MULTIMODAL CLUSTERING FOR WEAK ANNOTATION OF HUMAN ACTIVITY RECOGNITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Annotating Inertial Measurement Unit (IMU) data for Human Activity Recognition (HAR) is challenging and time-consuming, as raw inertial signals are difficult for humans to directly interpret and assign activity labels. Recent studies therefore leverage synchronized ego-centric video for weak annotation, using visual representations to cluster activity samples and requiring manual labels only for representative samples. However, visual data can raise privacy concerns and its representations may encode scene and recording context unrelated to the performed activity, influencing the resulting feature space, clustering structure, and activity labels. To incorporate additional activity-relevant information, we propose **WAFER** (**W**earable **A**nd visual **F**oundation **E**mbeddings for weak annotation of HA**R**), a multimodal weak annotation method that incorporates representations from a pretrained wearable foundation model alongside visual and optical-flow features. Temporally aligned feature embeddings from video, wearable IMU signals, and optical flow are independently standardised, weighted, and fused before clustering. Human annotation is required only for few representative samples, after which activity labels are propagated to the remaining samples within each cluster. The resulting weak annotations are then used to train downstream HAR models using only IMU data, such that video is not required during downstream training. Across three benchmark datasets, WAFER improves weak annotation accuracy by up to 6.7. Using these weak annotations for downstream HAR training improves accuracy by up to 6.7 and macro F1 by up to 10.5, achieving performance approaching the performance of fully supervised training.

1 INTRODUCTION

Recent advances in deep learning have enabled Human Activity Recognition (HAR) systems to achieve increasingly accurate activity recognition performance from wearable sensor data Singh et al. (2020); Bock et al. (2021); Ordóñez & Roggen (2016); Zhang et al. (2022); Zhou et al. (2022). Despite this progress, developing such systems still depends heavily on the availability of large, accurately labeled datasets Roggen et al. (2010). Annotating Inertial Measurement Unit (IMU) recordings is particularly challenging, as raw inertial signals are difficult for humans to directly interpret and associate with the activities being performed, making manual annotation tedious and time-consuming. For instance, annotating a 30-minute recording of the Opportunity++ dataset required approximately 7-10 hours of manual annotation Ciliberto et al. (2021).

These challenges have motivated research on reducing the annotation burden for HAR through approaches such as active learning Adaimi & Thomaz (2019), cross-modal learning Xia et al. (2024a), and weak annotation Bock et al. (2024b); Schönfeldt et al. (2025). In particular, synchronized video has been used as an auxiliary modality for weakly annotating IMU data, as it provides an interpretable view of the performed activities. For example, the clustering-based approach of Bock et al. (2024b) extracts CLIP-based Radford et al. (2021) representations from synchronized ego-centric video, groups them into clusters, and requires a human annotator to label only representative samples before propagating these labels within each cluster. However, visual data can introduce privacy concerns, particularly in human-centered, everyday activities, while visual representations may also encode scene and background information unrelated to the performed activity Zhou et al. (2025). Consequently, such environmental context may influence how activity samples are represented and grouped, potentially affecting clustering and propagated annotations.

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

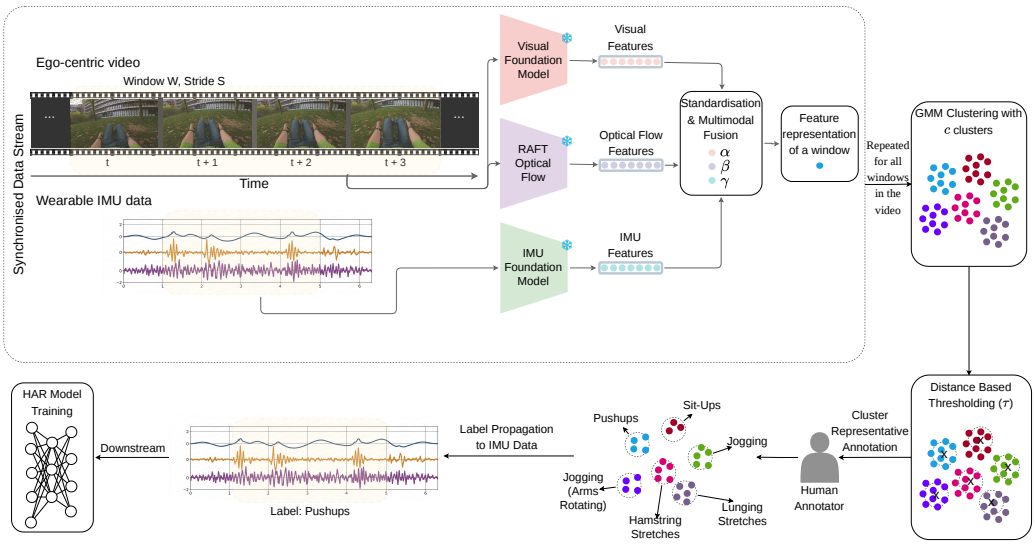


Figure 1: **Overview of WAFER.** Temporally aligned video, optical-flow and IMU representations are fused and clustered. Representative samples are manually annotated, and the propagated labels are used for downstream HAR training.

To address this limitation, we propose **WAFER** (**W**earable **A**nd visual **F**oundation **E**mbeddings for weak annotation of HA**R**), a multimodal approach for weak annotation of IMU data for HAR. WAFER complements visual and optical-flow representations with features from a pretrained wearable foundation model, providing an additional activity-related signal for clustering. The temporally aligned representations are independently standardised and fused using modality-specific weights before clustering. Human annotation is required only for few representative cluster samples, after which their labels are propagated to the remaining samples. The resulting weak labels are used to train downstream HAR models using only IMU data, to evaluate whether the generated annotations provide effective supervision for activity recognition. Therefore, synchronized video is required only during weak annotation and not during downstream HAR training. The overview of WAFER is shown in Figure 1.

To evaluate our approach, we conduct experiments on three benchmark HAR datasets: WEAR Bock et al. (2024a), ActionSense DelPreto et al. (2022), and Wetlab Scholl et al. (2015). WAFER achieves higher annotation accuracy than the weak annotation approach of Bock et al. (2024b). Furthermore, when the weak annotations are used to train TinyHAR Zhou et al. (2022) and DeepConvLSTM Bock et al. (2021), the resulting activity recognition performance approaches to fully supervised training.

Our main contributions are:

- **The WAFER Approach.** We introduce WAFER, a multimodal clustering-based approach that addresses the labeling bottleneck of difficult to interpret IMU recordings by integrating visual, wearable, and optical-flow representations while requiring manual annotation only for few representative cluster samples.
- **Complementary Wearable Representations.** We demonstrate the benefit of incorporating pretrained wearable representations as complementary activity information during weak annotation. Across three benchmark datasets, WAFER improves annotation accuracy by up to 6.7, with the resulting labels improving downstream HAR accuracy by up to 6.7 and macro F1 by up to 10.5.
- **Visual Context Analysis.** We analyse the learned feature spaces of CLIP Radford et al. (2021) and show that visual representations are influenced by scene and background context, motivating the incorporation of complementary wearable representations that provide additional activity-related information and improve the alignment of the feature space with activity semantics.

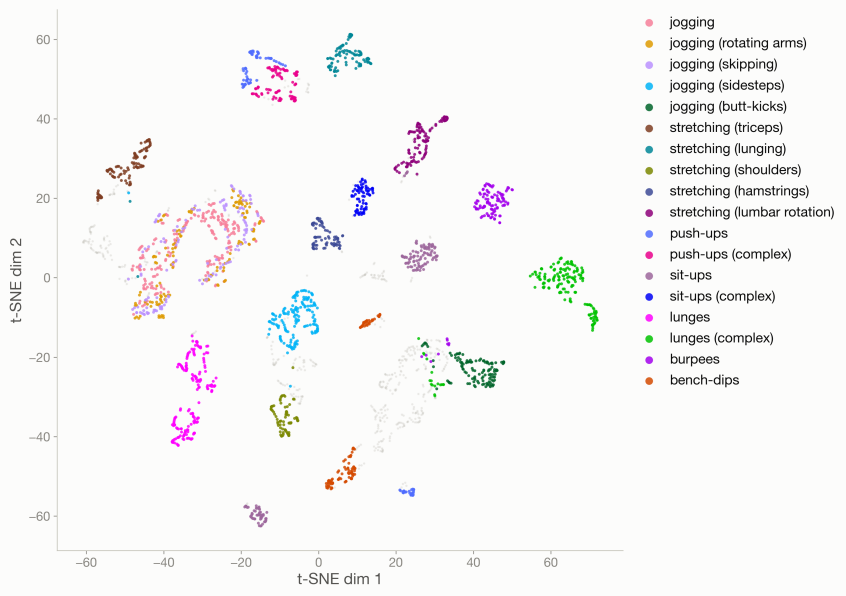


Figure 2: **CLIP Latent Space by Activity.** t-SNE visualization of CLIP embeddings for subject 0 of WEAR, colored by ground-truth activity labels.

2 Related Work

Reducing the annotation burden in HAR has been explored through active learning, few-shot learning, and cross-modal supervision. Active-learning approaches selectively identify informative sensor samples for manual annotation Adaimi & Thomaz (2019), while cross-modal approaches exploit complementary information from other modalities, such as visual or semantic representations, to support activity recognition with limited labelled sensor data Tong et al. (2021); Xia et al. (2024a;b). These approaches primarily reduce the amount of labelled data required for model training rather than generating annotations for the underlying sensor dataset.

More recent work has directly investigated visual information for weak annotation. The most related to our work is Bock et al. (2024b), which extracts visual and optical-flow representations from synchronized egocentric video, clusters these representations, and requires manual annotation only for representative cluster samples. The assigned labels are then propagated within each cluster to generate weak labels for the corresponding IMU data. Schönfeldt et al. (2025) further investigate vision-language models for reducing annotation effort in HAR datasets by automatically annotating activity-intensity categories from egocentric data. WAFER incorporates representations from a pretrained wearable foundation model alongside visual and optical-flow features, providing complementary information directly from the wearable signals for clustering-based weak annotation.

3 Methodology

3.1 Motivation: Latent Space Analysis of Visual Context

Existing weak annotation approach uses visual representations to cluster synchronized activity samples and propagate labels from a small number of manually annotated representatives to the remaining IMU samples Bock et al. (2024b). For this label propagation to be reliable, the visual feature space should group samples based primarily on the performed activity, such that samples within the same cluster are likely to share the same activity label. To examine how activity and visual context are reflected in the CLIP Radford et al. (2021) feature space, we analyze how its representations are organized using t-SNE Van der Maaten & Hinton (2008). Specifically, we examine whether samples are grouped according to the performed activity or are also influenced by scene and recording con-

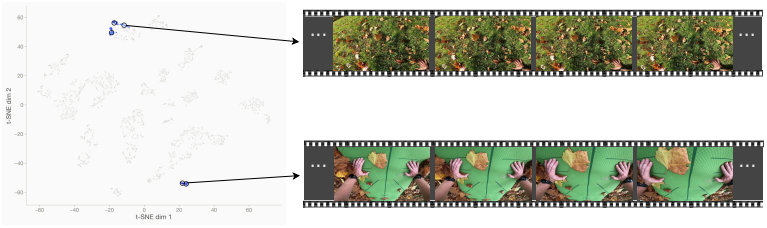


Figure 3: **Within-Activity Separation in the CLIP Latent Space: Push-Ups.** t-SNE visualization of CLIP embeddings corresponding to Push-Ups for subject 0 of WEAR.

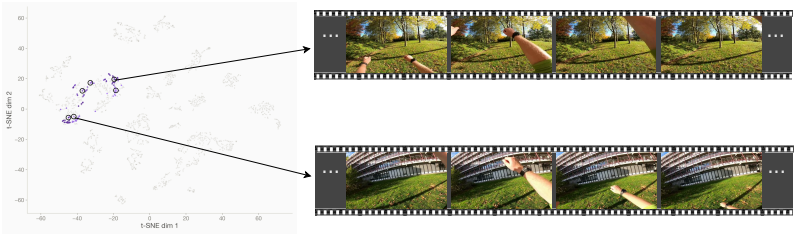


Figure 4: **Within-Activity Separation in the CLIP Latent Space: Jogging (Skipping).** t-SNE visualization of CLIP embeddings corresponding to Jogging (Skipping) for subject 0 of WEAR.

text. We conduct this analysis using the CLIP embeddings of subject 0 from the WEAR Bock et al. (2024a) dataset. Figure 2 visualizes these embeddings colored by ground-truth activity labels.

First, we examine the distribution of windows belonging to the same ground-truth activity. Figure 3 shows that Push-Ups performed on an exercise mat and directly on the ground occupy spatially separated regions of the feature space, despite representing the same activity. Similarly, Figure 4 shows that Jogging (Skipping) samples recorded with and without a building in the background form distinct regions. We define *within-activity separation* as cases where samples from the same activity form distinct t-SNE regions associated with different visual contexts. This separation is observed for 12 of the 19 activities, suggesting that scene appearance can influence visual representations even when the performed activity remains the same.

Secondly, we apply a Gaussian Mixture Model (GMM) with 100 components to these embeddings and identify clusters containing windows from multiple ground-truth activities. We define a *mixed-activity cluster* as a cluster containing samples from more than one ground-truth activity. Figure 5 shows that Lunges (Complex), Burpees, and Jogging (Butt-Kicks) are grouped within the same cluster, with all activities recorded with similar ground and outdoor background. Similarly, Figure 6 shows Jogging (Butt-Kicks) and Burpees occupying the same cluster, where representative windows again exhibit similar surrounding visual context despite different activity labels. Overall, 21 of the 100 clusters are mixed-activity clusters with visually similar recording contexts.

Together, these observations indicate that *the CLIP features are influenced by surrounding visual context and are not fully aligned with activity semantics.* This motivates incorporating wearable representations, which provide complementary activity-related information directly from the IMU signals rather than the visual scene.

3.2 Problem Formulation

We consider a HAR dataset consisting of synchronized video and wrist-worn inertial recordings for each subject. Let $X = \{x_1, x_2, \ldots, x_T\}$ denote the video stream containing T frames recorded at f_v frames per second, and $I = \{i_1, i_2, \ldots, i_M\}$ denote the corresponding IMU signal containing M tri-axial accelerometer samples $i_m \in \mathbb{R}^3$ acquired at a sampling rate of f_s Hz.

To capture the temporal context of human activities, the synchronized streams are partitioned into overlapping windows of duration W with a stride of S seconds. This yields N temporally aligned

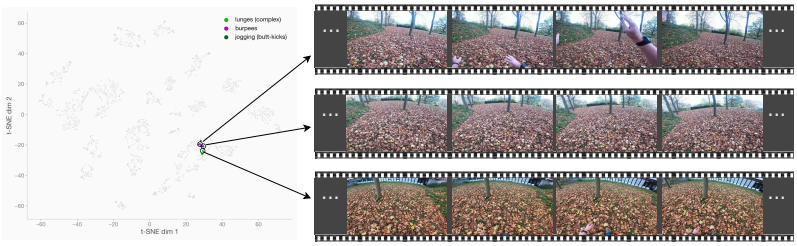


Figure 5: **Mixed-Activity Cluster.** t-SNE visualization of a GMM cluster for subject 0 of WEAR containing windows from multiple ground-truth activities.

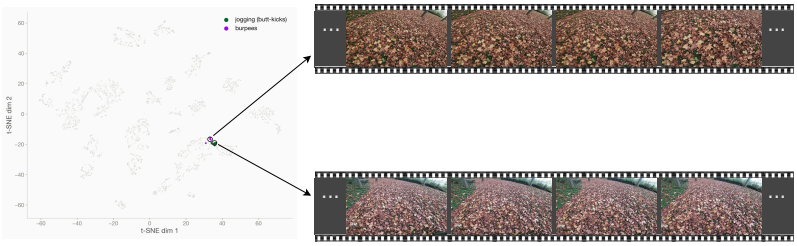


Figure 6: **Mixed-Activity Cluster.** t-SNE visualization of a GMM cluster for subject 0 of WEAR containing windows from multiple ground-truth activities.

video $\mathcal{X} = \{X_1, \ldots, X_N\}$ and IMU windows $\mathcal{I} = \{I_1, \ldots, I_N\}$, where X_n and I_n contain the video frames and IMU samples, respectively, corresponding to the same window n.

The objective is to assign activity labels to the IMU samples while minimizing the amount of manual annotation. WAFER therefore uses multimodal information during weak annotation and these annotations are subsequently used to train downstream HAR models that operate only on IMU data.

3.3 WAFER Weak Annotation

Multimodal Feature Extraction. For each temporally aligned window, feature representations are independently extracted from the corresponding video frames, wearable IMU signals, and optical-flow information using their respective pretrained models.

- **Visual Features.** Each frame $x_t,\ t = 1, \ldots, T$, is independently passed through a visual foundation model $V(\cdot)$ to extract frame-level feature embeddings, $e_t^v = V(x_t) \in \mathbb{R}^{d_v}$. The frame-level embeddings corresponding to the frames in window X_n are then mean pooled to obtain a window-level representation, $E_n^v = \frac{1}{|X_n|} \sum_{x_t \in X_n} e_t^v \in \mathbb{R}^{d_v}$.
- **Wearable IMU Features.** For each IMU window $I_n \in \mathcal{I}$, the inertial sequence is provided to a wearable foundation model $F(\cdot)$, which directly produces a window-level feature embedding, $E_n^s = F(I_n) \in \mathbb{R}^{d_s}$.
- **Optical Flow Features.** To capture motion dynamics, dense optical flow is computed between consecutive frames of each video window X_n and encoded by $R(\cdot)$ into a window-level motion embedding, $E_n^f = R(X_n) \in \mathbb{R}^{d_f}$.

Feature Standardisation and Fusion. Since feature embeddings are produced on different numeric scales, each modality $m \in \{v, s, f\}$ is standardised independently for each subject before fusion, $\tilde{E}_n^m = (E_n^m - \mu^m)/\sigma^m$, where μ^m and σ^m are the per-dimension mean and standard deviation computed. Without standardisation, modalities with larger feature magnitudes would dominate the fused representation irrespective of their semantic relevance.

The standardised modalities are then fused at each window n by weighted concatenation, $z_n = [\,\alpha\,\tilde{E}_n^v \,\|\, \beta\,\tilde{E}_n^s \,\|\, \gamma\,\tilde{E}_n^f\,]$, where $\|$ denotes concatenation, with $\alpha + \beta + \gamma = 1$ and $\alpha, \beta, \gamma \geq 0$, so that the weights control the relative influence of each modality.

Table 1: **Threshold to Retention Mapping.** Percentage of windows retained for the absolute distance thresholds used by Bock et al. (2024b). WAFER uses these percentages rather than directly applying the absolute distance thresholds.

Dataset	Threshold $T = 6$	Threshold $T = 4$
WEAR Bock et al. (2024a)	77.6%	35.5%
ActionSense DelPreto et al. (2022)	45.3%	13.9%
Wetlab Scholl et al. (2015)	83.0%	49.0%

Table 2: **Average Annotation Accuracies.** Average percentage of IMU samples whose propagated weak label matches the ground-truth activity label. Best in Bold.

Dataset	**Bock et al. (2024b) ↑**	**WAFER (ours) ↑**	**Gain**
WEAR Bock et al. (2024a)	83.96	**90.64**	**+6.68**
ActionSense DelPreto et al. (2022)	57.29	**60.84**	**+3.55**
Wetlab Scholl et al. (2015)	66.23	**71.60**	**+5.37**

Weak Annotation. We apply GMM-based clustering using multimodal embeddings z_n. A GMM with C components is fitted independently for each subject, assigning each window to a cluster $c_n \in \{1, \ldots, C\}$. Because GMMs model probability distributions rather than explicitly defining cluster centroids, the representative window of each cluster is selected as the sample with the highest likelihood under the corresponding Gaussian component. A human annotator labels only this representative window with an activity label $a \in \{1, \ldots, A\}$. This annotation is subsequently propagated to all remaining windows assigned to the same cluster.

Downstream HAR Training. To evaluate the quality of the generated weak labels, the propagated window-level labels are mapped back to the underlying IMU samples and used to train a downstream HAR model M_{dl} using only the raw IMU signals.

4 Experimental Results

4.1 Experimental Setup

Datasets. We evaluate WAFER on three publicly available multimodal HAR datasets: WEAR Bock et al. (2024a), ActionSense DelPreto et al. (2022), and Wetlab Scholl et al. (2015), containing 19, 20, and 9 activity classes, respectively. These datasets provide synchronized video recordings and wrist-worn IMU measurements together with frame-level activity annotations. Visual information is used only during the weak annotation stage, while the downstream HAR models are trained and evaluated exclusively on IMU data.

Implementation Details. The visual modality is represented using CLIP Radford et al. (2021), while wearable representations are extracted using the pretrained OxWearables foundation model Yuan et al. (2024). For the optical flow features, dense optical flow is estimated between consecutive video frames using RAFT Teed & Deng (2020), and the resulting flow sequences are encoded into window-level motion representations using I3D Carreira & Zisserman (2017). Video streams are sampled at 12 fps and IMU signals at 50 Hz. Temporal windows of 4 seconds with a stride of 1 second are used to construct synchronized representations. The multimodal fusion weights are set to $(\alpha, \beta, \gamma) = (0.50, 0.25, 0.25)$, and a GMM with 100 components is fitted independently for each subject.

Evaluation Protocol. We evaluate WAFER at both the annotation and downstream HAR levels. Annotation quality is measured as the percentage of IMU samples whose propagated weak activity label matches the corresponding ground-truth activity label. To evaluate the usefulness of the generated weak labels, downstream HAR models are trained using only the raw IMU signals under a Leave-One-Subject-Out (LOSO) evaluation protocol. We evaluate two representative HAR architectures: TinyHAR Zhou et al. (2022) and Shallow DeepConvLSTM Bock et al. (2021), reporting classification accuracy and macro-averaged F1 score.



Figure 7: **Comparison of CLIP and WAFER representations for Bench-Dips.** t-SNE visualization of CLIP (left) and WAFER (right) features corresponding to Bench-Dips for subject 0 of WEAR.

Table 3: **Downstream HAR performance.** Accuracy and macro F1 (%) for TinyHAR and DeepConvLSTM trained using weak annotations. T-6 and T-4 denote thresholded training for WAFER. Bold indicates the better result between Bock et al. (2024b) and WAFER whereas yellow highlights the best weak-training result in each column for each dataset.

Dataset	Training Type	Accuracy (%) ↑				F1 ↑			
		TinyHAR		DeepConvLSTM		TinyHAR		DeepConvLSTM	
		Bock et al. (2024b)	WAFER (ours)	Bock et al. (2024b)	WAFER (ours)	Bock et al. (2024b)	WAFER (ours)	Bock et al. (2024b)	WAFER (ours)
WEAR Bock et al. (2024a)	Fully Supervised	77.83	77.83	85.14	85.14	71.89	71.89	82.23	82.23
	Few-Shot-CE	65.61	**68.58**	65.61	**68.57**	53.51	**58.34**	53.51	**57.65**
	Random-CE	66.27	**68.89**	65.91	**66.33**	55.00	**57.94**	53.38	**54.73**
	Weak-CE	71.10	**76.16**	73.38	**80.12**	59.65	**67.57**	63.23	**73.77**
	Weak-PHGCE	73.37	**76.88**	76.15	**81.02**	61.68	**68.14**	63.43	**72.39**
	Weak-CE-T-6	72.03	**75.51**	76.28	**79.70**	60.31	**68.18**	66.13	**74.59**
	Weak-PHGCE-T-6	74.05	**77.03**	77.55	**81.64**	61.67	**68.86**	64.77	**73.71**
	Weak-CE-T-4	73.93	**75.50**	77.81	**78.63**	63.42	**68.16**	68.22	**72.93**
	Weak-PHGCE-T-4	74.83	**75.93**	76.64	**78.81**	63.94	**67.48**	64.81	**71.48**
ActionSense DelPreto et al. (2022)	Fully Supervised	22.19	22.19	20.73	20.73	18.04	18.04	14.82	14.82
	Few-Shot-CE	11.15	**11.74**	**9.96**	8.57	6.72	**6.92**	**6.08**	5.71
	Random-CE	**11.44**	11.41	10.02	**10.19**	6.50	**6.67**	5.77	**6.28**
	Weak-CE	17.16	**18.02**	**17.35**	16.26	13.45	**14.39**	**12.20**	11.30
	Weak-PHGCE	14.52	**15.62**	11.62	**13.02**	10.09	**11.27**	7.02	**8.23**
	Weak-CE-T-6	17.53	**18.32**	15.29	**15.68**	13.15	**13.52**	10.69	**11.77**
	Weak-PHGCE-T-6	14.53	**14.77**	**11.78**	11.03	9.01	**9.75**	6.79	**7.23**
	Weak-CE-T-4	15.03	**17.07**	12.08	**12.81**	11.16	**12.82**	8.80	**9.46**
	Weak-PHGCE-T-4	12.11	**13.57**	9.91	**10.90**	7.35	**7.64**	5.56	**6.03**
Wetlab Scholl et al. (2015)	Fully Supervised	38.75	38.75	45.27	45.27	28.85	28.85	38.64	38.64
	Few-Shot-CE	25.92	**30.09**	22.78	**28.20**	18.27	**20.33**	17.38	**20.84**
	Random-CE	**29.70**	26.11	**26.50**	23.02	**19.37**	18.41	**20.05**	17.55
	Weak-CE	**36.30**	36.01	36.51	**39.63**	25.76	**25.94**	29.72	**32.45**
	Weak-PHGCE	**35.53**	35.07	33.78	**34.10**	**25.53**	25.38	**27.62**	27.13
	Weak-CE-T-6	36.20	**36.39**	38.41	**40.15**	25.76	**26.24**	29.77	**32.08**
	Weak-PHGCE-T-6	34.93	**35.09**	32.29	**33.79**	25.17	**25.47**	26.71	**27.37**
	Weak-CE-T-4	34.64	**35.76**	35.25	**38.06**	25.04	**25.46**	28.95	**30.11**
	Weak-PHGCE-T-4	33.66	**34.29**	31.20	**33.39**	**24.70**	24.58	25.64	**27.30**

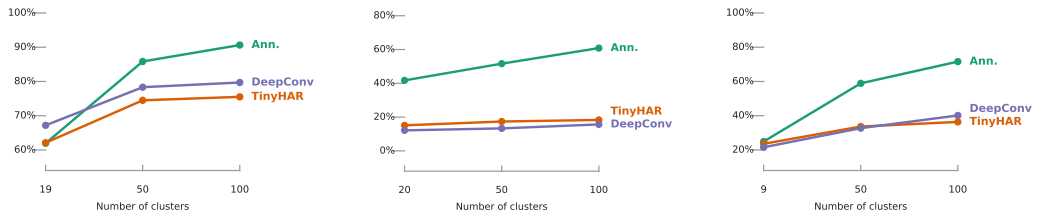
We compare the following five training types:

- *Fully-Supervised*: trained using ground-truth labels with the standard cross-entropy (CE) classification loss.
- *Few-Shot-CE*: trained using only the manually annotated representative samples.
- *Random-CE*: trained using an equal number of randomly selected labelled samples.
- *Weak-CE*: trained using the retained weakly labelled samples.
- *Weak-PHGCE*: trained using the retained weakly with partially Huberised generalised cross-entropy (PHGCE) Menon et al. (2020).

Thresholding. Windows within each cluster are ranked by their Euclidean distance to the representative sample, and distant windows are discarded before downstream training. The authors of Bock et al. (2024b) use absolute distance thresholds $T = 6$ and $T = 4$. Because feature standardisation and multimodal fusion change the scale of WAFER's embedding space, directly applying these absolute thresholds would not provide a comparable filtering level. We therefore match the retention percentage induced by each baseline threshold: for each dataset and cluster configuration, WAFER retains the same proportion of windows as retained by $T = 6$ or $T = 4$ in the baseline. We denote these matched retention settings as T-6 and T-4 for consistency with the baseline. The corresponding retention percentages are reported in Table 1 and more detailed retention percentages in Appendix.

Table 4: **Modality-Specific Weight Space Search.** Annotation accuracy (%) on WEAR for different fusion weights (α, β, γ) corresponding to CLIP, IMU, and RAFT, respectively, using 100 GMM clusters. Best in Bold.

	Fusion Weight Configuration														
CLIP α	0.00	0.00	0.25	0.00	0.25	0.50	0.00	0.25	**0.50**	0.75	0.00	0.25	0.50	0.75	1.00
IMU β	1.00	0.75	0.75	0.50	0.50	0.50	0.25	0.25	**0.25**	0.25	0.00	0.00	0.00	0.00	0.00
RAFT γ	0.00	0.25	0.00	0.50	0.25	0.00	0.75	0.50	**0.25**	0.00	1.00	0.75	0.50	0.25	0.00
Annotation Acc. (%) ↑	56.20	64.77	70.46	83.67	84.73	87.12	81.28	87.45	**90.64**	86.57	79.25	83.17	85.14	83.69	82.81



(a) WEAR (b) ActionSense (c) Wetlab

Figure 8: **Effect of the Number of GMM Clusters.** Annotation accuracy (Ann.) and downstream HAR performance of TinyHAR and DeepConvLSTM for different numbers of GMM clusters on WEAR, ActionSense, and Wetlab. Downstream results use Weak-CE-T-6 with matched retention ratios for each cluster configuration.

4.2 Weak Annotation Results

Table 2 reports annotation accuracy showing that WAFER consistently improves accuracy across all three datasets, achieving 90.64% on WEAR Bock et al. (2024a), 60.84% on ActionSense DelPreto et al. (2022), and 71.60% on Wetlab Scholl et al. (2015), corresponding to absolute improvements of 6.68, 3.55, and 5.37 respectively. Activity-wise confusion matrix is provided in Appendix.

To further assess the effect of wearable representations, we repeat the latent space analysis described in Section 3.1 using WAFER representations. While within-activity separation is observed for 12 of the 19 activities in CLIP, WAFER reduces this to 8 of the 19 activities. Figure 7 shows an example of Bench-Dips performed in two different visual contexts, one with an exercise mat and building in the background and another without them. While these samples appear separated in the CLIP representation space, they are positioned closer together in the WAFER representation space. Furthermore, CLIP produces 21 mixed-activity clusters out of 100, compared with 15 out of 100 for WAFER. These observations suggest that *incorporating wearable representations provides complementary activity-related information and can reduce the influence of visual context in the learned feature space, contributing to improved weak-annotation quality.*

4.3 Downstream HAR Results

Table 3 reports the downstream HAR performance using WAFER weak labels. Across the three datasets and two HAR architectures, WAFER generally outperforms Bock et al. (2024b), with the largest improvements observed on WEAR Bock et al. (2024a). For example, *Weak-PHGCE* with DeepConvLSTM improves accuracy from 76.15% to 81.02% and macro F1 from 63.43 to 72.39. Notably, this performance approaches the fully supervised DeepConvLSTM accuracy of 85.14%, despite training on propagated weak labels. Thresholding the weak annotations further improves performance in several settings by retaining samples that are closer to their cluster representatives. Among the evaluated configurations, *Weak-CE-T-6* provides comparatively consistent performance across datasets and architectures. On ActionSense, for example, WAFER with *Weak-CE-T-6* improves TinyHAR accuracy from 17.53% to 18.32% and DeepConvLSTM macro F1 from 10.69 to 11.77. On Wetlab, the largest gains are observed with DeepConvLSTM, where *Weak-CE-T-6* reaches 40.15% accuracy and 32.08 macro F1, compared with 38.41% and 29.77% using the annotations of Bock et al. (2024b). Overall, WAFER improves the strongest weak-training results on each dataset,

although gains are not uniform across all training configurations. Therefore, *incorporating wearable representations alongside visual and optical-flow features improves downstream HAR performance across the evaluated datasets. These results indicate that improvements in weak annotation quality can translate into stronger IMU-based activity recognition.*

4.4 Ablation Study

We conduct ablation studies to investigate how key design choices influence weak annotation quality and downstream HAR performance.

Number of Clusters. Figure 8 shows the effect of the number of GMM clusters on annotation quality and downstream HAR performance using the consistently strong training configuration *Weak-CE-T-6*. Annotation accuracy consistently decreases as the number of clusters is reduced. On WEAR, WAFER achieves 90.64% with 100 clusters, decreasing to 85.84% with 50 clusters and 61.95% with 19 clusters. Similar trends are observed on ActionSense (60.84% $\rightarrow$ 51.63% $\rightarrow$ 41.73%) and Wetlab (71.60% $\rightarrow$ 58.92% $\rightarrow$ 24.92%). The reduction in annotation quality is generally accompanied by lower downstream TinyHAR Zhou et al. (2022) and DeepConvLSTM Bock et al. (2021) performance. Complete results for all cluster settings and training configurations are provided in Appendix. This suggests that *increasing the number of clusters improves annotation quality at the cost of requiring more representative samples to be manually labelled.*

Modality-Specific Weights. We analyse the modality weights using the WEAR dataset. Since the objective of this experiment is to identify the weight configuration that produces the highest quality weak annotations, we evaluate only annotation accuracy rather than downstream HAR performance. We evaluate the multimodal fusion weights (α, β, γ) in Table 4. The best annotation accuracy of 90.64% is obtained with $(0.50, 0.25, 0.25)$, assigning the largest weight to visual representations while equally weighting wearable and optical-flow information. For datasets with bilateral wrist sensors, the effect of different wrist fusion weights is evaluated in the Appendix. This indicates that *visual information remains the primary cue for weak annotation, while complementary wearable and motion representations substantially improve the resulting annotation quality.*

Visual Foundation Model. To assess whether WAFER depends on a specific visual backbone, we replace CLIP Radford et al. (2021) with DINOv2 Oquab et al. (2023) and DINOv3 Siméoni et al. (2025), while keeping the remaining WAFER configuration fixed, including the wearable foundation model, RAFT features, clustering, and fusion weights. On the WEAR dataset, CLIP achieves the highest annotation accuracy of 90.64%, outperforming DINOv3 with 87.57% and DINOv2 with 86.65%. Complete results for different visual backbone for different weights are provided in Appendix. This shows that *annotation performance depends on the choice of visual backbone.*

5 Conclusion

In this work, we addressed the problem of manual annotation effort for IMU-based HAR datasets. Existing weak annotation approaches leverage visual representations to reduce this annotation burden. However, our latent space analysis shows that these representations are influenced by surrounding visual context. We therefore propose **WAFER**, a multimodal weak annotation framework that complements visual and optical-flow representations with representations from a pretrained wearable foundation model. WAFER consistently improves annotation quality over baselines across three benchmark datasets. Furthermore, downstream HAR models trained using the generated weak labels achieve performance nearing that of fully supervised training. Our experiments indicate that wearable representations provide additional information for weak annotation, while fine-grained clustering is important for generating reliable weak annotations. Annotation quality also depends on the choice of visual representation and the relative contribution of each modality. Although visual context effects can remain in the multimodal representation, incorporating wearable and motion information improves weak annotation quality. While WAFER uses synchronized video during annotation, downstream HAR training requires only IMU data. Future work could therefore investigate adaptive multimodal fusion strategies that further reduce visual context bias during annotation.

AI DISCLOSURE

Generative AI tools were used in this work to assist with code development and implementation, interpretation of experimental results, and preparation of the manuscript. For manuscript preparation, generative AI was used for grammar checking, improving the clarity and readability of scientific writing, assisting with the presentation of mathematical equations, LaTeX formatting, and providing feedback on the structure of the paper. All AI-assisted content was reviewed by the authors, and AI-generated code was manually verified and tested for correctness. Generative AI was not used to conduct the literature review, generate research hypotheses, or generate experimental data. The authors take full responsibility for the final content of this work, including any text, claims, code, or other artifacts produced with the assistance of generative AI.

REFERENCES

Rebecca Adaimi and Edison Thomaz. Leveraging active learning and conditional mutual information to minimize data annotation in human activity recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 3(3):1–23, 2019.

Marius Bock, Alexander Hölzemann, Michael Moeller, and Kristof Van Laerhoven. Improving deep learning for har with shallow lstms. In *Proceedings of the 2021 ACM International Symposium on Wearable Computers*, pp. 7–12, 2021.

Marius Bock, Hilde Kuehne, Kristof Van Laerhoven, and Michael Moeller. Wear: An outdoor sports dataset for wearable and egocentric activity recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(4):1–21, 2024a.

Marius Bock, Kristof Van Laerhoven, and Michael Moeller. Weak-annotation of har datasets using vision foundation models. In *Proceedings of the 2024 ACM International Symposium on Wearable Computers*, pp. 55–62, 2024b.

Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 4724–4733. IEEE, 2017.

Mathias Ciliberto, Vitor Fortes Rey, Alberto Calatroni, Paul Lukowicz, and Daniel Roggen. Opportunity++: A multimodal dataset for video-and wearable, object and ambient sensors-based human activity recognition. *Frontiers in Computer Science*, 3:792065, 2021.

Joseph DelPreto, Chao Liu, Yiyue Luo, Michael Foshey, Yunzhu Li, Antonio Torralba, Wojciech Matusik, and Daniela Rus. Actionsense: A multimodal dataset and recording framework for human activities using wearable sensors in a kitchen environment. *Advances in Neural Information Processing Systems*, 35:13800–13813, 2022.

Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 249–256. JMLR Workshop and Conference Proceedings, 2010.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

Aditya Krishna Menon, Ankit Singh Rawat, Sashank J Reddi, and Sanjiv Kumar. Can gradient clipping mitigate label noise? 2020.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Francisco Javier Ordóñez and Daniel Roggen. Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition. *Sensors*, 16(1):115, 2016.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Daniel Roggen, Alberto Calatroni, Mirco Rossi, Thomas Holleczek, Kilian Förster, Gerhard Tröster, Paul Lukowicz, David Bannach, Gerald Pirkl, Alois Ferscha, et al. Collecting complex activity datasets in highly rich networked sensor environments. In *2010 Seventh international conference on networked sensing systems (INSS)*, pp. 233–240. IEEE, 2010.

Philipp M Scholl, Matthias Wille, and Kristof Van Laerhoven. Wearables in the wet lab: a laboratory system for capturing and guiding experiments. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pp. 589–599, 2015.

Abram Schönfeldt, Benjamin Maylor, Xiaofang Chen, Ronald Clark, and Aiden Doherty. Reducing annotation burden in physical activity research using vision language models. *Scientific Reports*, 15(1):37253, 2025.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Satya P Singh, Madan Kumar Sharma, Aimé Lay-Ekuakille, Deepak Gangwar, and Sukrit Gupta. Deep convlstm with self-attention for human activity decoding using wearable sensors. *IEEE Sensors Journal*, 21(6):8575–8582, 2020.

Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*, pp. 402–419. Springer, 2020.

Catherine Tong, Jinchen Ge, and Nicholas D Lane. Zero-shot learning for imu-based activity recognition using video embeddings. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 5(4):1–23, 2021.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Kang Xia, Wenzhong Li, Shiwei Gan, and Sanglu Lu. Ts2act: Few-shot human activity sensing with cross-modal co-learning. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 7(4):1–22, 2024a.

Kang Xia, Wenzhong Li, Yimiao Shao, and Sanglu Lu. Vi2act: Video-enhanced cross-modal co-learning with representation conditional discriminator for few-shot human activity recognition. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 1848–1856, 2024b.

Hang Yuan, Shing Chan, Andrew P Creagh, Catherine Tong, Aidan Acquah, David A Clifton, and Aiden Doherty. Self-supervised learning for human activity recognition using 700,000 person-days of wearable data. *NPJ digital medicine*, 7(1):91, 2024.

Shibo Zhang, Yaxuan Li, Shen Zhang, Farzad Shahabi, Stephen Xia, Yu Deng, and Nabil Alshurafa. Deep learning in human activity recognition with wearable sensors: A review on advances. *Sensors*, 22(4):1476, 2022.

Ellie Zhou, Jihoon Chung, and Olga Russakovsky. Seeing beyond the scene: Analyzing and mitigating background bias in action recognition. *arXiv preprint arXiv:2512.17953*, 2025.

Yexu Zhou, Haibin Zhao, Yiran Huang, Till Riedel, Michael Hefenbrock, and Michael Beigl. Tinyhar: A lightweight deep learning model designed for human activity recognition. In *Proceedings of the 2022 ACM International Symposium on Wearable Computers*, pp. 89–93, 2022.

A Appendix

A.1 Threshold to Retention Mapping

Following Bock et al. (2024b), downstream HAR models are trained using threshold settings $T = 6$ and $T = 4$, which retain only windows closest to the representative sample within each cluster. Since feature standardisation and multimodal fusion alter the scale of the embedding space, these absolute Euclidean distance thresholds are not directly transferable to WAFER. Instead, we reproduce each threshold setting by retaining the same proportion of windows as in Bock et al. (2024b). Table A1 reports the dataset-specific retention percentages for additional cluster setting used in our experiments.

Table A1: **Threshold to Retention Mapping Across Cluster Settings.** Percentage of windows retained for the absolute distance thresholds $T = 6$ and $T = 4$.

Dataset	Clusters	Threshold $T = 6$	Threshold $T = 4$
WEAR Bock et al. (2024a)	50	63.8%	24.3%
	19	39.5%	12.7%
ActionSense DelPreto et al. (2022)	50	30.4%	7.8%
	20	17.0%	3.5%
Wetlab Scholl et al. (2015)	50	71.3%	36.3%
	9	37.9%	14.3%

A.2 Wrist Fusion Weights

For datasets with bilateral wrist sensors, the standardised right- and left-wrist embeddings are combined using weighted feature concatenation, $\tilde{E}_n^s = [\, p\, \tilde{E}_n^{s,r} \,\|\, q\, \tilde{E}_n^{s,l} \,]$, with $p + q = 1$ and $p, q \geq 0$, where $\|$ denotes concatenation, so that the weights control the relative contribution of each wrist.

As shown in Table A2, equally weighting the right and left wrist representations ($p = q = 0.5$) achieves the highest annotation accuracy of 56.20%. This suggests that both wrist representations provide complementary information for the exercise activities in WEAR, many of which involve movements of both arms. However, the optimal wrist weighting may vary across datasets depending on the activities and sensor configuration.

A.3 Detailed Number of Cluster Ablation

Table A3 presents the complete annotation accuracy and downstream HAR results for all evaluated cluster settings and training regimes. Across all three datasets, WAFER consistently achieves higher annotation accuracy than Bock et al. (2024b), demonstrating the benefit of incorporating wearable foundation model representations into the weak annotation pipeline. These improvements generally translate to stronger downstream HAR performance for both TinyHAR Zhou et al. (2022) and DeepConvLSTM Bock et al. (2021) across different training regimes.

A.4 Additional Visual Foundation Model Analysis

Table A4 reports the complete multimodal weight-space search for CLIP Radford et al. (2021), DINOv2 Oquab et al. (2023), and DINOv3 Siméoni et al. (2025) on WEAR Bock et al. (2024a). The optimal fusion configuration varies across visual backbones: CLIP achieves its highest annotation accuracy of 90.64% with $(\alpha, \beta, \gamma) = (0.50, 0.25, 0.25)$, while DINOv2 reaches 88.60% with $(0.25, 0.25, 0.50)$. DINOv3 achieves its best accuracy of 87.57% with $(0.50, 0.25, 0.25)$. Even after allowing backbone-specific fusion weights, CLIP provides the highest annotation accuracy and is therefore used as the default visual foundation model in WAFER.

A.5 Activity-wise Annotation Analysis

Figure A1 presents the activity-wise confusion matrix of the generated weak annotations for subject 0 of WEAR Bock et al. (2024a). Most activities show high diagonal values, indicating that the

Table A2: **Inertial Weight Space Search.** Annotation accuracy (%) on WEAR for different right- and left-wrist weights (p, q), using 100 GMM clusters. Best in Bold.

	Inertial Weight Configuration				
Right wrist p	1.00	0.75	**0.50**	0.25	0.00
Left wrist q	0.00	0.25	**0.50**	0.75	1.00
Annotation Acc. (%) ↑	48.79	50.49	**56.20**	52.80	50.02

Table A3: **Effect of the number of GMM clusters on annotation accuracy and downstream HAR performance.** Annotation accuracy and downstream accuracy and macro F1 (%) for TinyHAR and DeepConvLSTM across different numbers of GMM clusters. Bold indicates the better result between Bock et al. (2024b) and WAFER, whereas yellow highlights the best weak-training result in each column within each cluster setting.

Dataset	Clusters	Annotation Acc. ↑		Training Type	Accuracy ↑				F1 ↑			
					TinyHAR		DeepConvLSTM		TinyHAR		DeepConvLSTM	
		Bock et al. (2024b)	WAFER		Bock et al. (2024b)	WAFER	Bock et al. (2024b)	WAFER	Bock et al. (2024b)	WAFER	Bock et al. (2024b)	WAFER
WEAR Bock et al. (2024a)	50	77.43	**85.84**	Fully-Supervised	77.83	77.83	85.14	85.14	71.89	71.89	82.23	82.23
				Few-Shot-CE	59.58	**62.65**	59.58	**62.80**	46.25	**51.26**	46.25	**51.82**
				Random-CE	59.72	**61.17**	59.46	**60.25**	46.34	**50.60**	46.98	**48.84**
				Weak-CE	66.68	**72.43**	64.17	**76.18**	54.05	**62.72**	54.59	**68.52**
				Weak-PHGCE	68.53	**73.83**	70.34	**77.00**	55.40	**63.72**	55.27	**66.61**
				Weak-CE-T-6	68.19	**74.50**	73.06	**78.33**	55.56	**66.16**	60.60	**71.77**
				Weak-PHGCE-T-6	69.47	**75.08**	73.17	**78.44**	56.29	**65.71**	58.84	**69.88**
				Weak-CE-T-4	71.31	**75.29**	74.72	**77.50**	60.16	**67.08**	63.93	**71.06**
				Weak-PHGCE-T-4	72.25	**75.02**	74.45	**76.41**	60.84	**65.76**	60.61	**67.42**
	19	56.28	**61.95**	Fully-Supervised	77.83	77.83	85.14	85.14	71.89	71.89	82.23	82.23
				Few-Shot-CE	37.41	**42.33**	37.41	**50.24**	26.55	**27.43**	24.76	**37.34**
				Random-CE	**23.73**	36.63	**45.90**	42.32	**23.73**	23.06	**31.13**	29.75
				Weak-CE	49.45	**54.68**	42.55	**56.36**	38.75	**42.71**	34.09	**42.05**
				Weak-PHGCE	51.46	**56.08**	48.62	**59.86**	39.23	**43.79**	35.45	**45.79**
				Weak-CE-T-6	57.22	**62.10**	59.70	**67.19**	46.71	**51.92**	46.63	**56.19**
				Weak-PHGCE-T-6	58.78	**63.71**	59.39	**66.69**	46.27	**51.55**	44.97	**52.33**
				Weak-CE-T-4	65.68	**66.93**	**68.86**	68.51	55.64	**58.05**	57.33	**58.27**
				Weak-PHGCE-T-4	62.90	**65.56**	61.35	**65.55**	50.46	**54.32**	47.00	**52.48**
ActionSense DelPreto et al. (2022)	50	**52.76**	51.63	Fully-Supervised	22.19	22.19	20.73	20.73	18.04	18.04	14.82	14.82
				Few-Shot-CE	**8.50**	7.77	**8.14**	7.72	**4.00**	3.85	**4.92**	4.80
				Random-CE	**7.80**	7.75	**9.42**	9.10	4.19	**4.48**	**5.09**	4.75
				Weak-CE	15.58	**16.73**	13.91	**15.59**	11.55	**12.35**	9.94	**10.12**
				Weak-PHGCE	13.88	**14.75**	**12.34**	11.37	9.25	**10.63**	**7.43**	7.37
				Weak-CE-T-6	15.70	**17.39**	**15.95**	13.32	11.22	**12.72**	**10.07**	9.91
				Weak-PHGCE-T-6	**13.72**	13.62	**10.57**	8.77	8.16	**8.77**	**5.44**	5.20
				Weak-CE-T-4	13.65	**14.47**	**12.31**	10.38	9.24	**10.23**	7.10	**7.24**
				Weak-PHGCE-T-4	11.07	**12.09**	**8.92**	8.51	6.09	**6.61**	4.35	**4.64**
	20	38.51	**41.73**	Fully-Supervised	22.19	22.19	20.73	20.73	18.04	18.04	14.82	14.82
				Few-Shot-CE	**6.19**	5.51	**6.47**	6.19	**2.32**	2.20	**2.58**	2.13
				Random-CE	5.01	**5.53**	4.69	**5.58**	1.70	**2.19**	1.63	**2.08**
				Weak-CE	13.74	**15.30**	12.32	**13.56**	9.75	**10.78**	7.67	**8.11**
				Weak-PHGCE	**14.35**	13.97	**11.65**	11.08	9.43	**9.54**	**6.34**	6.22
				Weak-CE-T-6	13.35	**15.12**	**14.67**	12.15	9.32	**11.09**	**8.81**	8.78
				Weak-PHGCE-T-6	**12.17**	11.80	**9.91**	9.84	**7.52**	7.20	4.92	**5.28**
				Weak-CE-T-4	10.80	**12.23**	10.40	**10.93**	8.02	**8.43**	**7.15**	6.74
				Weak-PHGCE-T-4	9.24	**10.79**	8.34	**9.16**	**5.82**	5.68	3.86	**3.92**
Wetlab Scholl et al. (2015)	50	50.24	**58.92**	Fully-Supervised	38.75	38.75	45.27	45.27	28.85	28.85	38.64	38.64
				Few-Shot-CE	22.43	**28.15**	21.89	**27.61**	16.14	**17.84**	16.46	**20.35**
				Random-CE	**27.74**	22.07	**26.48**	22.05	**17.79**	15.47	**18.62**	16.23
				Weak-CE	**34.23**	33.00	32.57	**33.05**	**24.38**	23.36	25.78	**25.95**
				Weak-PHGCE	**33.79**	32.82	27.64	**28.24**	**24.35**	23.23	**23.17**	22.20
				Weak-CE-T-6	**33.79**	33.68	**33.78**	32.74	**24.21**	24.07	**27.08**	26.15
				Weak-PHGCE-T-6	33.34	**33.54**	**28.77**	28.10	**24.25**	23.68	**23.90**	22.40
				Weak-CE-T-4	32.85	**34.41**	32.02	**33.83**	23.84	**23.96**	26.37	**26.68**
				Weak-PHGCE-T-4	32.39	**32.77**	28.33	**30.33**	**23.72**	23.19	**23.91**	23.72
	9	21.05	**24.92**	Fully-Supervised	38.75	38.75	45.27	45.27	28.85	28.85	38.64	38.64
				Few-Shot-CE	15.18	**15.95**	15.60	**18.12**	**11.62**	11.31	11.39	**12.26**
				Random-CE	**34.23**	15.18	**16.33**	13.68	**24.38**	10.83	8.95	**9.69**
				Weak-CE	**23.90**	20.85	16.97	**17.16**	14.70	**15.45**	5.53	**12.08**
				Weak-PHGCE	**24.06**	20.59	**18.62**	15.00	**15.19**	15.06	10.48	**11.19**
				Weak-CE-T-6	**26.14**	23.60	**23.01**	21.57	**18.63**	16.79	**18.12**	16.88
				Weak-PHGCE-T-6	**25.22**	23.63	**20.26**	18.10	**18.41**	16.78	**13.72**	11.71
				Weak-CE-T-4	25.15	**25.54**	23.54	**25.64**	17.91	**17.34**	18.71	**19.49**
				Weak-PHGCE-T-4	23.82	**25.20**	21.82	**22.97**	17.29	17.29	15.81	**15.93**

corresponding activity labels are generally preserved. The main confusions occur between similar activity variants, such as Jogging (Rotating Arms) and Jogging (Skipping), and between Push-Ups and Push-Ups (Complex).

A.6 Implementation and Reproducibility Details

Pretrained feature extractors. Visual representations are extracted using openai/clip-vit-large-patch14 for CLIP, facebook/dinov2-large for DINOv2, and facebook/dinov3-vitl16-pretrain-

Table A4: **Visual foundation model weight-space search.** Annotation accuracy (%) on WEAR for CLIP, DINOv2, and DINOv3 under different fusion weights (α, β, γ), corresponding to the visual, IMU, and RAFT representations, respectively. All experiments use 100 GMM clusters. Best in Bold.

		Fusion Weight Configuration														
Visual α		0.00	0.00	0.25	0.00	0.25	0.50	0.00	0.25	0.50	0.75	0.00	0.25	0.50	0.75	1.00
IMU β		1.00	0.75	0.75	0.50	0.50	0.50	0.25	0.25	0.25	0.25	0.00	0.00	0.00	0.00	0.00
RAFT γ		0.00	0.25	0.00	0.50	0.25	0.00	0.75	0.50	0.25	0.00	1.00	0.75	0.50	0.25	0.00
Annotation Acc. (%) ↑	**CLIP**	56.20	64.77	70.46	83.67	84.73	87.12	81.28	87.45	**90.64**	86.57	79.25	83.17	85.14	83.69	82.81
	DINOv2	56.20	64.77	74.06	83.67	85.24	87.99	81.28	**88.60**	86.65	82.04	79.25	81.55	84.54	80.39	79.11
	DINOv3	56.20	64.77	71.93	83.67	84.90	87.19	81.28	86.72	**87.57**	81.51	79.25	82.15	84.50	79.70	78.53

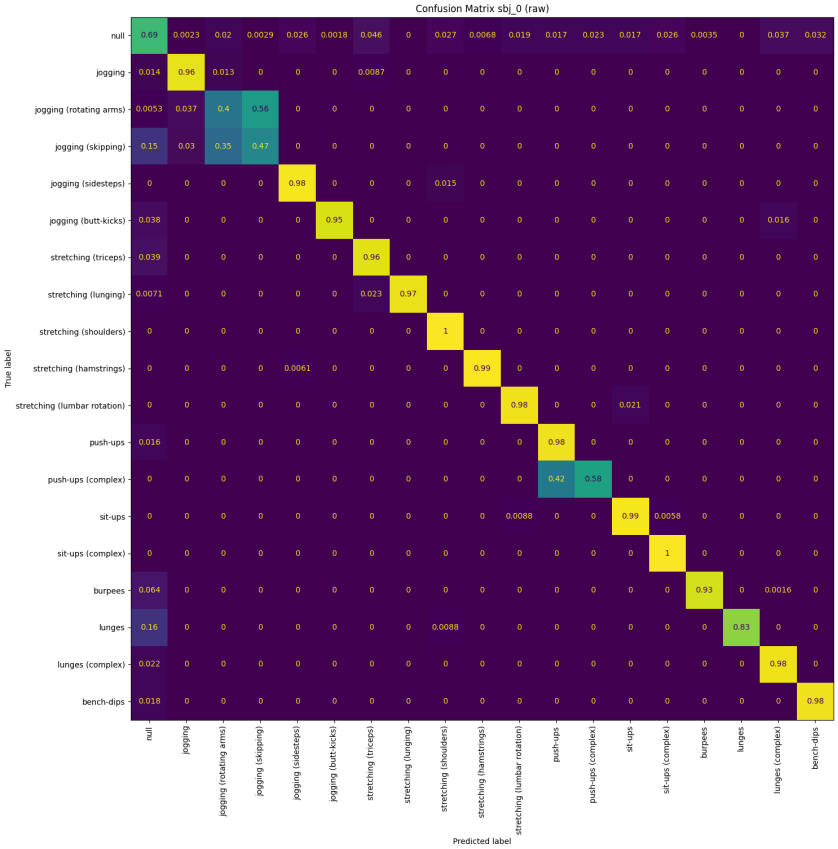


Figure A1: **Activity-wise confusion matrix of generated weak annotations for subject 0 of WEAR.** Values are row-normalized, with rows representing ground-truth labels and columns representing predicted labels.

lvd1689m for DINOv3. Wearable representations are extracted using the pretrained harnet5 model from OxWearables/ssl-wearables. The corresponding embedding layers and output dimensionalities are reported in Table A5.

GMM clustering. We use a GMM with 100 mixture components, full covariance, a tolerance of 10^{-3}, a maximum of 100 iterations, regularization of 10^{-6}, and k-means initialization. Results are averaged over three random seeds of 1, 2, and 3.

Downstream HAR training. TinyHAR Zhou et al. (2022) and DeepConvLSTM Bock et al. (2021) are trained for 30 epochs using the Adam optimizer Kingma & Ba (2015) with an initial learning rate of 10^{-4}, weight decay of 10^{-6}, and a batch size of 100. Model parameters are initialized using Xavier normal initialization Glorot & Bengio (2010).

Table A5: **Embedding layers and output dimensionalities of the pretrained models used for feature extraction.**

Model	Embedding Layer	Dimension
CLIP Radford et al. (2021)	Last projection layer (vision encoder)	$\mathbb{R}^{768}$
DINOv2 Oquab et al. (2023)	Last hidden state (CLS token)	$\mathbb{R}^{1024}$
DINOv3 Siméoni et al. (2025)	Last hidden state (CLS token)	$\mathbb{R}^{1024}$
OxWearables (HarNet5) Yuan et al. (2024)	last average pooling layer	$\mathbb{R}^{512}$
I3D (Optical Flow)	Last average pooling layer	$\mathbb{R}^{1024}$

Table A6: **Manual annotation effort across GMM cluster settings.** Percentage of windows requiring manual annotation and percentage receiving propagated labels. Values are reported as mean $\pm$ standard deviation across subjects.

Dataset	Clusters	Manual Annotation (%)	Propagated Labels (%)
WEAR Bock et al. (2024a)	100	3.44 ± 0.72	96.56 ± 0.72
	50	1.72 ± 0.36	98.28 ± 0.36
	19	0.65 ± 0.14	99.35 ± 0.14
ActionSense DelPreto et al. (2022)	100	3.49 ± 0.89	96.51 ± 0.89
	50	1.74 ± 0.45	98.26 ± 0.45
	20	0.70 ± 0.18	99.30 ± 0.18
Wetlab Scholl et al. (2015)	100	4.04 ± 0.63	95.96 ± 0.63
	50	2.02 ± 0.31	97.98 ± 0.31
	9	0.36 ± 0.06	99.63 ± 0.06

A.7 Manual Annotation Effort

Table A6 reports the percentage of windows requiring manual annotation and the complementary percentage receiving propagated labels. With 100 clusters, only 3.44%, 3.49%, and 4.04% of windows require manual annotation for WEAR, ActionSense, and Wetlab, respectively, while more than 95% of labels are obtained through propagation. Reducing the number of clusters further decreases the annotation effort, although, as shown in the cluster ablation study, this is accompanied by lower weak annotation accuracy.