A Comparison of Fusion Techniques for Multi-Modal Human Activity Recognition on the HARMES Dataset

Ahmed Mohamady*, Robin Burchard*, and Kristof Van Laerhoven

University of Siegen, Germany
engahmedmohamady23@gmail.com
https://ubicomp.eti.uni-siegen.de/

*These authors contributed equally to this work.

Abstract. Recent advances in Human Activity Recognition (HAR) from wearable sensors have shown that multi-modal deep learning models consistently outperform their uni-modal counterparts. Modalities can include IMUs, RGB cameras, audio signals, and others. One important aspect of multi-modal deep learning is the sensor fusion approach we apply. Over recent years, multiple fusion paradigms have been proposed for multi-modal HAR. However, to the best of our knowledge, no head-to-head comparison of these paradigms exists on a common multi-modal HAR benchmark dataset. To partially address this research gap, we systematically compare seven state-of-the-art sensor fusion methods on the recently released HARMES dataset, which comprises 61 hours of fully labeled IMU, audio, and ambient humidity data. The chosen dataset focuses on 15 household and personal hygiene activities of daily living (ADLs). By applying the seven different fusion techniques to a state-of-the-art multi-modal model architecture, we show that Gated Multi-modal Fusion and late (concatenation) fusion achieves the highest macro F1-score (0.82), surpassing the concatenation-based late fusion HARMES paper baseline of 0.76 by +6pp under leave-one-participant-out evaluation. All code used in our experiments is made publicly available on GitHub[1].

Keywords: Modality Fusion · Multi-Modal · Gated Fusion · Human Activity Recognition · HARMES · IMU · Audio

1 Introduction

Human Activity Recognition (HAR) from wearable and ambient sensors is a foundational task for healthcare monitoring, assistive living, smart-home automation, and activity-aware computing [6]. The past decade has seen rapid progress driven by deep learning architectures applied to inertial measurement

[1] https://github.com/AhmedMohamady98/A-Comparison-of-Fusion-Techniques-for-Multi-Modal-Human-Activity-Recognition-on-the-HARMES-Dataset

unit (IMU) data [41,30]. These models now routinely achieve high classification accuracy on scripted activity corpora, but they operate on a single sensory channel and remain fragile to real-world confounds such as handedness, sensor placement variability, and activities with similar motion signatures.
A growing body of work shows that the limitations of unimodal HAR can be overcome by combining multiple sensing modalities [36,3]. For example, audio captures characteristic acoustic signatures of water-based and mechanical activities (e.g., dishwashing, vacuuming, brushing teeth) that are difficult to disambiguate from IMU sensors alone. Environmental sensors such as humidity provide contextual priors: high humidity near a sink is indicative of water-related tasks [10,7]. The HARMES dataset [8] offers a public, multi-participant, long-duration benchmark with synchronised IMU, audio, and humidity streams recorded from activities of daily living (ADLs), enabling multi-modal fusion research on ecologically valid data. The literature on multi-modal fusion is diverse. Architectures range from simple late concatenation of modality embeddings to gated units that learn per-sample modality weights [2], bottleneck transformers routing information through shared latent tokens [29], directional cross-attention [38], low-rank tensor factorization [23], and decision-level weighted averaging. While these approaches beat the respective baselines, each approach has been validated on different datasets with different encoder backbones and evaluation protocols, making direct comparison across many fusion methods impossible. A researcher choosing a fusion strategy for a new multi-modal HAR system currently has no direct empirical guidance. In this work, we provide such a comparison. Using HARMES as a shared testbed, we train and evaluate **seven fusion strategies** under identical conditions: the same three encoder backbones, the same 10-second window size, the same training hyperparameters, and the same evaluation metrics. We report results under 3-fold group cross-validation and 20-fold LOPO evaluation, matching the HARMES benchmark protocol. We further conduct a modality ablation and a subgroup analysis of left-handed participants.

Our contribution is threefold.

1. We present a systematic head-to-head comparison of seven prominent multi-modal fusion paradigms on a uniform HAR benchmark, providing an empirical ranking independent of backbone or protocol variation.
2. We identify the practical drivers of fusion performance on the HARMES dataset by identifying the per-class-performance contribution of each modality.
3. We show that multi-modal fusion partially mitigates the handedness gap by relying on handedness-invariant audio cues when IMU signals become unreliable.

2 Related Work

To position this work in the context of current research, we discuss existing HAR model architectures, fusion methods, and fusion method comparison studies.

2.1 Encoder Architectures for Multi-Modal Wearable HAR

Deep learning for IMU-based activity recognition has progressed through several generations of architecture. Ordóñez et al. [30] introduced DeepConvLSTM, pairing convolutional feature extraction with LSTM temporal modelling into the baseline most later work is measured against, while Ha and Choi [17] apply CNNs that can generalise across sensor placements via partial weight sharing. The trend has since shifted toward lightweight, deployable designs. Zhou et al. [46] proposed TinyHAR, which chains depthwise separable convolutions with transformer encoded blocks, LSTM, and temporal attention, while remaining small enough for execution on mobile devices. Bian et al. [4] published an even smaller, transformer-free model with TinierHAR, which is built around bidirectional GRUs and temporal aggregation instead, reducing parameter count and computational cost significantly while maintaining competitive accuracy.

Audio is a powerful complementary channel for a multitude of activities. Gong et al. [16] adapted the Vision Transformer to audio by treating log-mel spectrograms as patch sequences, producing the Audio Spectrogram Transformer (AST), which remains a strong and widely used audio classification backbone. Mollyn et al. [26] showed that IMU-triggered, privacy-preserving audio sampling recognizes 26 daily activities without continuous recording, and Lee et al. [21] confirmed that audio is highly effective for detecting activities with minimal limb motion. For general timeseries signals, Ekambaram et al. [13] proposed TSMixer, an all-MLP time-series architecture.

Several multi-modal HAR systems combine these channels in different ways: E.g., AttnSense's multi-level attention [25], Cosmo's quality-guided contrastive fusion [31], and Mamba-MHAR's two-branch selective state space models [20].

2.2 Multi-Modal Fusion Strategies

Multi-modal fusion can be placed along a spectrum defined by how early in the processing pipeline modalities are combined [3,36]. At one extreme, early fusion merges raw or low-level features before any modality-specific learning has taken place. At the other extreme, decision-level fusion combines the outputs of fully independent per-modality classifiers. Between these two there exists a multitude of intermediate approaches that operate on learned modality embeddings. Ramachandram and Taylor [36] identified at least five distinct architectural families within this space, and the field has continued to diversify since. Below, we list and discuss the most prominent fusion approaches in current literature.

Early fusion concatenates raw sensor readings into a single input vector before any model processing. It is simple but assumes all modalities share the same temporal resolution and scale, which rarely holds for heterogeneous sensors such as IMUs, microphones, and environmental probes. Yilmaz. et al. [44] show that audio and IMU data can be processed into images, which leads to beneficial results with their early fusion approach. In practice, early fusion is often outperformed by methods that encode each modality separately [36,3,28].

Late embedding fusion, oftentimes called feature-level or representation-level fusion, concatenates the output embeddings of per-modality encoders and feeds the combined vector to a shared classifier. We refer to embedding-level feature fusion as late embedding fusion to distinguish it from decision-level late fusion. It is often applied in multi-modal HAR because it decouples encoder training from fusion: each encoder can be pre-trained independently, and the fusion head stays lightweight. Münzner et al. [28] found embedding concatenation outperformed decision-level fusion on PAMAP2, and it has since been used for HAR tasks, e.g., household activity recognition [8] and exercise counting [21].

Gated fusion addresses the fact that not all modalities are equally informative for every sample: a static concatenation head cannot down-weight a noisy input. Arevalo et al. [2] introduced the Gated Multi-modal Unit (GMU), where a sigmoid gate computed from all modalities controls how much each contributes to the fused vector. Because the gate is conditioned on the combined context, the model can learn to suppress modalities adaptively, when they are less informative, e.g., due to background noise. As the weights for each modality can be inspected, GMU produces interpretable modality "importances".

Tensor fusion models cross-modal interactions explicitly via the outer product of modality embeddings, capturing pairwise and higher-order combinations in one operation. The Tensor Fusion Network (TFN) of Zadeh et al. [45] showed this improves over additive baselines on sentiment analysis, but the outer-product tensor grows multiplicatively with the number and dimensionality of modalities, making it expensive. Liu et al. [23] addressed this with Low-Rank multi-modal Fusion (LMF), which decomposes the weight tensor into modality-specific low-rank factors combined by an element-wise product. LMF matches or beats TFN with far fewer parameters, showing the full-rank tensor is largely redundant.

Cross-modal attention builds on the Transformer [40], using one modality as the query and another as the value so the query can attend to the most relevant parts of the other modality. Tsai et al. [38] systematised this as the multi-modal Transformer (MulT), applying it in both directions for every modality pair, which gives six cross-attention streams per layer with three modalities. MulT is commonly used for language-vision-acoustic fusion. The same idea shows up in ViLBERT's co-attentional layers [24] and in RGB-D detection [42]. Within HAR, related attention-based fusion ideas appear in AttnSense [25], which applies multi-level attention over inertial sensors, implicitly weighting both on-body placement channels and time steps

Bottleneck transformer fusion is a compute-efficient alternative to full cross-attention. Instead of letting every token attend to every other, Nagrani et al. [29] route all cross-modal information through a small set of shared bottleneck tokens. In the multi-modal Bottleneck Transformer (MBT), each layer processes a modality's tokens together with the bottleneck, but never mixes two modalities directly. The per-modality bottleneck copies are then averaged into a shared state. On AudioSet and VGGSound, it beat late fusion and full cross-attention with fewer parameters, and the bottleneck may also act as an implicit regularizer by constraining cross-modal information flow to a limited set of shared tokens.

Bralina et al. [5] extended it with the Adaptive Bottleneck Transformer (AMBT), adding factorised contrastive learning to prevent modality collapse.

Decision-level fusion never builds a shared representation. Each modality has its own classifier producing a class distribution, and these get combined by a fixed rule or a learned weighted sum. Atrey et al. [3] formalised this as one of the three canonical fusion levels. The approach is modular and transparent, since the weights directly quantify per-modality confidence, but it loses interactions at the representation level: the model has no way to discover that a humid environment combined with repetitive arm motion is more diagnostic than either signal on its own.

CLS-token fusion treats each per-modality embedding as a token in a shared sequence with a prepended, learnable `[CLS]` token, processed by standard transformer encoder layers [40] following the classification convention of BERT (Devlin et al. [11]) and ViT (Dosovitskiy et al. [12]). The `[CLS]` state aggregates all modality tokens through unrestricted self-attention. There is no bottleneck constraint as in MBT, and no directional structure as in MulT. All tokens take part in one unified attention operation.

Contrastive self-supervised fusion is a term for self-supervised fusion methods that align modalities using the natural pairing structure of the data rather than merging them into one representation. Tian et al. [37] introduced Contrastive Multiview Coding (CMC), pulling together views of the same scene and pushing apart different scenes, and Radford et al. [34] scaled this to image-text pairs with CLIP, enabling zero-shot transfer via similarity retrieval. Girdhar et al. [15] extended it to six modalities, IMU included, with ImageBind, showing that pairing each modality with images alone suffices to align them all. Closest to our setting, Moon et al. [27] introduced IMU2CLIP, aligning wrist-worn IMU with video and text, and Ouyang et al. [31] applied contrastive fusion to wearable HAR with Cosmo. These approaches suit wearable sensing well, where labels are expensive but paired unlabeled recordings are cheap to collect, but do not apply to our fully supervised modality fusion problem.

Feature conditioning uses one modality to modulate the internal computation of a network processing another, instead of merging them at a fixed layer. Perez et al. [32] introduced Feature-wise Linear Modulation (FiLM), where a conditioning network predicts per-channel scale and shift parameters applied to a target network's feature maps. It was proposed for visual question answering and has since been used wherever one signal provides context for another. Rahman et al. [35] adapted this to pretrained transformers with the multi-modal Adaptation Gate (MAG), injecting a learned shift into BERT's representations from visual and acoustic signals while leaving the backbone intact.

Latent bottleneck architectures generalise the MBT bottleneck into a modality-agnostic framework. Jaegle et al. [18] introduced the Perceiver, which maps inputs of arbitrary modality and size into a fixed-size latent array via asymmetric cross-attention, then processes that array with standard Transformer layers. Decoupling compute cost from input size lets one architecture handle images, audio, point clouds, and video with minimal modality-specific architectural

changes. Where MBT still feeds per-modality token sequences in separately, the Perceiver compresses the entire input into a shared latent space from the first step.

Feature recalibration lets each modality adjust the internal features of the others while the per-modality streams stay separate. Joze et al. [39] introduced the multi-modal Transfer Module (MMTM), which uses squeeze-and-excitation to recalibrate the channel-wise features of each convolutional stream from a joint summary of all modalities. Because it slots between existing branches, each one can keep its pretrained weights. Within HAR, Gao et al. [14] proposed MMTSA, which encodes inertial signals as Gramian Angular Field images and fuses RGB and IMU streams via inter-segment attention, reaching strong cross-participant accuracy.

2.3 Fusion Comparison Studies

The literature on multi-modal fusion is large, but systematic head-to-head comparisons of multiple fusion strategies under controlled conditions are surprisingly rare. Most papers introduce a single new method and compare it against a concatenation baseline, making it difficult to draw conclusions about the relative merit of different architectural families.

At the survey level, Atrey et al. [3] provide a foundational taxonomy of fusion approaches, distinguishing feature-level, decision-level, and hybrid fusion and categorising methods by their combination rule. Ramachandram and Taylor [36] later surveyed the deep multi-modal learning landscape and identified five architectural families, but their treatment is conceptual rather than empirical. In the HAR domain specifically, Qiu et al. [33] reviewed over 300 multi-sensor papers and mapped out dominant design patterns, and Yadav et al. [43] concluded that multi-modal fusion consistently outperforms unimodal approaches. Neither study, however, evaluates two or more fusion methods under the same experimental conditions.

One of the few HAR papers to directly compare fusion strategies is Münzner et al. [28], who tested sensor-level CNN fusion against decision-level fusion on the PAMAP2 dataset and reported higher performance for feature-level fusion than decision-level fusion. Aguileta et al. [1] similarly surveyed multi-sensor fusion for activity recognition and noted that no consensus had emerged on the best strategy, with results varying substantially across datasets and sensor combinations.

Outside HAR, more controlled comparisons do exist but remain tied to their own domains and encoder choices. Tsai et al. [38] evaluated MulT against late fusion and tensor fusion on language-vision-acoustic sentiment datasets. Nagrani et al. [29] compared MBT against late fusion and full cross-attention on video-audio classification, finding that the bottleneck constraint does not hurt performance while reducing computational cost. Koutoupis et al. [19] evaluated contrastive fusion objectives for higher-order multi-modal alignment across several non-HAR settings. These studies are valuable but their findings do not transfer directly to wearable sensor streams, where modality characteristics, sampling rates, and

temporal structure differ fundamentally from video, language, and audio-visual pairs.

To the best of our knowledge, we found no prior study that compares such a broad range of fusion paradigms under identical encoder backbones on a public multi-modal HAR benchmark.

3 Methodology

3.1 Dataset

We use the HARMES dataset [9,8], a public multi-modal corpus for wearable HAR. It contains recordings from 20 participants (10 female, 10 male, ages 18–67, mean 37.8, SD 14.4) performing 15 activities of daily living in their own homes. Each participant contributed about three hours of fully labeled data across three recording sessions. We exclude the fourth, free-form, session, leaving 61 hours of labelled data in total.

Modalities Three synchronized streams are provided. The *IMU* stream includes dual-wrist accelerometer and gyroscope data at 50 Hz, from a Puck.js device on the left wrist and a smartwatch on the right, totaling 12 channels (3-axis acceleration and 3-axis angular velocity per wrist). The *audio* stream is an ambient recording at 44.1 kHz from a wrist-worn microphone (in the smartwatch), with no speech contained in the dataset for privacy reasons. The *atmospheric* stream is humidity from a wrist-mounted BME280 sensor at 1 Hz, up-sampled to the IMU grid as a 1D time series.

Activity classes The 15 labelled activities fall into three groups: self-care (washing hands, brushing teeth, applying hand cream, disinfecting hands), household (floor cleaning, window cleaning, vacuum cleaning, washing dishes, putting away dishes, cleaning table, cleaning out dishwasher), and feeding (making tea, cutting vegetables, drinking, watering plants). The Null class covers background and transition periods between activities, leading to a total of 16 classes. At 10 s granularity, the Null class is the most common label (20.5% of all windows).

Left-handed participants Three participants (P07, P10, P14) are left-handed, a 15% prevalence in line with the real-world estimate of 9–18% [8]. This leads to a generalization challenge: IMU signals depend on which hand performs the dominant motion. While better IMU sensor placement generalization can likely be achieved with data augmentation, audio is expected to be mostly handedness-invariant.

3.2 Encoder Architectures

We process each modality with its own encoder that maps the raw stream to a 128-dimensional embedding. We fix this embedding size across all three encoders so that every fusion method operates on a common representation, which

lets us swap fusion strategies without changing the encoders. The IMU and humidity encoders are trained from scratch, while the audio encoder uses a frozen pretrained backbone with only a small trainable head.

IMU encoder For the dual-wrist inertial stream, we use TinyHAR [46]. It first applies lightweight convolutions across the sensor channels to extract local motion patterns, then a temporal self-attention block to capture how those patterns relate over the length of a window. The classic baseline here is DeepConvLSTM [30], which pairs convolutions with a recurrent layer, but its LSTM processes time steps sequentially and is comparatively heavy to train. TinyHAR reaches similar or better accuracy on standard HAR benchmarks at a fraction of the parameters, and its attention block handles long-range temporal structure without the sequential bottleneck of a recurrent network. Recent state-space models such as HARMamba [22] are promising but less established, so we favour TinyHAR as a proven, compact backbone well suited to wrist-worn data.

Audio encoder For the ambient audio stream, we use the Audio Spectrogram Transformer (AST) [16], which treats a spectrogram as a sequence of patches and applies transformer attention over them. We keep its backbone frozen and train only a small projection head. The reason is transfer learning: AST is pretrained on AudioSet, a very large and diverse sound corpus, so it already carries a strong general audio representation. Our labelled HAR audio is far too small to train a comparable model from scratch, and fine-tuning the whole backbone on it would risk overfitting. Convolutional audio models such as VGGish or PANNs were possible alternatives, but AST's attention gives a stronger pretrained starting point and a clean way to freeze the backbone and adapt with a single linear layer.

Humidity encoder For the atmospheric stream, we use TSMixer [13], an all-MLP architecture that alternates between mixing information along the time axis and along the feature axis. Humidity is a slow, single-channel signal, so it does not need the heavy machinery of attention or recurrence to be modelled well. A transformer would be overkill and prone to overfitting on such a simple stream, while an LSTM adds sequential cost for little benefit. TSMixer stays lightweight and trains quickly, which keeps the humidity branch cheap relative to the IMU and audio encoders, an appropriate balance given that humidity likely carries the least activity-discriminative information of the three [8].

3.3 Fusion Architectures

We compare seven fusion methods drawn from the strategies surveyed in Section 2.2. Our aim is not to propose a new fusion method but to compare existing ones under identical conditions, so the selection follows a single principle: every method must operate on the fixed 128-dimensional embeddings produced by our

encoders, treat the three modalities symmetrically, and require no modality-specific pretraining or input format. This allows us to use the same encoders for every fusion method and vary only the fusion step, so any difference in performance is attributable to the fusion mechanism alone.

This criterion rules out several families discussed in Section 2.2. Contrastive and joint-embedding methods such as CMC, CLIP, ImageBind, and IMU2CLIP need large-scale paired data and a separate pretraining stage, which does not fit a controlled supervised comparison on a single dataset. Methods tied to a particular input structure, such as the Perceiver's raw-input bottleneck or MMTM and MMTSA's convolutional streams, would bind the comparison to a specific encoder format. Conditioning methods such as FiLM and MAG assume an asymmetric relationship in which one modality modulates another, which does not match three co-equal sensor streams.

The seven retained methods cover the remaining intermediate and decision-level paradigms that do fit this setting (cf. Section 2.2 for the exact descriptions):

- **Late Fusion**: embedding concatenation, followed by a simple MLP classification head.
- **Gated Multi-modal Fusion (GMF)**: learned per-modality gating.
- **Low-Rank multi-modal Fusion (LMF)**: tensor interaction with a low-rank factorisation.
- **Cross-Modal Attention (CMA)**: directional attention between every modality pair.
- **Multi-modal Bottleneck Transformer (MBT)**: cross-modal exchange through shared bottleneck tokens.
- **CLS-Token Transformer**: a learnable token aggregating all modalities under unrestricted self-attention.
- **Decision Fusion**: a learned weighted sum of per-modality class distribution predictions.

Together, these span the breadth of the fusion taxonomy, from the simplest concatenation to tensor, gated, attention-based, and bottleneck designs. Six operate at the intermediate (embedding) level and one, Decision Fusion, at the decision level, so the set also covers two of the three canonical fusion stages. Early (raw-signal) fusion is excluded by construction, since all methods operate on encoder embeddings. The general structure of our multi-modal classification pipeline is displayed in Fig. 1. Full architectural details for each method are given in the respective original publications and in graph-based visual representations in Appendix A.

3.4 Training and Evaluation Protocol

Windowing We segment the continuous recordings into 10-second windows (leading to 500 samples per window at 50 Hz). For 3-fold cross-validation, we use a step of 250 samples (50% overlap), which gives roughly 43,760 windows. For LOPO we, use non-overlapping windows (step = 500, matching the HARMES

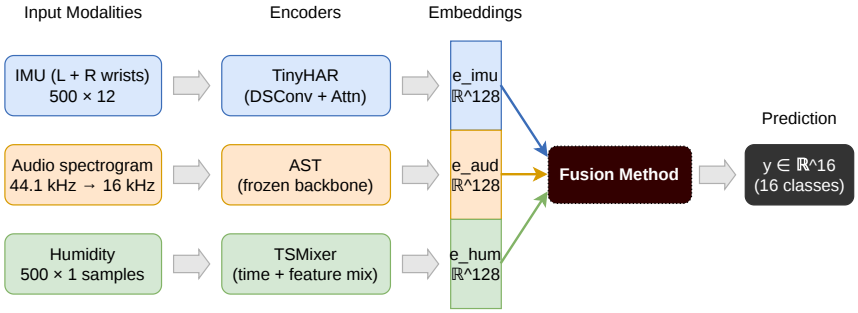


Fig. 1. Overview of the multi-modal pipeline. Each modality is encoded into a 128-dimensional embedding by a dedicated encoder. The three embeddings are then combined by one of seven interchangeable fusion methods to predict the activity class. The fusion block is held generic here. The seven concrete architectures are detailed in Section 3.3 and Appendix A.

paper [8]), yielding 21,897 windows. Each window takes the activity label that covers most of its span (simple majority voting). Because the sliding window runs across the entire recording, including background and transition periods, windows without majority activity labels are assigned to the "Null" class.

Cross-validation splits We use two cross-validation protocols: 3-fold group cross-validation for all fusion methods and 20-fold Leave-One-Participant-Out for the best method from the 3-fold CV evaluation, for comparison with the HARMES baseline. In **3-fold group cross-validation**, participants are split into three disjoint folds (see Table 1). Within each fold, the smallest-ID participant outside the test set is held out for validation, used for early stopping and checkpoint selection, and the rest form the training pool. Train, validation, and test sets are disjoint at the participant level in every fold. In **20-fold Leave-One-Participant-Out (LOPO)**, each participant is held out once as the test set. For each test participant P_k, the smallest-ID participant among the remaining 19 is used for validation, and the other 18 are used for training. The LOPO paradigm matches the evaluation in the original HARMES paper [8], allowing a direct comparison. We use three folds rather than a larger fold count for the main comparison for experiment runtime reasons. As every fusion method is trained and evaluated on each fold, the total training cost scales with the number of methods times the number of folds. Three folds still leave roughly seven held-out participants per fold, enough for a stable estimate, while keeping the seven-way comparison tractable. We reserve the more expensive LOPO protocol for the single best-performing method, where a direct comparison with the original HARMES paper is most informative.

Table 1. Participant assignment for the three cross-validation folds.

Fold	Test	Val	Train
0	P1–P7	P8	P9–P20
1	P8–P14	P1	P2–P7, P15–P20
2	P15–P20	P1	P2–P14

Training configuration We apply an almost identical training configuration to all model configurations. We use Adam as the optimizer with cosine annealing (no restarts), a maximum of 50 epochs, and a batch size of 32. The learning rate is 10^{-3} for every method except LMF, which uses 5×10^{-3} to compensate for vanishing gradients through its three-way tensor product. Early stopping (patience 10 epochs) monitors validation macro F1, and the best checkpoint is loaded for a single final evaluation on the held-out test set, which is never used for model selection. TinyHAR and TSMixer are trained from scratch, while the AST backbone, which is available pre-trained, stays frozen. We train all methods using cross-entropy loss except Decision Fusion, for which we use negative log-likelihood on log-probabilities.

Evaluation For our comparison, we compute metric scores on ground truth labels and predictions by each model. The primary metric is macro-averaged F1, which weights all 16 classes equally regardless of their frequency. This is relevant, as the class distribution in the original dataset is imbalanced. We also report accuracy for comparability with prior work, along with per-class and per-participant results, and confusion matrices for diagnostic analysis of confounded classes/samples.

To assess whether the observed performance differences between fusion methods are statistically significant, we perform paired significance tests on the per-participant F1 scores. We conduct a pairwise comparison of all fusion methods with each other, leaving us with 21 pairwise significance tests. Each participant serves as one paired observation, resulting in 20 paired F1 scores for every comparison. We run a Shapiro-Wilk test to test for normality on the differences of the 20 participants' F1 scores for each pair of methods. When normality of the paired differences was not rejected, we report the paired t-test. We additionally report the Wilcoxon signed-rank test, which is a non-parametric alternative. Because 21 pairwise comparisons are performed, Holm-Bonferroni correction is applied to control the family-wise error rate.

4 Results

We compare the seven fusion methods against the unimodal baselines and the HARMES reference using the protocols of Section 3. Our analysis, and thus, the results are organised around two questions: how much does combining modalities help, and which fusion strategy makes the best use of them. Table 2 summarises

Table 2. Results across modality combination, fusion strategies, and cross-validation modes. F1 is the macro F1 score, averaged over folds (3-fold CV) or participants (LOPO), Acc. is the prediction accuracy, averaged in the same way. All fusion methods use the three modalities jointly (Modality = All). **Bold** font marks the best result within each section. The HARMES baseline is the multimodal leave-one-participant-out macro F1 reported by Burchard et al. [8]

Model / Method	Modality	F1	Acc
Unimodal baselines (3-fold CV)			
TinyHAR	IMU	0.696	0.724
AST	Audio	**0.734**	**0.777**
TSMixer	Humidity	0.088	0.210
Fusion methods (3-fold CV)			
GMF	All	**0.827**	**0.854**
Late Fusion	All	0.817	0.845
CMA	All	0.795	0.832
CLS Transformer	All	0.793	0.832
MBT	All	0.787	0.821
LMF	All	0.747	0.786
Decision Fusion	All	0.747	0.783
Leave-one-participant-out (LOPO)			
GMF	All	**0.819**	**0.856**
HARMES baseline [8]	All	0.760	0.794

the aggregated results for each fusion strategy, and the subsections that follow break them down by class, by participant, and by handedness.

Fusion strategies improve over the strongest single modality across the board. The best unimodal model is AST on audio at 0.734 macro F1, and the best fusion method, GMF, reaches 0.827, an improvement of 9.3 points. Audio is by far the most informative single stream, while humidity on its own is close to the chance level at 0.09 macro F1, which is consistent with the modality ablation reported in the HARMES paper [8]. The 3-fold and leave-one-participant-out protocols agree closely for GMF (0.827 and 0.819), so the comparison is stable across evaluation schemes, and GMF under LOPO also exceeds the HARMES multi-modal baseline of 0.760 by 5.9 points.

4.1 Fusion Method Comparison

Figure 2 ranks the seven fusion methods on 3-fold cross-validation. Every method clears the best unimodal baseline (AST, 0.734 macro F1), confirming that combining modalities helps regardless of the tested fusion strategy. The margin, however, varies. GMF leads at 0.827, with Late Fusion close behind at 0.817. The two simplest mechanisms, gated summation and plain concatenation, are also the two strongest. The attention- and transformer-based methods form a middle cluster (CMA 0.795, CLS-Token Transformer 0.793, MBT 0.787) that lands several points below GMF. LMF and Decision Fusion trail at 0.747, beating the audio-based unimodal baseline only by a smaller margin.

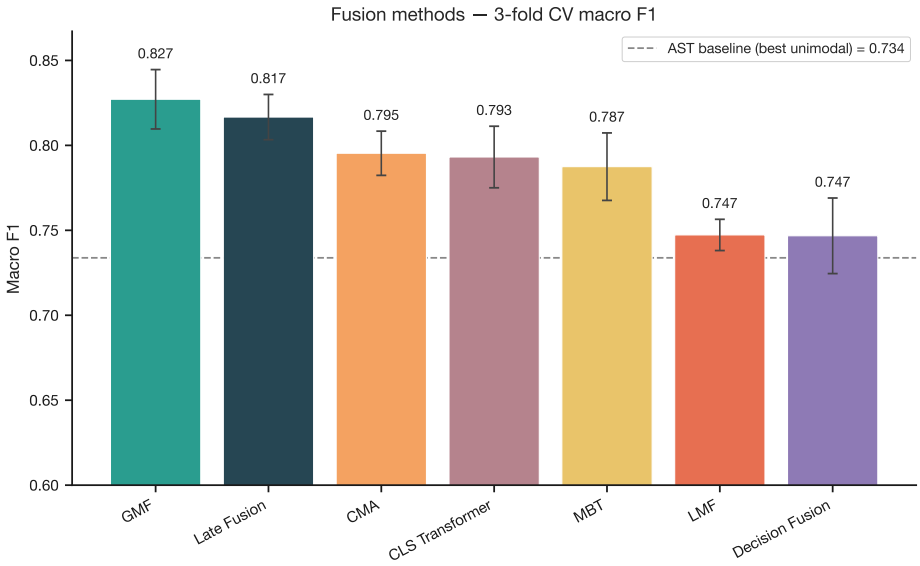


Fig. 2. Fusion method comparison on 3-fold group cross-validation (macro F1). All methods use the three modalities jointly. The dashed line marks the best unimodal baseline (AST).

As explained in Section 3.4, we conduct statistical tests on the prediction results. As a primary comparison, we look at the differences between GMF and concatenation late fusion. The Shapiro-Wilk test does not reject normality of the pairwise differences (W=0.0962, p=0.592). The paired t-test shows no significant difference (p=0.0519) at $\alpha = 0.05$. The Wilcoxon signed-rank test (p=0.0826) also does not yield a significant difference between GMF and late fusion. As the normality was rejected for three comparisons, we report the results of the applied Wilcoxon signed-rank test. Figure 3 displays the Holm-adjusted Wilcoxon signed-rank test results for all 21 pairwise method comparisons. While GMF and late fusion are not separable, they can be grouped into a leading group, which can be separated from all lower-scoring methods with small (≤ 0.001) corrected p-values. Another group can be formed out of CMA, CLS transformer, and MBT, which are not separable from each other. The trailing group, separated from the middle group, is formed by LMF and Decision Fusion. These findings from the statistical tests confirm the clustering we deduce from Figure 2.

4.2 Class performance analysis

To analyze on which classes the performance is improved by multimodel models, Figure 4 shows the per-class confusion-matrix difference between the best multimodal model (GMF) and the best unimodal one (AST). Each diagonal entry shows how much more often GMF classifies that activity correctly, and the blue,

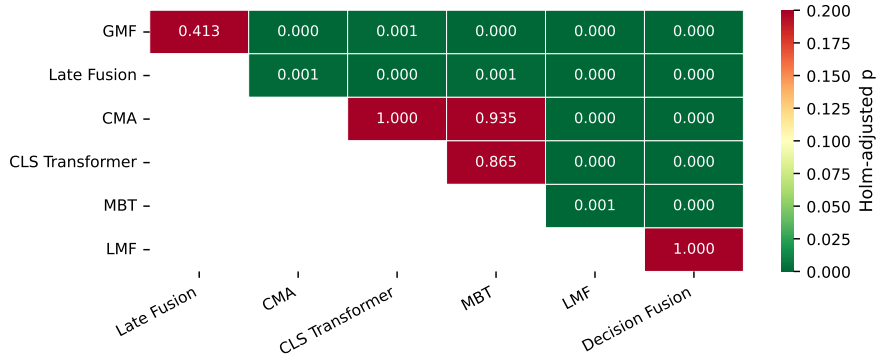


Fig. 3. Pairwise separability - Wilcoxon signed-rank over 20 participants, Holm-Bonferroni-adjusted p-values (green = separable at $\alpha = 0.05$, red = indistinguishable). Methods are ordered by mean F1 scores (left to right and top to bottom).

negative off-diagonal entries are the confusions GMF removes in comparison. The improvement is concentrated in activities that are quiet or acoustically ambiguous but involve distinctive hand motion. Notably, *Putting away dishes* is recognized correctly +25pp more often, disinfecting hands +24pp more, and drinking +19pp more. These are exactly the classes where audio alone is weak, as there is little characteristic sound. The overall confounding pair of *Apply hand cream* and *Disinfecting hands* is handled better by GMF as well, although notably, while the confusion of *Disinfecting hands* with *Apply hand cream* is reduced by -24pp, vice-versa, it is increased by +8pp.

We additionally report per-modality absolute confusion matrices, together with those of GMF and the GMF-minus-TinyHAR (multimodal-minus-IMU-only) difference, in Appendix B.1.

In Fig. 5, we display the per-class F1-scores of all fusion strategies in a bar chart. For each class on the x-axis, the fusion strategies are ordered by their previously calculated macro performance. With minor exceptions, the global ordering of fusion strategy performances also applies to the per-class performance, where GMF and Late Fusion by concatenation perform best. Disinfecting hands is consistently the most challenging class for all fusion strategies.

4.3 Generalisation to unseen participants

Leave-one-participant-out The 3-fold protocol mixes participants across folds. To approximate the performance on a single, unseen participant and to be comparable to the HARMES baseline [8], we re-evaluate the best model, GMF, under the stricter leave-one-participant-out (LOPO) scheme. GMF reaches 0.819 macro F1 under LOPO, within one point of its 3-fold score of 0.827. The drop is negligible despite LOPO being the more demanding test. GMF under LOPO also exceeds the multi-modal HARMES baseline of 0.760 by 5.9 points, on the

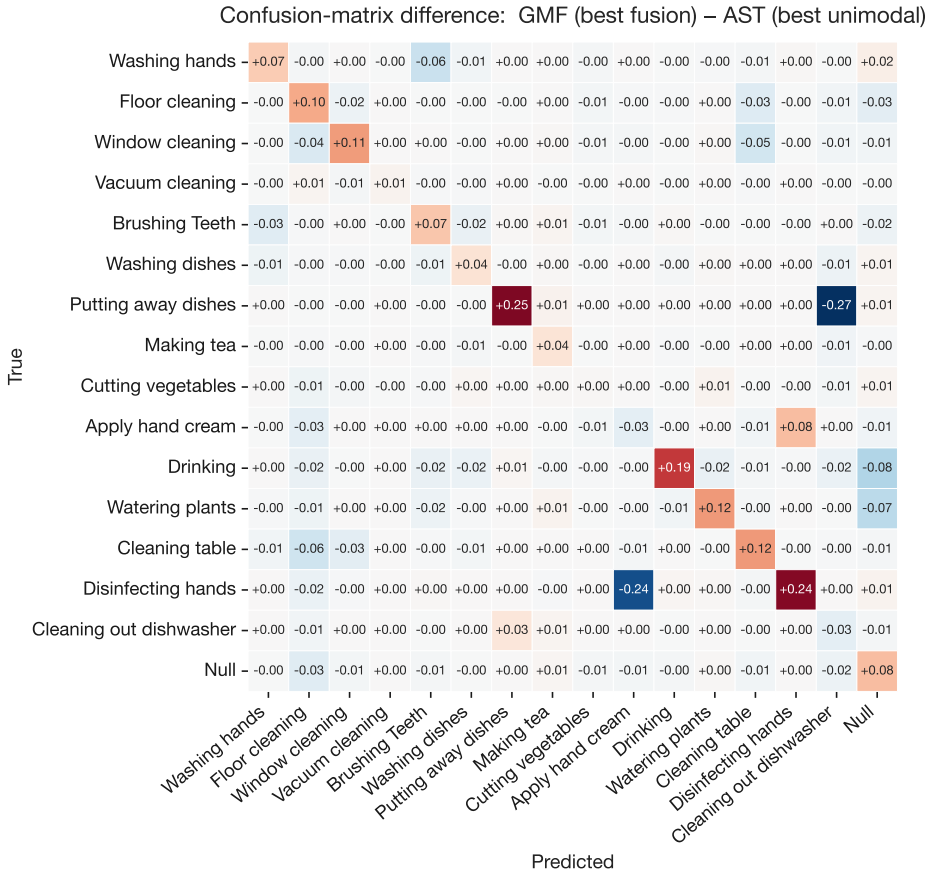


Fig. 4. Per-class confusion-matrix difference, GMF minus AST (best unimodal). Note the interpretation: Red, positive values on the diagonal mark classes recognised more reliably under fusion. Blue, negative off-diagonal entries mark confusions that fusion removes. Blue, negative values on the diagonal, or off-diagonal red values mean that the GMF model performed worse than the unimodal AST for the specific confusion matrix entry.

exact same evaluation protocol. To the best of our knowledge, our multi-modal classification model with GMF is therefore also the currently highest performing model on the HARMES dataset.

The per-class picture reinforces this. We display the GMF LOPO confusion matrix in Fig. 6. When comparing the LOPO confusion matrix with the 3-fold matrix in Appendix B.1 (Figure B.1), LOPO matches or slightly exceeds the 3-fold result on most activities, gaining a little on vacuum cleaning, brushing teeth, and window cleaning. The model is almost a 100 % accurate on the activities with a clear acoustic and motion signature: vacuum cleaning (0.97), cutting vegetables (0.94), making tea and brushing teeth (0.91), and washing dishes and window

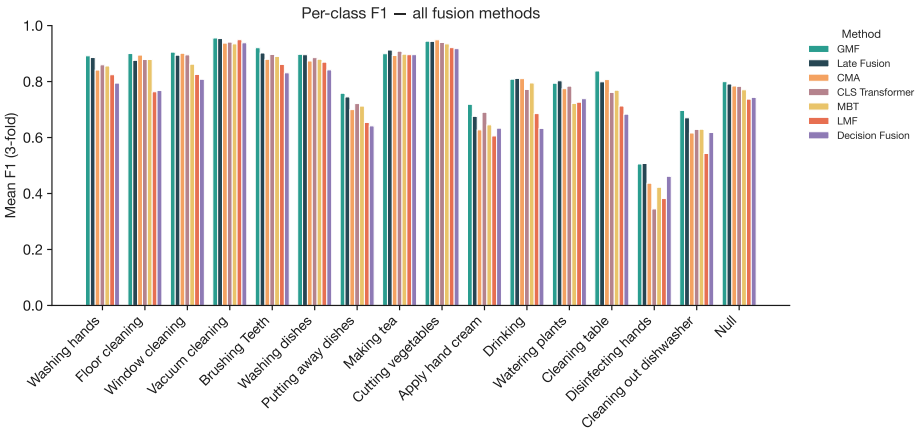


Fig. 5. Per-class macro F1 scores for each fusion strategy. Results presented are averaged over the three-fold CV. Models are sorted in descending order by their global performance, from left to right.

cleaning (0.90). Where LOPO falls slightly behind, the gap is confined to the low-support self-care classes, disinfecting hands and applying hand cream, which have the fewest windows and the least distinctive signatures and are therefore the most sensitive to an unseen participant's personalized motion. These differences are small, and the bulk of the activities are recognized cleanly, confirming that the multi-modal model generalizes well to new users rather than overfitting to the training participants.

Per-participant results and left-handed participants The results are mostly stable across participants, but deviations exist. Fig. 7 shows a heatmap of the macro F1 scores per model and per participant. The largest variations exist in the unimodal IMU-model TinyHAR, with left-handed participants scoring significantly lower. No model strictly beats all other models over all participants. Rather, the results are close, and GMF performs best for most participants, closely followed by Late fusion by concatenation.

Three participants (P07, P10, P14) are left-handed, which negatively affects the IMU modality because the dominant-hand motion appears on the opposite wrist from the right-handed majority. Audio and humidity are largely unaffected by handedness, so any modality-induced robustness gap likely comes from the IMU. We show the gap between the groups of left-handers and right-handers in Fig. 8, for each fusion strategy and for the unimodal models. The IMU-based unimodal TinyHAR by far shows the largest gap of 0.18 (0.54 vs 0.72).

Humidity is excluded from this comparison because it carries almost no discriminative signal for any participant.

Audio behaves in the opposite way. AST is essentially handedness-invariant and in fact performs marginally better on the left-handers (0.74 versus 0.73),

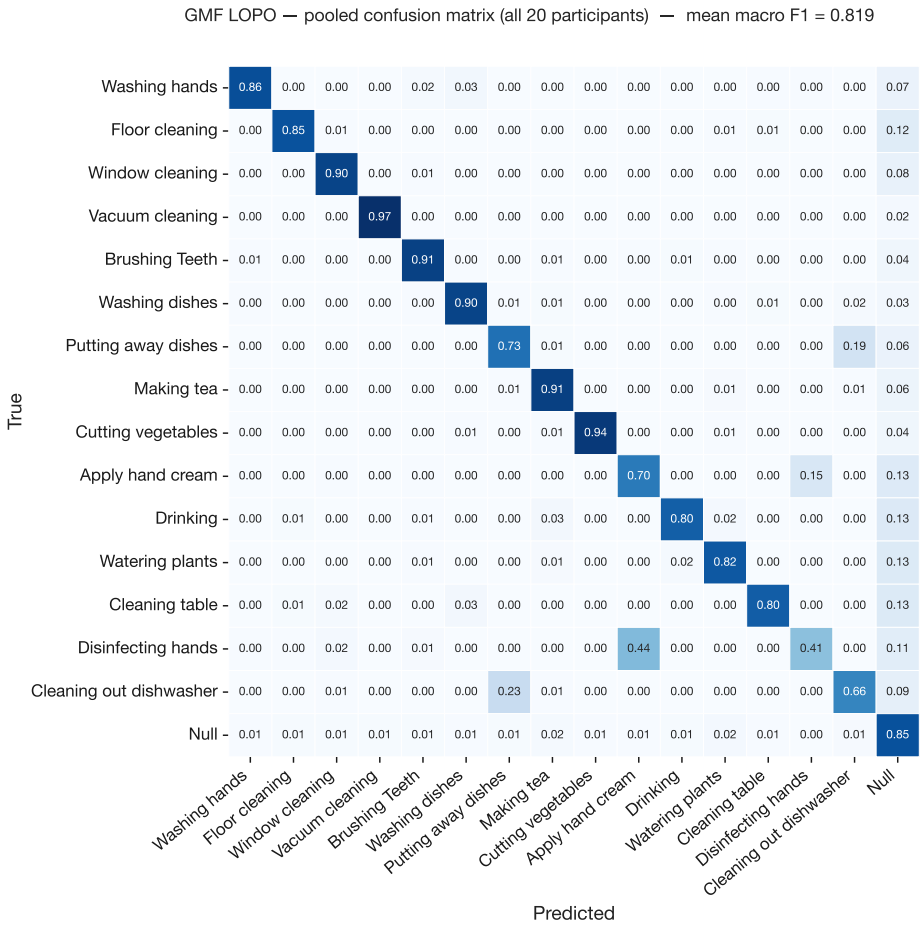


Fig. 6. Pooled confusion matrix for GMF under leave-one-participant-out evaluation, aggregated over all 20 held-out participants.

since a microphone picks up the same activity regardless of which hand performs it. Fusion inherits this robustness and recovers what the IMU alone loses. For every left-hander, the most informative model is a fusion method.

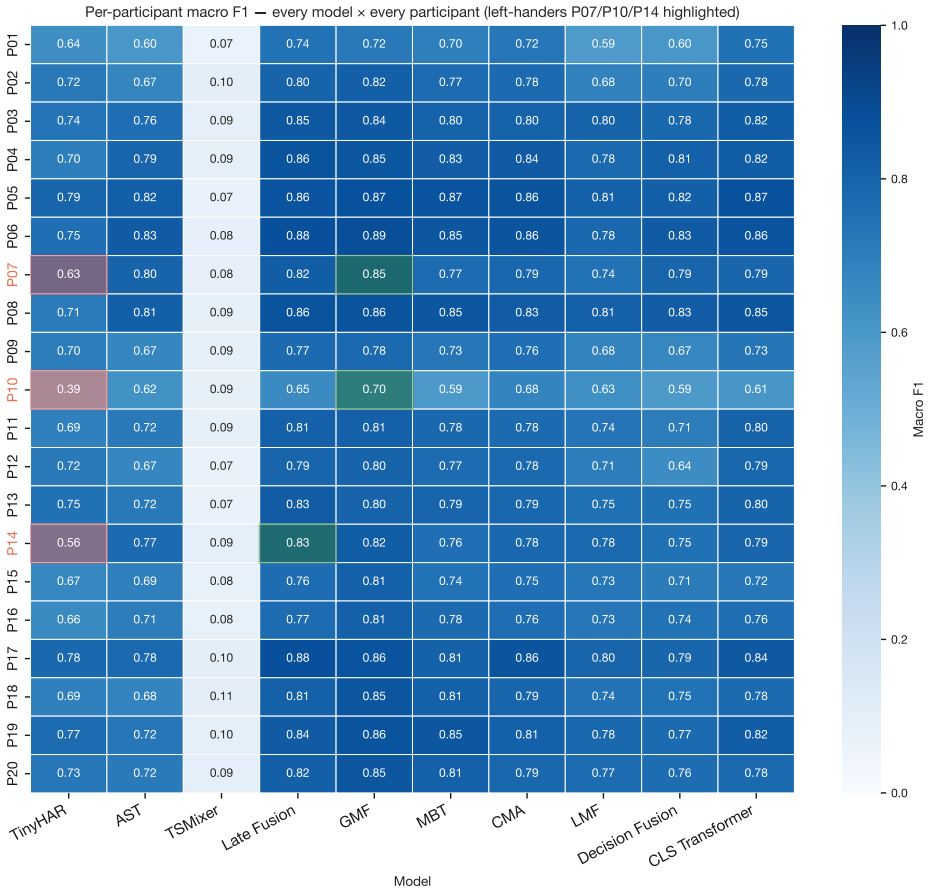


Fig. 7. Heatmap table showing per-participant macro F1 scores for each participant and model (3-fold CV). Left-handed participants are marked in red font. The three leftmost models are unimodal: TinyHAR (IMU), AST (Audio), TSMixer (humidity). The worst results on each left-handed participant are marked in red, and the best results on them are marked in green.

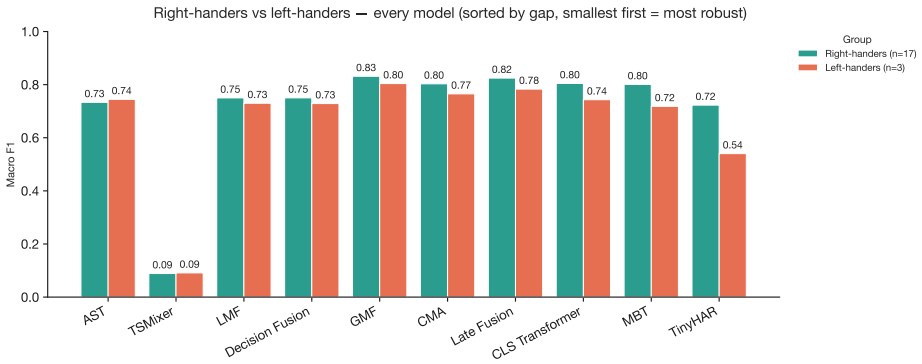


Fig. 8. 3-Fold CV macro F1 performance per model, split by dominant hand into left-handers and right-handers. The plot shows both unimodal (AST: audio, TSMixer: humidity, TinyHAR: IMU) and multi-modal methods (all others), sorted by performance gap between groups of left-handers and right-handers, descending from left to right.

5 Discussion

Simple fusion outperforms complex fusion in untuned setting From the performance results, we infer that two of the simplest fusion methods score highest. GMF (0.827) and Late Fusion (0.817) outrank the attention- and transformer-based methods (CMA, CLS Transformer, MBT, all near 0.79), and the tensor and decision-level methods (LMF, Decision Fusion) trail at 0.75. This runs against the intuition that richer cross-modal interaction should help. We attribute it to the setting rather than to a flaw in those methods. With only three modalities, one of which (humidity) is of limited use in 10s windows, and with a frozen, already-strong audio encoder, there is little structure left for elaborate attention to exploit, and the extra parameters are harder to fit on a dataset of 20 participants, despite the comparatively large size of 61 labeled hours. Especially for HAR, where datasets are often not that large, it seems that these simpler methods already deliver the best performance.

Fusion improves performance by resolving specific confusions, not by a uniform lift The per-class analysis (Section 4.2) shows the performance gain is concentrated in activities that are ambiguous in a unimodal IMU or audio setting, but have an additional distinctive acoustic or motion pattern. This applies, e.g., to putting away dishes, disinfecting hands, and drinking. These improve because fusion untangles confusion pairs the audio or IMU model could not separate on its own, for example, putting away dishes versus cleaning out the dishwasher, or disinfecting hands versus applying hand cream. Activities with a loud, unmistakable acoustic signature (vacuum cleaning, cutting vegetables, making tea) are already solved by the acoustic model alone, and their performance thus does not benefit from other modalities. The gain of the fusion models over the best

unimodal model (AST) is moderate in total, and fusion adds value precisely on the subset of activities where a microphone is blind. Overall, the results of this work strengthen pre-existing findings that showed the benefits of multi-modal HAR.

Humidity contribution is minimal In our 10 s-window model setting, across the fusion methods, removing the humidity stream changes macro F1 by a fraction of a percentage point, and TSMixer (0.088) performs near the chance level. Only watering plants and cleaning out the dishwasher, where moisture genuinely changes, benefit at the class level. The practical implication is that a two-sensor IMU-plus-audio system captures essentially all of the achievable performance on this dataset for 10 s windows, and the humidity sensor can be dropped without cost. We report this as a negative result worth recording rather than a limitation of the fusion methods. As acknowledged by Burchard et al. in the dataset publication [8], the humidity sensor reacts to the environment changes slowly and should thus be treated differently from the accurate high-frequency measurements of IMU and microphone.

The models generalise to unseen participants Under the stricter leave-one-participant-out protocol, where every test participant is entirely unseen and evaluated alone, GMF reaches 0.819, within one point of its 3-fold score and 5.9 points above the already multi-modal HARMES baseline of 0.760 on the same protocol. The per-class behaviour is stable between the two protocols, with the small residual drops confined to the low-support classes *Apply hand cream* and *disinfecing hands*. This indicates the multi-modal model learns activity signatures that transfer across individuals rather than fitting the particular participants it was trained on. While the literature sees LOPO-validation as the gold standard, the small performance difference between 3-fold CV and LOPO for GMF hints at the validity of the tradeoff between cross-validation meaningfulness and training time/energy consumption we chose.

Fusion improves robustness to handedness The IMU-only unimodal model (TinyHAR) degrades sharply for the three left-handed participants, since the dominant-hand motion appears on the opposite wrist, giving it the largest right-hand-versus-left-hand gap of any model (0.72 versus 0.54). Audio is handedness-invariant and is unaffected. Because fusion combines the two, the multimodal models inherit the audio model's robustness and recover most of what the IMU loses: GMF narrows the gap to 0.027 and is the best model for the left-handers in nearly every case. Beyond raw accuracy, this is a fairness argument for multimodal sensing: a single-modality IMU system would systematically underperform for a minority of users, and adding audio removes most of that disparity. Likely, this issue can also be overcome by unimodal IMU models to a certain degree, e.g., by applying sophisticated data augmentation techniques.

Limitations Several caveats bound these conclusions. The dataset has twenty participants and only three left-handers, so the handedness finding, while consistent, rests on a small sample. All results come from a single dataset of activities

of daily living, and the audio backbone is frozen, so the ceiling we observe partly reflects a fixed pretrained representation rather than end-to-end optimization. Due to the design of the encoder networks, the conclusions we draw are specific to this architecture and should be tested against other encoder architectures and datasets in the future.

The ordering of fusion methods may also shift with more modalities or larger data, where the more expressive architectures could have room to pay off. Finally, performance on the rare self-care classes remains modest under cross-participant evaluation, which is the most likely target for future improvement. Our selection of fusion methods is also limited to methods fusing at a late intermediary stage in the model, after passing the data through completely separate encoder heads. Early fusion approaches are excluded from this particular work, but should also be considered by practitioners when developing fusion models.

Limitations Several caveats bound these conclusions. First, all experiments are conducted on the HARMES dataset, which comprises 20 participants, including only three left-handed individuals. While the findings are consistent across both evaluation protocols, they should therefore be interpreted with appropriate caution and validated on additional multi-modal HAR datasets. Second, the conclusions are specific to the evaluated model setting. All fusion methods use the same encoder architectures and largely identical training hyperparameters to enable a controlled comparison, but different encoder designs or method-specific hyperparameter tuning could alter the relative ranking of individual fusion approaches. Likewise, the audio encoder is kept frozen throughout training, meaning that the observed results reflect fusion of strong pre-trained representations rather than fully end-to-end optimized models.

Consequently, our findings should not be interpreted as evidence that simple fusion strategies universally outperform more expressive attention- or transformer-based approaches. Rather, they indicate that, for the evaluated setting (three sensing modalities, a frozen audio backbone, 10 s windows, and a dataset of this scale) simple fusion methods provide the strongest performance. The ordering of fusion methods may therefore change when additional sensing modalities, larger datasets, different encoder architectures, or end-to-end training are considered.

Finally, our comparison focuses exclusively on intermediate and late fusion methods operating on separate modality-specific encoders. Early fusion approaches were intentionally excluded from this study and therefore fall outside the scope of the presented comparison.

6 Conclusion

We compared seven fusion strategies for multi-modal activity recognition on the HARMES dataset under a leakage-free, participant-grouped protocol. Every fusion method improved on the best unimodal model, while gated multimodal fusion (GMF) and late concatenation fusion achieved the highest performance. Although GMF obtained the highest average macro F1 score (0.827), its advantage over late fusion (0.817) was not statistically significant, indicating that

both represent competitive design choices for this experimental setting. The more complex attention- and tensor-based methods did not provide additional benefits under the evaluated conditions. The difference can possibly be attributed to the specific setting, with a relatively dominant frozen audio-encoder present in the fusion system.

The multi-modal gains came from resolving confusions between activities that sound alike but display different movement patterns and vice versa, with audio and IMU providing complementary information. The best-performing fusion models generalized well to unseen participants (macro F1 up to 0.819 under leave-one-participant-out, 5.9 percentage points above the HARMES baseline) and were substantially more robust to handedness than the IMU-only model. Overall, our findings suggest that lightweight fusion of IMU and audio is an effective and practical choice for activity recognition on the HARMES dataset, while more complex fusion strategies do not necessarily provide additional benefits in this particular setting.

Future work should investigate whether these findings generalize beyond HARMES by evaluating additional multi-modal HAR datasets with different sensing modalities, participant populations, and activity domains. Likewise, exploring alternative encoder architectures, end-to-end training, and method-specific hyperparameter optimization will help determine whether the observed similarity between simple and more complex fusion strategies persists under different experimental conditions.

From an application perspective, a natural next step is wearable deployment through model quantization and pruning for on-device inference, combined with streaming inference for continuous rather than window-based recognition. Additional comparisons with early fusion approaches would further broaden the design space. Finally, longer temporal contexts or multi-rate fusion strategies may better exploit slow-changing environmental modalities such as humidity, while larger datasets and additional sensing modalities could reveal whether more expressive fusion mechanisms become advantageous.

Acknowledgments. The authors would like to thank the participants of the HARMES dataset and all researchers involved in creating it.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

Contribution Statement. AM and RB contributed equally to this work. AM co-designed the experiments, executed the experiments, collected all results, and co-wrote the manuscript. RB co-designed the experiments and co-wrote the manuscript. All authors reviewed the manuscript.

References

1. Aguileta, A.A., Brena, R.F., Mayora, O., Molino-Minero-Re, E., Trejo, L.A.: Multi-Sensor Fusion for Activity Recognition—A Survey. Sensors **19**(17), 3808 (Jan 2019). https://doi.org/10.3390/s19173808

2. Arevalo, J., Solorio, T., Montes-y-Gómez, M., González, F.A.: Gated multimodal networks. Neural Computing and Applications **32**(14), 10209–10228 (Jul 2020). https://doi.org/10.1007/s00521-019-04559-1
3. Atrey, P.K., Hossain, M.A., El Saddik, A., Kankanhalli, M.S.: Multimodal fusion for multimedia analysis: A survey. Multimedia Systems **16**(6), 345–379 (Nov 2010). https://doi.org/10.1007/s00530-010-0182-0
4. Bian, S., Liu, M., Rey, V.F., Geißler, D., Lukowicz, P.: TinierHAR: Towards Ultra-Lightweight Deep Learning Models for Efficient Human Activity Recognition on Edge Devices. In: Proceedings of the 2025 ACM International Symposium on Wearable Computers. pp. 163–169. ACM, Espoo Finland (Oct 2025). https://doi.org/10.1145/3715071.3750410
5. Bralina, S., Yazici, A., Guan, C., Lee, M.H.: Adaptive bottleneck transformer for multimodal EEG, audio, and vision fusion. Expert Systems with Applications **312**, 131487 (May 2026). https://doi.org/10.1016/j.eswa.2026.131487
6. Bulling, A., Blanke, U., Schiele, B.: A tutorial on human activity recognition using body-worn inertial sensors. ACM Comput. Surv. **46**(3), 33:1–33:33 (Jan 2014). https://doi.org/10.1145/2499621
7. Burchard, R., Ali, H., Van Laerhoven, K.: Improved Strategies for Multi-modal Atmospheric Sensing to Augment Wearable IMU-Based Hand Washing Detection. In: Durmaz Incel, Ö., Qin, J., Bieber, G., Kuijper, A. (eds.) Sensor-Based Activity Recognition and Artificial Intelligence, vol. 16292, pp. 308–323. Springer Nature Switzerland, Cham (2026). https://doi.org/10.1007/978-3-032-13312-0_18
8. Burchard, R., Brückner, P.A., Bock, M., Gall, J., Laerhoven, K.V.: HARMES: A Multi-Modal Dataset for Wearable Human Activity Recognition with Motion, Environmental Sensing and Sound (May 2026). https://doi.org/10.48550/arXiv.2605.02596
9. Burchard, R., Brückner, P.A., Bock, M., Van Laerhoven, K.: HARMES: A Multi-Modal Dataset for Wearable Human Activity Recognition with Motion, Environmental Sensing and Sound (Apr 2026). https://doi.org/10.5281/zenodo.19425719
10. Burchard, R., Van Laerhoven, K.: Multi-modal Atmospheric Sensing to Augment Wearable IMU-Based Hand Washing Detection. In: Konak, O., Arnrich, B., Bieber, G., Kuijper, A., Fudickar, S. (eds.) Sensor-Based Activity Recognition and Artificial Intelligence. pp. 55–68. Springer Nature Switzerland, Cham (2025). https://doi.org/10.1007/978-3-031-80856-2_4
11. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (May 2019). https://doi.org/10.48550/arXiv.1810.04805
12. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale (Jun 2021). https://doi.org/10.48550/arXiv.2010.11929
13. Ekambaram, V., Jati, A., Nguyen, N., Sinthong, P., Kalagnanam, J.: TSMixer: Lightweight MLP-Mixer Model for Multivariate Time Series Forecasting. In: Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. pp. 459–469 (Aug 2023). https://doi.org/10.1145/3580305.3599533
14. Gao, Z., wang, Y., Chen, J., Xing, J., Patel, S., Liu, X., Shi, Y.: MMTSA: Multi-Modal Temporal Segment Attention Network for Efficient Human Activity Recognition. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies **7**(3), 96:1–96:26 (Sep 2023). https://doi.org/10.1145/3610872

15. Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K.V., Joulin, A., Misra, I.: ImageBind One Embedding Space to Bind Them All. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 15180–15190 (Jun 2023). https://doi.org/10.1109/CVPR52729.2023.01457
16. Gong, Y., Chung, Y.A., Glass, J.: AST: Audio Spectrogram Transformer. In: Interspeech 2021. pp. 571–575. ISCA (Aug 2021). https://doi.org/10.21437/Interspeech.2021-698
17. Ha, S., Choi, S.: Convolutional neural networks for human activity recognition using multiple accelerometer and gyroscope sensors. In: 2016 International Joint Conference on Neural Networks (IJCNN). pp. 381–388. IEEE, Vancouver, BC, Canada (Jul 2016). https://doi.org/10.1109/IJCNN.2016.7727224
18. Jaegle, A., Gimeno, F., Brock, A., Vinyals, O., Zisserman, A., Carreira, J.: Perceiver: General Perception with Iterative Attention. In: Proceedings of the 38th International Conference on Machine Learning. pp. 4651–4664. PMLR (Jul 2021)
19. Koutoupis, S., Zervou, M.A., Kontras, K., Vos, M.D., Tsakalides, P., Tsagkatakis, G.: The More, the Merrier: Contrastive Fusion for Higher-Order Multimodal Alignment (Apr 2026). https://doi.org/10.48550/arXiv.2511.21331
20. Le, T.H., Nguyen, T.K., Le, T.A., Delalandre, M., Trung, K.T., Tran, T.H., Pham, C.: Mamba-MHAR: An efficient multimodal framework for human action recognition. Journal of Computer Science and Cybernetics pp. 245–264 (Sep 2025). https://doi.org/10.15625/1813-9663/22770
21. Lee, S., Lim, Y., Lim, K.: Multimodal sensor fusion models for real-time exercise repetition counting with IMU sensors and respiration data. Information Fusion **104**, 102153 (Apr 2024). https://doi.org/10.1016/j.inffus.2023.102153
22. Li, S., Zhu, T., Duan, F., Chen, L., Ning, H., Nugent, C., Wan, Y.: HARMamba: Efficient and Lightweight Wearable Sensor Human Activity Recognition Based on Bidirectional Mamba. IEEE Internet of Things Journal **12**(3), 2373–2384 (Feb 2025). https://doi.org/10.1109/JIOT.2024.3463405
23. Liu, Z., Shen, Y., Lakshminarasimhan, V.B., Liang, P.P., Zadeh, A., Morency, L.P.: Efficient Low-rank Multimodal Fusion with Modality-Specific Factors (May 2018). https://doi.org/10.48550/arXiv.1806.00064
24. Lu, J., Batra, D., Parikh, D., Lee, S.: ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks (Aug 2019). https://doi.org/10.48550/arXiv.1908.02265
25. Ma, H., Li, W., Zhang, X., Gao, S., Lu, S.: AttnSense: Multi-level Attention Mechanism For Multimodal Human Activity Recognition pp. 3109–3115 (2019)
26. Mollyn, V., Ahuja, K., Verma, D., Harrison, C., Goel, M.: SAMoSA: Sensing Activities with Motion and Subsampled Audio. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies **6**(3), 1–19 (Sep 2022). https://doi.org/10.1145/3550284
27. Moon, S., Madotto, A., Lin, Z., Saraf, A., Bearman, A., Damavandi, B.: IMU2CLIP: Language-grounded Motion Sensor Translation with Multimodal Contrastive Learning. In: Bouamor, H., Pino, J., Bali, K. (eds.) Findings of the Association for Computational Linguistics: EMNLP 2023. pp. 13246–13253. Association for Computational Linguistics, Singapore (Dec 2023). https://doi.org/10.18653/v1/2023.findings-emnlp.883
28. Münzner, S., Schmidt, P., Reiss, A., Hanselmann, M., Stiefelhagen, R., Dürichen, R.: CNN-based sensor fusion techniques for multimodal human activity recognition. In: Proceedings of the 2017 ACM International Symposium on Wearable Computers. pp. 158–165. ACM, Maui Hawaii (Sep 2017). https://doi.org/10.1145/3123021.3123046

29. Nagrani, A., Yang, S., Arnab, A., Jansen, A., Schmid, C., Sun, C.: Attention Bottlenecks for Multimodal Fusion (Nov 2022). https://doi.org/10.48550/arXiv.2107.00135
30. Ordóñez, F., Roggen, D.: Deep Convolutional and LSTM Recurrent Neural Networks for Multimodal Wearable Activity Recognition. Sensors **16**(1), 115 (Jan 2016). https://doi.org/10.3390/s16010115
31. Ouyang, X., Shuai, X., Zhou, J., Shi, I.W., Xie, Z., Xing, G., Huang, J.: Cosmo: Contrastive fusion learning with small data for multimodal human activity recognition. In: Proceedings of the 28th Annual International Conference on Mobile Computing And Networking. pp. 324–337. MobiCom '22, Association for Computing Machinery, New York, NY, USA (Oct 2022). https://doi.org/10.1145/3495243.3560519
32. Perez, E., Strub, F., De Vries, H., Dumoulin, V., Courville, A.: FiLM: Visual Reasoning with a General Conditioning Layer. Proceedings of the AAAI Conference on Artificial Intelligence **32**(1) (Apr 2018). https://doi.org/10.1609/aaai.v32i1.11671
33. Qiu, S., Zhao, H., Jiang, N., Wang, Z., Liu, L., An, Y., Zhao, H., Miao, X., Liu, R., Fortino, G.: Multi-sensor information fusion based on machine learning for real applications in human activity recognition: State-of-the-art and research challenges. Information Fusion **80**, 241–265 (Apr 2022). https://doi.org/10.1016/j.inffus.2021.11.006
34. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning Transferable Visual Models From Natural Language Supervision. In: Proceedings of the 38th International Conference on Machine Learning. pp. 8748–8763. PMLR (Jul 2021)
35. Rahman, W., Hasan, M.K., Lee, S., Bagher Zadeh, A., Mao, C., Morency, L.P., Hoque, E.: Integrating Multimodal Information in Large Pretrained Transformers. In: Jurafsky, D., Chai, J., Schluter, N., Tetreault, J. (eds.) Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 2359–2369. Association for Computational Linguistics, Online (Jul 2020). https://doi.org/10.18653/v1/2020.acl-main.214
36. Ramachandram, D., Taylor, G.W.: Deep Multimodal Learning: A Survey on Recent Advances and Trends. IEEE Signal Processing Magazine **34**(6), 96–108 (Nov 2017). https://doi.org/10.1109/MSP.2017.2738401
37. Tian, Y., Krishnan, D., Isola, P.: Contrastive Multiview Coding. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M. (eds.) Computer Vision – ECCV 2020. pp. 776–794. Springer International Publishing, Cham (2020). https://doi.org/10.1007/978-3-030-58621-8_45
38. Tsai, Y.H.H., Bai, S., Liang, P.P., Kolter, J.Z., Morency, L.P., Salakhutdinov, R.: Multimodal Transformer for Unaligned Multimodal Language Sequences (Jun 2019). https://doi.org/10.48550/arXiv.1906.00295
39. Vaezi Joze, H.R., Shaban, A., Iuzzolino, M.L., Koishida, K.: MMTM: Multimodal Transfer Module for CNN Fusion. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13286–13296. IEEE, Seattle, WA, USA (Jun 2020). https://doi.org/10.1109/CVPR42600.2020.01330
40. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., ukasz Kaiser, Ł., Polosukhin, I.: Attention is All you Need. In: Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017). https://doi.org/10.48550/arXiv.1706.03762

41. Wang, J., Chen, Y., Hao, S., Peng, X., Hu, L.: Deep learning for sensor-based activity recognition: A survey. Pattern Recogn. Lett. **119**(C), 3–11 (Mar 2019). https://doi.org/10.1016/j.patrec.2018.02.010
42. Wang, K., Liu, C., Zhang, R.: CMA-SOD: Cross-modal attention fusion network for RGB-D salient object detection. The Visual Computer **41**(7), 5135–5151 (May 2025). https://doi.org/10.1007/s00371-024-03712-9
43. Yadav, S.K., Tiwari, K., Pandey, H.M., Akbar, S.A.: A review of multimodal human activity recognition with special emphasis on classification, applications, challenges and future directions. Knowledge-Based Systems **223**, 106970 (Jul 2021). https://doi.org/10.1016/j.knosys.2021.106970
44. Yılmaz, T.A., Yatbaz, H.Y., Ever, E., Yazici, A.: Hierarchical human activity recognition with fusion of audio and multiple inertial sensor modalities. Scientific Reports **16**(1), 382 (Dec 2025). https://doi.org/10.1038/s41598-025-29801-w
45. Zadeh, A., Chen, M., Poria, S., Cambria, E., Morency, L.P.: Tensor Fusion Network for Multimodal Sentiment Analysis. In: Palmer, M., Hwa, R., Riedel, S. (eds.) Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. pp. 1103–1114. Association for Computational Linguistics, Copenhagen, Denmark (Sep 2017). https://doi.org/10.18653/v1/D17-1115
46. Zhou, Y., Zhao, H., Huang, Y., Riedel, T., Hefenbrock, M., Beigl, M.: 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. ACM, Cambridge United Kingdom (Sep 2022). https://doi.org/10.1145/3544794.3558467

A Fusion Strategy Visualization

The following diagrams depict the seven fusion strategies we employed for this comparison study. They replace the block "Fusion Method" in Fig. 1.

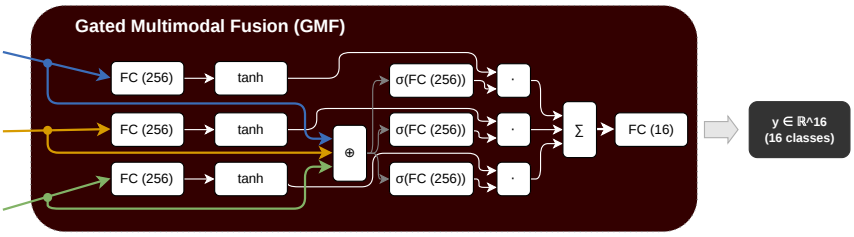


Fig. 9. Gated Multi-modal Fusion (GMF)

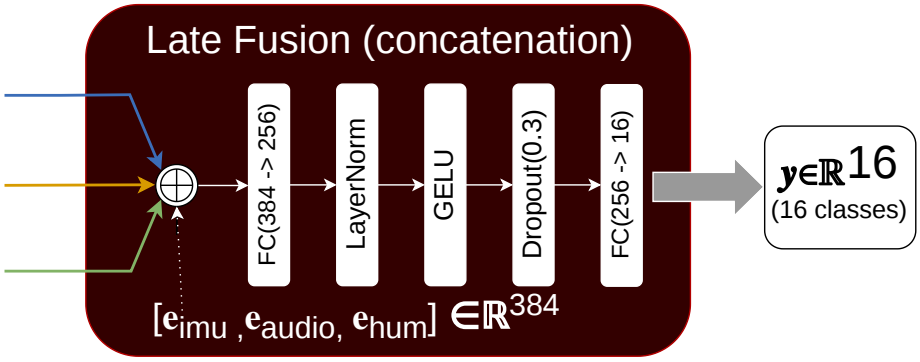


Fig. 10. Late Fusion (concatenation)

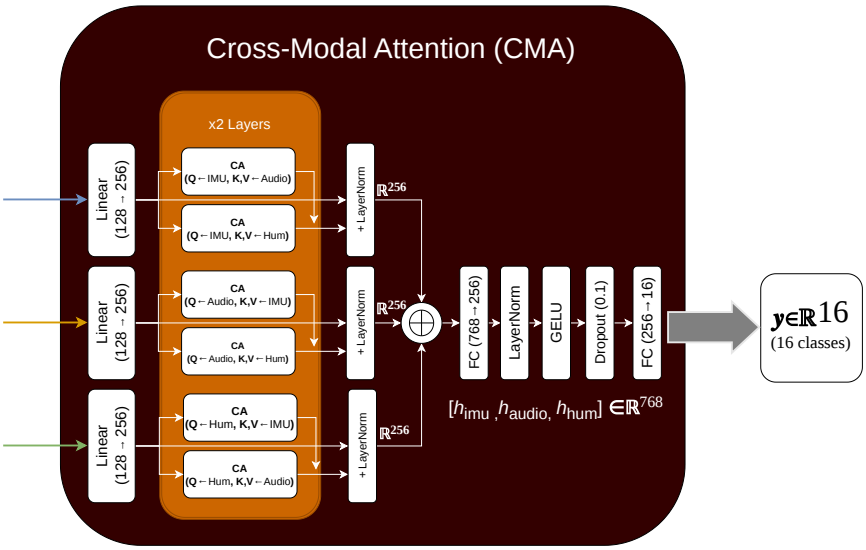


Fig. 11. Cross-Modal Attention (CMA)

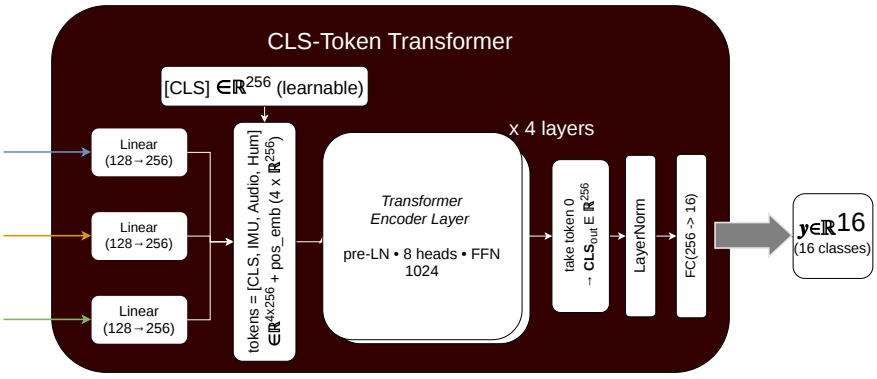


Fig. 12. CLS-Token Transformer

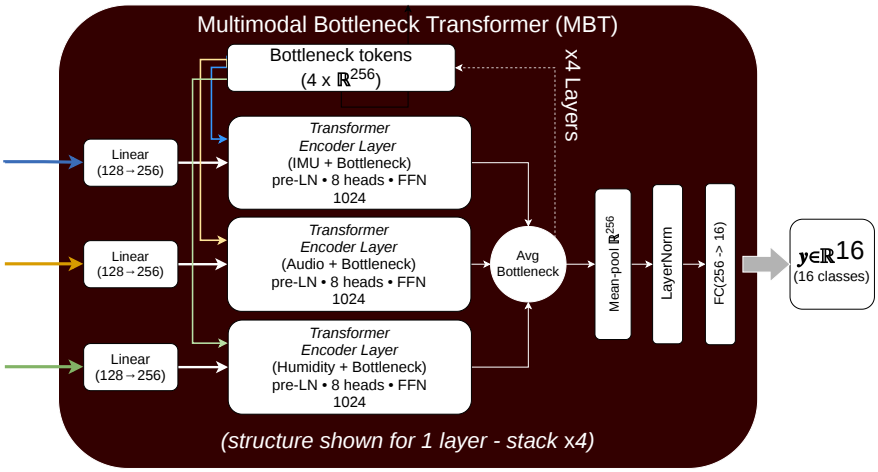


Fig. 13. Multi-modal Bottleneck Transformer

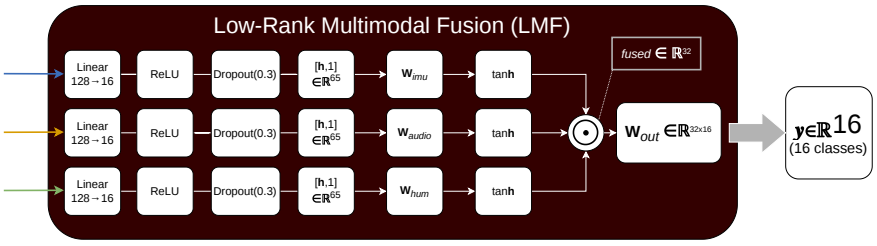


Fig. 14. Low-Rank multi-modal Fusion

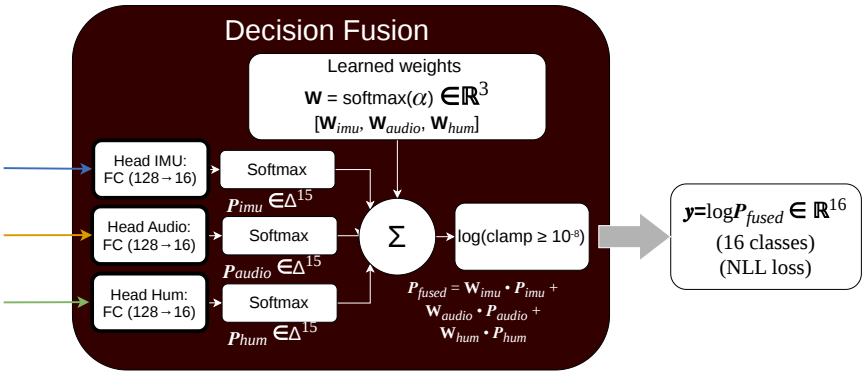
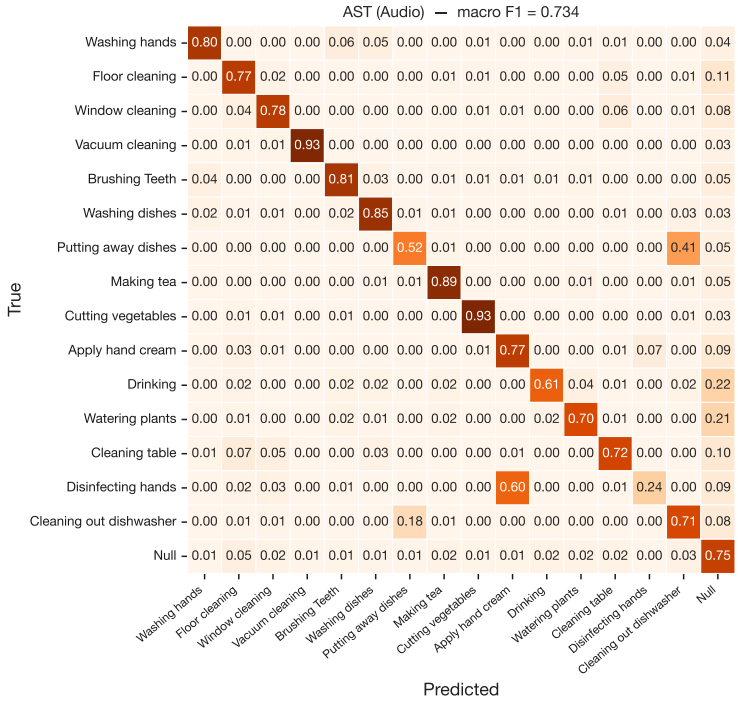


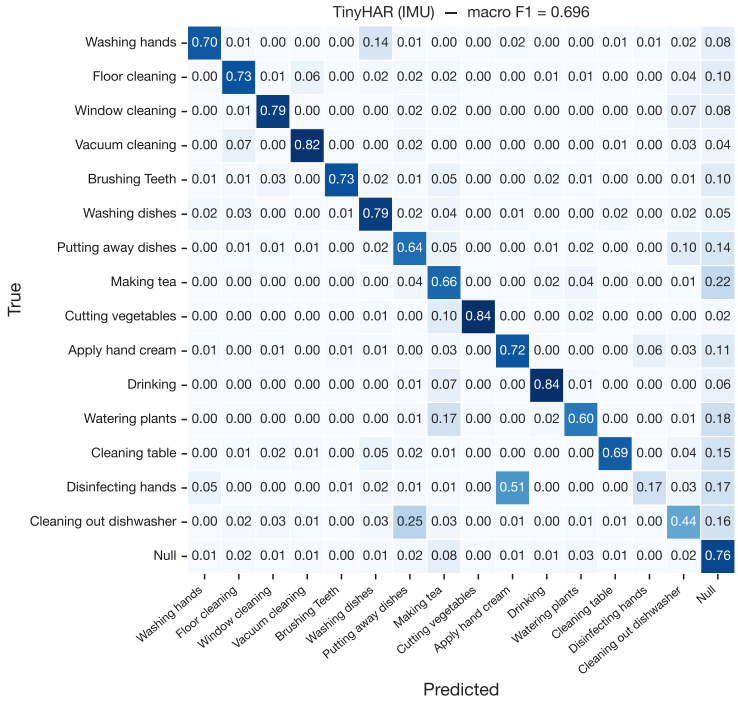
Fig. 15. Decision Fusion

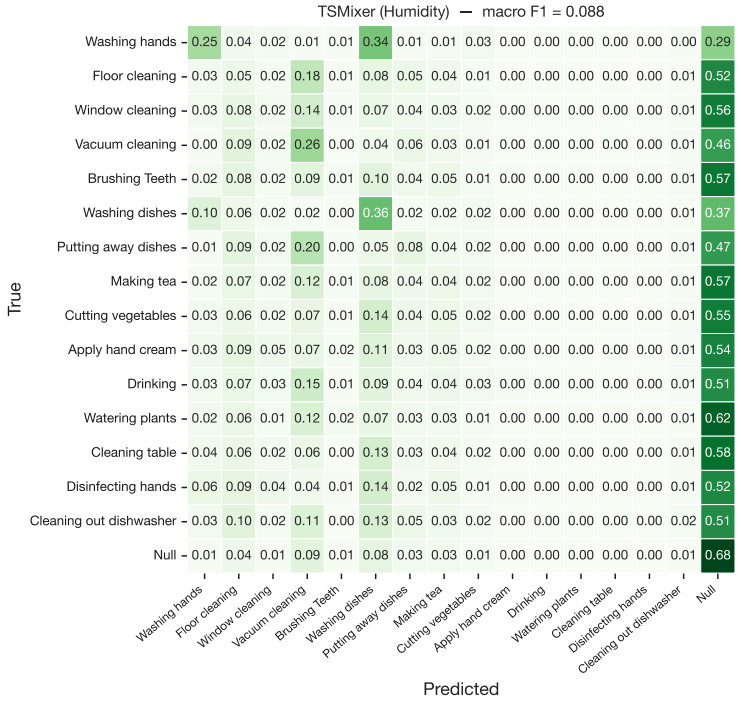
B Additional Machine Learning Results

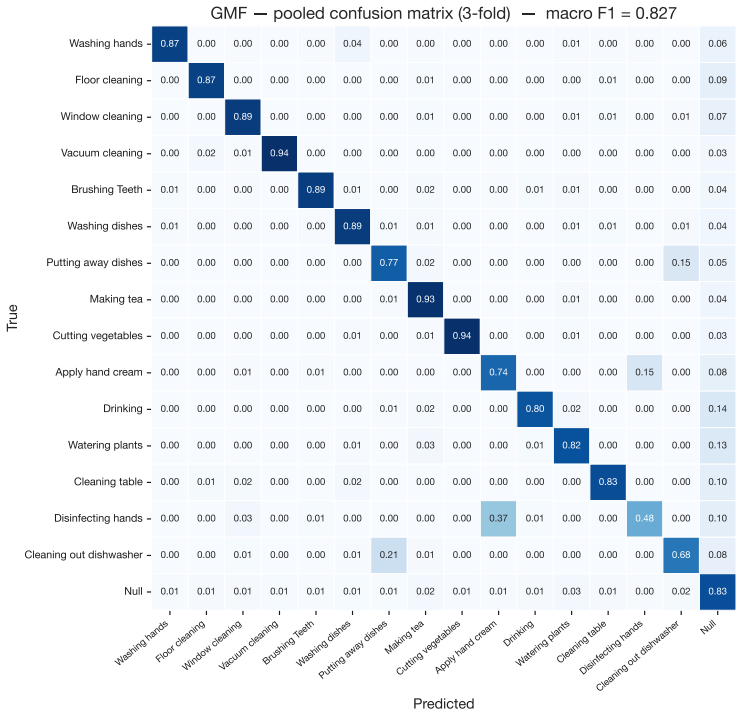
B.1 Confusion Matrices

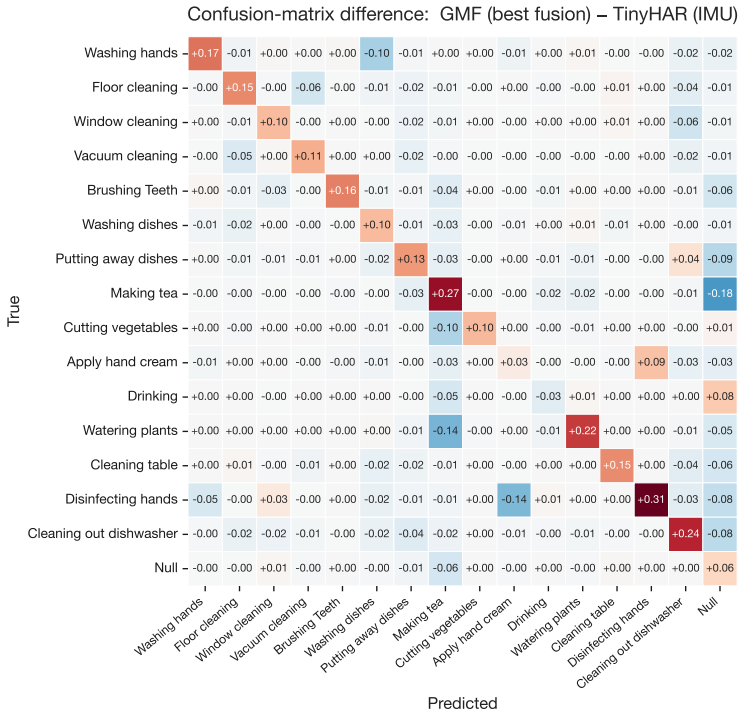
In this section, we show additional confusion matrices. We include one for each unimodal model (AST, TinyHAR, TSMixer), as well as the 3-Fold confusion matrix of the best performing model (GMF), and the confusion difference between TinyHAR and GMF


AST (Audio) — macro F1 = 0.734
True
Predicted
Washing hands 0.80 0.00 0.00 0.00 0.06 0.05 0.00 0.00 0.01 0.00 0.00 0.01 0.01 0.00 0.00 0.04
Floor cleaning 0.00 0.77 0.02 0.00 0.00 0.00 0.00 0.01 0.01 0.00 0.00 0.00 0.05 0.00 0.01 0.11
Window cleaning 0.00 0.04 0.78 0.00 0.00 0.00 0.00 0.00 0.01 0.01 0.00 0.00 0.06 0.00 0.01 0.08
Vacuum cleaning 0.00 0.01 0.01 0.93 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.03
Brushing Teeth 0.04 0.00 0.00 0.00 0.81 0.03 0.00 0.01 0.01 0.01 0.01 0.01 0.00 0.00 0.00 0.05
Washing dishes 0.02 0.01 0.01 0.00 0.02 0.85 0.01 0.01 0.00 0.00 0.00 0.00 0.01 0.00 0.03 0.03
Putting away dishes 0.00 0.00 0.00 0.00 0.00 0.00 0.52 0.01 0.00 0.00 0.00 0.00 0.00 0.00 0.41 0.05
Making tea 0.00 0.00 0.00 0.00 0.00 0.01 0.01 0.89 0.00 0.00 0.00 0.01 0.00 0.00 0.01 0.05
Cutting vegetables 0.00 0.01 0.01 0.00 0.01 0.00 0.00 0.00 0.93 0.00 0.00 0.00 0.00 0.00 0.01 0.03
Apply hand cream 0.00 0.03 0.01 0.00 0.00 0.00 0.00 0.00 0.01 0.77 0.00 0.00 0.01 0.07 0.00 0.09
Drinking 0.00 0.02 0.00 0.00 0.02 0.02 0.00 0.02 0.00 0.00 0.61 0.04 0.01 0.00 0.02 0.22
Watering plants 0.00 0.01 0.00 0.00 0.02 0.01 0.00 0.02 0.00 0.00 0.02 0.70 0.01 0.00 0.00 0.21
Cleaning table 0.01 0.07 0.05 0.00 0.00 0.03 0.00 0.00 0.00 0.01 0.00 0.00 0.72 0.00 0.00 0.10
Disinfecting hands 0.00 0.02 0.03 0.00 0.01 0.00 0.00 0.00 0.00 0.60 0.00 0.00 0.01 0.24 0.00 0.09
Cleaning out dishwasher 0.00 0.01 0.01 0.00 0.00 0.00 0.18 0.01 0.00 0.00 0.00 0.00 0.00 0.00 0.71 0.08
Null 0.01 0.05 0.02 0.01 0.01 0.01 0.01 0.02 0.01 0.01 0.02 0.02 0.02 0.00 0.03 0.75
Washing hands
Floor cleaning
Window cleaning
Vacuum cleaning
Brushing Teeth
Washing dishes
Putting away dishes
Making tea
Cutting vegetables
Apply hand cream
Drinking
Watering plants
Cleaning table
Disinfecting hands
Cleaning out dishwasher
Null


TinyHAR (IMU) — macro F1 = 0.696
True
Predicted
Washing hands 0.70 0.01 0.00 0.00 0.00 0.14 0.01 0.00 0.00 0.02 0.00 0.00 0.01 0.01 0.02 0.08
Floor cleaning 0.00 0.73 0.01 0.06 0.00 0.02 0.02 0.02 0.00 0.00 0.01 0.01 0.00 0.00 0.04 0.10
Window cleaning 0.00 0.01 0.79 0.00 0.00 0.00 0.02 0.02 0.00 0.00 0.00 0.00 0.00 0.00 0.07 0.08
Vacuum cleaning 0.00 0.07 0.00 0.82 0.00 0.00 0.02 0.00 0.00 0.00 0.00 0.00 0.01 0.00 0.03 0.04
Brushing Teeth 0.01 0.01 0.03 0.00 0.73 0.02 0.01 0.05 0.00 0.00 0.02 0.01 0.00 0.00 0.01 0.10
Washing dishes 0.02 0.03 0.00 0.00 0.01 0.79 0.02 0.04 0.00 0.01 0.00 0.00 0.02 0.00 0.02 0.05
Putting away dishes 0.00 0.01 0.01 0.01 0.00 0.02 0.64 0.05 0.00 0.00 0.01 0.02 0.00 0.00 0.10 0.14
Making tea 0.00 0.00 0.00 0.00 0.00 0.00 0.04 0.66 0.00 0.00 0.02 0.04 0.00 0.00 0.01 0.22
Cutting vegetables 0.00 0.00 0.00 0.00 0.00 0.01 0.00 0.10 0.84 0.00 0.00 0.02 0.00 0.00 0.00 0.02
Apply hand cream 0.01 0.00 0.01 0.00 0.01 0.01 0.00 0.03 0.00 0.72 0.00 0.00 0.00 0.06 0.03 0.11
Drinking 0.00 0.00 0.00 0.00 0.00 0.00 0.01 0.07 0.00 0.00 0.84 0.01 0.00 0.00 0.00 0.06
Watering plants 0.00 0.00 0.00 0.00 0.00 0.00 0.01 0.17 0.00 0.00 0.02 0.60 0.00 0.00 0.01 0.18
Cleaning table 0.00 0.01 0.02 0.01 0.00 0.05 0.02 0.01 0.00 0.00 0.00 0.00 0.69 0.00 0.04 0.15
Disinfecting hands 0.05 0.00 0.00 0.00 0.01 0.02 0.01 0.01 0.00 0.51 0.00 0.00 0.00 0.17 0.03 0.17
Cleaning out dishwasher 0.00 0.02 0.03 0.01 0.00 0.03 0.25 0.03 0.00 0.01 0.00 0.01 0.01 0.00 0.44 0.16
Null 0.01 0.02 0.01 0.01 0.00 0.01 0.02 0.08 0.00 0.01 0.01 0.03 0.01 0.00 0.02 0.76
Washing hands
Floor cleaning
Window cleaning
Vacuum cleaning
Brushing Teeth
Washing dishes
Putting away dishes
Making tea
Cutting vegetables
Apply hand cream
Drinking
Watering plants
Cleaning table
Disinfecting hands
Cleaning out dishwasher
Null


TSMixer (Humidity) — macro F1 = 0.088
True
Predicted
Washing hands 0.25 0.04 0.02 0.01 0.01 0.34 0.01 0.01 0.03 0.00 0.00 0.00 0.00 0.00 0.00 0.29
Floor cleaning 0.03 0.05 0.02 0.18 0.01 0.08 0.05 0.04 0.01 0.00 0.00 0.00 0.00 0.00 0.01 0.52
Window cleaning 0.03 0.08 0.02 0.14 0.01 0.07 0.04 0.03 0.02 0.00 0.00 0.00 0.00 0.00 0.01 0.56
Vacuum cleaning 0.00 0.09 0.02 0.26 0.00 0.04 0.06 0.03 0.01 0.00 0.00 0.00 0.00 0.00 0.01 0.46
Brushing Teeth 0.02 0.08 0.02 0.09 0.01 0.10 0.04 0.05 0.01 0.00 0.00 0.00 0.00 0.00 0.01 0.57
Washing dishes 0.10 0.06 0.02 0.02 0.00 0.36 0.02 0.02 0.02 0.00 0.00 0.00 0.00 0.00 0.01 0.37
Putting away dishes 0.01 0.09 0.02 0.20 0.00 0.05 0.08 0.04 0.02 0.00 0.00 0.00 0.00 0.00 0.01 0.47
Making tea 0.02 0.07 0.02 0.12 0.01 0.08 0.04 0.04 0.02 0.00 0.00 0.00 0.00 0.00 0.01 0.57
Cutting vegetables 0.03 0.06 0.02 0.07 0.01 0.14 0.04 0.05 0.02 0.00 0.00 0.00 0.00 0.00 0.01 0.55
Apply hand cream 0.03 0.09 0.05 0.07 0.02 0.11 0.03 0.05 0.02 0.00 0.00 0.00 0.00 0.00 0.01 0.54
Drinking 0.03 0.07 0.03 0.15 0.01 0.09 0.04 0.04 0.03 0.00 0.00 0.00 0.00 0.00 0.01 0.51
Watering plants 0.02 0.06 0.01 0.12 0.02 0.07 0.03 0.03 0.01 0.00 0.00 0.00 0.00 0.00 0.01 0.62
Cleaning table 0.04 0.06 0.02 0.06 0.00 0.13 0.03 0.04 0.02 0.00 0.00 0.00 0.00 0.00 0.01 0.58
Disinfecting hands 0.06 0.09 0.04 0.04 0.01 0.14 0.02 0.05 0.01 0.00 0.00 0.00 0.00 0.00 0.01 0.52
Cleaning out dishwasher 0.03 0.10 0.02 0.11 0.00 0.13 0.05 0.03 0.02 0.00 0.00 0.00 0.00 0.00 0.02 0.51
Null 0.01 0.04 0.01 0.09 0.01 0.08 0.03 0.03 0.01 0.00 0.00 0.00 0.00 0.00 0.01 0.68
Washing hands
Floor cleaning
Window cleaning
Vacuum cleaning
Brushing Teeth
Washing dishes
Putting away dishes
Making tea
Cutting vegetables
Apply hand cream
Drinking
Watering plants
Cleaning table
Disinfecting hands
Cleaning out dishwasher
Null


GMF — pooled confusion matrix (3-fold) — macro F1 = 0.827
True
Predicted
Washing hands 0.87 0.00 0.00 0.00 0.00 0.04 0.00 0.00 0.00 0.00 0.00 0.01 0.00 0.00 0.00 0.06
Floor cleaning 0.00 0.87 0.00 0.00 0.00 0.00 0.00 0.01 0.00 0.00 0.00 0.00 0.01 0.00 0.00 0.09
Window cleaning 0.00 0.00 0.89 0.00 0.00 0.00 0.00 0.01 0.00 0.00 0.00 0.01 0.01 0.00 0.01 0.07
Vacuum cleaning 0.00 0.02 0.01 0.94 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.03
Brushing Teeth 0.01 0.00 0.00 0.00 0.89 0.01 0.00 0.02 0.00 0.00 0.01 0.01 0.00 0.00 0.00 0.04
Washing dishes 0.01 0.00 0.00 0.00 0.00 0.89 0.01 0.01 0.00 0.00 0.00 0.01 0.01 0.00 0.01 0.04
Putting away dishes 0.00 0.00 0.00 0.00 0.00 0.00 0.77 0.02 0.00 0.00 0.00 0.00 0.00 0.00 0.15 0.05
Making tea 0.00 0.00 0.00 0.00 0.00 0.00 0.01 0.93 0.00 0.00 0.00 0.01 0.00 0.00 0.00 0.04
Cutting vegetables 0.00 0.00 0.00 0.00 0.00 0.01 0.00 0.01 0.94 0.00 0.00 0.01 0.00 0.00 0.00 0.03
Apply hand cream 0.00 0.00 0.01 0.00 0.01 0.00 0.00 0.00 0.00 0.74 0.00 0.00 0.00 0.15 0.00 0.08
Drinking 0.00 0.00 0.00 0.00 0.00 0.00 0.01 0.02 0.00 0.00 0.80 0.02 0.00 0.00 0.00 0.14
Watering plants 0.00 0.00 0.00 0.00 0.00 0.01 0.00 0.03 0.00 0.00 0.01 0.82 0.00 0.00 0.00 0.13
Cleaning table 0.00 0.01 0.02 0.00 0.00 0.02 0.00 0.00 0.00 0.00 0.00 0.00 0.83 0.00 0.00 0.10
Disinfecting hands 0.00 0.00 0.03 0.00 0.01 0.00 0.00 0.00 0.00 0.37 0.01 0.00 0.00 0.48 0.00 0.10
Cleaning out dishwasher 0.00 0.00 0.01 0.00 0.00 0.01 0.21 0.01 0.00 0.00 0.00 0.00 0.00 0.00 0.68 0.08
Null 0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.02 0.01 0.01 0.01 0.03 0.01 0.00 0.02 0.83
Washing hands
Floor cleaning
Window cleaning
Vacuum cleaning
Brushing Teeth
Washing dishes
Putting away dishes
Making tea
Cutting vegetables
Apply hand cream
Drinking
Watering plants
Cleaning table
Disinfecting hands
Cleaning out dishwasher
Null


Confusion-matrix difference: GMF (best fusion) – TinyHAR (IMU)
True
Predicted
Washing hands
Floor cleaning
Window cleaning
Vacuum cleaning
Brushing Teeth
Washing dishes
Putting away dishes
Making tea
Cutting vegetables
Apply hand cream
Drinking
Watering plants
Cleaning table
Disinfecting hands
Cleaning out dishwasher
Null