WEARABLEBRIDGE: TASK-GUIDED LATENT TRANSPORT FOR FREE-LIVING ACTIVITY RECOGNITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Wrist-worn accelerometers make activity monitoring feasible at scale, but incidental arm motion, device orientation, and shifting context blur the signal that separates one activity from another. Calibration studies supply higher-quality reference measurements such as concurrent thigh recordings or supervised laboratory wrist sessions, but they are expensive, cover few participants, and are rarely available at deployment. How can limited reference data improve recognition for participants who wear only a wrist device? We propose WearableBridge, an algorithm that learns a task-guided transport map from the wrist-feature distribution toward the reference-feature distribution. WearableBridge holds the reference encoder and classifier fixed and learns a map, the bridge, that carries each wrist feature into the region the classifier associates with its activity. Reference features or the reference classifier's predictions indicate the destination, and a classification loss penalizes wrist features that land in the region of another activity. We instantiate WearableBridge with three bridges that differ in how the reference guides the map: Mapper regresses onto reference features, Distillation matches the reference classifier's class probabilities, and MeanFlow learns a one-step flow from wrist toward reference features. On 40 held-out free-living PAAWS participants, all three bridges improve macro-F1 in every training run, and the strongest, MeanFlow, raises mean macro-F1 from 35.1% to 56.0% with concurrent thigh references (Thigh2Wrist), and from 54.0% to 63.5% with group-matched laboratory wrist references (Lab2Freeliving). Fitted only on PAAWS, MeanFlow also raises macro-F1 on the CAPTURE-24 cohort from 42.8% to 54.7% without using its labels. The gains replicate on SELFback, hold across four reference encoders, and grow further with temporal context. Reference recordings from a calibration subset can therefore improve wrist-only recognition for new participants while the reference model stays unchanged.

1 INTRODUCTION

Wrist-worn accelerometers can record activity over many days with a single device, which has made population studies with more than 100,000 participants feasible (Doherty et al., 2017). The same placement makes the signal hard to interpret: a wrist recording mixes arm gestures, changes in device orientation, transitions, and daily context with the movement that distinguishes cycling, sitting, standing, and walking. These sources of variation are strongest in free-living (FL) recordings, collected as participants follow their daily routines without scripted activities or supervision. Classifiers trained under controlled conditions therefore lose accuracy on FL recordings (Gyllensten & Bonomi, 2011; Sasaki et al., 2016; Pavey et al., 2017), which is where wrist-worn monitoring is deployed.

Other recordings reveal activity more clearly. A thigh sensor gives a direct view of lower-body movement, and a supervised laboratory wrist session holds recording conditions steady and supplies observed labels. We call these *higher-quality references*, meaning that they carry more information about activity. Both are expensive at scale: a thigh recording requires a second device and participants' compliance, a laboratory session requires staff time and visits, and FL labels can require prolonged observation or video review (Sasaki et al., 2016; Potter et al., 2025). A realistic study

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.

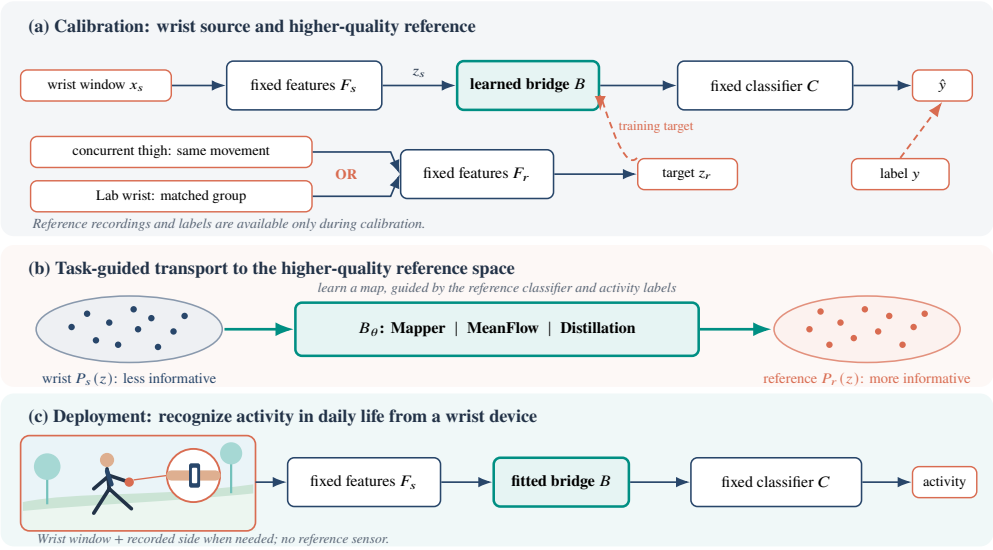

Figure 1: WearableBridge overview. (a) Calibration pairs each wrist window with a higher-quality reference: a concurrent thigh window (Thigh2Wrist) or a Lab wrist window matched by participant, activity, side, and placement (Lab2Freeliving). (b) A bridge (Mapper, MeanFlow, or Distillation) maps wrist features, with distribution P_s, toward the reference feature distribution P_r, with classification loss through the fixed classifier. (c) Deployment uses only a free-living wrist window and its recorded side.

therefore collects references for a calibration subset and deploys wrist-only monitoring to everyone else. We ask: *how can limited reference data improve wrist-only activity recognition for new participants who have no reference recording?*

Two calibration designs make this question concrete, each named by the direction in which activity information transfers. In Thigh2Wrist, a wrist window and a thigh window record the same movement at the same time, so each wrist window has its own reference, an instance-level target. In Lab2Freeliving, the wrist windows come from simulated free-living (SimFL) sessions, which imitate everyday conditions, and the references are laboratory (Lab) wrist windows from a separate session. Matching by participant, activity, wrist side, and device placement gives a group-level target. In both designs, the deployed input is a ten-second FL wrist window with its recorded side, and neither its label nor a reference is available.

Existing approaches either use the reference at prediction time, as in sensor fusion (Ordóñez & Roggen, 2016), or learn a new wrist model through feature adaptation across placements (Chang et al., 2020; Chen et al., 2019), a wrist-only student (Hinton et al., 2015; Lopez-Paz et al., 2016), or fine-tuning on labeled wrist data. Fusion needs the reference sensor at deployment, and the other approaches replace the classifier that the reference recordings trained. We propose WearableBridge, which reuses both the reference encoder and this classifier. Both stay fixed, and a small learned map with 0.13 to 2.4 million parameters, which we call the bridge, transforms wrist features within the reference feature space. In this space, reference features and the reference classifier's predictions serve directly as training targets, and the identity map reproduces the unadapted reference model.

The key observation is that a classifier trained on reference recordings defines regions of feature space associated with each activity, and those regions remain meaningful when no reference accompanies a deployed window. WearableBridge therefore trains the bridge as a task-guided transport map that carries each wrist feature into the region of its activity. During calibration, reference features or predictions indicate where wrist features should move, and a classification loss through the frozen classifier penalizes wrist features that land in the region of another activity. At deployment, the fixed classifier reads the mapped FL wrist feature, so neither a reference recording nor a label is needed (Figure 1). We compare three forms of reference guidance, instantiated as three bridges: Mapper matches reference features, Distillation matches the reference classifier's class probabilities, and MeanFlow learns a one-step flow from wrist toward reference features (Geng et al., 2025; 2026).

On the Physical Activity Assessment Using Wearable Sensors (PAAWS) dataset (Potter et al., 2025), with 40 held-out FL participants and more than 1.5 million wrist windows per design, every bridge improves macro-F1 in every run of both designs. The strongest, MeanFlow, raises macro-F1 from 35.07% to 55.98% in Thigh2Wrist and from 54.00% to 63.54% in Lab2Freeliving. Class-level results show that each reference transfers the distinctions it observes best: thigh posture separates sitting from standing, and laboratory wrist sessions separate cycling from walking. Fitted only on PAAWS, MeanFlow also raises macro-F1 on the CAPTURE-24 cohort from 42.75% to 54.73% without using its labels. The gains replicate on SELFback, hold across four reference encoders, and grow further with temporal context.

Our contributions are:

- **A reference-reuse formulation** of wrist recognition with limited reference data, as learning a map from wrist features into the feature space of a fixed classifier trained on reference recordings, covering instance-level (Thigh2Wrist) and group-level (Lab2Freeliving) references.
- **WearableBridge**, one objective that combines a classification loss through the frozen classifier with a reference-guidance term, instantiated as feature matching (Mapper), class-probability matching (Distillation), or a one-step flow (MeanFlow).
- **Consistent gains for participants without references** on PAAWS with four reference encoders, on SELFback, and on CAPTURE-24, and comparisons showing that reusing the reference classifier reaches similar or higher macro-F1 than retraining it on the same wrist labels with the encoder frozen.

2 Related Work

Transfer across positions and protocols. Domain adaptation aligns feature distributions across domains (Ganin et al., 2016; Courty et al., 2017), including wearable devices and placements (Chang et al., 2020; Chen et al., 2019), and ColloSSL learns representations from simultaneous unlabeled recordings across worn devices (Jain et al., 2022). These methods train a new encoder or classifier for the target placement, which replaces the classifier fitted on the more informative placement together with the activity boundaries it learned. WearableBridge keeps that classifier and its encoder fixed and learns where wrist features should land among its activity regions.

Privileged information and cross-modal distillation. Knowledge distillation and learning with privileged information train a student from a teacher that sees more at training time (Hinton et al., 2015; Lopez-Paz et al., 2016), including a teacher on a paired modality (Gupta et al., 2016). The student is a new model, and each training input needs a paired teacher input. WearableBridge trains only a map into the teacher's own classifier and also learns from group-matched references recorded in separate sessions, as in Lab2Freeliving.

Adapting frozen models. Model reprogramming learns an input transformation that repurposes a frozen network for a new task (Elsayed et al., 2019), adapters train small modules inside a fixed model (Houlsby et al., 2019), and test-time adaptation updates normalization parameters from unlabeled test data (Wang et al., 2021). These methods adapt a fixed model from task labels or unlabeled inputs alone, with no target for where a new input should go. WearableBridge supplies that target: paired or group-matched reference features and predictions indicate where each wrist feature should land in the frozen classifier's feature space. The activity-only bridge in Section 4.5 is the label-only counterpart in feature space, and reference guidance improves on it on PAAWS.

Flow-based transport. Flow matching and rectified flow learn velocity fields that carry one distribution to another (Lipman et al., 2023; Liu et al., 2023b), bridge models learn maps between paired domains (Liu et al., 2023a), and MeanFlow learns average velocities for one-step generation (Geng et al., 2025; 2026). These models are trained to reproduce a target distribution. For recognition, what matters is that each transported wrist feature lands in the decision region of its activity. WearableBridge conditions MeanFlow on the wrist feature and adds a classification loss through the frozen classifier, so the transport is guided by the task and takes one network evaluation at deployment.

3 METHOD

3.1 CALIBRATION AND WRIST-ONLY PREDICTION

WearableBridge learns a transformation of wrist features for a frozen reference classifier. We call the wrist input the source, marked by subscript s, and the higher-quality recording the reference, marked by subscript r. Let x_s be a ten-second wrist accelerometer window, $y \in \mathcal{Y}$ its label in the activity set $\mathcal{Y}$, and x_r an associated reference window. In Thigh2Wrist, x_r is the thigh window recorded simultaneously with x_s. In Lab2Freeliving, x_s is a SimFL wrist window and x_r a Lab wrist window of the same participant, activity, and wrist side, and both sessions use the device worn in the top position on the wrist.

Let F_s and F_r be fixed source and reference feature maps into $\mathbb{R}^d$, and let $C : \mathbb{R}^d \to \mathbb{R}^{|\mathcal{Y}|}$ be a classifier fitted on labeled reference features and then frozen. The two maps share an encoder within each comparison, and every wrist window first has its first axis reversed if it comes from a right wrist (Appendix A).

We standardize features with the coordinatewise mean μ and population standard deviation σ (floored at 10^{-4}) of training reference features, counting each reference window once. For a raw feature $h \in \mathbb{R}^d$, write $S(h) = (h-\mu) \oslash \sigma$ and its inverse $D(z) = \mu + \sigma \odot z$, with coordinatewise $\oslash$ and $\odot$, and define the standardized wrist and reference features $z_s = S(F_s(x_s))$ and $z_r = S(F_r(x_r))$. The bridge $B_\theta : \mathbb{R}^d \to \mathbb{R}^d$, with parameters θ, maps z_s within this space, and the classifier produces logits $a \in \mathbb{R}^{|\mathcal{Y}|}$ and a prediction $\hat{y}$ through

$$a = C(D(B_\theta(z_s))), \qquad \hat{y} = \arg\max_{k \in \mathcal{Y}} a_k. \tag{1}$$

With the reference model fixed, only the bridge is fitted, and training passes gradients through the frozen classifier without changing it. Setting $B_\theta(z) = z$ gives the unadapted reference, our baseline. Every bridge is initialized at this identity map by zero-initializing its output layer, so training begins from the baseline predictions. Bridge fitting uses no data from held-out FL participants.

Attainable predictions. A natural question is whether freezing C limits what the bridge can predict. It does not. Take any classifier $C_s : \mathbb{R}^d \to \mathbb{R}^{|\mathcal{Y}|}$ that could be trained directly on standardized wrist features. When $C(h) = Wh + c$ is linear with $W \in \mathbb{R}^{|\mathcal{Y}| \times d}$ of full row rank, as in Thigh2Wrist, the bridge $B(z) = S\{W^+(C_s(z) - c)\}$, with W^+ the Moore–Penrose pseudoinverse, reproduces the logits of C_s exactly: $C(D(B(z))) = WW^+(C_s(z) - c) + c = C_s(z)$ because $WW^+ = I$. When C is nonlinear, as in Lab2Freeliving, a bridge reproduces the predicted labels of C_s by sending each wrist feature into the decision region of the label that C_s assigns, which is possible whenever every activity has a nonempty decision region. A sufficiently expressive bridge can therefore match any wrist classifier, and freezing C serves to fix the feature space in which the bridge operates.

3.2 REFERENCE-GUIDED BRIDGE OBJECTIVES

Let Π denote the calibration distribution of (x_s, x_r, y): simultaneous windows in Thigh2Wrist, and pairs drawn within participant–activity–side groups g in Lab2Freeliving. Every bridge minimizes a weighted empirical version of

$$\mathcal{L}(\theta) = \mathbb{E}_\Pi[\mathrm{CE}(a, y) + \rho_\theta(z_s, z_r)], \qquad \mathrm{CE}(a, y) = -\log \mathrm{softmax}(a)_y, \tag{2}$$

with a from (1). The cross-entropy term passes gradients through the frozen classifier and moves each wrist feature toward the decision region $\{z : \arg\max_k C(D(z))_k = y\}$ of its activity. The classifier learned these regions from reference features, and the reference-guidance term ρ_θ indicates where within them a wrist feature should land. The three bridges specify ρ_θ through reference features (Mapper), reference class probabilities (Distillation), and a path between paired features (MeanFlow). Setting $\rho_\theta \equiv 0$ gives the activity-only bridge compared in Section 4.5. Mapper and Distillation use the same residual multilayer perceptron.

Mapper. The reference feature supplies an endpoint target:

$$\ell_{\mathrm{map}} = \frac{1}{d}\|B_\theta(z_s) - z_r\|_2^2 + \mathrm{CE}(a, y). \tag{3}$$

In Thigh2Wrist, z_r is the concurrent thigh feature. In Lab2Freeliving, the Lab window paired with each SimFL window is redrawn within its group g at every epoch, so z_r is approximately independent of z_s given g. Because $B_\theta(z_s)$ depends only on z_s, the expected squared error then equals $\mathbb{E}\|B_\theta(z_s) - m(z_s)\|_2^2$ plus a term that does not depend on θ, with $m(z_s) = \mathbb{E}[\mathbb{E}(z_r \mid g) \mid z_s]$. Mapper therefore moves each wrist feature toward reference group means, which act as activity prototypes.

Distillation. The reference classifier supplies a probability target. At temperature $T = 2$, let $q_r = \mathrm{softmax}(C(D(z_r))/T)$ and $q_s = \mathrm{softmax}(a/T)$. The loss is

$$\ell_{\mathrm{dist}} = T^2 \,\mathrm{KL}(q_r \,\|\, q_s) + \mathrm{CE}(a, y), \tag{4}$$

where KL denotes Kullback–Leibler divergence. This objective matches reference predictions and lets the mapped wrist feature differ from the reference feature whenever their class probabilities agree. Under group matching, the expected loss depends on q_r only through $\mathbb{E}[q_r \mid z_s]$, a group-averaged soft label.

Source-conditioned MeanFlow. We adapt the MeanFlow formulation (Geng et al., 2025; 2026) to learn a one-step update conditioned on the wrist feature. For each calibration pair, define a straight path from the reference feature at time zero to the wrist feature at time one, with constant velocity v:

$$z(\tau) = (1-\tau)z_r + \tau z_s, \qquad v = z_s - z_r. \tag{5}$$

For $0 \le t_0 \le t \le 1$, the network $u_\theta(z(t), t_0, t \mid z_s)$ predicts the average velocity over the interval $[t_0, t]$, so that an exact prediction gives $z(t_0) = z(t) - (t - t_0)\, u_\theta$. Conditioning on z_s gives the network the original wrist feature at every point on the path. At deployment, one network evaluation carries the wrist feature from time one back to time zero:

$$B_\theta(z_s) = z_s - u_\theta(z_s, 0, 1 \mid z_s). \tag{6}$$

Following improved MeanFlow (Geng et al., 2026), training uses the identity that links average and instantaneous velocity: for the exact average velocity u, the sum $u + (t - t_0)\,\frac{d}{dt}u$ equals the instantaneous velocity, which is v on the path (5). Write $u = u_\theta(z(t), t_0, t \mid z_s)$ and $u_{\mathrm{inst}} = u_\theta(z(t), t, t \mid z_s)$ for the predicted average and instantaneous velocities. With t_0 and z_s fixed, define

$$V_\theta = u + (t - t_0)\,\mathrm{sg}[\partial_t u_\theta + (\nabla_z u_\theta) u_{\mathrm{inst}}], \tag{7}$$

$$e = \|V_\theta - v\|_2^2, \qquad \ell_{\mathrm{path}} = \frac{e}{\mathrm{sg}(e) + \epsilon}. \tag{8}$$

The bracket is the time derivative of u_θ along the path, with u_{inst} in place of the path velocity. The stop-gradient operator sg blocks gradients through this correction and through the adaptive denominator, and $\epsilon > 0$ stabilizes the denominator. Training also includes $t_0 = t$ examples, for which $V_\theta = u_{\mathrm{inst}}$ is regressed directly onto v. Appendix B gives the time sampler and ϵ settings.

The complete objective adds source preservation and activity supervision at the one-step output:

$$\ell_{\mathrm{flow}} = \ell_{\mathrm{path}} + \frac{0.01}{d}\|B_\theta(z_s) - z_s\|_2^2 + \mathrm{CE}(a, y). \tag{9}$$

The source-preservation term keeps updates small, and the cross-entropy term trains the one-step update for the fixed classifier. In the notation of (2), ρ_θ is the scaled squared error for Mapper, the scaled KL divergence for Distillation, and the path loss ℓ_{path}, averaged over sampled times, plus source preservation for MeanFlow.

Each method minimizes a batch mean of weighted per-pair losses (Appendices B and C).

4 Experiments

4.1 Experimental setup

Data. PAAWS calibration uses Lab and SimFL recordings from 170 training and 43 validation participants, and FL testing uses 40 held-out participants (Potter et al., 2025). Models classify cycling, sitting, standing, and walking from ten-second PAAWS wrist windows at 50 Hz. Thigh2Wrist

Table 1: PAAWS current-window results with frozen MoCA references, in percent (mean ± sample SD over three runs). The unadapted reference is the baseline: it applies the frozen reference classifier directly to wrist features. Mapper, MeanFlow, and Distillation are the WearableBridge variants, which map wrist features before the same classifier. Measured thigh, listed below the last rule, applies the same classifier to the concurrent thigh window, so it uses the reference sensor at test time and shows the accuracy of the reference itself. Both designs test the same 40 participants, and bold marks the highest wrist-input mean per column.

(a) Thigh2Wrist. 1,588,369 test windows.

	Overall			Class F1			
Method	Macro-F1	Bal. acc.	Accuracy	Cycling	Sitting	Standing	Walking
Unadapted reference (baseline)	35.07±0.21	42.45±0.13	51.80±0.30	2.77±0.03	70.27±0.53	18.13±0.24	49.12±1.11
Mapper	54.13±0.54	67.00±1.91	73.83±1.93	17.14±3.81	83.82±1.72	35.69±0.48	79.87±0.40
MeanFlow	**55.98**±0.51	**69.87**±0.73	**76.11**±1.64	**19.01**±2.56	**85.66**±1.41	**38.07**±0.50	**81.17**±0.11
Distillation	55.06±0.32	66.35±0.89	75.93±1.85	17.84±2.43	85.52±1.55	36.78±0.16	80.12±0.48
Measured thigh (reference sensor)	82.50±0.33	89.62±0.10	94.97±0.01	61.81±1.45	97.97±0.02	82.87±0.12	87.37±0.08

(b) Lab2Freeliving. 1,595,771 test windows. Lab references come from separate laboratory sessions and have no FL counterpart to score.

	Overall			Class F1			
Method	Macro-F1	Bal. acc.	Accuracy	Cycling	Sitting	Standing	Walking
Unadapted reference (baseline)	54.00±1.29	62.34±1.93	78.00±2.18	9.42±2.76	87.49±1.64	**41.25**±1.15	77.85±1.37
Mapper	60.90±0.57	65.88±1.15	75.63±2.01	43.60±0.72	85.22±1.61	34.48±1.10	80.28±0.39
MeanFlow	**63.54**±1.23	**67.93**±0.94	**80.02**±1.47	**47.21**±2.21	**88.58**±1.06	38.09±1.16	**80.29**±0.79
Distillation	62.30±1.26	66.80±0.89	79.05±1.55	44.77±2.86	87.91±1.13	36.30±1.00	80.20±0.52

uses concurrent wrist and thigh windows from both sessions. SELFback covers nine activities with 20 training, six validation, and seven test participants (UCI Machine Learning Repository, 2020). CAPTURE-24 supplies 134,658 windows from 50 participants across the four PAAWS activities (Chan et al., 2024), and its labels are used only for evaluation (Appendices A, E, and F).

Models. The primary encoder is the Multi-modal Cross-masked Autoencoder (MoCA) (Ryu et al., 2025), a masked-autoencoder model for multichannel wearable time series with public pretrained weights. We use it because its features are learned without activity labels and because one frozen network embeds wrist, thigh, and Lab windows in a shared space, as our formulation assumes. It supplies features for a linear thigh classifier in Thigh2Wrist and for a Lab-trained projection and classifier in Lab2Freeliving (Appendix B). To test whether gains depend on MoCA, we repeat both designs with DeepConvLSTM, InceptionTime, and HART reference encoders trained from scratch on PAAWS training participants (Ordóñez & Roggen, 2016; Ismail Fawaz et al., 2020; Ek et al., 2023). For CAPTURE-24, we use frozen HARNet10 features because they were pretrained on free-living wrist accelerometry from UK Biobank (Yuan et al., 2024).

Comparators. Besides the unadapted reference, we compare MeanFlow with three models trained on the same labeled calibration wrist windows. The direct wrist classifier updates the reference classifier with the encoder frozen, joint fine-tuning also updates the encoder (PAAWS Thigh2Wrist only), and the activity-only bridge trains the MeanFlow architecture with cross-entropy alone ($\rho_\theta \equiv 0$ in (2)).

Evaluation. Macro-F1 is the primary metric because sitting accounts for 79% of PAAWS FL windows. Balanced accuracy, accuracy, and class F1 are secondary, and differences are in percentage points (pp), computed before rounding. We report the mean and sample SD over three runs with different random seeds and a fixed participant split, except for the single-run PAAWS comparisons (Table 3). Paired participant-bootstrap 95% intervals for differences resample participants jointly for both methods (Appendix C). Predictions use the current window unless stated otherwise.

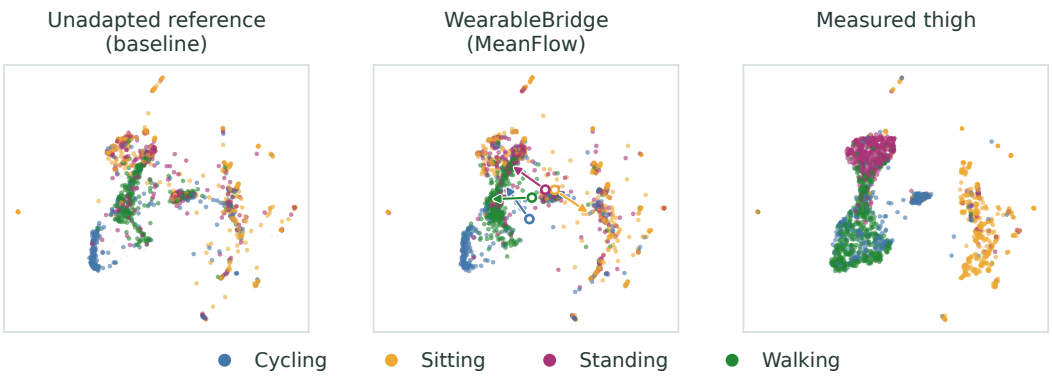

Figure 2: Thigh2Wrist features before and after MeanFlow, with concurrent measured thighs. A fixed UMAP (McInnes et al., 2018) fitted on SimFL-validation thigh features projects 2,048 held-out FL wrist–thigh pairs, sampled equally by activity and wrist side, with validation references in faint gray; class means are computed over 8,192 pairs (Appendix C). Open circles mark each activity's mean wrist input, and arrows point toward its mean output, with lengths scaled for readability (Appendix C).

4.2 Bridging improves wrist-only recognition in both designs

Every bridge improves macro-F1 over the unadapted-reference baseline in every run of both calibration designs (Table 1). MeanFlow has the highest mean macro-F1, balanced accuracy, and accuracy in each design.

In Thigh2Wrist, MeanFlow raises macro-F1 from 35.07% to 55.98%, a gain of 20.91 percentage points (95% participant interval, 17.31 to 24.61). This closes 44% of the gap to the measured thigh (82.50%), which applies the same classifier to the concurrent thigh window. The gain extends to all four activities, with walking F1 rising from 49.12% to 81.17% and standing F1 from 18.13% to 38.07%. Cycling F1, the rarest class, rises nearly sevenfold, from 2.77% to 19.01%. In Figure 2, MeanFlow moves the mean wrist features of sitting, standing, and walking toward the regions of the corresponding thigh features. The thigh space separates activities mainly by leg posture. Sitting and standing, which the wrist often confuses because arm movements are similar in both postures, lie far apart there, and sitting F1 rises from 70.27% to 85.66%. Standing and walking are both upright and differ mainly in thigh motion, which also varies while standing, so they lie close together and account for most remaining walking errors (Appendix D).

In Lab2Freeliving, MeanFlow raises macro-F1 from 54.00% to 63.54%, a gain of 9.54 percentage points from group-matched Lab references. Most of the gain comes from cycling, whose F1 rises from 9.42% to 47.21%, and walking F1 also improves. In Figure 3, all three bridges move each activity's mean wrist feature in similar directions. Cycling windows, which the unadapted reference places near walking, move toward the Lab cycling region, consistent with the group-mean target in Section 3, and the share of cycling windows predicted as walking falls from 46% to 12% with MeanFlow (Appendix D). This pattern is consistent with Lab wrist sessions recording arm posture and motion cleanly, which separates cycling, with the hands on the handlebars, from walking with its arm swing. Sitting and standing look alike at the wrist, so the Lab reference offers little to separate them, and standing F1 dips from 41.25% to 38.09% with MeanFlow. Each reference thus transfers the distinctions it observes best.

4.3 Gains hold across reference encoders and with temporal context

Reference encoders. The gains hold with DeepConvLSTM, InceptionTime, and HART reference encoders trained from scratch (Appendix Table 7). Every bridge improves mean macro-F1 over its unadapted-reference baseline by 9.62 to 20.15 percentage points in Thigh2Wrist, with improvements in every run, and by 8.13 to 12.76 in Lab2Freeliving. In Lab2Freeliving, the three bridges are within 0.7 percentage points of each other for every encoder. These encoders are trained only on PAAWS training participants, so the gains hold without a pretrained encoder.

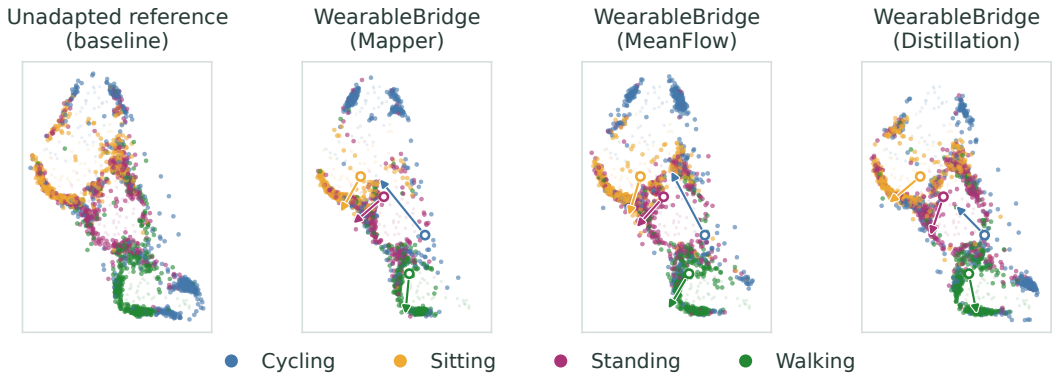

Figure 3: Lab2Freeliving features before and after each bridge. A fixed UMAP (McInnes et al., 2018) fitted on Lab-validation references (faint points) projects 2,048 held-out FL wrist windows, sampled equally by activity and wrist side, and their mapped outputs. Circles and arrows are as in Figure 2. Supplementary animations of both figures interpolate between input and output positions.

Temporal context. Averaging class probabilities over the current window and up to four preceding windows (five-window averaging) further improves every bridge in both designs, adding 5.26 to 6.15 percentage points in macro-F1 (Appendix Table 9). MeanFlow reaches 61.70% in Thigh2Wrist and 68.90% in Lab2Freeliving. In Thigh2Wrist, five-window averaging adds 0.46 percentage points without a bridge and 5.72 with MeanFlow, so bridging and temporal context are complementary.

4.4 Gains replicate on SELFback and transfer to CAPTURE-24

On SELFback, with a fixed nine-class thigh classifier on frozen MoCA features, Thigh2Wrist calibration again improves recognition for the seven test participants, whose recordings follow structured sessions (Table 2). MeanFlow raises macro-F1 from 25.73% to 66.74%, a gain of 41.01 percentage points (95% participant interval, 31.70 to 50.59), and all three bridges exceed 65%.

CAPTURE-24 tests transfer to a new cohort. MeanFlow is fitted on PAAWS in the Lab2Freeliving design, with the HARNet10 encoder and PAAWS Lab classifier fixed, and applied to CAPTURE-24 without target labels or refitting. It raises macro-F1 from 42.75% to 54.73%, a gain of 11.98 percentage points (95% participant interval, 8.33 to 14.60), and cycling F1 rises from 5.38% to 50.64% (Appendix Table 6).

Table 2: SELFback replication in the Thigh2Wrist design, in percent (mean ± sample SD over three runs), on 963 concurrent wrist–thigh test windows from seven participants, with nine activities and frozen MoCA. The unadapted reference is the baseline. The activity-only bridge trains the MeanFlow architecture with cross-entropy alone, and the direct wrist classifier updates the reference classifier on the same labeled wrist windows (§4.5). Measured thigh, listed below the last rule, applies the thigh classifier to the concurrent thigh window. Bold marks the highest wrist-input mean per column.

Method	Macro-F1	Bal. acc.	Accuracy
Unadapted reference (baseline)	25.73±0.19	32.96±0.09	35.31±0.10
Mapper	65.80±1.04	67.44±0.41	70.61±1.28
MeanFlow	66.74±0.38	68.11±0.19	71.37±0.42
Distillation	65.43±1.06	67.18±0.71	70.54±1.17
Activity-only bridge	**67.76**±1.34	**68.93**±1.58	**71.89**±1.53
Direct wrist classifier	63.59±0.07	64.62±0.14	68.78±0.60
Measured thigh (reference sensor)	64.42±0.37	65.51±0.21	67.88±0.22

4.5 Reference reuse is competitive with classifier retraining

All methods here train on the same labeled wrist windows, so the comparisons hold the data budget fixed and vary only the adaptation strategy.

Direct wrist classifier. MeanFlow exceeds the direct wrist classifier on PAAWS Thigh2Wrist, 56.55% versus 54.34% in one matched run (Table 3), and on SELFback by 3.16 percentage points (95% interval, -2.10 to 7.89, Table 2). On CAPTURE-24 (Appendix Table 6), where both methods are fitted on labeled PAAWS SimFL windows, the two differ by -0.85 percentage points (-3.90 to 2.31), and the direct wrist classifier has higher balanced accuracy and accuracy by 7.67 and 2.77 percentage points.

Activity-only bridge. On PAAWS, where the measured thigh reaches 82.50%, MeanFlow exceeds the activity-only bridge by 2.46 percentage points in the current window (1.15 to 4.96) and by 1.99 with five-window averaging (-0.04 to 5.79). On SELFback, where the measured thigh reaches 64.42%, the activity-only bridge is highest (67.76%), so the bridge itself carries the gain.

Joint fine-tuning. Joint fine-tuning also updates the encoder, which replaces the shared reference model, and reaches 63.09% macro-F1 on PAAWS (Table 3). Adapting the encoder therefore offers further headroom, a natural place to add reference guidance.

Table 3: PAAWS Thigh2Wrist comparisons on the same FL test windows as Table 1. Values are percentages from one matched run per method, and all methods train on the same labeled wrist windows. The direct wrist classifier updates the thigh classifier with MoCA frozen, the activity-only bridge trains the MeanFlow architecture with cross-entropy alone, and joint fine-tuning updates both MoCA and the thigh classifier. Bold marks the highest score among the three frozen-encoder rows.

	Current window			Five-window averaging		
Method	Macro-F1	Bal. acc.	Accuracy	Macro-F1	Bal. acc.	Accuracy
Direct wrist classifier	54.34	69.69	**76.39**	58.90	72.87	81.14
Activity-only bridge	54.10	65.36	75.29	60.50	69.36	**81.39**
MeanFlow	**56.55**	**70.14**	76.01	**62.49**	**74.68**	80.90
Joint fine-tuning	63.09	71.27	80.40	69.37	74.26	85.06

5 Discussion and Conclusion

Summary. We asked how limited reference data can improve wrist-only recognition for participants who have no reference recording. WearableBridge answers by reusing the reference model: a small bridge carries wrist features into the activity regions of a fixed classifier trained on reference recordings. Every bridge improves macro-F1 in every run of both PAAWS designs, and the gains hold across four reference encoders, replicate on SELFback, and transfer to CAPTURE-24 without target labels. With the encoder frozen, reusing the fixed classifier reaches similar or higher macro-F1 than retraining it on the same wrist labels. Existing laboratory sessions improve free-living recognition without new reference data, and a bridge fitted in one study carries to another cohort.

What each reference contributes. Each reference transfers the distinctions it observes best. Thigh posture separates sitting from standing, and laboratory wrist sessions separate cycling from walking. Reference guidance improved on the activity-only bridge on PAAWS, where the thigh classifier is far more accurate than the bridged wrist, and the bridge itself carries the gains on SELFback, where the two are comparable. The three bridges score within a few points of each other, consistent with their shared activity supervision through the fixed classifier.

Limitations and future work. We evaluate one calibration cohort of 213 participants, so gains with fewer reference participants remain to be measured. Per-term ablations with repeated runs would separate the effects of bridge architecture and loss. Combining thigh and laboratory references, which separate different activity pairs, could improve all four activities, and reference guidance could extend to encoder fine-tuning and to several devices sharing one classifier.

AI USE STATEMENT

The research idea and method are the authors' own. Generative AI tools were used to polish writing, help locate related work, discuss some candidate experiments, and assist with code implementation, debugging, experiment orchestration, figure preparation, and manuscript checks, including citation checks and verification of reported values against preserved run artifacts. The authors designed the experiments and take full responsibility for the paper's content.

REPRODUCIBILITY STATEMENT

Appendices A–C give preprocessing, pairing, weighting, architectures, optimization, checkpoint selection, and the participant bootstrap. Appendices E and F give the SELFback and CAPTURE-24 protocols. The supplement provides code to train Mapper, MeanFlow, and Distillation bridges on features from any frozen encoder and classifier, with instructions and the interactive figures.

REFERENCES

Shing Chan, Yuan Hang, Catherine Tong, Aidan Acquah, Abram Schonfeldt, Jonathan Gershuny, and Aiden Doherty. CAPTURE-24: A large dataset of wrist-worn activity tracker data collected in the wild for human activity recognition. *Scientific Data*, 11:1135, 2024. doi: 10.1038/s41597-024-03960-3.

Youngjae Chang, Akhil Mathur, Anton Isopoussu, Junehwa Song, and Fahim Kawsar. A systematic study of unsupervised domain adaptation for robust human-activity recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 4(1):39:1–39:30, 2020. doi: 10.1145/3380985.

Yiqiang Chen, Jindong Wang, Meiyu Huang, and Han Yu. Cross-position activity recognition with stratified transfer learning. *Pervasive and Mobile Computing*, 57:1–13, 2019. doi: 10.1016/j.pmcj.2019.04.004.

Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(9):1853–1865, 2017. doi: 10.1109/TPAMI.2016.2615921.

Aiden Doherty, Dan Jackson, Nils Hammerla, Thomas Plötz, Patrick Olivier, Malcolm H. Granat, Tom White, Vincent T. van Hees, Michael I. Trenell, Christopher G. Owen, Stephen J. Preece, Rob Gillions, Simon Sheard, Tim Peakman, Soren Brage, and Nicholas J. Wareham. Large scale population assessment of physical activity using wrist worn accelerometers: The UK Biobank study. *PLOS ONE*, 12(2):e0169649, 2017. doi: 10.1371/journal.pone.0169649.

Sannara Ek, François Portet, and Philippe Lalanda. Transformer-based models to deal with heterogeneous environments in human activity recognition. *Personal and Ubiquitous Computing*, 27 (6):2267–2280, 2023. doi: 10.1007/s00779-023-01776-3.

Gamaleldin F. Elsayed, Ian Goodfellow, and Jascha Sohl-Dickstein. Adversarial reprogramming of neural networks. In *International Conference on Learning Representations*, 2019.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016.

Zhengyang Geng, Mingyang Deng, Xingjian Bai, J. Zico Kolter, and Kaiming He. Mean flows for one-step generative modeling. In *Advances in Neural Information Processing Systems*, volume 38, pp. 75460–75482, 2025. doi: 10.52202/085713-2534.

Zhengyang Geng, Yiyang Lu, Zongze Wu, Eli Shechtman, J. Zico Kolter, and Kaiming He. Improved mean flows: On the challenges of fastforward generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 30467–30476, 2026.

Saurabh Gupta, Judy Hoffman, and Jitendra Malik. Cross modal distillation for supervision transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2827–2836, 2016.

Illapha Cuba Gyllensten and Alberto G. Bonomi. Identifying types of physical activity with a single accelerometer: Evaluating laboratory-trained algorithms in daily life. *IEEE Transactions on Biomedical Engineering*, 58(9):2656–2663, 2011. doi: 10.1109/TBME.2011.2160723.

Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. arXiv:1503.02531, 2015.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 2790–2799, 2019.

Hassan Ismail Fawaz, Benjamin Lucas, Germain Forestier, Charlotte Pelletier, Daniel F. Schmidt, Jonathan Weber, Geoffrey I. Webb, Lhassane Idoumghar, Pierre-Alain Muller, and François Petitjean. InceptionTime: Finding AlexNet for time series classification. *Data Mining and Knowledge Discovery*, 34(6):1936–1962, 2020. doi: 10.1007/s10618-020-00710-y.

Yash Jain, Chi Ian Tang, Chulhong Min, Fahim Kawsar, and Akhil Mathur. ColloSSL: Collaborative self-supervised learning for human activity recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 6(1):17:1–17:28, 2022. doi: 10.1145/3517246.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

Guan-Horng Liu, Arash Vahdat, De-An Huang, Evangelos A. Theodorou, Weili Nie, and Anima Anandkumar. I^2SB: Image-to-image Schrödinger bridge. In *Proceedings of the 40th International Conference on Machine Learning*, 2023a.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations*, 2023b.

David Lopez-Paz, Léon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. Unifying distillation and privileged information. In *International Conference on Learning Representations*, 2016. URL `https://arxiv.org/abs/1511.03643`.

Leland McInnes, John Healy, and James Melville. UMAP: Uniform manifold approximation and projection for dimension reduction. arXiv:1802.03426, 2018.

Francisco Javier Ordóñez and Daniel Roggen. Deep convolutional and LSTM recurrent neural networks for multimodal wearable activity recognition. *Sensors*, 16(1):115, 2016. doi: 10.3390/s16010115.

Toby G. Pavey, Nicholas D. Gilson, Sjaan R. Gomersall, Bronwyn Clark, and Stewart G. Trost. Field evaluation of a random forest activity classifier for wrist-worn accelerometer data. *Journal of Science and Medicine in Sport*, 20(1):75–80, 2017. doi: 10.1016/j.jsams.2016.06.003.

Veronika Potter, Hoan Tran, Daniel Mobley, Suzanne M. Bertisch, Dinesh John, and Stephen Intille. The physical activity assessment using wearable sensors (PAAWS) dataset: Labeled laboratory and free-living accelerometer data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 9(4):204:1–204:32, 2025. doi: 10.1145/3770639.

Howon Ryu, Yuliang Chen, Yacun Wang, Andrea Z. LaCroix, Chongzhi Di, Loki Natarajan, Yu Wang, and Jingjing Zou. MoCA: Multi-modal cross-masked autoencoder for time series in digital health. arXiv:2506.02260, 2025.

Jeffer Eidi Sasaki, Amanda M. Hickey, John W. Staudenmayer, Dinesh John, Jane A. Kent, and Patty S. Freedson. Performance of activity classification algorithms in free-living older adults. *Medicine & Science in Sports & Exercise*, 48(5):941–950, 2016. doi: 10.1249/MSS.0000000000000844.

UCI Machine Learning Repository. selfBACK [dataset], 2020. URL https://doi.org/10.24432/C5SK65.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations*, 2021.

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*, 7:91, 2024. doi: 10.1038/s41746-024-01062-3.

A PAAWS Data and Calibration Pairs

Participants and signals. Calibration uses laboratory (Lab) and simulated free-living (SimFL) recordings from 170 training and 43 validation participants. Free-living (FL) testing uses 40 other participants. All comparisons classify cycling, sitting, standing, and walking.

Acceleration is resampled from 80 to 50 Hz with a polyphase filter. Each nonoverlapping ten-second window contains 3×500 values in units of standard gravity. Lab and SimFL inputs use the device worn in the top position on each wrist. FL inputs use the available left and right wrist devices, whose top or bottom position is uncertain. All methods reverse the first axis for right-wrist windows and retain the released axes for left-wrist windows.

Thigh2Wrist pairs. Each wrist window is paired with its concurrent right-thigh window. The 224,697 training pairs contain 113,091 unique thigh windows, each of which can accompany both wrists.

Lab2Freeliving pairs. Independently timed SimFL and Lab windows are matched within participant–activity–wrist-side groups. Within each group, both lists are independently shuffled and paired by position, cycling the shorter list. The 1,226 groups yield 137,299 training pairs. Pairs are reshuffled each epoch, with the same schedule for all bridge variants within a run.

Feature standardization. Coordinatewise means and population standard deviations are computed without pair weights over 132,279 Lab wrist windows of training participants for Lab2Freeliving and 113,091 unique thigh windows of training participants for Thigh2Wrist. Standard deviations are floored at 10^{-4}. Statistics remain fixed during fitting and evaluation.

B Models and Optimization

Reference models. Frozen MoCA produces 768-dimensional features. In Thigh2Wrist, a linear classifier reads these features, so $d = 768$. In Lab2Freeliving, a linear $768 \to 128$ projection is trained with the Lab classifier and then frozen as part of the feature map, so $d = 128$. The classifier applies layer normalization, GELU, and a linear four-class layer to the projected features.

The three additional encoders produce 128-dimensional features, so $d = 128$. DeepConvLSTM uses four 64-channel temporal convolutions with kernel size 5 and two 128-unit LSTM layers with dropout 0.1. InceptionTime uses six modules with 32-channel branches, kernels 41, 21, and 11, and residual connections every three modules. The accelerometer-only HART uses stride-20 patches, 25 tokens of dimension 128, and four blocks combining four-head attention with a depthwise convolution of kernel size 5 and a 256-unit feed-forward layer.

Bridge architectures. All bridges use two SiLU hidden layers, with width 768 in primary MoCA Thigh2Wrist and 256 elsewhere. Mapper and Distillation use residual multilayer perceptrons. MeanFlow receives the path point $z(t)$, the wrist feature z_s, and three 32-dimensional sinusoidal embeddings of t_0, t, and $t - t_0$. Output layers are zero-initialized. The encoder, classifier, and standardization remain frozen. Bridge comparisons combine architecture and objective differences.

MeanFlow time sampling. For each pair, we apply a sigmoid to two independent $\mathcal{N}(-0.4, 1^2)$ draws and sort them into $t_0 \le t$. For a uniformly selected half of each batch, we set $t_0 = t$ at the sampled upper time. The loss stabilizer is $\epsilon = 0.01$ in Lab2Freeliving and 0.06 in primary MoCA Thigh2Wrist.

Lab2Freeliving training. Reference class weights are $132{,}279/(4N_k)$, where N_k counts Lab training windows of activity k. Each bridge pair in participant–activity–wrist-side group g receives weight $137{,}299/(8G_g n_g)$. Here n_g is the larger of the group's SimFL and Lab counts, and G_g counts the participant groups that share the activity and wrist side of g. Training averages the weighted per-pair losses in each batch. Additional encoders train jointly with their classifiers on Lab data and are frozen before bridge fitting. Table 4 gives optimization settings.

Table 4: Lab2Freeliving optimization settings. All methods use AdamW and batches of 128, and bridges clip gradient norms at 1.

Setting	MoCA projection/classifier	Other encoders	Bridges
Learning rate	10^{-4}	10^{-3}	10^{-4}
Weight decay	0.05	10^{-4}	10^{-4}
Epochs	30	30	5–30
Precision	Mixed	Mixed	Full

Primary Thigh2Wrist training. The thigh classifier uses weights that equalize session–activity groups, where the session is Lab or SimFL. AdamW uses learning rate 10^{-4} and weight decay 0.05. Each bridge-training pair in session–activity–wrist-side group g receives weight $224{,}697/(16n_g)$, where n_g is the group count. Bridge learning rate and weight decay are 10^{-4}, with gradient clipping at 1. Reference and bridge training use batches of 128, full float32 precision, and 30 epochs.

Matched comparisons. Single-run PAAWS comparisons (Table 3) share the primary Thigh2Wrist data, weights, optimizer, and 30-epoch schedule. The direct wrist classifier and joint fine-tuning both initialize from the thigh classifier. The direct wrist classifier updates the classifier alone, and joint fine-tuning also updates the encoder. The activity-only bridge trains the MeanFlow architecture with activity cross-entropy alone. The MeanFlow row uses the matched primary run.

Additional Thigh2Wrist encoders. DeepConvLSTM, InceptionTime, and HART train jointly with linear thigh classifiers from scratch on the 113,091 unique training-thigh windows, then remain frozen. Reference fitting uses AdamW with learning rate 10^{-3}, weight decay 10^{-4}, batches of 128, and 30 epochs. Bridge fitting follows the primary Thigh2Wrist protocol with hidden width 256 and MeanFlow stabilizer $\epsilon = 0.01$. All training uses float32 precision.

C Evaluation and Uncertainty

Checkpoint selection. Lab2Freeliving selects the earliest maximum macro-F1 on 33,502 Lab-validation windows for the reference and 22,742 SimFL-validation windows for the bridge. Bridges train for 5–30 epochs with early-stopping patience of five epochs and a minimum improvement of 10^{-4}. All Thigh2Wrist comparisons use the final checkpoint after 30 epochs. Checkpoints are fixed before FL scoring.

Metrics and class support. Metrics pool windows across held-out participants. Measured-thigh scores use the same wrist-pair weighting as wrist predictions. Table 5 gives class support for both PAAWS designs.

Table 5: PAAWS test wrist windows. Every method and run within a calibration design uses the same windows. Thigh2Wrist requires a concurrent thigh recording.

Activity	Thigh2Wrist	Lab2Freeliving
Cycling	9,552	9,552
Sitting	1,261,789	1,267,245
Standing	176,233	177,148
Walking	140,795	141,826
Total	1,588,369	1,595,771

Five-window averaging. Five-window averaging combines class probabilities from the current and up to four preceding windows. History uses consecutive ten-second windows within each participant and wrist stream, before four-class scoring. Gains are computed within each run before averaging.

Participant intervals. For each of 5,000 bootstrap draws, we sample participants with replacement and retain all their windows. The same resample is applied across methods and runs. We

recompute pooled macro-F1 within each run and average paired differences across runs. Pointwise 95% percentile intervals describe participant variation conditional on the fitted models. Single-run comparisons use the same procedure. Run-to-run sample standard deviation is reported separately.

Feature visualization. For Thigh2Wrist (Figure 2), UMAP is fitted on 10,712 unique SimFL-validation thigh features. The plot displays 128 validation references and computes class means over 8,192 held-out wrist–thigh pairs. For Lab2Freeliving (Figure 3), UMAP is fitted on 128 Lab-validation references, and class means use the 2,048 displayed wrist recordings. Each view uses one run.

D Class-Level Errors on PAAWS

Cells show mean row percentages over three runs on the same test windows within each comparison. Rows are true activities and columns are current-window predictions.

MeanFlow raises Thigh2Wrist walking recall from 35.3% to 78.6% (Figure 4). In Lab2Freeliving, fewer cycling windows are predicted as walking (Figure 5). Sitting and standing remain the closest pair at the wrist, consistent with Section 4.2.

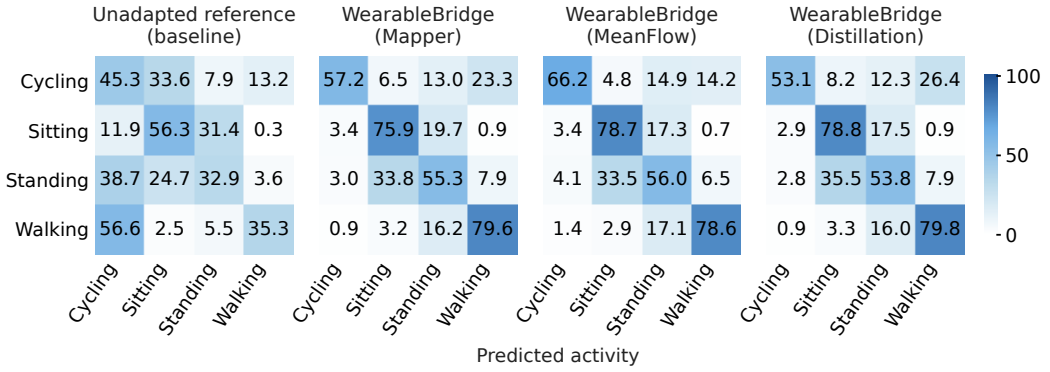

Figure 4: PAAWS Thigh2Wrist with MoCA on the same 1,588,369 wrist windows from 40 participants. Methods are defined as in Table 1, and Table 5 gives class counts.

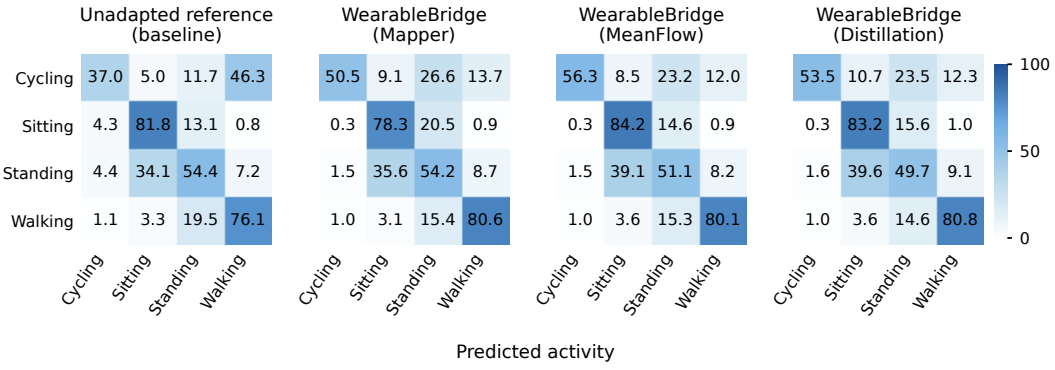

Figure 5: PAAWS Lab2Freeliving with MoCA on the same 1,595,771 windows from 40 participants. The unadapted reference applies the frozen Lab classifier directly to wrist features. Table 5 gives class counts.

E SELFBACK PAIRED REPLICATION

Data and pairing. SELFback supplies simultaneous wrist–thigh recordings for nine activities: lying, sitting, standing, jogging, ascending and descending stairs, and walking at three speeds. The participant split is 20/6/7 for training/validation/test. Streams are aligned on absolute timestamps, interpolated onto a 100 Hz grid using gaps of at most 30 ms, and resampled into complete ten-second windows at 50 Hz. This gives 2,828/874/963 paired windows. Wrist handedness is unspecified in the released metadata, so inputs retain the released axes.

Training and evaluation. Each run fits a nine-class thigh classifier on frozen 768-dimensional MoCA features, then freezes it for bridge training. The direct wrist classifier initializes from the same thigh classifier. All methods share training windows and class-balanced weights. Architectures, losses, and optimization follow primary Thigh2Wrist (Appendix B), with the final checkpoint after 30 epochs. Current-window evaluation uses three runs and the participant-bootstrap procedure in Appendix C.

F CAPTURE-24 TRANSFER WITHOUT TARGET LABELS

Data and preprocessing. We evaluate 134,658 windows from 50 CAPTURE-24 participants, P102–P151 (Chan et al., 2024). Fully annotated ten-second windows receive their modal label and are restricted to cycling, sitting, standing, and walking, with bicycling mapped to cycling. After 50 Hz preprocessing, signals are interpolated to 30 Hz for frozen HARNet10 features (1,024 dimensions) pretrained on UK Biobank (Yuan et al., 2024). CAPTURE-24 devices are worn on the dominant wrist, and all methods treat them as right wrists.

Training on PAAWS and evaluation. The reference, MeanFlow, and direct wrist classifier use PAAWS training and validation participants for fitting, standardization, and checkpoint selection. CAPTURE-24 labels define the evaluation set and metrics. The linear direct wrist classifier initializes from the PAAWS Lab classifier and updates it on labeled PAAWS SimFL features while HARNet10 stays frozen. MeanFlow keeps both the encoder and the Lab classifier fixed and learns the bridge. Both share SimFL examples, weights, optimizer, maximum budget, and selection rule, following Lab2Freeliving (Appendices B and C) with method-specific early stopping. MeanFlow uses two 256-unit hidden layers and $\epsilon = 0.01$. Evaluation uses current-window predictions, three runs, and the participant-bootstrap procedure in Appendix C. Table 6 gives overall and class-level scores. MeanFlow improves cycling and walking F1, the distinctions the Lab reference observes best, and standing F1 dips from 26.23% to 22.57%, as in PAAWS Lab2Freeliving.

Table 6: CAPTURE-24 transfer in the Lab2Freeliving design with frozen HARNet10, in percent (mean ± sample SD over three runs), on 134,658 windows from 50 participants (2,409 cycling, 100,733 sitting, 9,592 standing, and 21,924 walking). Fitting and selection use PAAWS only. The unadapted reference is the baseline: it applies the frozen PAAWS Lab classifier to CAPTURE-24 wrist features. The direct wrist classifier updates that classifier on labeled PAAWS SimFL windows. Bold marks the highest mean per column.

	Overall			Class F1			
Method	Macro-F1	Bal. acc.	Accuracy	Cycling	Sitting	Standing	Walking
Unadapted reference (baseline)	42.75±0.01	47.09±0.04	68.78±0.06	5.38±0.14	84.12±0.05	**26.23**±0.04	55.27±0.05
MeanFlow	54.73±1.58	57.50±1.18	69.40±0.94	**50.64**±4.94	83.16±0.84	22.57±0.45	**62.56**±0.63
Direct wrist classifier	**55.58**±0.42	**65.17**±0.54	**72.17**±0.48	50.30±2.63	**85.81**±0.39	25.15±0.14	61.08±0.83

G ADDITIONAL PERFORMANCE METRICS

Tables 7 and 9 give macro-F1 for the additional reference encoders and for five-window averaging. Tables 8 and 10 give the corresponding balanced accuracy and accuracy. All values use the same fitted models, test windows, and three runs.

Table 7: Additional reference encoders: current-window macro-F1 (%), mean ± sample SD over three runs. The unadapted reference is the baseline: it applies each encoder's frozen reference classifier to wrist features. Mapper, MeanFlow, and Distillation map wrist features before that classifier. Bold marks the highest wrist-input mean per row.

(a) Thigh2Wrist. 1,588,369 wrist windows from 40 participants. Measured thigh applies the same classifier to the concurrent thigh window and uses the reference sensor at test time.

	Baseline	WearableBridge			Reference sensor
Reference encoder	Unadapted reference	Mapper	MeanFlow	Distillation	Measured thigh
DeepConvLSTM	33.73±4.30	43.93±0.93	43.36±1.22	**44.20**±0.92	83.30±2.05
InceptionTime	32.49±5.47	52.04±1.71	52.37±2.66	**52.63**±1.38	86.67±2.24
HART	31.67±2.78	48.21±0.73	**49.20**±0.72	47.90±0.59	88.01±0.28

(b) Lab2Freeliving. 1,595,771 wrist windows from the same 40 participants. All columns use wrist input.

	Baseline	WearableBridge		
Reference encoder	Unadapted reference	Mapper	MeanFlow	Distillation
DeepConvLSTM	57.10±0.96	69.47±0.41	**69.76**±0.26	69.15±1.02
InceptionTime	60.65±0.24	68.82±1.76	68.78±0.76	**68.97**±0.64
HART	56.37±0.28	68.91±0.33	68.98±0.50	**69.14**±0.81

Table 8: Additional reference encoders: balanced accuracy (BA) and accuracy (Acc.), in percent, mean ± sample SD over three runs. Methods are defined as in Table 7. Bold marks the highest wrist-input mean for each encoder and metric.

	DeepConvLSTM		InceptionTime		HART	
Method	BA	Acc.	BA	Acc.	BA	Acc.
Thigh2Wrist						
Unadapted reference (baseline)	39.26±8.44	54.59±2.03	34.20±3.97	56.73±1.85	33.40±2.86	56.17±1.52
Mapper	58.36±3.21	64.93±6.05	67.16±0.82	74.74±3.69	63.49±0.52	71.02±2.12
MeanFlow	57.72±2.64	65.44±2.19	**67.72**±0.72	75.25±4.68	**64.02**±0.13	**73.99**±1.82
Distillation	**59.02**±2.75	**66.11**±2.50	67.34±0.60	**75.79**±2.04	63.47±0.72	70.60±0.80
Measured thigh (reference sensor)	89.92±0.48	94.94±0.12	89.74±0.44	95.26±0.39	89.40±0.21	95.31±0.20
Lab2Freeliving						
Unadapted reference (baseline)	64.76±0.55	79.35±1.23	65.56±0.79	82.26±0.94	60.46±0.58	81.18±0.36
Mapper	71.30±1.03	83.52±0.38	71.11±1.94	**84.06**±0.96	70.78±1.00	82.40±1.25
MeanFlow	71.64±1.06	83.60±1.03	**73.61**±0.10	82.58±0.60	70.35±1.11	83.73±1.62
Distillation	**72.18**±0.67	**84.26**±0.73	73.47±0.52	83.50±0.59	**71.26**±0.56	**84.13**±0.61

Table 9: Macro-F1 with five-window averaging, which averages class probabilities over the current and up to four preceding windows, mean ± sample SD over three runs. Δ is the within-run gain over the same method in Table 1, in percentage points (pp). Methods are defined as in Table 1. Bold marks the highest macro-F1 per design.

	Thigh2Wrist		Lab2Freeliving	
Method	Macro-F1 (%)	Δ macro-F1 (pp)	Macro-F1 (%)	Δ macro-F1 (pp)
Unadapted reference (baseline)	35.53±0.25	+0.46	58.16±1.70	+4.16
Mapper	60.28±0.78	+6.15	66.49±0.24	+5.59
MeanFlow	**61.70**±0.75	+5.72	**68.90**±1.42	+5.36
Distillation	60.93±0.43	+5.87	67.55±1.32	+5.26

Table 10: Primary MoCA references with five-window averaging: balanced accuracy (BA) and accuracy (Acc.), in percent, mean ± sample SD over three runs. Methods are defined as in Table 1, whose current-window scores these complement. Bold marks the highest mean per column.

	Thigh2Wrist		Lab2Freeliving	
Method	BA	Acc.	BA	Acc.
Unadapted reference (baseline)	42.25±0.12	53.47±0.39	64.78±2.78	82.90±2.04
Mapper	71.45±2.10	79.93±2.23	69.05±0.70	80.12±1.85
MeanFlow	**73.64**±1.36	81.27±1.73	**70.51**±0.90	**84.10**±1.39
Distillation	70.15±1.15	**81.40**±1.97	69.36±0.66	83.17±1.39