR datasets, respectively.

Among the three datasets, the OpenPack dataset presents the lower overall F1-scores, which is above 60% with both wrists

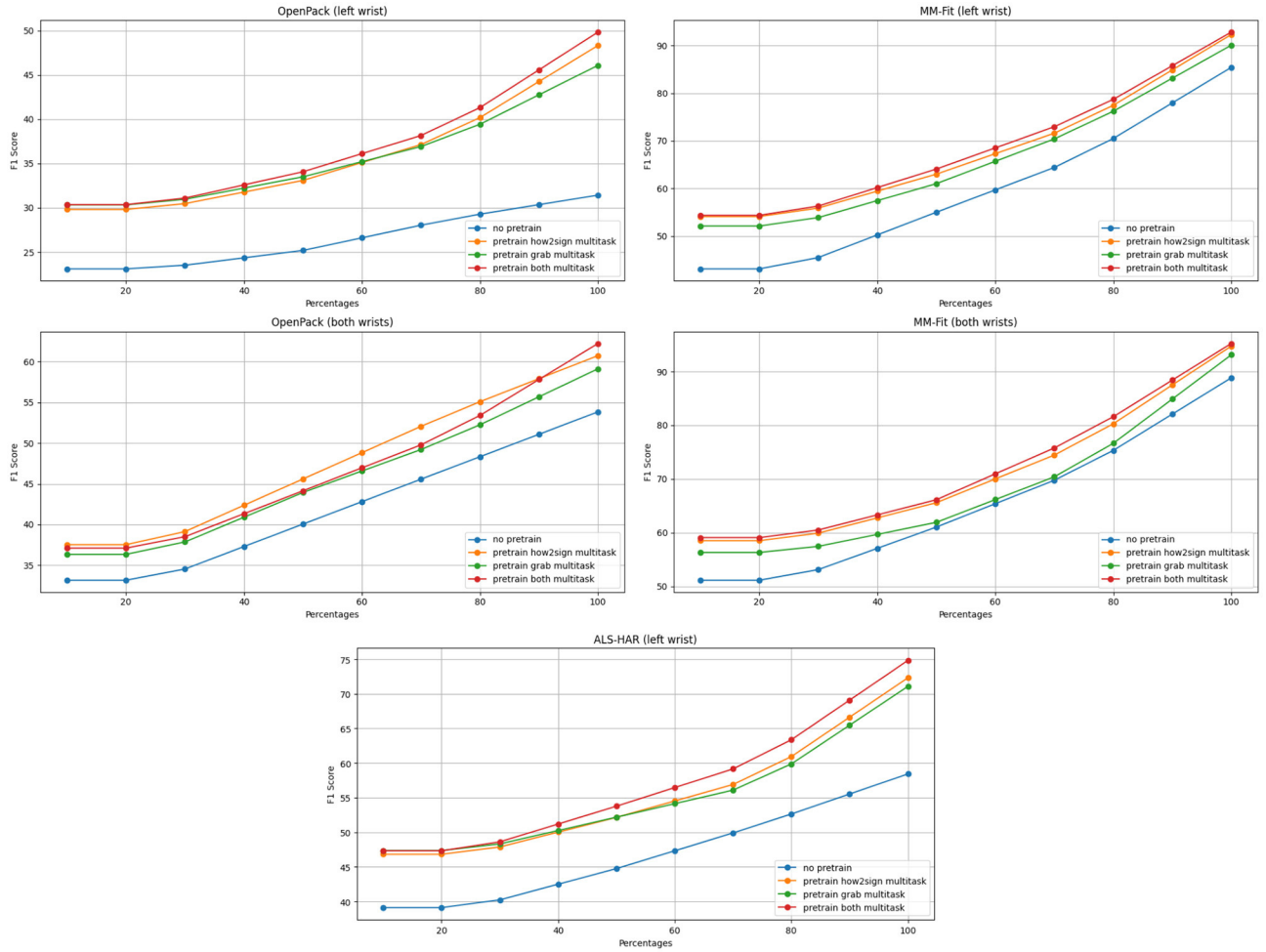

FIGURE 6
Macro F1-scores for three downstream HAR tasks under various pretraining settings as the percentage of available training data increases. The results demonstrate the effectiveness of contrastive pretraining and the MoE classifier in improving model performance, particularly under low-data regimes.

when using all large video datasets for pre-training. The fine-grained segmentation of industrial activities makes distinguishing between different classes more difficult, leading to smaller but still meaningful improvements from pretraining. While all pretrained models outperform the non-pretrained baseline. Interestingly, pretraining on the how2sign dataset alone provides comparable or slightly better results than the other strategies at lower percentages of pre-training data, both pre-training with both datasets ultimately achieves the best performance. This suggests that the benefits of multitask pretraining take effect more gradually for highly complex industrial tasks.

The MM-Fit and ALS-HAR datasets benefit considerably from multitask pre-training, with all pretrained models surpassing the baseline at all data percentages. This demonstrates the effectiveness of the amount of pre-training video datasets for downstream HAR tasks. For the MM-Fit dataset, which includes relatively simpler activities (as fitness-related activities are usually repetitive), the improvements become more obvious as more pre-training video data is available, increasing by more than 30% when the proportion of pre-training data is raised from 10% to 100%. This indicates that the learned features of the pretrained models are more suitable for simpler activities.

As for the OpenPack and MM-Fit datasets, the gap between the no pre-training method and other methods using video datasets for pre-training is smaller when using IMU data from both wrists for the downstream tasks. This result suggests that as the amount of real IMU data increases, the impact of the sythetic IMU data on the results gradually diminishes.

Overall, pre-training using video datasets is highly beneficial across all three HAR datasets, particularly with the both-multitask approach. However, the magnitude of improvement varies based on dataset complexity. For more structured activities (MM-Fit), pre-training provides a consistent boost. For highly fine-grained industrial activities (OpenPack), pretraining helps but does not completely bridge the complexity gap, indicating room for more domain-specific adaptations.

6.3.2 Impact of number of MoE experts

The impact of varying the number of experts on the macro F1-Score was evaluated across target HAR datasets, as summarized in Table 5. In the OpenPack (Left Wrist) dataset, the macro F1-Score rose from 48.4% with one expert to 49.6% with 16 experts, representing an improvement of ∼2.48%. In the OpenPack (Both Wrists) dataset, the score increased from 61.1% with one expert to 62.8% with 16 experts, reflecting an improvement of about 2.79%. For the MM-Fit dataset, the score from one expert to 16 experts improved 1.87 and 2.55%, respectively. In the ALS-HAR (Left Wrist) dataset, the score increased from 70.8% with one expert to 74.4% with 16 experts, representing an improvement of about 5.68%. Overall, increasing the number of experts consistently enhances model accuracy across all datasets, with the most significant performance gains observed when using data from both wrists. This suggests that both wrist data is particularly effective for downstream task classification when a MoE classifier is applied. While the trend of improvement continues with more experts, the gains appear to diminish as the number of experts increases, especially for datasets that already have high initial scores (e.g., MM-Fit dataset). This suggests a point of saturation where adding more experts may yield diminishing returns, indicating an optimal number of experts that balances costs with model performance.

6.4 Limitations

Despite the demonstrated benefits of pretraining and MoE classifiers, the proposed approach has several limitations. First, the cost of generating large-scale synthetic IMU data remains high, which may limit scalability for broader applications. Second, while pretraining improves performance across all datasets, its impact on highly complex industrial HAR tasks, such as OpenPack, is relatively modest, suggesting the need for more domain-specific adaptations. Lastly, the performance of models trained on synthetic IMU data is still lower compared to real IMU data, indicating a domain gap that requires further refinement in data generation or adaptation techniques.

TABLE 5 Impact of number of experts on the performance for different datasets.

Dataset	Experts	Macro F1 score
OpenPack (Left wrist)	1	48.4 ± 0.18
	2	48.9 ± 0.21
	4	49.2 ± 0.31
	8	49.5 ± 0.17
	16	**49.6 ± 0.31**
OpenPack (Both wrists)	1	61.1 ± 0.39
	2	62.0 ± 0.11
	4	62.5 ± 0.31
	8	62.7 ± 0.27
	16	**62.8 ± 0.28**
MM-Fit (Left wrist)	1	91.2 ± 0.26
	2	91.8 ± 0.15
	4	92.5 ± 0.25
	8	**92.9 ± 0.28**
	16	**92.9 ± 0.31**
MM-Fit (Both wrists)	1	93.4 ± 0.07
	2	94.1 ± 0.15
	4	94.9 ± 0.52
	8	95.5 ± 0.16
	16	**95.8 ± 0.28**
ALS-HAr (Left Wrist)	1	70.8 ± 0.19
	2	72.5 ± 0.25
	4	73.3 ± 0.09
	8	74.1 ± 0.28
	16	**74.4 ± 0.25**

Bold value depicts best performing model instance.

7 Conclusion

In this study, we proposed Multi^{3}Net+ a multimodal framework that generates high-fidelity synthetic IMU data and learns semantically rich representations by aligning text, pose, and IMU through contrastive pretraining. We addressed the limitations of prior approaches by integrating SMPL-based motion modeling, skeleton normalization, and human attribute augmentation to improve the quality and generalizability of synthetic data. Furthermore, our use of a MoE classifier enables adaptive feature selection, resulting in improved downstream HAR performance.

Our experiments demonstrate that leveraging large-scale video-text datasets for pretraining allows the model to effectively learn transferable representations, even in the absence of real IMU data. This approach achieves state-of-the-art results across multiple public benchmarks (OpenPack, MM-Fit, ALS-HAR), especially when applied to structured activity domains. We found that multitask contrastive pretraining consistently yields robust representations, and that the impact of pretraining is most prominent on structured datasets, while still providing meaningful gains in more complex domains. Performance scales with the number of MoE experts, though benefits plateau as complexity increases.

Despite these achievements, challenges remain in bridging the domain gap between synthetic and real IMU signals and in optimizing pretraining for highly diverse activity sets. Future work will explore domain adaptation techniques and further refinements to simulation fidelity to push HAR performance even further.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Author contributions

LR: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. QX: Conceptualization, Formal analysis, Investigation, Methodology, Supervision, Writing – original draft, Writing – review & editing. VR: Investigation, Methodology, Supervision, Visualization, Writing – review & editing. KW: Funding acquisition, Writing – review & editing. PL: Funding acquisition, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This work was supported by the German Federal Ministry of Education and Research (BMBF) under the CrossAct project (01IW25001), Guangdong Provincial Key Lab of Integrated Communication, Sensing and Computation for Ubiquitous Internet of Things (No. 2023B1212010007), China NSFC Grant (U2001207 and 62472366), the Project of DEGP (Nos. 2024GCZX003, 2023KCXTD042, and 2021ZDZX1068), and G01RF000200.

Acknowledgments

We acknowledge the use of ChatGPT, for aiding in text editing and rephrasing during the writing of the paper while following the guidelines provided by Frontiers.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Generative AI statement

The author(s) declare that Gen AI was used in the creation of this manuscript. We verify and take full responsibility for the use of generative AI in the preparation of this manuscript which is used for aiding in text editing and rephrasing during the writing.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Bian, S., Liu, M., Zhou, B., and Lukowicz, P. (2022). The state-of-the-art sensing techniques in human activity recognition: a survey. *Sensors* 22:4596. doi: 10.3390/s22124596

Czekaj, L., Kowalewski, M., Domaszewicz, J., Kitlowski, R., Szwoch, M., and Duch, W. (2024). Real-time sensor-based human activity recognition for efitness and ehealth platforms. *Sensors* 24:3891. doi: 10.3390/s24123891

Duarte, A., Palaskar, S., Ventura, L., Ghadiyaram, D., DeHaan, K., Metze, F., et al. (2021). "How2sign: a large-scale multimodal dataset for continuous American sign language," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Nashville, TN: IEEE), 2735–2744. doi: 10.1109/CVPR46437.2021.00276

Fortes Rey, V., Ray, L. S. S., Xia, Q., Wu, K., and Lukowicz, P. (2024). "Enhancing inertial hand based har through joint representation of language, pose and synthetic

IMUS," in *Proceedings of the 2024 ACM International Symposium on Wearable Computers* (New York, NY: ACM), 25–31. doi: 10.1145/3675095.3676609

Guo, C., Zou, S., Zuo, X., Wang, S., Ji, W., Li, X., et al. (2022). "Generating diverse and natural 3D human motions from text," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (New Orleans, LA: IEEE), 5152–5161. doi: 10.1109/CVPR52688.2022.00509

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* 770–778.

Inoue, S., Lago, P., Hossain, T., Mairittha, T., and Mairittha, N. (2019). Integrating activity recognition and nursing care records: The system, deployment, and a verification study. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3:1244. doi: 10.1145/3351244

Keyvanpour, M. R., Mehrmolaei, S., Shojaeddini, S. V., and Esmaeili, F. (2024). HAR-CO: a comparative analytical review for recognizing conventional human activity in stream data relying on challenges and approaches. *Multimedia Tools Appl.* 83, 40811–40856. doi: 10.1007/s11042-023-16795-8

Kwon, H., Tong, C., Haresamudram, H., Gao, Y., Abowd, G. D., Lane, N. D., et al. (2020). Imutube: automatic extraction of virtual on-body accelerometry from video for human activity recognition. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4:11841. doi: 10.1145/3411841

Leng, Z., Bhattacharjee, A., Rajasekhar, H., Zhang, L., Bruda, E., Kwon, H., et al. (2024). Imugpt 2.0: language-based cross modality transfer for sensor-based human activity recognition. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8:545. doi: 10.1145/3678545

Leng, Z., Jain, Y., Kwon, H., and Ploetz, T. (2023a). "On the utility of virtual on-body acceleration data for fine-grained human activity recognition," in *Proceedings of the 2023 ACM International Symposium on Wearable Computers, ISWC '23* (New York, NY: Association for Computing Machinery), 55–59. doi: 10.1145/3594738.3611364

Leng, Z., Kwon, H., and Ploetz, T. (2023b). "Generating virtual on-body accelerometer data from virtual textual descriptions for human activity recognition," in *Proceedings of the 2023 ACM International Symposium on Wearable Computers, ISWC '23* (New York, NY, USA: Association for Computing Machinery), 39–43. doi: 10.1145/3594738.3611361

Leng, Z., Kwon, H., and Plötz, T. (2023c). "Generating virtual on-body accelerometer data from virtual textual descriptions for human activity recognition," in *Proceedings of the 2023 ACM International Symposium on Wearable Computers*, 39–43.

Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., and Black, M. J. (2015). Smpl: a skinned multi-person linear model. *ACM Trans. Graph.* 34:2818013. doi: 10.1145/2816795.2818013

Lyu, S., Chen, Y., Duan, D., Jia, R., and Xu, W. (2024). "EarDA: towards accurate and data-efficient earable activity sensing," in *2024 IEEE Coupling of Sensing & Computing in AIoT Systems (CSCAIoT)* (Hong Kong: IEEE), 1–7. doi: 10.1109/CSCAIoT62585.2024.00005

Moon, S., Madotto, A., Lin, Z., Saraf, A., Bearman, A., and Damavandi, B. (2023). "IMU2CLIP: language-grounded motion sensor translation with multimodal contrastive learning," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, eds. H. Bouamor, J. Pino, and K. Bali (Singapore: Association for Computational Linguistics), 13246–13253. doi: 10.18653/v1/2023.findings-emnlp.883

Oord, A., Li, Y., and Vinyals, O. (2018). Representation learning with contrastive predictive coding. *arXiv* [Preprint] arXiv:1807.03748. doi: 10.48550/arXiv.1807.03748

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., et al. (2021). "Learning transferable visual models from natural language supervision," in *Proceedings of the 38th International Conference on Machine Learning, Volume 139 of Proceedings of Machine Learning Research*, eds. M. Meila, and T. Zhang (PMLR), 8748–8763.

Ray, L. S. S., Geißler, D., Liu, M., Zhou, B., Suh, S., and Lukowicz, P. (2025). "Als-har: harnessing wearable ambient light sensors to?enhance imu-based human activity recognition," in *Pattern Recognition*, eds. A. Antonacopoulos, S. Chaudhuri, R. Chellappa, C.-L. Liu, S. Bhattacharya, and U. Pal (Cham: Springer Nature Switzerland), 133–147. doi: 10.1007/978-3-031-78110-0_9

Ray, L. S. S., Zhou, B., Suh, S., Krupp, L., Rey, V. F., and Lukowicz, P. (2024). "Text me the data: generating ground pressure sequence from textual descriptions for har," in *2024 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops)* (Biarritz: IEEE), 461–464. doi: 10.1109/PerComWorkshops59983.2024.10503379

Ray, L. S. S., Zhou, B., Suh, S., and Lukowicz, P. (2023). "Pressim: an end-to-end framework for dynamic ground pressure profile generation from monocular videos using physics-based 3D simulation," in *2023 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops)* (Atlanta, GA: IEEE). doi: 10.1109/PerComWorkshops56833.2023.10150221

Rey, V. F., Hevesi, P., Kovalenko, O., and Lukowicz, P. (2019). "Let there be imu data: generating training data for wearable, motion sensor based activity recognition from monocular rgb videos," in *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers, UbiComp/ISWC '19 Adjunct* (New York, NY, USA: Association for Computing Machinery), 699–708. doi: 10.1145/3341162.3345590

Santhalingam, P. S., Pathak, P., Rangwala, H., and Kosecka, J. (2023). Synthetic smartwatch imu data generation from in-the-wild asl videos. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 74, 1–34. doi: 10.1145/3596261

Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G., et al. (2017). Outrageously large neural networks: the sparsely-gated mixture-of-experts layer. *arXiv preprint* arXiv:1701.06538.

Strömbäck, D., Huang, S., and Radu, V. (2020). Mm-fit: multimodal deep learning for automatic exercise logging across sensing devices. *Proc. ACM Inter. Mobile, Wear. Ubiquit. Technol.* 4, 1–22. doi: 10.1145/3432701

Su, H., Shi, W., Kasai, J., Wang, Y., Hu, Y., Ostendorf, M., et al. (2022). One embedder, any task: instruction-finetuned text embeddings. *arXiv* [Preprint]. arXiv:2212.09741. doi: 10.4855/arXiv.2212.09741

Taheri, O., Ghorbani, N., Black, M. J., and Tzionas, D. (2020). "Grab: a dataset of whole-body human grasping of objects," in *Computer Vision-ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16* (cham: Springer), 581–600. doi: 10.1007/978-3-030-58548-8_34

Xia, Q., Korpela, J., Namioka, Y., and Maekawa, T. (2020). Robust unsupervised factory activity recognition with body-worn accelerometer using temporal structure of multiple sensor data motifs. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4:3411836. doi: 10.1145/3411836

Xiao, F., Pei, L., Chu, L., Zou, D., Yu, W., Zhu, Y., et al. (2021). "A deep learning method for complex human activity recognition using virtual wearable sensors," in *Spatial Data and Intelligence*, eds. X. Meng, X. Xie, Y. Yue, and Z. Ding (Cham: Springer International Publishing), 261–270. doi: 10.1007/978-3-030-69873-7_19

Yang, X., Yao, C., and Ban, X. (2024). "Spatial-related sensors matters: 3D human motion reconstruction assisted with textual semantics," in *Proceedings of the AAAI Conference on Artificial Intelligence* (New York, NY: ACM), 10225–10233. doi: 10.1609/aaai.v38i9.28888

Yoshimura, N., Morales, J., Maekawa, T., and Hara, T. (2024). "Openpack: a large-scale dataset for recognizing packaging works in iot-enabled logistic environments," in *2024 IEEE International Conference on Pervasive Computing and Communications (PerCom)* (Biarritz: IEEE), 90–97. doi: 10.1109/PerCom59722.2024.10494448

Young, A. D., Ling, M. J., and Arvind, D. K. (2011). "Imusim: a simulation environment for inertial sensing algorithm design and evaluation," in *Proceedings of the 10th ACM/IEEE International Conference on Information Processing in Sensor Networks* (Nw York, NY: ACM), 199–210.

Zhang, J., Zhang, Y., Cun, X., Zhang, Y., Zhao, H., Lu, H., et al. (2023). "Generating human motion from textual descriptions with discrete representations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Vancouver, BC: IEEE), 14730–14740. doi: 10.1109/CVPR52729.2023.01415

Zhang, M., Cai, Z., Pan, L., Hong, F., Guo, X., Yang, L., et al. (2024). Motiondiffuse: text-driven human motion generation with diffusion model. *IEEE Trans. Pattern Anal. Mach. Intell.* 46, 4115–4128. doi: 10.1109/TPAMI.2024.3355414

Zhang, M., Guo, X., Pan, L., Cai, Z., Hong, F., Li, H., et al. (2023). "Remodiffuse: retrieval-augmented motion diffusion model," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (Paris: IEEE), 364–373. doi: 10.1109/ICCV51070.2023.00040

Zheng, C., Zhu, S., Mendieta, M., Yang, T., Chen, C., and Ding, Z. (2021). "3D human pose estimation with spatial and temporal transformers," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (Montreal, QC: IEEE), 11656–11665. doi: 10.1109/ICCV48922.2021.01145