2025 34th IEEE International Conference on Robot and
Human Interactive Communication (RO-MAN)
August 25-29, 2025. Eindhoven, The Netherlands

2025 34th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN) | 979-8-3315-8771-0/25/$31.00 ©2025 IEEE | DOI: 10.1109/RO-MAN63969.2025.11217600

Multimodal Transfer Learning for Privacy in Human Activity Recognition

Sigmund Rolfsjord*[16]†, Safia Fatima*[25]†, Hugh Alexander von Arnim*[34]† and Adel Baselizadeh[35]†

Abstract— **Human Activity Recognition (HAR) models often rely on small, specialized datasets, limiting their generalizability. In addition, many systems rely on privacy-invasive RGB video as their primary sensing modality. This choice raises ethical concerns, especially in health- and home-care robotics, where patient privacy is paramount. In this study, we evaluate transfer learning as a method to improve HAR generalizability across RGB, IMU, and depth imaging modalities while assessing how privacy-preserving modalities can compensate for a lack of RGB video in multimodal learning contexts.**

We train a feature-fusion model on aggregated HAR datasets, leveraging pretrained backbones for each modality, and compare it to a general multimodal model pretrained on non-HAR datasets. Evaluating on the PriMA-Care privacy-focused dataset across combinations of modalities, we find that the general model outperforms the HAR model, with best accuracies of 98.29% for the general model with RGB + IMU and 94.97% for the HAR-specific model with RGB + depth. Analysis shows that the general model more readily identifies individuals from RGB input, while IMU and depth better preserve privacy with a small accuracy loss (5%).

I. Introduction

Human Activity Recognition (HAR) aims to classify activities using a diverse range of sensor data [1]. Employing both machine learning and signal processing methods, HAR is widely used in a wide range of applications such as healthcare monitoring, fitness tracking, smart environments, security and human-computer interaction [2]. However, HAR faces challenges with generalizability [3], as well as potential privacy concerns in health- and homecare settings due to an over-reliance on RGB (visual spectrum imaging) imaging [4], [5], as RGB video is generally rated as more privacy invasive compared to e.g. depth sensors [6]. The sensor type, however, may need to be conveyed e.g., through labelling.

In this study, we explore transfer learning – the employment of a model trained on data from one domain for use in another [7] – as a method to improve HAR generalizability across RGB, IMU (inertial measurement unit) and depth imaging modalities, while evaluating how the privacy-preserving sensors compare to RGB imaging. We evaluate the performance of a feature-level fusion HAR-specific model, leveraging pretrained backbones for each modality fine-tuned on aggregated HAR datasets, against a general multimodal model on the *PriMA-Care* dataset, a privacy-oriented dataset for HAR in human-robot interaction [4]. With this, we contribute:

1) The proposal of a systematic methodology for evaluating the generalizability of HAR datasets and apply this framework to assess the cross-dataset performance of several aggregated HAR collections, providing insights into their real-world applicability.
2) A comprehensive evaluation of large pretrained multimodal models adapted for HAR tasks, demonstrating their potential to leverage transfer learning from broader domains to improve activity recognition performance.
3) An analysis of the fusion of privacy-preserving modalities as alternatives to RGB video data in HAR applications, offering practical alternatives that balance recognition accuracy with privacy protection.

II. Background and Related work

A. HAR and Multimodal Learning

Multimodal learning can enhance HAR systems by integrating different sensor modalities, leveraging their strengths to compensate for individual limitations [8]. It has been widely adopted for HAR due to the prevalence of multimodal HAR datasets and its real-world applicability [9].

RNNs and CNNs have traditionally been commonly employed for HAR [9], but recently architectures such as Temporal Convolutional Networks (TCNs) [10] have gained traction. A recent study on eight multimodal architectures for HAR (MSENet, TST, TCN, CNN-LSTM, ConvLSTM, XGBoost, decision tree, and k-nearest neighbor) found that MSENet offered the highest accuracy for both datasets [11].

At larger scale, transformer architectures are becoming increasingly prevalent. State-of-the-art general models such as LanguageBind [12] and ImageBind [13] create a shared embedding space for multiple input modalities, enabling zero-shot and few-shot transfer across modalities. These models can extract useful representations even when fewer modalities are available at inference. Given their ability to generalize across diverse data sources, studying their ability to generalize to HAR tasks is a promising prospect.

B. HAR and Transfer Learning

Transfer learning is a method in which a model trained on data from one domain is employed in another [14]. This can be achieved through methods such as transferring classifier weights, learned features, or instances of data [15], [16], [17].

*These authors contributed equally to this work.
[1]Department of Technology Systems, University of Oslo, Oslo, Norway
[2]Department of Data-Driven Software Engineering, Simula Research Laboratory, Oslo, Norway
[3]RITMO Centre for Interdisciplinary Studies in Rhythm, Time and Motion, University of Oslo, Oslo, Norway
[4]Department of Musicology, University of Oslo, Oslo, Norway
[5]Department of Informatics, University of Oslo, Oslo, Norway
[6]Norwegian Defence Research Establishment, Lillestrøm, Norway
†Corresponding authors' email: sigmunjr@uio.no, safia@simula.no, hughav@uio.no, adelb@ifi.uio.no

Authorized licensed use limited to: UNIVERSITY OF OSLO. Downloaded on January 21,2026 at 19:40:49 UTC from IEEE Xplore. Restrictions apply.

Transfer learning has been widely employed for HAR [18], [19]. However, a significant challenge facing HAR is the lack of large, labeled datasets and corresponding models for many of the sensor modalities employed [18]. This is especially the case for modalities which better preserve privacy such as IMU and depth imaging. This challenge is further compounded in multimodal HAR, where aligned HAR datasets across modalities are scarce.

To address this, we evaluate the ability of recent developments in large, general, multimodal models to transfer to HAR tasks. Specifically, we compare the performance of a general model consisting of ImageBind and LanguageBind against a HAR model, consisting of a feature-level fusion of three single modality pretrained models fine-tuned on an aggregation of HAR datasets.

III. Methods and Materials

A. Overview

To evaluate the need for HAR-specific pretraining, we employ backbones of different levels of specificity (HAR and general). We train a two-layer classification head on top of the frozen backbones with the *PriMA-Care* dataset. The results are evaluated using cross-validation, where one subject is left out of training for each iteration.

The classification head comprises a hidden layer of 512, and an output value for each action (6). We used a standard cross-entropy loss and an Adam optimizer with a learning rate of 1e-3 and weight decay of 1e-4. The classification head was trained from scratch 14 times for 10 epochs, with different subject hold-outs.

Modalities were fused through concatenating the output vector for each backbone. The classification head was trained on that representation using the same procedure.

B. Datasets

Several datasets have been developed for benchmarking and evaluating multimodal HAR models. For training, we selected three datasets: *UTD-MHAD*, *NTU RGB+D 120*, and *GOTOV*, as they overlap with the modalities we have chosen from the PriMA-Care dataset. Each backbone is trained on a combination of two datasets to diversify the range of participants and activities and to test dataset generalizability in otherwise controlled environments. PriMA-Care is used exclusively for evaluation as our primary benchmark.

UTD-MHAD [20] contains RGB, depth video, and six-channel IMU data from 8 participants performing 27 gestures in a controlled setting. While classification models perform well in this constrained environment, we anticipate limited transferability to more dynamic real-world scenarios.

NTU RGB+D 120 [21] is a large multimodal dataset with RGB, depth, IR videos, and skeleton pose data. Though more diverse than *UTD-MHAD*, it uses staged scenarios, making its real-world transferability unclear.

GOTOV [22], [23] is a multimodal dataset of 35 adults over the age of 60 performing 16 daily activities, using wearable sensors (accelerometers, physiological monitors, and spirometry). Due to its elderly-specific focus, the dataset may have limited generalizability.

PriMA-Care [4] is our primary evaluation dataset: a privacy-preserving multimodal dataset for healthcare human–robot interaction. It employs RGB, depth, thermal, 3D/2D lidar, ultra-wideband, wearable IMU, force/torque sensors, and TIAGo arm-joint encoders to record 17 participants performing 27 activities. The original PriMA-Care collection was approved by the Norwegian Center for Research Data (NSD, Ref. No. 863469), and all participants gave informed consent for data collection and public release. No new human-subject data were collected in this study, and we used the dataset strictly under the scope of that consent.

C. Models

1) HAR-Specific Backbones

For RGB and depth, we use the near state-of-the-art backbones from *UMDR* [24], which are trained on the *NTU RGB+D 120* dataset. For RGB we also tested a 3D ResNet-18 (*R3D*), pretrained on *Kinetics-400* [25]. With *R3D* we can evaluate whether training on a larger and more diverse, but still HAR specific dataset, can be beneficial. For this, we employ the same preprocessing used for the models during training, but tune the depth clipping range to fit the *PriMA-Care* dataset, setting a $0.5\,\mathrm{m}$ to $5\,\mathrm{m}$ cutoff to mimic the masking on the *NTU RGB+D 120* dataset.

For the IMU, we did not find any HAR pretrained network, so we trained a backbone on other open datasets before training the classification head on *PriMA-Care*.

a) HAR-Specific IMU Backbone

A challenge facing IMU data is a large degree of heterogeneity between datasets [26], for example employing different sensor arrays with varying sensors, differing sampling rates and placement on the body. Therefore, we fine-tuned our own IMU network, with a focus on an architecture that could be used across datasets.

We fine-tuned the network on a merged *GOTOV + UTD-MHAD* set, collapsing shared labels into three intensity levels (low, medium, high). Because *GOTOV* lacks orientation, we kept only acceleration signals and applied a Butterworth low-pass filter to remove dataset-specific high-frequency noise.

The sequences were split into segments (length: $2\,\mathrm{s}$, hop: $100\,\mathrm{ms}$), and time-frequency spectrogram representations of each segment for each Cartesian axis were derived with a Short-Time Fourier Transform (STFT). This resulted in an image sequence, for which the X, Y, and Z axis spectrograms were placed in the R, G, and B channels, respectively. Each image was resampled to 224×224.

The IMU network employs a *ResNet-18* backbone, appended with a bi-directional LSTM, an attention layer and a fully connected classification layer.

2) General Backbones

Our general backbones are *ImageBind* and *LanguageBind*. These models are not trained specifically on HAR datasets nor for classification tasks, but rather to produce representations that are similar across modalities through self-supervised learning. All the models for all modalities

Authorized licensed use limited to: UNIVERSITY OF OSLO. Downloaded on January 21,2026 at 19:40:49 UTC from IEEE Xplore. Restrictions apply.

TABLE I: Activity recognition accuracy for all models and modalities.

Modality	HAR-Specific Model	General Model
RGB (R3D)	82.12%	–
RGB	88.27%	96.89%
Depth	86.59%	92.18%
IMU	50.00%	65.71%
RGB+Depth	94.97%	97.76%
RGB+IMU	86.51%	**98.29%**
Depth+IMU	79.99%	93.14%
RGB+Depth+IMU	75.57%	96.57%

are similar Transformer architectures, with only minimal preprocessing. For depth and video, we use *LanguageBind*, while for IMU we use *ImageBind*.

The depth backbone takes a depth image with values clipped and scaled between 0 and 1. The RGB backbone takes a series of 8 consecutive frames. The IMU backbone takes a series of up to 2000 samples with 6 values: tri-axial acceleration and orientation.

D. Evaluations

1) Classification Accuracy

We evaluate the overall classification accuracy of our models by calculating the accuracy for both model conditions across all modality combinations.

2) Generalizability

To test generalizability, we employ zero-shot tests. This is a technique in which a model attempts to classify inputs into classes unseen during training. For *ImageBind* and *LanguageBind*, which include text as an input/output modality, these tests are fairly straightforward. We use a text string starting with “A person is” and then append the class name, e.g. “A person is sitting”. We choose the label from the text-string that best corresponds with the sensor model representation. For the HAR-specific models, we employ overlapping or partially overlapping class labels.

Additionally, we added a setting where we attempted to obtain a more targeted and naturally formulated description of the classes that better coincided with the videos. Since all the subjects are sitting while talking on the phone, for example, we made the class string “A person is sitting and talking on the phone”. As a similar alternative for the HAR models, we selected the model output that empirically correlated the most with each of the classes instead of choosing the labels that are most semantically similar.

3) Privacy Preservation

To evaluate the extent to which the models preserve user privacy we tested fine-tuning the classification head to classify by subject instead of activity. We used a similar leave-one-out cross-validation for evaluation.

IV. Results

A. Classification Accuracy

Table I shows that the general models outperform HAR-specific models. Fusing two modalities increases performance, whereas fusing all three modalities decreases performance. IMU alone resulted in quite poor performance but yielded a larger accuracy gain when paired with RGB rather than depth. Since the general model’s results are better, we focus on those.

TABLE II: Zero-shot accuracy on the PriMA-Care dataset.

Model	Zero-shot	Prompt/Label-engineered
LanguageBind (RGB)	82.85%	90.86%
LanguageBind (Depth)	64.57%	85.71%
NTU-UMDR (RGB)	43.42%	72.00%
NTU-UMDR (Depth)	20.04%	29.61%

TABLE III: Individual participant identification accuracy.

Model	Identification Accuracy
LanguageBind (RGB)	74.19%
LanguageBind (Depth)	45.16%
ImageBind (IMU)	33.89%
R3D (RGB)	50.00%
NTU-UMDR (RGB)	33.87%
NTU-UMDR (Depth)	37.09%
IMU-Net (IMU)	15.63%
ImageBind (IMU) + LanguageBind (Depth)	49.15%

Fig. 1a shows that IMU underperformed RGB (Fig. 1b) in all categories. However, the fused results (Fig. 1c) show an overall improvement. As expected, IMU input helps most on classes distinguished mainly by hand motion.

The depth model (Fig. 1d) underperforms on more fine-grained actions such as brushing hair and brushing teeth. Depth and RGB (Fig. 1f) improves performance over RGB alone. IMU also shows a slight increase in performance (Fig. 1e), however not to the extent of RGB.

B. Generalizability

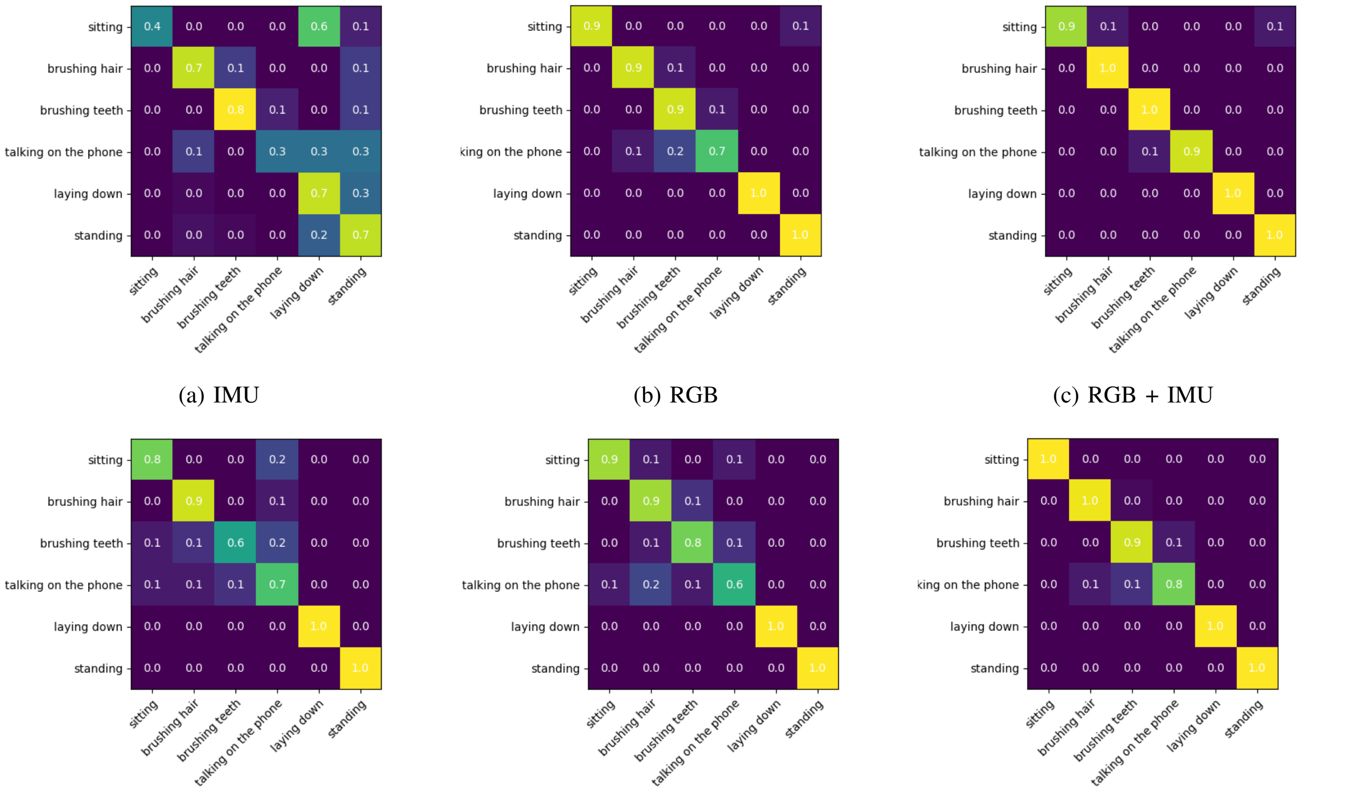
Table II shows a clearly better performance for the language models, especially when combined with RGB. All labels in *PriMA-Care* have similar correspondences in *NTU RGB+D 120* except “laying down”, for which we chose “falling” as the closest alternative. In general, the labels for the *NTU-UMDR* model did not correspond well with PriMA-Care. The labels “brushing hair” and “brushing teeth”, however, did work to a certain degree for RGB.

For *LanguageBind*, we saw that sitting was often confused with the other activities performed while participants were also sitting. Therefore adding this to the text prompt — e.g. “Person sitting and talking on the phone” — improved the results significantly. We refer to this as the “prompt-engineered” approach in Table II.

C. Privacy Preservation

Having the model classify participants is far from a perfect test of privacy preservation. Nonetheless, it provides an indication of how much information related to individuals is learned by the model. Table III unsurprisingly shows that RGB modalities across all models and general models are far better at identifying individuals. However, for all modalities and models, there seems to be some learning of identifying characteristics, as all outperform the random baseline of 7%.

Authorized licensed use limited to: UNIVERSITY OF OSLO. Downloaded on January 21,2026 at 19:40:49 UTC from IEEE Xplore. Restrictions apply.



(a) IMU (b) RGB (c) RGB + IMU

(d) Depth (e) Depth + IMU (f) Depth + RGB

Fig. 1: Confusion matrices for general model modality combinations

V. Discussion

In alignment with the contributions, we focus our discussion on four key issues emerging from our results:

1) The lack of generalization between HAR datasets (**Contribution 1**)
2) The utilization of aggregated datasets to train new HAR models (**Contribution 1**).
3) The potential to enhance privacy in HAR through the application of multimodal learning (**Contribution 3**).
4) The transferability of large pretrained models to HAR tasks (**Contribution 2**).

A. Generalizability of HAR datasets

Supporting contribution 1, we compared the performance of HAR-specific models versus general models. Our results show that the general model outperforms the HAR model and that these effects are substantial, though the small size of the PriMA-Care evaluation set (17 participants) makes it difficult to draw definitive conclusions. The zero-shot tests provide a greater measure of generalizability, as the *PriMA-Care* dataset is another dataset to which the model can overfit in training.

LanguageBind shows the greatest ability to generalize, which points towards a similar performance in real-world applications. Its errors seem to follow logically, such as the ambiguous activity labels shown in Fig. 2.

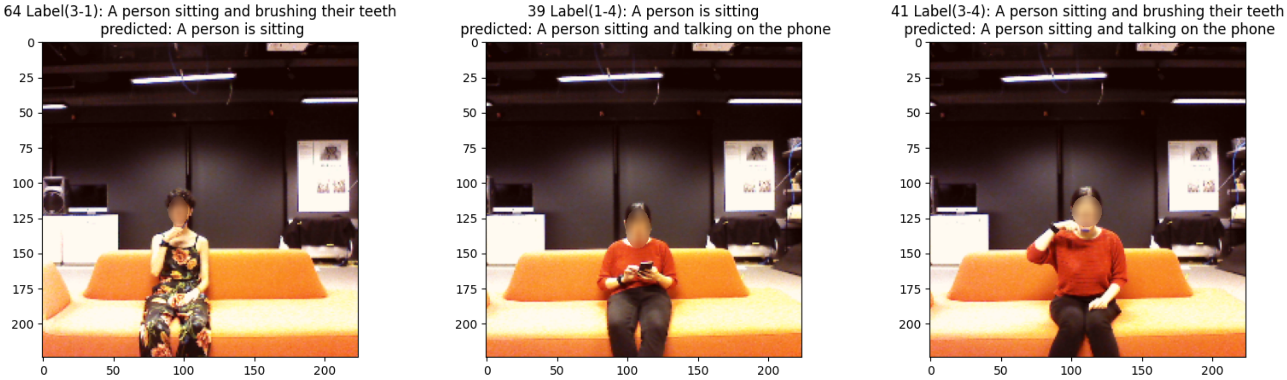
Despite the large size of the *NTU RGB+D 120* dataset, it is not intended for the training of generalizable models but rather as a baseline for evaluating HAR models and methodological approaches. The action sequences of the dataset are highly artificial, with all participants performing activities in a similar manner in the same environment. Models can therefore learn extraneous information not related to the activity, such as "all people talking on the phone are also sitting in that chair".

B. Aggregation of Datasets

One reason for the underperformance of the HAR model seems to lie in the heterogeneity and specificity of HAR datasets, which complicates aggregation and training on multiple sets. Our study met significant challenges in combining multiple HAR datasets, particularly for IMU and depth data. These difficulties arose from variations in sensor configurations, including different sampling rates, depth ranges, and data formats. Additionally, accelerometers and IMUs produce diverse outputs based on their placement, and the aggregated dataset included acceleration data from various body positions. Despite implementing mitigation strategies, fully reconciling these differences proved challenging.

Exemplifying challenges in aggregating datasets is the variation in activity recording lengths and the resulting differences in spectrogram image sequences extracted from acceleration data. For instance, the *UTD-MHAD* dataset

Authorized licensed use limited to: UNIVERSITY OF OSLO. Downloaded on January 21,2026 at 19:40:49 UTC from IEEE Xplore. Restrictions apply.


64 Label(3-1): A person sitting and brushing their teeth
predicted: A person is sitting
39 Label(1-4): A person is sitting
predicted: A person sitting and talking on the phone
41 Label(3-4): A person sitting and brushing their teeth
predicted: A person sitting and talking on the phone
0
25
50
75
100
125
150
175
200
0
50
100
150
200

(a) Confusing activity with posture (b) Confusing posture and device interaction (c) Confusing activities in similar postures

Fig. 2: Errors for the prompt-engineered *LanguageBind*

produced significantly shorter image sequences than the *GOTOV* dataset. Consequently, even common activities such as walking had vastly different representations. This disparity hindered the learning of generalized representations, as evidenced by the superior performance of acceleration encoders trained on individual datasets rather than aggregate. It also explained the subpar performance on the *PriMA-Care* dataset, where image sequence lengths differed considerably from both *GOTOV* and *UTD-MHAD*.

Although creating an aggregated dataset comprising a large number of individual data sets could help address these differences, the current landscape lacks sufficient datasets containing acceleration data. This shortage prevents the creation of a comprehensive dataset on the scale of, for example, *Kinetics-400* for HAR RGB videos or *ImageNet* [27] for more general RGB images. As a result, general models, which benefit from larger and more diverse training data, tend to outperform HAR-specific models in our comparative analysis.

C. Privacy Preservation

Addressing contribution 3, we focused on privacy-preserving sensors such as IMU and depth cameras. While video and image sequences are crucial data modalities in HAR, with most datasets offering a video component, their use, especially in combination with large, pretrained image classification models, raises privacy concerns for both dataset participants and end-users. These models primarily identify general image features rather than specific human activities.

This privacy issue is evident in Experiment IV-C, where we find that even stored feature vectors potentially leak private data. This highlights a specific risk when using transfer learning for HAR instead of training task-specific models, especially in sensitive environments such as healthcare and elderly care. Specific models can reduce this risk to a certain extent, but the mere presence of a camera is still a concern.

While RGB cameras outperform depth sensors for action classification on PriMA-Care, this advantage varies by task [28]. By fusing less identity-revealing modalities—like depth and IMU—we can close some of that performance gap.

D. Transferability of Large, Pretrained Models

Supporting contribution 2, we outline how the large, pre-trained models can be leveraged for multimodal HAR tasks. We found that large transformer models trained on large datasets performed the best. Pretrained models, while focusing more on individual characteristics, still distinguished between activities effectively. The ImageBind model was previously trained on IMU data from head-mounted cameras, yet the features proved useful for wrist-worn sensors, providing a mitigation strategy for the challenges discussed in section V-B. Even for IMU models we see that it seems that large and varied datasets are more important than similarities across activities. However, the lack of orientation data in the HAR-specific model can also be a source of reduced performance. General models trained through contrastive learning outperformed the *UMDR* model despite the similar datasets. This suggests that HAR datasets' limited nature outweighs specialized training benefits. The distance from camera at which actions were performed, as well as differences in posture (sitting vs standing), can be a large contributor to incorrect prediction.

The *R3D* network's underperformance, despite its large HAR dataset training, can indicate that the contrastive method improves the ability of models to transfer. However, the low resolution of 112 × 112 can also be the source of confusion for more fine-grained actions.

The zero-shot and prompt/label-engineered approaches, while not achieving accuracy scores approaching the fine-tuned results, show a promising avenue for future work to meet the challenges of heterogeneous datasets. This enables the classification of activities beyond those seen during training.

VI. Conclusion

In this study, we evaluated transfer learning across modalities for Human Activity Recognition (HAR), with a particular focus on privacy preservation. Our findings support the conclusion that large general models pretrained on diverse

Authorized licensed use limited to: UNIVERSITY OF OSLO. Downloaded on January 21,2026 at 19:40:49 UTC from IEEE Xplore. Restrictions apply.

datasets consistently outperform specialized HAR models trained on aggregated HAR datasets, challenging the conventional assumption that domain-specific models work best for specialized tasks.

Multimodal Fusion Benefits: Combining two modalities enhanced performance (RGB+IMU: 98.29%), though adding a third modality did not provide additional benefit. **Privacy-Preserving Alternatives:** Privacy-focused modalities (Depth+IMU) achieved 93.14% accuracy, demonstrating their viability for healthcare applications. **Transfer Learning Efficacy:** General models surpassed HAR-specific ones, reaching 82.85% accuracy in zero-shot classification using *LanguageBind* Video. **Dataset Aggregation Challenges:** Variations in sensor configurations and protocols highlighted the need for standardized HAR data collection methods, especially in relation to IMU data.

This study has several limitations, primarily the limited size of the *PriMA-Care* dataset and task-specific focus, which constrain our ability to make broad generalizations. Future research should explore zero-shot classification for HAR by evaluating privacy-preserving modalities with various general multimodal models across different datasets. This direction shows promise for the development of HAR systems that balance accuracy with privacy concerns, a crucial requirement for healthcare and homecare robotics applications.

This research advances our understanding of how transfer learning from large, general models can enhance specialized tasks while prioritizing critical ethical concerns such as user privacy.

ACKNOWLEDGMENT

This work is partially supported by the Research Council of Norway as part of the Vulnerability in the Robot Society (VIROS) project (Grant No. 288285), the Predictive and Intuitive Robot Companion (PIRC) project (Grant No. 312333), and through its Centers of Excellence scheme, RITMO (Project No. 262762).

REFERENCES

[1] M. Vrigkas, C. Nikou, and I. A. Kakadiaris, "A review of human activity recognition methods," Frontiers in Robotics and AI, vol. 2, p. 28, 2015.

[2] G. Bhola and D. K. Vishwakarma, "A review of vision-based indoor har: state-of-the-art, challenges, and future prospects," Multimedia Tools and Applications, vol. 83, no. 1, pp. 1965–2005, 2024.

[3] X. Qin, J. Wang, Y. Chen, W. Lu, and X. Jiang, "Domain generalization for activity recognition via adaptive feature fusion," ACM Transactions on Intelligent Systems and Technology, vol. 14, Nov. 2022.

[4] A. Baselizadeh, M. Z. Uddin, W. Khaksar, D. S. Lindblom, and J. Torresen, "PriMA-Care: Privacy-Preserving Multi-modal Dataset for Human Activity Recognition in Care Robots," in Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction, HRI '24, (New York, NY, USA), p. 233–237, Association for Computing Machinery, 2024.

[5] Z. Sun, Q. Ke, H. Rahmani, M. Bennamoun, G. Wang, and J. Liu, "Human action recognition from various data modalities: A review," IEEE transactions on pattern analysis and machine intelligence, vol. 45, no. 3, pp. 3200–3225, 2022.

[6] A. Baselizadeh, D. S. Lindblom, W. Khaksar, M. Z. Uddin, and J. Torresen, "Comparative analysis of vision-based sensors for human monitoring in care robots: Exploring the utility-privacy trade-off," in 2024 33rd IEEE International Conference on Robot and Human Interactive Communication (ROMAN), pp. 1794–1801, 2024.

[7] S. G. Dhekane and T. Ploetz, "Transfer learning in human activity recognition: a survey," 2024. arXiv: 2401.10185 [cs.LG].

[8] D. Lahat, T. Adali, and C. Jutten, "Multimodal data fusion: An overview of methods, challenges, and prospects," Proceedings of the IEEE, vol. 103, no. 9, p. 1449–1477, 2015. Citation Key: 7214350.

[9] S. K. Yadav, K. Tiwari, H. M. Pandey, and S. A. Akbar, "A review of multimodal human activity recognition with special emphasis on classification, applications, challenges and future directions," Knowledge-Based Systems, vol. 223, p. 106970, July 2021.

[10] X. Wei and Z. Wang, "Tcn-attention-har: human activity recognition based on attention mechanism time convolutional network," Scientific Reports, vol. 14, p. 7414, Mar. 2024.

[11] F. Luo, S. Khan, Y. Huang, and K. Wu, "Activity-based person identification using multimodal wearable sensor data," IEEE Internet of Things Journal, vol. 10, p. 1711–1723, Jan. 2023.

[12] B. Zhu, B. Lin, M. Ning, Y. Yan, J. Cui, H. Wang, Y. Pang, W. Jiang, J. Zhang, Z. Li, et al., "Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment," arXiv preprint arXiv:2310.01852, 2023.

[13] R. Girdhar, A. El-Nouby, Z. Liu, M. Singh, K. V. Alwala, A. Joulin, and I. Misra, "ImageBind One Embedding Space to Bind Them All," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2023-June, 2023.

[14] S. J. Pan and Q. Yang, "A survey on transfer learning," IEEE Transactions on Knowledge and Data Engineering, vol. 22, no. 10, pp. 1345–1359, 2010.

[15] W. Dai, Q. Yang, G.-R. Xue, and Y. Yu, "Boosting for transfer learning," in Proceedings of the 24th international conference on Machine learning, pp. 193–200, 2007.

[16] T. Evgeniou and M. Pontil, "Regularized multi–task learning," in Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 109–117, 2004.

[17] R. Raina, A. Battle, H. Lee, B. Packer, and A. Y. Ng, "Self-taught learning: transfer learning from unlabeled data," in Proceedings of the 24th international conference on Machine learning, pp. 759–766, 2007.

[18] S. G. Dhekane and T. Ploetz, "Transfer learning in human activity recognition: A survey," arXiv preprint arXiv:2401.10185, 2024.

[19] D. Cook, K. D. Feuz, and N. C. Krishnan, "Transfer learning for activity recognition: A survey," Knowledge and information systems, vol. 36, pp. 537–556, 2013.

[20] C. Chen, R. Jafari, and N. Kehtarnavaz, "Utd-mhad: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor," in 2015 IEEE International Conference on Image Processing (ICIP), pp. 168–172, 2015.

[21] J. Liu, A. Shahroudy, M. Perez, G. Wang, L.-Y. Duan, and A. C. Kot, "Ntu rgb+d 120: A large-scale benchmark for 3d human activity understanding," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 42, p. 2684–2701, Oct. 2020.

[22] S. Paraschiakos, C. R. de Sá, J. Okai, P. E. Slagboom, M. Beekman, and A. Knobbe, "A recurrent neural network architecture to model physical activity energy expenditure in older people," Data Mining and Knowledge Discovery, vol. 36, p. 477–512, Jan. 2022.

[23] S. Paraschiakos, R. Cachucho, M. Moed, D. van Heemst, S. Mooijaart, E. P. Slagboom, A. Knobbe, and M. Beekman, "Activity recognition using wearable sensors for tracking the elderly," User Modeling and User-Adapted Interaction, vol. 30, p. 567–605, July 2020.

[24] B. Zhou, P. Wang, J. Wan, Y. Liang, and F. Wang, "A unified multimodal de- and re-coupling framework for rgb-d motion recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 45, no. 10, pp. 11428–11442, 2023.

[25] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pp. 6450–6459, 2018.

[26] H. Yoon, H. Cha, H. C. Nguyen, T. Gong, and S.-J. Lee, "Img2imu: Translating knowledge from large-scale images to imu sensing applications," 2024.

[27] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in 2009 IEEE conference on computer vision and pattern recognition, p. 248–255, 2009.

[28] A. Baselizadeh, W. Khaksar, M. Z. Uddin, D. Saplacan, and J. Torresen, "Privacy-preserving user pose prediction for safe and efficient human-robot interaction," in 2023 IEEE 19th International Conference on Automation Science and Engineering (CASE), pp. 1–8, 2023.

Authorized licensed use limited to: UNIVERSITY OF OSLO. Downloaded on January 21,2026 at 19:40:49 UTC from IEEE Xplore. Restrictions apply.