ψ-ViT: Human Activity Recognition using Auxiliary Tasks-Enhanced Video Transformers

by

Kirill Kirillov

Submitted to the Department of Computer Science
in partial fulfillment of the requirements for the degree of

Master of Science in Data Science

at the

NAZARBAYEV UNIVERSITY

June 2025

Author .....
Department of Computer Science
11/05/2025

Certified by .....
Adnan Yazici
Chair, Department of Computer Science
Thesis Supervisor

Accepted by .....
Yelyzaveta Arkhangelsky
Dean, School of Engineering and Digital Sciences

ψ-ViT: Human Activity Recognition using Auxiliary Tasks-Enhanced Video Transformers

by

Kirill Kirillov

Submitted to the Department of Computer Science
on 11/05/2025, in partial fulfillment of the
requirements for the degree of
Master of Science in Data Science

Abstract

Human Activity Recognition (HAR) is a critical task in healthcare, enabling the goal of emergency detection and prevention without human supervision by employing IoT devices and machine learning techniques. While traditional unimodal approaches to HAR often fall short in accurately recognizing complex or subtle activities, multimodal systems integrating data from sensors such as accelerometers, gyroscopes, video, and audio provide richer context and higher accuracy. This work introduces Pose- and Sensor-Induced Video Transformer (ψ-ViT) framework that enhances HAR performance by inducing motion sensor data through auxiliary learning tasks during training, while maintaining vision-only inference efficiency. Building on the principles of the Pose Induced Video Transformer (π-ViT), our methodology extends auxiliary task learning to gyroscope and accelerometer modalities by introducing induction modules. Experiments demonstrate that combining these modules with a video transformer backbone improves recognition of fine-grained human activities by up to 7%, particularly for subtle motions, thus advancing HAR systems toward practical healthcare deployment without requiring wearable sensors during real-world use.

Thesis Supervisor: Adnan Yazici
Title: Chair, Department of Computer Science

Acknowledgments

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Advances in commodity sensors, embedded systems, portable and wearable devices, all encompassed under the term of Internet of Things (IoT), are translating into transformations in the healthcare sector. There is an especially big potential for the IoT devices and machine learning models to ensure the well-being of elderly, disabled, or otherwise vulnerable people, without the need for constant human attention. Higher efficiency of care can be achieved through early disease detection [25], enhanced drug discovery [4], or human activity recognition.

Human Activity Recognition (HAR) is one of the fundamental challenges in this domain. Human activity spans a wide range of activities, yet in the context of healthcare, and, specifically, emergency detection, actions like falling or seizures are of utmost interest. Earlier, unimodal HAR approaches, which rely on a single type of sensor data, such as video or gyroscope data, are less powerful at recognizing complex activities or differentiate between very similar activities. Fine-grained activity detection benefits from multiple sources of data due to different correlations between data and respective actions [5].

Multimodal HAR combines data from various sources – accelerometers, gyroscopes, video, audio, etc. – to capture as much subtleties of the actions performed. At the same time, deep learning models, particularly convolutional neural networks (CNNs) and recurrent neural networks (RNNs), have proven their power at extracting important, abstract features of human activity from various modalities of data [23].

Recent advancements transformer architectures have demonstrated even better performance at the task of HAR without sacrificing on the computational power [31].

However, traditional approaches to multimodal HAR incorporate fusion-based methods, which use the available modalities both during training and inference. In most deployment settings, collection of all modalities of data in deployment is not feasible, and the fusion-based approaches are not suitable for real-time applications. Recent works, such as [26], explore the idea of training the video vision transformer with 2D and 3D pose information to learn better linear mappings from video to tokens that capture features that are more insightful for HAR. This thesis work extends this idea by incorporating sensor data from accelerometer and gyroscope sensors into the video transformer.

1.1 Motivation

The application of multimodal HAR in healthcare, especially for the monitoring of vulnerable populations, can be life-saving. For instance, accurately detecting falls in elderly patients can lead to rapid medical intervention, while monitoring for seizures in patients with epilepsy can prevent serious injury. However, currently existing video-based approaches are not feasible for deployment due to relatively low reliability when compared to multimodal architectures. At the same time, higher accuracy achieved with multimodality comes at additional computational costs, rendering real-time deployment impractical.

This research explores the effectiveness of auxiliary tasks utilizing sensor data for vision transformer fine-tuning, improving both the reliability and usability of HAR systems.

Chapter 2

Related works

The earliest promising works on HAR involved using data from sensors, such as accelerometers, gyroscopes, and orientation sensors, using traditional machine learning (ML) methods. Approaches that showed the best results at the time include autoregressive (AR) models, Support Vector Machines (SVMs), small Artificial Neural Networks (ANNs) [13, 16], Random Forests (RFs), Hidden Markov Models (HMMs) [5], and hierarchical combinations of those [15].

Unimodal approaches, data processing, and feature extraction methods have evolved with time. Deep learning (DL) has become the dominant paradigm in the field of ML, and DL-based approaches have shown promising results in HAR. Convolutional Neural Networks (CNNs), for example, have been adapted for use with sensor data that has been pre-processed in a form of multi-channel images [23], achieving better results than the previous, simpler ML methods could. On the other hand, [32] shows the importance of using data augmentation and processing techniques: by leveraging windowing and feature generation, a RF model achieves an impressive 99.97% accuracy at the HAR task using inertial sensor data. Although the number of activities and their complexity in this study is relatively small, the results show that sensor data conveys highly relevant information when it comes to certain human activities.

As the volume and variety of data increased, multimodal approaches have grown in popularity, especially in a field related to HAR - human emotion recognition (HER). One of the most common tasks in this field is emotion recognition in video, and the

effectiveness of using audio and text modalities, albeit both extracted from video, has been shown to be powerful in Recurrent Neural Networks (RNNs)-based approaches [6, 21]. However, the recent developments in transformer-based models lead the researchers to adopt the use of transformers, which are less computationally expensive than CNNs for feature extraction from images and RNNs for sequence processing, while keeping the same performance characteristics-based tasks [31].

Similarly, transformer-based approaches have been adapted to HAR tasks and have outperformed CNN-RNN-based approaches in terms of accuracy and speed, making them suitable for deployment on modern mobile devices [18, 11].

Transformers have also been extended to use multiple modalities of data. A very recent work [12] proposes a novel human activity recognition method based on Vision Transformer (ViT). The proposed method integrates enhanced Adaptive Graph Convolutional Layer (eAGCL) in a Two-Stream Adaptive Graph Convolutional Network (2s-AGCN) to ViT, which allows the model to process spatio-temporal data (3D skeleton) effectively [12]. Another prominent approach is π-ViT (Pose-Induced Video Transformer) [26], which is similarly designed to utilize one extra modality other than video. π-ViT augments RGB-based video transformers with 2D and 3D pose information, addressing the challenge of visually similar activities and viewpoint variations in Activities of Daily Living (ADL). However, there is a significant improvement from the model proposed in [12]: the derived 2D and 3D skeletons of humans are incorporated into video transformers, refining the learned representations during training, while removing the pose-induction modules during inference, since extracting the pose information adds computational overhead [26].

Table 2.1 summarizes the prior literature in terms of the model and data modality used. Moreover, it illustrates that no work has tried to include different kinds of modalities - accelerometer and gyroscope sensor data are similar in nature, and so are the video and pose modalities.

The next section describes how we incorporate sensor and visual data modalities into a single video transformer.

Methods	**MM†**	**Modality**			
		Acc	**Gyro**	**Video**	**Pose**
Sensor-based Approaches					
NB [29]		✓			
SVM [13]		✓			
ANN [16]		✓			
Hierarchical (AR, ANN) [15]		✓			
CNN [23]		✓			
Multimodal Sensor Approaches					
RF, HMM [5]	✓	✓	✓		
RF, CNN [32]		✓	✓		
CNN [27]	✓	✓			
Vision-based Approaches					
ViT-ReT [31]				✓	
ViT [18]				✓	
Multimodal Vision Approaches					
ViT [26]	✓			✓	✓
ViT [12]	✓			✓	✓
Proposed ViT	✓	✓	✓	✓	✓

Table 2.1: Comparison of HAR methods across different modalities and approaches. MM† indicates multimodal approach (utilizes two or more modalities concurrently). Bold text indicates the proposed method. ✓indicates modality is used, empty cells indicate modality is not used.

Chapter 3

Methodology

Building upon the recently proposed Pose-Induced Video Transformer (π-ViT) [26] framework, we develop new auxiliary modules that infuse motion sensor information into video transformers without requiring these signals during inference. Our experimental results demonstrate significant improvements in recognizing fine-grained activities of daily living through this approach.

3.1 Background: Video Transformers and Auxiliary Learning

3.1.1 TimeSformer and Divided Space-Time Attention

Video transformers have emerged as powerful architectures for action recognition, with TimeSformer [3] demonstrating that pure transformer-based models can outperform convolutional architectures. The key innovation in TimeSformer is the divided space-time attention mechanism, shown in Figure 3-2, which factorizes self-attention into separate spatial and temporal operations.

In this approach, each transformer block first computes temporal attention by comparing each token with all tokens at the same spatial location across different frames. This is followed by spatial attention, where each token attends to all tokens from the same frame. This factorization reduces complexity from $\mathcal{O}((n_t \cdot n_h \cdot n_w)^2)$

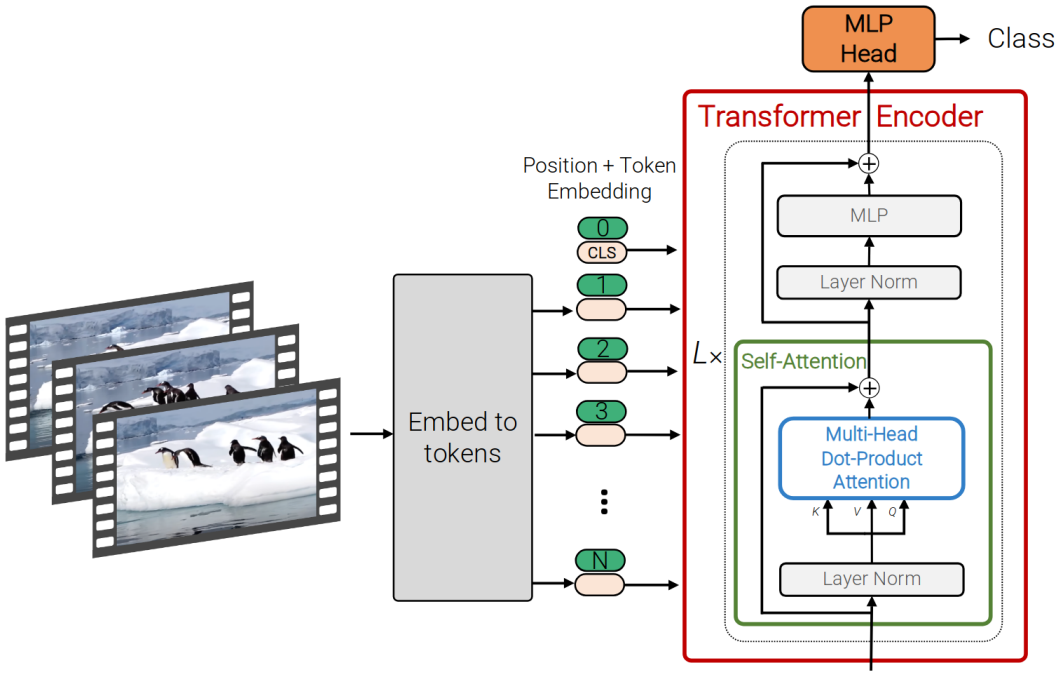

Figure 3-1: Video transformer architecture example (ViViT [2]). This architecture processes video by dividing it into spatio-temporal tokens that are then passed through transformer layers.

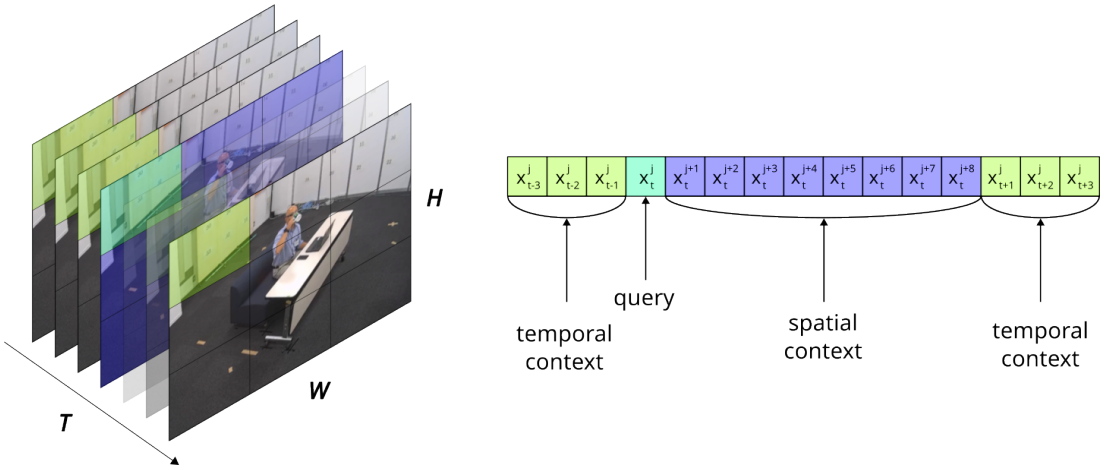

Figure 3-2: Divided space-time attention mechanism from TimeSformer [3], where temporal and spatial attention are computed separately within each transformer block. This design maintains computational efficiency while effectively modeling spatio-temporal relationships.

to $\mathcal{O}((n_h \cdot n_w)^2 + n_t^2)$, where n_t, n_h, and n_w represent the temporal, height, and width dimensions of the input tokens, respectively. This enables efficient processing of spatio-temporal data while still capturing complex relationships across both space and time.

Our work adopts the TimeSformer architecture with its divided space-time attention as the backbone for our model, providing a strong foundation for video understanding that we enhance through our auxiliary learning approach.

3.1.2 The π-ViT Framework

The π-ViT framework addresses key challenges in Activities of Daily Living (ADL) recognition by augmenting RGB-based video transformers with human pose information through an innovative approach that requires poses only during training.

ADL recognition presents unique challenges compared to general action recognition tasks due to its requirements for fine-grained appearance discrimination, view-invariance, and fine-grained motion discrimination. Many ADL actions appear visually similar but involve subtle differences, such as pouring from a bottle versus pouring from a kettle, necessitating systems that can detect these nuanced distinctions. Additionally, ADL actions must be recognized accurately regardless of camera viewpoint, adding complexity to the recognition process. Furthermore, many ADL actions involve similar overall movements with subtle differences in motion patterns, requiring systems to detect the fine motor distinctions that differentiate one activity from another.

The π-ViT framework addresses these challenges by creating a video transformer that induces both 2D and 3D pose information through auxiliary tasks during training. The key innovation lies in the design of two specialized modules:

1. 2D Skeleton Induction Module (2D-SIM): This module creates a mapping between video tokens and 2D skeleton joints, providing explicit supervision to RGB regions containing relevant anatomical information. It operates by:

    - Constructing a token-skeleton map that defines correspondences between

RGB regions and 2D skeleton joints

- Performing an auxiliary task of predicting the presence of skeleton joints from visual tokens
- Training with a binary cross-entropy loss to supervise this mapping

2. 3D Skeleton Induction Module (3D-SIM): This module addresses view-invariance and motion pattern recognition by aligning visual features with 3D skeleton features. It:

   - Leverages a pre-trained 3D skeleton model
   - Performs feature alignment between visual tokens and skeleton features
   - Uses both alignment and classification losses during training

The core innovation in π-ViT is that after training, both modules are removed during inference, resulting in a standard video transformer that has learned to encode pose information implicitly without requiring pose estimation at deployment time.

3.2 Proposed Methodology: Pose- and Sensor-Induced Video Transformer (ψ-ViT)

Building upon the strengths and limitations of the π-ViT framework described above, we propose an extension that incorporates additional modalities beyond just pose information. While π-ViT effectively induces human pose information into visual representations, we hypothesize that further performance gains can be achieved by leveraging the rich motion data captured by inertial sensors that are commonly available in wearable and mobile devices.

Our Pose- and Sensor-Induced Video Transformer (PSI-ViT, ψ-ViT) extends the auxiliary learning paradigm to incorporate gyroscope and accelerometer data, which capture rotational and linear motion patterns, respectively. By designing specialized induction modules for these sensor modalities, we aim to enhance the video transformer's ability to recognize subtle movements that are characteristic of fine-grained

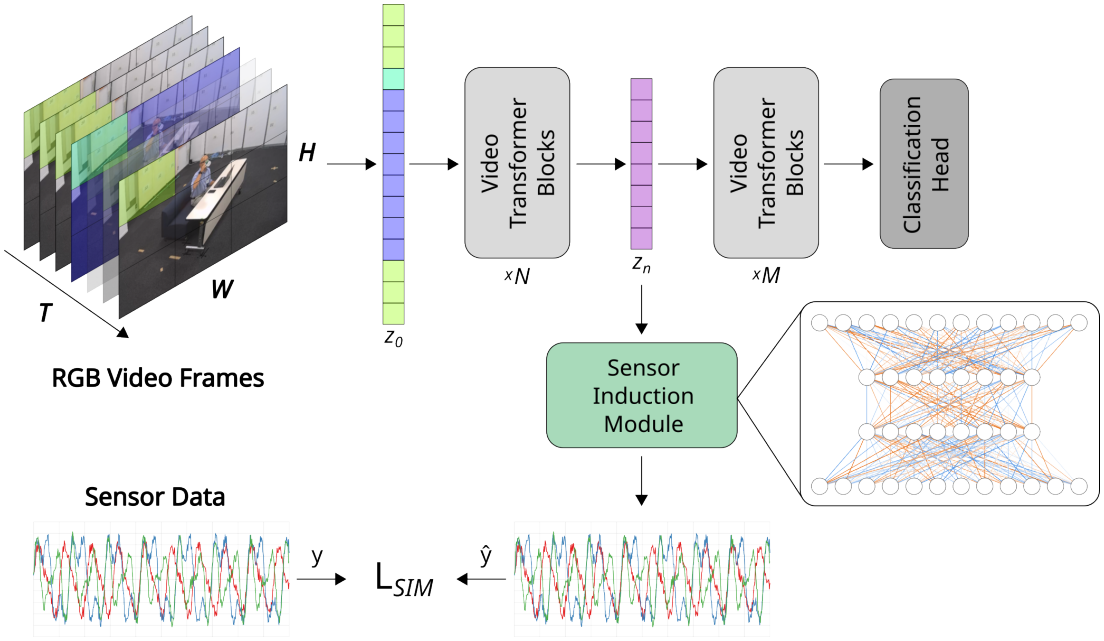

Figure 3-3: Sensor Induction Module insertion in ψ-ViT. The figure illustrates how our auxiliary sensor modules are integrated into the TimeSformer backbone with divided space-time attention, allowing sensor information to guide the learning of visual representations.

activities, particularly in healthcare applications where fall detection or unusual gait patterns may indicate medical emergencies.

The urge for leveraging sensor data comes from an observation that different sensor modalities capture complementary aspects of human activity, with visual data recording appearance and context while motion sensors directly measure physical movement patterns. However, rather than using sensor fusion during inference, we leverage sensor data through auxiliary tasks that guide the learning process during training. By discarding auxiliary modules during inference, we maintain the computational efficiency and deployment simplicity of vision-only models.

As shown in Figure 3-3, the ψ-ViT architecture builds on the TimeSformer video transformer backbone with divided space-time attention [3], augmented with two new auxiliary modules designed to induce information from gyroscope and accelerometer sensors:

1. Gyroscope Induction Module (GIM): Induces rotational movement patterns captured by gyroscope sensors.

2. Accelerometer Induction Module (AIM): Induces linear acceleration patterns captured by accelerometer sensors.

These modules can be inserted after any transformer layer, providing supervision that encourages the transformer to encode motion patterns implicitly in its visual representations.

3.2.1 Theoretical Foundation for Sensor Induction Modules

Information Complementarity Principle

The design of the sensor induction modules in ψ-ViT is fundamentally motivated by the Information Complementarity Principle, which posits that multiple data modalities often provide distinct, non-redundant information about the target variable. In the context of human activity recognition, visual data (e.g., video) and motion sensor data (e.g., accelerometer, gyroscope) capture different aspects of human motion: while video excels at spatial and appearance cues, sensor data offers precise temporal and dynamic movement signals. From an information-theoretic perspective, the complementary information provided by two modalities X and Z for a target Y can be formalized as $\Gamma_{X,Y} = I(X; Y \mid Z)$ and $\Gamma_{Z,Y} = I(Z; Y \mid X)$, where $I(\cdot;\cdot|\cdot)$ denotes conditional mutual information [1]. Larger values of $\Gamma_{X,Y}$ and $\Gamma_{Z,Y}$ indicate higher complementarity, suggesting that effective fusion or induction mechanisms can leverage the unique contributions of each modality to improve model robustness and accuracy [19, 1]. By integrating sensor induction modules, ψ-ViT explicitly encourages the learning of such complementary representations, enhancing the model's ability to recognize subtle or ambiguous activities that may be underrepresented in any single modality.

Learning Using Privileged Information (LUPI)

The architectural strategy of incorporating sensor data only during training in ψ-ViT is inspired by the Learning Using Privileged Information (LUPI) paradigm introduced by Vapnik and Izmailov [30]. In the LUPI framework, an "Intelligent Teacher"

provides additional information x^* (privileged information) alongside standard training pairs (x, y), resulting in triplets (x, x^*, y) available only during training. This privileged information, which may not be accessible at inference, can significantly accelerate learning by correcting the student's concept of similarity between examples and facilitating direct knowledge transfer from the teacher's space to the student's space. In the context of ψ-ViT, sensor signals serve as privileged information: they are used to guide the learning of the video transformer during training, but are not required for inference. This approach enables the model to benefit from richer supervision and improved generalization, while maintaining the practical efficiency of unimodal (video-only) inference in deployment [30, 22].

3.2.2 Gyroscope Induction Module (GIM)

The GIM is designed to induce rotational motion patterns into the video transformer's representations. Gyroscope sensors measure angular velocity across three axes (x, y, z), providing valuable information about rotational movements that are often difficult to discern from visual data alone.

Raw gyroscope signals require preprocessing to extract meaningful patterns, including temporal alignment, noise reduction, normalization, and segmentation. Synchronizing gyroscope data with video frames requires timestamp matching and interpolation, while a low-pass filter removes high-frequency noise from the signals. Standardizing signal values to zero mean and unit variance ensures consistent input scaling, and dividing signals into segments corresponding to video clips enables proper alignment with visual data for subsequent analysis.

Our initial approach involved direct prediction of the processed gyroscope signals from video transformer features. We explored three neural network architectures for this task:

Multilayer Perceptron (MLP)

A simple fully-connected network that projects visual token features to predict gyroscope signals:

$$GI - MLP(z_l) = W_2 * ReLU(W_1 * z_l + b_1) + b_2 \tag{3.1}$$

where z_l represents the visual tokens from layer l of the video transformer.

Recurrent Neural Network (RNN)

Standard RNN cells that process the sequence of visual tokens to capture temporal dependencies:

$$\begin{aligned} h_t &= tanh(W_h x * z_{l_t} + W_h h * h_{t-1} + b_h) \\ GI - RNN(z_l) &= W_o * h_t + b_o \end{aligned} \tag{3.2}$$

Gated Recurrent Unit (GRU)

A more sophisticated recurrent architecture with gating mechanisms:

$$\begin{aligned} z_t &= \sigma(W_z * [z_{l_t}, h_{t-1}] + b_z) \\ r_t &= \sigma(W_r * [z_{l_t}, h_{t-1}] + b_r) \\ \tilde{h}_t &= tanh(W * [z_{l_t}, r_t * h_{t-1}] + b) \\ h_t &= (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t \\ GI - GRU(z_l) &= W_o * h_t + b_o \end{aligned} \tag{3.3}$$

The loss function for direct prediction uses mean squared error between predicted and actual gyroscope signals:

$$L_{GI_{direct}} = MSE(GI - X(z_l), G) \tag{3.4}$$

where G represents the ground-truth gyroscope signals and $GI - X$ is one of the prediction networks ($GI - MLP$, $GI - RNN$, or $GI - GRU$).

Convolutional Neural Network (CNN)

We have also developed a more sophisticated approach treating gyroscope signals as time-frequency patterns. To facilitate this, we performed Mel spectrogram conversion, a common technique in audio signal processing that we adapt for sensor data analysis.

This process transforms gyroscope signals into a visual representation of frequency content over time, similar to how spectrograms are used to analyze audio signals. The conversion involves the following steps:

1. **Segmentation**: Dividing the continuous gyroscope signal into short, overlapping time windows.

2. **Fourier Transform**: Applying Short-Time Fourier Transform (STFT) to each window to convert from time domain to frequency domain, revealing the frequency components present in each segment.

3. **Mel Scale Conversion**: Transforming the linear frequency scale to the Mel scale, which approximates the human auditory system's non-linear perception of frequency. While originally designed for audio, this scale helps emphasize the most perceptually relevant frequency bands in our motion signals.

4. **Multi-channel Integration**: Combining spectrograms from all three axes (x, y, z) of the gyroscope data to create a comprehensive representation of rotational movement patterns.

This conversion translates the time-domain gyroscope signals into frequency-domain representations that highlight the most meaningful motion patterns while compressing the data into a format that neural networks can effectively process. Repetitive movements, such as walking or cycling, produce distinctive periodic patterns in the mel-spectrogram, while sudden movements like falling create characteristic transient frequency signatures.

In this approach, we used a convolutional neural network as the module network to predict these mel-spectrograms:

$$GI - CNN(z_l) = CNN(z_l) \tag{3.5}$$

The loss function for the mel-spectrogram approach uses a combination of mean squared error and structural similarity:

$$L_{GI_{mel}} = \alpha * MSE(GI - CNN(z_l), G_{mel}) + (1 - \alpha) * (1 - SSIM(GI - CNN(z_l), G_{mel})) \tag{3.6}$$

where G_{mel} represents the ground-truth mel-spectrograms and α is a weighting factor.

3.2.3 Accelerometer Induction Module (AIM)

The AIM follows a similar design philosophy to the GIM but focuses on linear acceleration patterns. Accelerometer sensors measure linear acceleration across three axes, capturing information about translational movements and forces applied during activities. The preprocessing pipeline for accelerometer data follows similar steps to the gyroscope data, including temporal alignment to synchronize with video frames, gravity removal using a high-pass filter to separate device acceleration from gravitational acceleration, noise reduction by applying a low-pass filter to remove high-frequency noise, normalization to standardize signal values, and segmentation to divide signals into segments corresponding to video clips.

This module uses exactly the same networks, albeit with different hyperparameters, that the GIM does. For brevity, we omit repeating the implementation details already described in the previous section.

The overall training objective combines the primary action classification loss with the auxiliary losses from both modules:

$$L_{total} = L_{cls} + \lambda_{GI} * L_{GI} + \lambda_{AI} * L_{AI} \tag{3.7}$$

where:

- L_{cls} is the cross-entropy loss for action classification
- L_{GI} is the loss from the Gyroscope Induction Module (either direct or mel-spectrogram based)
- L_{AI} is the loss from the Accelerometer Induction Module (either direct or mel-spectrogram based)
- λ_{GI} and λ_{AI} are weighting factors that control the contribution of each auxiliary task

For training ψ-ViT, which includes pose induction modules, the total loss function becomes

$$L_{total} = L_{cls} + \lambda_{GI} * L_{GI} + \lambda_{AI} * L_{AI} + \lambda_{2D} * L_{2D} + \lambda_{3D} * L_{3D} \tag{3.8}$$

where L_{2D} and L_{3D} are loss components for the 2D and 3D pose induction modules from π-ViT, respectively.

3.2.4 Training Details

We employ a progressive training schedule to balance the primary and auxiliary tasks through initial warm-up, auxiliary task introduction, and fine-tuning. First, we train the base video transformer for a few epochs without auxiliary modules to establish baseline performance. Then, we gradually introduce and increase the weight of auxiliary tasks to incorporate their guidance without overwhelming the primary objective. Finally, we adjust weights to find the optimal balance between primary and auxiliary tasks, ensuring complementary learning. This schedule helps prevent the auxiliary tasks from dominating the learning process while ensuring they provide meaningful guidance to enhance the model's overall performance.

Cross-Dataset Transfer with Frozen Auxiliary Modules

When transferring the ψ-ViT model from MMAct (which contains sensor data) to datasets lacking sensor modalities such as Toyota-Smarthome, we employ a specialized

fine-tuning strategy that preserves the sensor-induced knowledge while adapting to the new domain.

Sensor Loss Weight Nullification For datasets without sensor ground truth, we modify the total training objective by setting the auxiliary loss weights to zero:

$$L_{total} = L_{cls} + \mathbf{0} * L_{GI} + \mathbf{0} * L_{AI} + \lambda_{2D} * L_{2D} + \lambda_{3D} * L_{3D} \tag{3.9}$$

This effectively reduces the training objective to that of π-ViT, eliminating the need for sensor ground truth while maintaining the architectural structure of the model.

Selective Layer Freezing Strategy To preserve the sensor-induced representations learned during MMAct pre-training, we implement a selective freezing strategy. All transformer layers at which auxiliary modules were inserted (layers 4, 8, and 12 in our implementation) are frozen during fine-tuning. This preserves the sensor-enhanced feature representations learned during the initial training phase. The classification head remains trainable to adapt to new action classes, while remaining transformer layers without auxiliary module insertion points remain trainable to allow domain adaptation. Positional embeddings remain trainable to accommodate potential differences in video resolution or temporal dynamics. The GIM and AIM modules attached to frozen layers remain inactive during both forward and backward passes, effectively becoming dormant while preserving their learned parameters for potential future use.

Fine-tuning Protocol The cross-dataset fine-tuning follows a three-phase protocol. In the first phase (Architecture Preparation, 1 epoch), we load the pre-trained ψ-ViT checkpoint from MMAct training, freeze transformer layers 4, 8, and 12 corresponding to auxiliary module insertion points, set $\lambda_{GI} = \lambda_{AI} = 0$ in the loss function, and initialize a new classification head for target dataset classes. The second phase (Domain Adaptation, 10 epochs) trains only unfrozen components using the target

dataset with a reduced learning rate of 1e-5 to prevent catastrophic forgetting while applying standard data augmentation techniques appropriate for the target domain. The final phase (Fine-tuning Refinement, 5 epochs) unfreezes layers 4, 8, and 12 for minimal adaptation using a very low learning rate of 5e-6 for careful parameter adjustment while monitoring validation performance to prevent overfitting.

Theoretical Justification This approach is grounded in the principle that the sensor-induced features learned during MMAct training have already enhanced the model's ability to recognize motion patterns through visual cues alone. By freezing the layers where this knowledge is most concentrated, we preserve these learned representations while allowing the model to adapt to domain-specific characteristics of the new dataset. The selective freezing strategy ensures that motion understanding acquired through sensor supervision is retained, domain adaptation can occur through unfrozen layers, computational efficiency is maintained by avoiding recomputation of auxiliary losses, and model stability is preserved by preventing degradation of sensor-induced knowledge. This methodology enables effective transfer learning from multimodal datasets to unimodal target domains while maintaining the benefits of auxiliary task learning acquired during initial training.

Chapter 4

Experimental Results and Analysis

We evaluate our proposed ψ-ViT framework on the MMAct dataset [17], which contains synchronized video and sensor data for 35 different action classes across 20 subjects in various environments. This multimodal dataset is particularly suitable for our research as it provides synchronized RGB videos alongside accelerometer and gyroscope data captured from smartphone sensors.

4.1 Implementation Details

Our implementation uses the TimeSformer [3] as the base video transformer architecture with a patch size of 16×16 pixels. The model was pre-trained on Kinetics-400 [14] before being fine-tuned on MMAct. For optimization, we use AdamW with an initial learning rate of 1e-4 and weight decay of 0.05. We apply a cosine learning rate schedule with linear warmup for 5 epochs. The auxiliary task weights λ_{GI} and λ_{AI} were empirically set to 0.2 and 0.3 respectively based on validation performance.

For the GI and AI modules, we evaluated multiple neural network architectures as described in our methodology section. Tables 4.1, 4.2, and 4.3 present the performance metrics for different module networks.

Method	Accuracy	Precision	Recall	F1-score
TimeSformer [3]	81.4	80.9	82.0	81.4
π-ViT [26]	85.6	85.1	86.3	85.7
TimeSformer + GIM	85.8	85.2	86.1	85.6
TimeSformer + AIM	85.9	85.4	86.3	85.8
ψ-ViT	**87.2**	**86.8**	**87.6**	**87.2**

Table 4.1: Performance comparison of ViT variants with MLP as Module network. Our proposed methods (GIM, AIM, and their combination in ψ-ViT) are highlighted, with best results in **bold**.

Method	Accuracy	Precision	Recall	F1-score
TimeSformer [3]	81.4	80.9	82.0	81.4
π-ViT [26]	85.6	85.1	86.3	85.7
TimeSformer + GIM	85.3	84.7	85.9	85.2
TimeSformer + AIM	85.6	85.2	86.4	85.7
ψ-ViT	**87.0**	**86.6**	**87.4**	**87.0**

Table 4.2: Performance comparison with RNN as Module network. Our proposed methods (GIM, AIM, and their combination in ψ-ViT) are highlighted, with best results in **bold**.

4.2 Performance Comparison

The baseline ViT model achieved 81.4% accuracy on the MMAct test set, representing a strong starting point for activity recognition using only visual data. Implementing the pose-induced approach (π-ViT) improved performance to 85.6%, consistent with findings from [26] that pose information enhances the model's ability to recognize subtle motion patterns.

The integration of sensor induction modules showed promising results across all network architectures. As shown in Table 4.1, the MLP-based module networks performed best overall, with the gyroscope induction module providing a 4.4% accuracy improvement over the baseline ViT and the accelerometer induction module yielding a 4.5% improvement. This suggests that even simple fully-connected layers can effectively transfer motion sensor patterns to visual representations.

The RNN and CNN architectures (Tables 4.2 and 4.3) showed slightly lower but

Method	Accuracy	Precision	Recall	F1-score
TimeSformer [3]	81.4	80.9	82.0	81.4
π-ViT [26]	85.6	85.1	86.3	85.7
TimeSformer + GIM	85.1	84.6	85.7	85.0
TimeSformer + AIM	85.4	84.9	86.1	85.3
ψ-ViT	**86.7**	**86.3**	**87.1**	**86.6**

Table 4.3: Performance comparison with CNN as Module network. Our proposed methods (GIM, AIM, and their combination in ψ-ViT) are highlighted, with best results in **bold**.

still substantial improvements, with the RNN-based modules achieving 85.3% and 85.6% accuracy for gyroscope and accelerometer induction respectively, while CNN-based modules reached 85.1% and 85.4%. The reduced performance might be attributed to these more complex architectures requiring additional training data to fully leverage their capacity for temporal pattern recognition.

The most significant finding is that combining pose, gyroscope, and accelerometer induction in our full ψ-ViT framework consistently outperformed all other variants across all module network architectures. The MLP-based ψ-ViT achieved the highest overall accuracy at 87.2%, representing a 7% relative improvement over the baseline ViT and a 1.9% relative improvement over π-ViT. This demonstrates that the different modalities provide complementary information that enhances the model's ability to recognize fine-grained activities.

4.2.1 Results on the Toyota-Smarthome Dataset

To evaluate the generalizability of our approach, we conducted experiments on the Toyota-Smarthome dataset [7], which features activities of daily living performed by elderly subjects in a real smart home environment. This dataset presents unique challenges due to varied camera viewpoints and the subtle nature of many daily living activities. Since the dataset does not feature any sensor data, a checkpoint of ψ-ViT pre-trained on MMAct was fine-tuned on video and poses from this dataset while preserving knowledge obtained from sensor induction. This procedure has been

Methods	Modality		CS	CV_1	CV_2
	Pose	Video			
Pose Only					
PoseC3D† [10]	✓	✗	50.6	20.0	28.2
Hyperformer† [33]	✓	✗	57.5	31.6	35.2
Video + Pose					
VPN [9]	✓	✓	65.1	43.5	53.9
VPN++ + 3D Poses [8]	✓	✓	71.2	-	58.2
π-ViT + 3D Poses† [26]	✓	✓	73.1	55.6	65.0
Video Only (at inference)					
VPN++ [8]	○	✓	69.0	-	54.9
Video Swin† [20]	✗	✓	69.8	36.6	48.6
MotionFormer† [24]	✗	✓	65.8	45.2	51.0
TimeSformer† [3]	✗	✓	68.4	50.0	60.6
π-ViT† [26]	○	✓	72.9	55.2	64.8
TimeSformer + GIM	○	✓	71.7	54.1	64.4
TimeSformer + AIM	○	✓	71.9	54.3	64.7
ψ-ViT	○	✓	**73.5**	**55.5**	**65.4**

Table 4.4: Comparison with SoTA on Toyota-Smarthome dataset. We report the mean class accuracy on cross-subject (CS) and cross-view (CV_1, CV_2) protocols. ○ indicates that the modality has been used only in training. Bold text indicates best performance. † indicates results produced by the authors.

explained in 3.2.4.

Table 4.4 presents our comparison with state-of-the-art methods on this dataset. We report mean class accuracy on the standard cross-subject (CS) and cross-view (CV_1, CV_2) evaluation protocols. The cross-subject protocol tests generalization to unseen subjects, while the cross-view protocols evaluate robustness to viewpoint changes, with CV_1 being particularly challenging as it uses a ceiling-mounted camera view not seen during training.

Our proposed ψ-ViT framework achieves state-of-the-art performance across all evaluation protocols. Specifically, ψ-ViT attains 73.5% accuracy on the CS protocol, outperforming previous methods including the π-ViT approach (72.9%). Similar improvements are observed on the challenging cross-view protocols, with ψ-ViT reaching

55.5% on CV_1 and 65.4% on CV_2.

The individual induction modules also demonstrate strong performance, with GIM and AIM achieving 71.7% and 71.9% on CS, respectively. This confirms that our sensor induction approach generalizes effectively to complex, real-world environments where actions may be subtle and challenging to distinguish from visual cues alone.

Notably, while the π-ViT + 3D Poses variant shows strong performance (73.1% CS, 65.0% CV_2), it requires pose information during inference. In contrast, our ψ-ViT model achieves slightly superior performance (73.5% CS, 65.4% CV_2) while only requiring video input at test time, making it more practical for real-world deployment where sensor or pose extraction may not be feasible during system operation.

These results, together with our experiments on MMAct, demonstrate that learning from multimodal data during training provides significant benefits for action recognition tasks across diverse settings, even when only a single modality is available at inference time. The consistent performance improvements across different datasets highlight the effectiveness of our approach in capturing complementary information from multiple modalities through the induction mechanism.

4.2.2 Inference Time

Efficient inference is crucial for deploying action recognition models in real-world applications, particularly where computational resources or latency are constrained. Table 4.5 summarizes the cross-subject accuracy, inference time, and computational complexity (in FLOPs) for representative models. For a fair comparison, we report FLOPs values based on standard input resolutions and sequence lengths as described in the respective papers. The results were collected on NVidia A100 GPUs and averaged over 100 video samples.

PoseC3D and Hyperformer (pose-only models) require 17.6G and 12.3G FLOPs per sample, respectively, while VPN++ and VPN++ + 3D Poses (video and hybrid models) require 18.2G and 35.1G FLOPs. In contrast, TimeSformer and models based off of it (π-ViT, and our proposed ψ-ViT) have a higher computational footprint of 590G FLOPs per sample, owing to the quadratic complexity of self-attention with

Methods	Modality		CS (%)	Time (s)	FLOPs (G)
	Pose	Video			
Pose Only					
PoseC3D [10]	✓	✗	50.6	11.45	17.6
Hyperformer [33]	✓	✗	57.5	7.45	12.3
Video + Pose					
VPN++ + 3D Poses [8]	✓	✓	71.2	8.99	35.1
Video Only (at inference)					
VPN++ [8]	∘	✓	69.0	3.09	18.2
TimeSformer [3]	✗	✓	68.4	**0.78**	590
π-ViT [26]	∘	✓	72.9	**0.78**	590
ψ-ViT	∘	✓	**73.5**	**0.78**	590

Table 4.5: Comparison of performance, inference time, and computational complexity. We report the cross-subject (CS) accuracy (%), inference time in seconds, and computational complexity in FLOPs. ∘ indicates that the modality has been used only in training. Bold text indicates best performance.

respect to input sequence length.

Despite the higher FLOPs, transformer-based models achieve significantly faster inference times on modern GPUs. All three video transformer variants - TimeSformer, π-ViT, and ψ-ViT-require only 0.78 seconds per sample, which is markedly lower than the pose-based and hybrid CNN models. This efficiency is attributable to the parallelizable nature of transformer architectures and the optimization of GPU kernels for attention operations. Importantly, the sensor induction modules in ψ-ViT are only active during training, so there is no additional inference overhead compared to the baseline TimeSformer. As a result, ψ-ViT achieves the highest accuracy (73.5% CS) among video-only models, with no increase in inference time or FLOPs.

These findings highlight the practical advantages of our approach: ψ-ViT leverages multimodal learning during training to achieve state-of-the-art accuracy, while maintaining the same inference-time efficiency and computational requirements as standard video transformers. In scenarios where only video is available at test time, ψ-ViT offers a compelling combination of accuracy and speed, outperforming both pose-based and hybrid models that require additional modalities or incur greater

computational cost at inference.

4.2.3 Ablation Studies

Configuration	Accuracy (%)
TimeSformer [3]	81.4
ψ-ViT without progressive training	85.9
ψ-ViT with GIM after TimeSformer layer 4	86.3
ψ-ViT with AIM after TimeSformer layer 4	86.4
ψ-ViT with GIM & AIM after TimeSformer layer 4	86.7
ψ-ViT with GIM & AIM after TimeSformer layers 4, 8	87.0
ψ-ViT (full, layers 4, 8, 12)	**87.2**

Table 4.6: Ablation study results showing the impact of different architectural choices on the ψ-ViT model performance. The baseline TimeSformer is separated from our proposed ψ-ViT variants, with the full model achieving the best result (in **bold**).

We conducted ablation studies to analyze the contribution of different components in our framework (Table 4.6). The progressive training schedule (described in Section 3.2.4) proved essential, providing a 1.5% accuracy improvement compared to training without progressive scheduling.

In the non-progressive approach, all losses (classification, GIM, and AIM) were weighted equally from the beginning of training, which led to suboptimal convergence as the model struggled to simultaneously optimize for multiple objectives. The three-phase progressive schedule (initial warm-up, gradual introduction of auxiliary tasks, and fine-tuning) allowed the model to first establish strong visual representations before incorporating sensor-related patterns, resulting in more effective knowledge transfer between modalities.

Our layer-wise ablation experiments reveal the importance of inducing sensor information at multiple levels of feature abstraction. Early layers (e.g., layer 4) focus on local motions and benefit from both gyroscope and accelerometer induction, with a combined improvement of 5.3% over the baseline. Adding modules at middle layers (layer 8) provides an additional 0.3% gain, suggesting that mid-level features benefit from motion pattern guidance. The full model with modules at early, middle, and

late layers (4, 8, and 12) achieves the highest accuracy, indicating that sensor induction is beneficial across multiple levels of feature abstraction within the transformer architecture.

4.2.4 GradCAM Visualization: Benefits of Sensor Induction for Subtle Actions

To further explore how sensor induction enhances classification - particularly for activities involving fine or subtle movements - we present GradCAM [28] visualizations for two representative classes: *Drinking* and *Pocket In*. These classes were among the most improved following the application of the sensor induction modules (45% and 41% relative performance improvement, respectively).

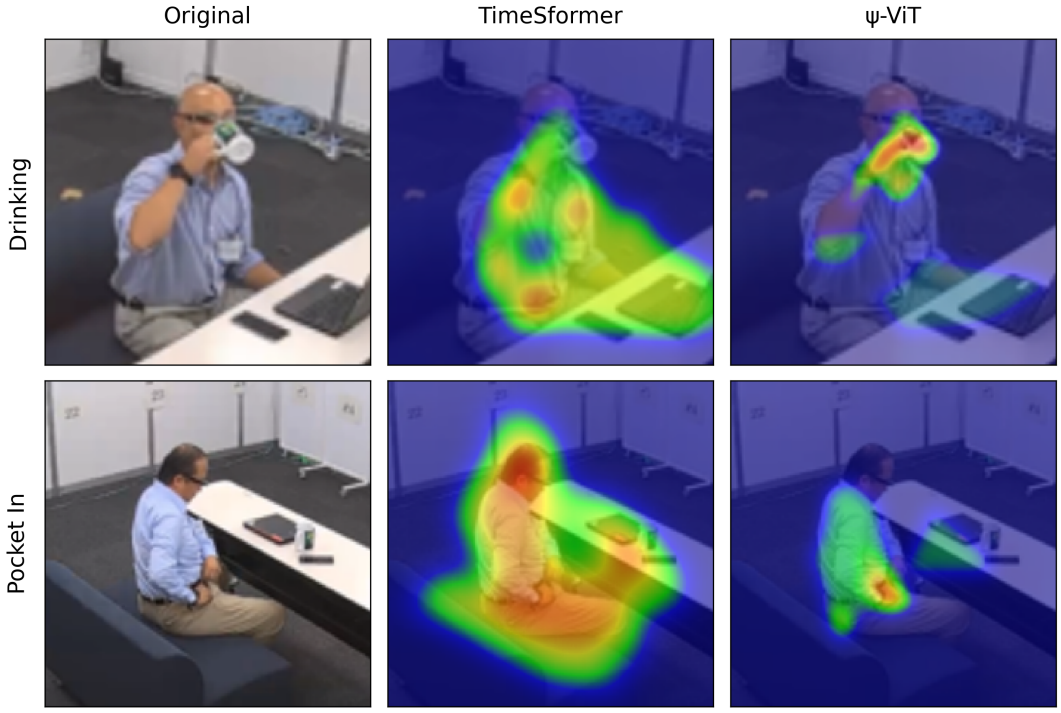

Figure 4-1: GradCAM visualizations comparing the baseline TimeSformer and the proposed ψ-ViT on two challenging examples: *Drinking* (top row) and *Pocket In* (bottom row). Both were originally misclassified by TimeSformer but correctly identified by ψ-ViT, which indicates sharper and more discriminative attention.

Figure 4-1 shows a comparison between original video frames, saliency maps from the baseline TimeSformer model, and those from the proposed ψ-ViT model.

Drinking → Misclassified as Using PC (TimeSformer)

In the top row of Figure 4-1, the subject is performing the action *Drinking*. The TimeSformer model produces a diffuse attention distribution, activating over the subject's torso, arms, and even unrelated areas like the desk. This misdirected focus likely led to the misclassification as *Using PC*, which shares spatial similarities such as a seated pose and interaction with nearby objects.

In contrast, the ψ-ViT model generates a much sharper and more localized class activation map. The focus is concentrated around the hand and cup—key visual cues for identifying the *Drinking* activity. This improved localization demonstrates how sensor-induced training enables the model to internalize subtle but semantically important motion patterns.

Pocket In → Misclassified as Sitting (TimeSformer)

In the bottom row, the subject is captured performing a *Pocket In* motion. Again, the TimeSformer model exhibits wide, unfocused attention around the torso and background. This lack of precision leads to the activity being incorrectly labeled as *Sitting*, a visually similar but semantically distinct action.

The ψ-ViT model, however, accurately centers its attention on the hand-to-pocket interaction, which is the defining feature of the *Pocket In* action. This suggests that the model has learned to prioritize minute localized movements through its exposure to accelerometer and gyroscope data during training.

Discussion

These visual comparisons provide strong qualitative evidence supporting the effectiveness of sensor induction. The improvements are especially evident in cases involving subtle hand movements, where conventional vision-only models often struggle. The

auxiliary supervision from inertial sensor data guides the model to develop more semantically meaningful feature representations during training.

Importantly, the ψ-ViT achieves this without requiring any sensor data at inference time. This ability to internalize cross-modal knowledge while preserving vision-only deployment makes it highly practical for real-world applications, especially in healthcare scenarios where wearable sensors may not be available.

4.3 Conclusion

This thesis has presented the Pose- and Sensor-Induced Video Transformer (ψ-ViT), a framework that advances Human Activity Recognition (HAR) by leveraging multimodal sensor data during training while maintaining vision-only inference. Building upon pose-induced transformers, our work extends auxiliary task learning to incorporate motion sensor modalities, demonstrating that gyroscope and accelerometer patterns can be integrated into a video transformer's latent space to enhance recognition of fine-grained activities.

The ψ-ViT framework addresses the gap between visual and sensor data without requiring wearable devices during deployment. Experiments on the MMAct dataset indicate that inducing both rotational and linear motion patterns through specialized modules improves activity recognition accuracy by 7% over baseline vision transformers, achieving an F1-score of 87.2%. This improvement is notable for subtle motions where traditional vision-based approaches typically demonstrate limitations, such as distinguishing between similar upper body movements or identifying early signs of instability in gait patterns. The framework's progressive training schedule provides a methodological approach for balancing primary and auxiliary objectives that may be adaptable to other sensor combinations.

An important methodological contribution of this work is the development of cross-dataset transfer techniques that enable the application of sensor-induced models to datasets lacking sensor modalities. Through selective layer freezing and loss weight nullification strategies, we demonstrate that knowledge acquired from auxiliary sensor tasks can be preserved when transferring to vision-only datasets such as Toyota-Smarthome. This approach achieved state-of-the-art performance (73.5% cross-subject accuracy) while maintaining computational efficiency, establishing a practical pathway for deploying multimodal-trained models in unimodal inference scenarios.

This work offers an alternative approach to the conventional distinction between multimodal and unimodal systems in HAR. By implementing sensor data as a training

signal rather than a deployment requirement, ψ-ViT presents an approach that differs from traditional fusion methods. The cross-dataset transfer methodology further extends this paradigm by enabling knowledge preservation across modality boundaries, potentially applicable to domains beyond activity recognition where training and deployment modalities may differ.

The results suggest that auxiliary task learning may serve as an effective method for encoding cross-modal correlations, with the selective freezing strategy providing a principled approach to knowledge transfer. Potential future research directions include exploring additional modalities like physiological signals, developing temporal alignment methods for asynchronous data streams, investigating the generalizability of frozen layer strategies across different transformer architectures, or adapting the framework for long-term activity sequences. As transformer architectures continue to develop, the principles established in this work of cross-modal knowledge transfer through auxiliary tasks and selective parameter preservation may remain relevant for developing systems that combine multimodal data analysis with efficient deployment requirements.

Bibliography

[1] Anonymous. Modality complementarity: Towards understanding multi-modal robustness. *OpenReview*, 2023. Under review as a conference paper at ICLR 2023.

[2] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer, 2021.

[3] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *Proceedings of the International Conference on Machine Learning (ICML)*, July 2021.

[4] Denise Catacutan, Jeremie Alexander, Autumn Arnold, and Jonathan Stokes. Machine learning in preclinical drug discovery. *Nature Chemical Biology*, 20, 07 2024.

[5] Liming Chen, Jesse Hoey, Chris D. Nugent, Diane J. Cook, and Zhiwen Yu. Sensor-based activity recognition. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42(6):790–808, 2012.

[6] Wenliang Dai, Zihan Liu, Tiezheng Yu, and Pascale Fung. Modality-transferable emotion embeddings for low-resource multimodal emotion recognition. In Kam-Fai Wong, Kevin Knight, and Hua Wu, editors, *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 269–280, Suzhou, China, December 2020. Association for Computational Linguistics.

[7] Srijan Das, Rui Dai, Michal Koperski, Luca Minciullo, Lorenzo Garattoni, Francois Bremond, and Gianpiero Francesca. Toyota smarthome: Real-world activities of daily living. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[8] Srijan Das, Rui Dai, Di Yang, and Francois Bremond. Vpn++: Rethinking video-pose embeddings for understanding activities of daily living, 2021.

[9] Srijan Das, Saurav Sharma, Rui Dai, Francois Bremond, and Monique Thonnat. Vpn: Learning video-pose embedding for activities of daily living, 2020.

[10] Haodong Duan, Yue Zhao, Kai Chen, Dahua Lin, and Bo Dai. Revisiting skeleton-based action recognition, 2022.

[11] Sannara Ek, François Portet, and Philippe Lalanda. Transformer-based models to deal with heterogeneous environments in human activity recognition. *Personal and Ubiquitous Computing*, pages 1–14, 2023.

[12] Huiyan Han, Hongwei Zeng, Liqun Kuang, Xie Han, and Hongxin Xue. A human activity recognition method based on vision transformer. *Scientific Reports*, 14(1):15310, Jul 2024.

[13] Zhen-Yu He and Lian-Wen Jin. Activity recognition from acceleration data using ar model representation and svm. In *2008 International Conference on Machine Learning and Cybernetics*, volume 4, pages 2245–2250, 2008.

[14] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset, 2017.

[15] A. M. Khan, Y. K. Lee, and S.Y. Lee. Accelerometer's position free human activity recognition using a hierarchical recognition model. In *The 12th IEEE International Conference on e-Health Networking, Applications and Services*, pages 296–301, 2010.

[16] Adil Mehmood Khan, Young-Koo Lee, Sungyoung Y Lee, and Tae-Seong Kim. A triaxial accelerometer-based physical-activity recognition via augmented-signal features and a hierarchical recognizer. *IEEE Trans Inf Technol Biomed*, 14(5):1166–1172, June 2010.

[17] Quan Kong, Ziming Wu, Ziwei Deng, Martin Klinkigt, Bin Tong, and Tomokazu Murakami. Mmact: A large-scale dataset for cross modal human action understanding. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[18] Rahul Kumar and Shailender Kumar. Effectiveness of vision transformers in human activity recognition from videos. In *2023 International Conference on Advancement in Computation & Computer Technologies (InCACCT)*, pages 593–597, 2023.

[19] Y. Li, Y. Zhang, Y. Liu, H. Li, and Y. Li. Complementary information mutual learning for multimodality medical image segmentation. *arXiv preprint arXiv:2401.02717*, 2022.

[20] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer, 2021.

[21] Trisha Mittal, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, and Dinesh Manocha. M3er: Multiplicative multimodal emotion recognition using facial, textual, and speech cues, 2019.

[22] Ahmadreza Momeni and Kedar Tatwawadi. Understanding lupi (learning using privileged information), 2017. Stanford University Technical Report.

[23] Henry Friday Nweke, Ying Wah Teh, Mohammed Ali Al-garadi, and Uzoma Rita Alo. Deep learning algorithms for human activity recognition using mobile and wearable sensor networks: State of the art and research challenges. *Expert Systems with Applications*, 105:233–261, 2018.

[24] Mandela Patrick, Dylan Campbell, Yuki M. Asano, Ishan Misra, Florian Metze, Christoph Feichtenhofer, Andrea Vedaldi, and João F. Henriques. Keeping your eye on the ball: Trajectory attention in video transformers, 2021.

[25] Kumbala Reddy, Mankala Satish, A. Prakash, S.Malli Babu, P.Pavan Kumar, and B. Devi. Machine learning revolution in early disease detection for healthcare: Advancements, challenges, and future prospects. pages 638–643, 10 2023.

[26] Dominick Reilly and Srijan Das. Just add π! pose induced video transformers for understanding activities of daily living, 2023.

[27] Yvxuan Ren, Dandan Zhu, Kai Tong, Lulu Xv, Zhengtai Wang, Lixin Kang, and Jinguo Chai. Pdchar: Human activity recognition via multi-sensor wearable networks using two-channel convolutional neural networks. *Pervasive and Mobile Computing*, 97:101868, 2024.

[28] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2):336–359, October 2019.

[29] Emmanuel Munguia Tapia, Stephen S. Intille, William Haskell, Kent Larson, Julie Wright, Abby King, and Robert Friedman. Real-time recognition of physical activities and their intensities using wireless accelerometers and a heart rate monitor. In *2007 11th IEEE International Symposium on Wearable Computers*, pages 37–40, 2007.

[30] Vladimir Vapnik and Rauf Izmailov. Learning using privileged information: Similarity control and knowledge transfer. *Journal of Machine Learning Research*, 16:2023–2049, 2015.

[31] James Wensel, Hayat Ullah, and Arslan Munir. Vit-ret: Vision and recurrent transformer neural networks for human activity recognition in videos. *IEEE Access*, 11:72227–72249, 2023.

[32] Adnan Yazici, Dana Zhumabekova, Aidana Nurakhmetova, Zhanggir Yergaliyev, Hakan Yekta Yatbaz, Zaida Makisheva, Michael Lewis, and Enver Ever. A smart e-health framework for monitoring the health of the elderly and disabled. *Internet of Things*, 24:100971, 2023.

[33] Yuxuan Zhou, Zhi-Qi Cheng, Chao Li, Yifeng Geng, Xuansong Xie, and Margret Keuper. Hypergraph transformer for skeleton-based action recognition. *arXiv preprint arXiv:2211.09590*, 2022.