PT Jurnal Cendekia Indonesia
Journal of Information Technology and Informatics Engineering
Homepage: https://journal.jci.co.id/jitie
Vol. 02 No. 04 (2026) Page: 77-89
E-ISSN: 3123-7827

Language-Guided Cross-Subject Wearable Activity Recognition with Self-Supervised Transfer, Selective Prediction, and Edge Coaching Explanations

[1]Oliver Thompson, [2]Rachel Lu, [4]Sophia Garcia, [5]Kevin Li

[1]Department of Computer Science, Drexel University, Philadelphia, PA, USA
[2]Department of Computer Science, University of South Carolina, Columbia, SC, USA
[3]Department of Computer Science, University of Kansas, Lawrence, KS, USA
[4]Department of Computer Science, University of California, Riverside, Riverside, CA, USA

Email: o.thompson_dev@yahoo.com

Abstract

Cross-subject wearable activity recognition often fails when a model trained on one person encounters another person's motion range, pace, or sensor placement. This paper evaluates a compact language-guided transfer pipeline on two complete participant archives from the OpenPack v1.1.0 Zenodo record. Four arm and wrist inertial sensors and two wrist physiological devices produced 184 one-second statistics; 4,072 nonoverlapping five-second sequences represented ten packaging operations. The evaluation reverses source and target participants. Source sessions S01–S03 train the models, S04 selects models and calibration parameters, and S05 provides an internal test. Target sessions S01–S02 form a label-blind correlation-alignment pool after operation-window filtering, while S03–S05 remain an external cross-subject test. The method combines a 20% masked linear reconstruction objective, a transparent TF-IDF operation-to-sensor gate, target-to-source CORAL, BiGRU and compact attention branches, temperature scaling, and confidence-based abstention. The LG-SSL-CORAL ensemble reached external macro-F1 scores of 0.490 and 0.524 in the two directions, compared with 0.450 and 0.409 for BiGRU. Box-block paired bootstrap gains were 0.040 (95% CI 0.012–0.068) and 0.115 (0.089–0.145). Extra Trees remained a strong comparator with mean macro-F1 0.519 versus 0.507 for the proposed ensemble. Masked reconstruction improved the forward Conformer-lite result from 0.385 to 0.443, whereas CORAL and source-only temperature scaling were direction-sensitive. At 50% coverage, selective accuracy rose to 0.669 and 0.677, but a source-validation 10% risk rule did not transfer to the target subject. The language branch required 55.6 KiB of FP32 weights and 1.13 ms per sequence in a TensorFlow.js host-CPU proxy. The results establish both the benefit and the limits of language-guided sensor selection, covariance transfer, and confidence rejection in a strict two-participant cross-subject setting.

Keywords: Wearable Activity Recognition; Cross-Subject Transfer; Self-Supervised Learning; Language-Guided Sensor Selection; Selective Prediction; Openpack; Edge Inference

Abstrak

Pengenalan aktivitas perangkat wearable lintas subjek seringkali gagal ketika model yang dilatih pada satu orang bertemu dengan rentang gerak, kecepatan, atau penempatan sensor orang lain. Makalah ini mengevaluasi pipeline transfer yang ringkas dan dipandu bahasa pada dua arsip partisipan lengkap dari catatan OpenPack v1.1.0 Zenodo. Empat sensor inersia lengan dan pergelangan tangan serta dua perangkat fisiologis pergelangan tangan menghasilkan 184 statistik satu detik; 4.072 urutan lima detik yang tidak tumpang tindih mewakili sepuluh operasi pengemasan. Evaluasi membalikkan partisipan sumber dan target. Sesi sumber S01–S03 melatih model, S04 memilih model dan parameter kalibrasi, dan S05 menyediakan pengujian internal. Sesi target S01–S02 membentuk kumpulan korelasi-penyelarasan buta label setelah penyaringan jendela operasi, sementara S03–S05 tetap menjadi pengujian lintas subjek eksternal. Metode ini menggabungkan tujuan rekonstruksi linier bertopeng 20%, gerbang operasi-ke-sensor TF-IDF transparan, CORAL target-ke-sumber, BiGRU dan cabang perhatian kompak, penskalaan suhu, dan abstensi berbasis kepercayaan. Ensemble LG-SSL-CORAL mencapai skor makro-F1 eksternal sebesar 0,490 dan 0,524 dalam dua arah, dibandingkan dengan 0,450 dan 0,409 untuk BiGRU. Perolehan bootstrap berpasangan box-block adalah 0,040 (95% CI 0,012–0,068) dan 0,115 (0,089–0,145). Extra Trees tetap menjadi pembanding yang kuat dengan rata-rata makro-F1 0,519 dibandingkan dengan 0,507 untuk ensemble yang diusulkan. Rekonstruksi bertopeng meningkatkan hasil Conformer-lite maju dari 0,385 menjadi 0,443, sedangkan CORAL dan penskalaan suhu hanya sumber sensitif terhadap arah. Pada cakupan 50%, akurasi selektif meningkat menjadi 0,669 dan 0,677, tetapi aturan risiko 10% validasi sumber tidak ditransfer ke subjek target. Cabang bahasa membutuhkan 55,6 KiB bobot FP32 dan 1,13 ms per urutan dalam proxy CPU host TensorFlow.js. Hasilnya menunjukkan manfaat dan batasan pemilihan sensor berbasis bahasa, transfer kovariansi, dan penolakan kepercayaan dalam pengaturan lintas subjek dua partisipan yang ketat.

Received: 01-08-2026 | Revised: 09-08-2026 | Published Online: 10-08-2026

Kata Kunci: *Pengenalan Aktivitas Wearable; Transfer Lintas Subjek; Pembelajaran Mandiri; Pemilihan Sensor Berbasis Bahasa; Prediksi Selektif; Openpack; Inferensi Tepi*

A. INTRODUCTION

Wearable activity recognition has moved from short laboratory gestures to long, structured work and training sequences. Deep recurrent and convolutional models can fuse multiple body locations [1], yet their accuracy often falls when a new person's movement amplitude, handedness, pace, or strap placement differs from the training distribution. OpenPack makes this problem concrete through synchronized packaging operations recorded with wearable, environmental, and IoT sensing [2]. Its challenge experience also shows that long workflows, class imbalance, and similar neighboring operations make conventional random-window evaluation overly optimistic [3]. Motif-aware acceleration modeling has improved packaging recognition [4], but robust transfer to an unseen worker remains unresolved. Recent public-benchmark work on UCI HAR and WISDM likewise found a substantial source-target gap when moving across sensing domains, even when alignment improved a source-only baseline [5]. Subject-independent motor-imagery decoding shows a related calibration-light problem in a different biosignal modality, where masked pretraining and a compact Conformer remained competitive but target-subject adaptation still mattered [6]. More generally, transfer analysis based on approximately shared features emphasizes separating content-like factors from environment-specific factors rather than assuming that all source structure transfers unchanged [7].

Self-supervised learning (SSL) offers a way to learn sensor structure before using class labels. Temporal and contextual contrastive learning [8], cross-modality agreement [9], and large-scale wearable pretraining [10] have all produced transferable representations. Careful assessments nevertheless show that SSL gains depend on the downstream split and the strength of supervised baselines [11]. CrossHAR further demonstrates that hierarchical pretraining can aid cross-dataset recognition [12], and a 2024 OpenPack study directly investigated SSL for complex industrial work [13]. The same design tension appears in adjacent time-series domains: lightweight cross-modal medical pretraining focuses on parameter-efficient few-shot transfer [14], cold-start workload forecasting evaluates limited target support [15], and masked log-sequence modeling tests what can be learned from normal sequences without anomaly labels [16]. These studies are not evidence about OpenPack, but they reinforce the need to isolate the pretext transformation, target-support assumptions, and downstream split rather than attributing every gain to SSL in the abstract.

Language supplies a second transfer signal. Natural-language supervision can bridge inconsistent activity vocabularies and datasets [17], mediate inertial cross-modal transfer [18], and align semantic and motion representations [19]. Wearable systems have also begun converting sensor evidence into natural-language summaries and feedback [20], [21]. Language-guided feature selection has been operationalized outside HAR by mapping natural-language attack descriptions to network features [22], while text-grounded design-rationale interfaces expose how metadata is converted into human-readable decision cards [23]. Work on dark-pattern explanation and design critique further shows that fluent text is most useful when tied to a transparent intermediate representation and judged against human-facing criteria [24], [25]. In the present work, language is used in a narrower and fully inspectable form: operation descriptions and sensor-location descriptions determine a soft sensor gate. It performs semantic feature selection that can be displayed directly to a worker or coach.

Reliable deployment requires more than a top-1 label. Neural confidence can be miscalibrated, and temperature scaling provides a simple validation-based correction [26]. Selective prediction instead allows a system to abstain when confidence is low and evaluates the resulting risk-coverage trade-off [27]. Recent studies in other domains have paired confidence calibration with learned fusion [28], retrieval coverage and abstention [29], selective refusal [30], uncertainty-driven rejection [31], distribution-free uncertainty quantification [32], conformal risk envelopes [33], and conformal alert control [34]. These are cross-domain design precedents rather than evidence that a given threshold will transfer to a new wearable user; they motivate measuring calibration and risk-coverage behavior explicitly. Lightweight HAR models [35] and JavaScript inference runtimes [36] make local feedback feasible, but accuracy, calibration, rejection behavior, and runtime must be measured under the same protocol. Sports studies using deep wearable recognition in beach volleyball [37] and a single wrist IMU for volleyball skill assessment [38] illustrate why a recognized action should be paired with a concise, sensor-grounded cue rather than an unexplained class label.

Compact deployment introduces a separate architecture question. BiGRU-attention and BiGRU-Transformer hybrids have been used for storage-health classification and GPU-memory forecasting [39], [40], while edge-oriented quality prediction, progressive rendering, IoT autoencoding, LiDAR-camera tracking, and TinyLLM-style intrusion detection emphasize latency, modularity, and failure-aware reporting [41], [42], [43], [44], [45]. The present study uses these works only as engineering analogies: its runtime claims remain limited to the measured TensorFlow.js host-CPU proxy.

Explanation quality also depends on provenance. Evidence-chain RAG, evidence-grounded operations QA, bilingual incident triage, and word-level source attribution all separate the evidence-producing branch from the text-generating branch [46], [47], [48], [49]. This distinction is directly relevant to coaching: a wearable cue should describe the sensors and confidence used by the branch

being explained, not a post hoc story assembled from unrelated signals.

Human-facing AI studies have therefore converged on compact evidence, risk, and recommendation cards for incident response, prompt-injection oversight, scientific search, patient communication, radiology, ultrasound, counseling support, e-commerce recommendations, capacity dashboards, and designer-editable briefs [50], [51], [52], [53], [54], [55], [56], [57], [58], [59], [60]. Reproducible vision-language prototyping additionally separates structural consistency from rendered visual consistency [61], while rubric-grounded formative feedback routes uncertain cases to human review rather than letting generated prose obscure the underlying decision trace [62]. These interface studies do not validate wearable coaching, but they provide a vocabulary for presenting probability, uncertainty, evidence, and escalation.

For sports-facing deployment, interface adaptation matters alongside recognition accuracy. A review-based study of volleyball learning apps emphasizes task-specific navigation and feedback presentation [63], complementing wearable volleyball recognition and wrist-skill assessment [37], [38] and the cross-dataset transfer results in [5]. This motivates operation-specific cues that are concise, inspectable, and easy for a coach to override.

This study makes four contributions. First, it audits the exact U0101 and U0102 archives and constructs a leakage-controlled, bidirectional session protocol. Second, it compares Extra Trees, BiGRU, Transformer-lite, and Conformer-lite models with masked reconstruction, language-guided gating, and correlation alignment. Third, it evaluates calibration, abstention, per-class behavior, and box-block uncertainty rather than reporting accuracy alone. Fourth, it measures an edge-oriented TensorFlow.js CPU profile and generates branch-faithful coaching cues from the recognized operation, confidence, and semantic sensor weights. Together, these analyses characterize recognition gains and failure modes under the two-participant protocol.

B. METHODS

Data Record and Integrity Audit

Experiments used U0101.zip and U0102.zip from the OpenPack Zenodo record v1.1.0, released on 24 April 2024 [64]. The record contains about 8.7 GB, while the two selected participant archives are 241,038,401 and 291,085,240 bytes. Their computed MD5 digests matched the record: 7c182630795ff0e93393a781a076be6f and 3c32acce47a27437bfb39eed8a0e0199. OpenPack as a whole contains 16 participants and 53.8 h [2]; the present results cover only U0101 and U0102 and are not estimates for all 16 participants.

Each archive contains five sessions, S0100 through S0500, with one-hertz operation labels. Ten operation IDs, 100 through 1000, were scored; null ID 8100 was excluded. Across both archives, 20,516 annotated seconds were present, of which 20,386 belonged to the ten scored operations. There were 1,997 operation intervals and 194 boxes. The selected windows yielded 1,808 U0101 sequences and 2,264 U0102 sequences (Table 1).

Table 1 Openpack Archive Integrity And Extracted Sequence Inventory.

Participant	Archive bytes	MD5	Annotated seconds	Selected 5-s sequences	Features/s	Median purity
U0101	241,038,401	7c182630795ff0e93393a781a076be6f	9,125	1,808	184	1.00
U0102	291,085,240	3c32acce47a27437bfb39eed8a0e0199	11,391	2,264	184	1.00

Four ATR TSND151 units occupied the right wrist, left wrist, right upper arm, and left upper arm. Their timestamps gave a median rate of 30.303 Hz. Acceleration and angular velocity were complete, but every quaternion value in all 2,487,627 ATR rows was zero; orientation was therefore excluded. Two Empatica E4 devices supplied wrist acceleration, blood-volume pulse, electrodermal activity, and temperature. Timestamp-derived median rates were 32.258, 62.5, 4, and 4 Hz, respectively. Table 2 reports measured rows and missing cells. Kinect keypoints were not used because the wearable pipeline was complete across all ten participant-sessions, whereas Kinect 3-D files were absent for one session per participant.

Table 2 Measured Wearable Stream Audit Across U0101 And U0102. ATR Quaternion Columns Were Excluded Because Every Value Was Zero.

Stream	Location	Rows	Rate (Hz)	Missing cells	Channels used
ATR01	Right wrist	621,908	30.303	0	3-axis acceleration + 3-axis gyroscope
ATR02	Left wrist	621,907	30.303	0	3-axis acceleration + 3-axis gyroscope
ATR03	Right upper arm	621,906	30.303	0	3-axis acceleration + 3-axis gyroscope
ATR04	Left upper arm	621,906	30.303	0	3-axis acceleration + 3-axis gyroscope
E4 ACC	Both wrists	1,312,172	32.258	0	3-axis acceleration
E4 BVP	Both wrists	2,624,365	62.500	0	BVP
E4 EDA	Both wrists	164,006	4.000	0	EDA
E4 TEMP	Both wrists	164,006	4.000	0	temperature

Class support was imbalanced (Table 3 and Fig. 1). Attach Box Label contained only 847 scored seconds and had a 4.10-s median interval, whereas Scan Label contained 3,225 seconds. The ten scored rows sum to 20,386 seconds and 1,967 intervals; the totals of 20,516 annotated seconds and 1,997 intervals additionally include null ID 8100. This distribution motivated macro-F1 as the primary metric and inverse-frequency class weights for neural training.

Table 3 Scored Operation Support And Interval Duration Across The Two Participants

ID	Operation	Scored seconds	Share (%)	Intervals	Median duration (s)
100	Picking	1,913	9.3	194	8.83
200	Relocate Item Label	3,134	15.3	200	11.20
300	Assemble Box	3,195	15.6	198	15.05
400	Insert Items	1,764	8.6	194	7.62
500	Close Box	2,060	10.0	193	10.53
600	Attach Box Label	847	4.1	194	4.10
700	Scan Label	3,225	15.7	212	14.15
800	Attach Shipping Label	1,371	6.7	194	7.05
900	Put on Back Table	1,135	5.5	194	5.70
1000	Fill out Order	1,742	8.5	194	7.81

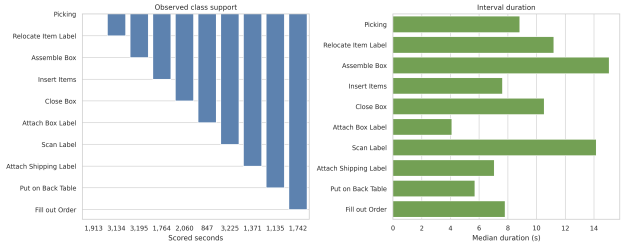

Figure 1 Observed Scored Seconds And Median Operation Duration In U0101 And U0102.

Signal Representation and Sequence Construction

All sensor timestamps were mapped to one-second bins. For each ATR axis, mean, standard deviation, root-mean-square value, and range were computed for three acceleration and three gyroscope channels. The same four statistics were applied to acceleration and angular-speed magnitudes, giving 32 features per ATR and 128 inertial features. Each E4 contributed 16 acceleration statistics, including magnitude, plus four statistics each for BVP, EDA, and temperature, giving 28 features per device. The final vector contained 184 features per second.

Within each session, gaps of at most two consecutive seconds were linearly interpolated. Remaining gaps used the nearest within-session observation in either direction and then the session median; the measured streams in Table 2 contained no missing raw cells, so this rule mainly guarded bin-boundary sparsity. This preprocessing is an offline session transformation rather than a causal streaming filter. Five consecutive one-second vectors formed a nonoverlapping sequence. A sequence was retained when at least three of its five one-hertz labels belonged to the same scored operation, and that majority operation became the target. Median purity was 1.00 for both participants. Figure 2 shows the four inertial magnitudes and aligned operation labels from an observed session interval.

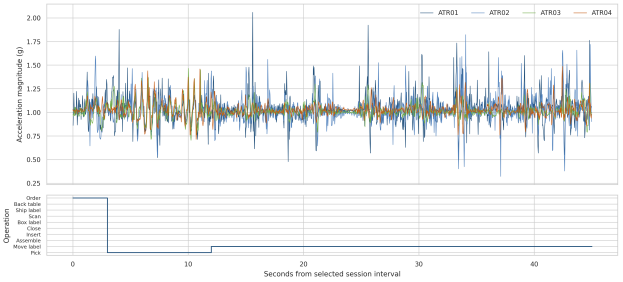

Figure 2 Four ATR Acceleration Magnitudes And Aligned One-Hertz Operation Labels From An Observed U0101 Interval.

Cross-Subject Protocol

The two transfer directions were evaluated separately. In U0101→U0102, U0101 sessions S01–S03 supplied 1,113 training sequences, S04 supplied 366 validation sequences, and S05 supplied 329 internal-test sequences. U0102 S01–S02 supplied 1,068 adaptation sequences, and U0102 S03–S05 supplied 1,196 external-test sequences. The reverse direction used 1,506 training, 397 validation, 361 internal-test, 767 adaptation, and 1,041 external-test sequences.

For the adaptation pool, the operation timeline was used to retain majority-scored five-second windows, after which all class identities were discarded before covariance estimation. Thus the pool was label-blind but annotation-filtered. No target S03–S05 label, feature statistic, calibration result, or threshold result influenced fitting or model selection. Standardization means and standard deviations were computed only from source S01–S03 and applied to every split. This protocol separates source-session generalization from external person transfer and prevents target-test labels from selecting hyperparameters.

Masked Reconstruction

The SSL stage used a compact masked linear denoiser. Twenty percent of standardized source-training entries were set to zero with random generator seed 20260805. If X is the unmasked matrix and Xm is its masked version, the reconstruction matrix was

$$W = (Xm^{T}Xm + I)^{-1}Xm^{T}X,$$

where the ridge coefficient was 1.0. The masked source matrix Xm was used only to fit W. At application time, each complete sequence X was transformed as 0.5X + 0.5XW, without masking validation, internal-test, adaptation, or external-test inputs. W was fitted without operation labels and applied identically to all held-out splits. This objective tests masked reconstruction rather than contrastive learning, avoiding an unsupported equivalence between the two SSL families. Its role parallels the broader goal of transferable time-series representation learning [8]-[12] while remaining small enough for controlled ablation. Masked pretraining in subject-independent BCI and log-sequence anomaly modeling provides adjacent evidence that the value of a reconstruction objective depends on the downstream shift and task [6], [16]; neither study is treated as a direct OpenPack baseline.

Language-Guided Sensor Gate

Ten short operation descriptions named the observed motion, for example bilateral flap folding for Assemble Box and dominant-wrist writing for Fill out Order. Six descriptions represented the four ATR locations and two E4 locations. A TF-IDF encoder with unigram and bigram features produced cosine similarities between operation and device text. Row-wise softmax with temperature 0.20 converted each similarity row to a distribution; the device multiplier was 0.75 + 1.50p, keeping every device active while emphasizing semantically related locations.

A balanced multinomial logistic model (C = 0.5) predicted operation probabilities from each sequence's temporal mean, standard deviation, and last-minus-first feature difference. Source-training gates used source-session grouped out-of-fold probabilities; the model refitted on all source-training sequences generated gates for validation, internal-test, and external-test inputs. Grouped out-of-fold macro-F1 was 0.776 for U0101 and 0.748 for U0102. Multiplying the class probabilities by the fixed operation-device matrix produced a dynamic six-device gate. Each feature group was scaled by its device multiplier before classification. Figure 3 displays the fixed semantic matrix. The gate is language-guided soft sensor selection, consistent with language-supervised HAR [17]-[19], not an end-to-end language-sensor foundation model. Its fixed text-to-feature mapping also echoes language-guided network-feature selection [22] and text-grounded design-rationale interfaces [23], but every wearable channel remains active so that a semantic mismatch cannot delete a sensor outright.

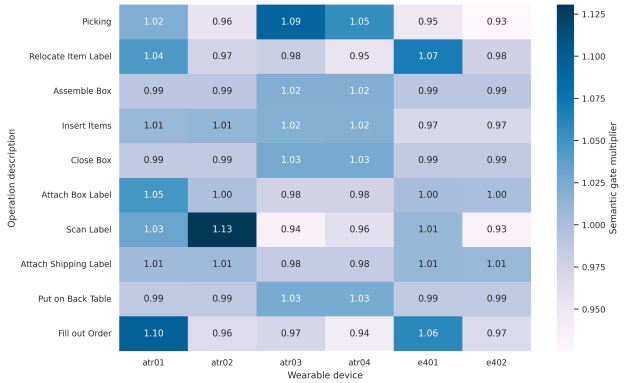

Figure 3 Text-Derived Operation-To-Device Gate Multipliers. Every Sensor Remains Active.

Correlation Alignment and Classifiers

CORAL aligned target adaptation features to source-training features using second-order statistics [65]. With source covariance Cs, target covariance Ct, and their means μs and μt, a target row x was transformed as

$$x' = (x - \mu t)Ct^{-1/2}Cs^{1/2} + \mu s.$$

An identity regularizer of 10^-3 stabilized both covariance matrices. The transform was fitted to target S01-S02 and applied only to target external features. It preceded masked reconstruction and the language gate in the combined branch. Adaptive spatial-temporal transfer has addressed related sensor and temporal mismatch across HAR datasets [66], and cross-dataset wearable transfer continues to show that alignment can be helpful without eliminating domain-specific error [5]. Approximately shared-feature theory [7], cross-cloud workload transfer [67], and cold-start tenant forecasting [15] further motivate treating source-target alignment as conditional rather than universally beneficial. CORAL was selected here because its closed form permits an exact, label-blind comparison.

Extra Trees used 300 estimators, square-root feature sampling, minimum leaf size two, balanced class weights, and the flattened 5 x 184 input [68]. The BiGRU baseline used 18 units in each direction, 0.10 dropout, a 24-unit ReLU layer, and a ten-class softmax, following the gated recurrent unit formulation [69]. Transformer-lite projected inputs to 24 dimensions, added sinusoidal position codes, and used one custom single-head self-attention block and a 48-unit feed-forward block [70]. Conformer-lite added a pre-attention feed-forward block and one-dimensional convolutions with kernels three and one; it is a compact Conformer-inspired network rather than the full speech architecture [71]. BiGRU-attention and BiGRU-Transformer combinations in storage-health and GPU-memory studies [39], [40] are cited as adjacent sequence-model examples, not as directly comparable HAR results.

Neural models used class-weighted cross-entropy, Adam at 0.0025, batch size 64, 18 epochs, and the weights with minimum source-validation loss. Initializers used seed 20260805, and each configuration was trained once. The final LG-SSL-CORAL ensemble averaged three probability vectors: raw BiGRU, masked-reconstruction-plus-language Conformer-lite, and masked-reconstruction-plus-CORAL Conformer-lite. Weights were selected on the source validation split over a 0.1 simplex grid with every component at least 0.1. Forward weights were 0.4/0.2/0.4; reverse weights were 0.5/0.1/0.4. Figure 4 summarizes the branch order and selective output.

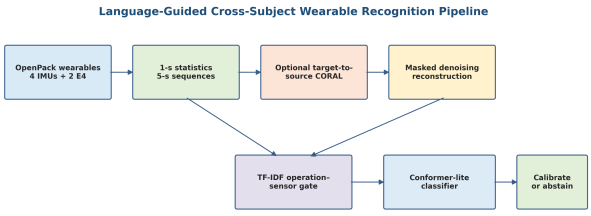

Figure 4 Language-Guided Cross-Subject Wearable Recognition Pipeline. CORAL Is Optional And Precedes Masked Reconstruction For Target Features.

Calibration, Selective Prediction, Explanations, and Statistics

Temperature was fitted by minimizing source-validation negative log likelihood [26]. Calibration was measured by negative log likelihood and 15-bin equal-count expected calibration error (ECE). Selective prediction ranked sequences by maximum probability [27]. Accuracy and risk were evaluated over the risk-coverage curve, with 50% and 100% coverage reported explicitly. A second operating point selected the lowest source-validation confidence

threshold whose accepted error was at most 10%; its target behavior was evaluated without retuning. This source-only policy mirrors the deployment questions raised by calibrated late fusion [28], abstention-aware RAG [29], selective recruiter screening [30], uncertainty-driven credit rejection [31], distribution-free model-risk workflows [32], conformal resource envelopes [33], and conformal alert control [34], while allowing the wearable experiment to test whether the operating point actually transfers.

The primary classification metric was macro-F1; accuracy, ECE, negative log likelihood, AURC, and per-class F1 provided complementary diagnostics. Ninety-five percent confidence intervals used 1,000 bootstrap samples of target session–box blocks. Paired differences reused the same sampled blocks for the ensemble and BiGRU. These intervals capture target box variability but not training-seed variability.

Coaching text was generated for the standalone language branch so that the stated sensor evidence matched the branch being explained. Each cue contained the predicted operation, calibrated probability, two highest semantic device multipliers, and an instruction to review timing and smoothness. This template follows the action-feedback goal of UbiPhysio and Sensor2Text [20], [21] without claiming that the semantic weights are causal feature attributions. The branch-faithful rule also follows the provenance separation used in evidence-chain RAG and operations QA [46], [47]-[49], while evidence and risk-card interfaces [50]-[57] motivate exposing confidence and escalation rather than presenting generated prose as an independent source of truth.

Host-CPU Proxy

Models ran with TensorFlow.js 4.22.0 under Node.js 24.14.0 on an x86-64 AMD EPYC 9V74 host with nine available CPU cores [36]. After 20 warm-up predictions, batch-one latency was averaged over 200 predictions and paired with raw FP32 weight bytes and forward-direction macro-F1. The measurement is an edge-oriented host proxy; it is not a phone, microcontroller, energy, quantized, or TensorFlow Lite benchmark. Edge-oriented proxy modeling, progressive rendering, compact IoT autoencoding, and TinyLLM-assisted detection provide engineering context for reporting model size and latency [41], [42], [43], [45], but do not change the scope of the measured hardware claim...

C. RESULTS AND DISCUSSION

Architecture and Cross-Subject Transfer

Table 4 and Fig. 5 compare the external tests. Extra Trees was the strongest overall comparator, with mean macro-F1 0.519. The LG-SSL-CORAL ensemble reached 0.507, outperforming every individual neural configuration on the mean of both directions. BiGRU averaged 0.430, SSL + language Conformer-lite 0.422, raw Conformer-lite 0.400, and Transformer-lite 0.341. The result cautions against assuming that a newer sequence architecture dominates a well-tuned tree ensemble when only two source participants are available. Related wearable transfer and calibration-light BCI studies likewise show that compact recurrent, convolutional, or classical baselines can remain competitive under limited target support [5], [6]; BiGRU-attention and BiGRU-Transformer successes in other sequence domains [39], [40] do not remove the need for task-specific comparison.

Table 4 External Cross-Subject Architecture Comparison. F1 Denotes Macro-F1.

Model	Parameters	F1 U0101→ U0102	Acc. U0101→ U0102	F1 U0102→ U0101	Acc. U0102→ U0101	Mean F1
Extra Trees	300 trees	0.461	0.521	0.577	0.608	0.519
BiGRU	23,062	0.450	0.515	0.409	0.415	0.430
Transformer-lite	9,466	0.291	0.308	0.390	0.419	0.341
Conformer-lite	14,242	0.385	0.425	0.414	0.467	0.400
SSL + language Conformer-lite	14,242	0.422	0.451	0.422	0.459	0.422
SSL + language + CORAL Conformer-lite	14,242	0.402	0.431	0.431	0.500	0.417
LG-SSL-CORAL ensemble	51,546	0.490	0.546	0.524	0.548	0.507

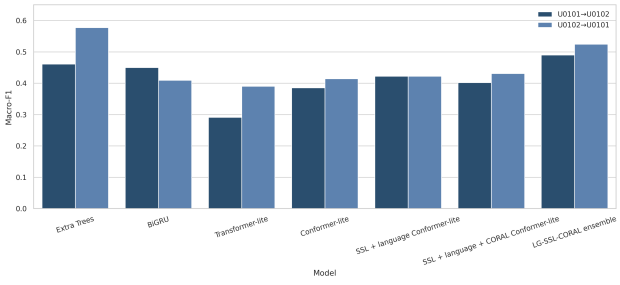

Figure 5 External Macro-F1 By Architecture And Transfer Direction.

The ensemble's internal macro-F1 was 0.864 for U0101 and 0.845 for U0102, but external F1 fell to 0.490 and 0.524 (Table 5). The corresponding drops of 0.374 and 0.321 quantify substantial participant shift. External accuracies were nearly identical at 0.546 and 0.548 even though class-wise behavior differed. Relative to BiGRU, paired box-block gains were 0.040 (95% CI 0.012–0.068) and 0.115 (0.089–0.145). Both intervals excluded zero within the measured target boxes. The reverse directions reuse the same two people and therefore represent two transfer

directions, not independent leave-one-subject-out replications.

Table 5 Directional Performance And 1,000-Sample Target Session–Box Bootstrap Intervals For The LG-SSL-CORAL Ensemble.

Direction	Internal F1	External F1	External acc.	F1 95% CI	ΔF1 vs. BiGRU	Δ 95% CI	Box blocks
U0101→U0102	0.864	0.490	0.546	0.452–0.526	0.040	0.012–0.068	54
U0102→U0101	0.845	0.524	0.548	0.487–0.559	0.115	0.089–0.145	60

Figure 6 places the directional macro-F1 intervals beside the paired improvement intervals. The separation from zero is wider in the reverse direction, where BiGRU was weakest, while the forward interval shows a smaller but still positive box-level gain. Per-class behavior is examined after the transfer-component and calibration analyses.

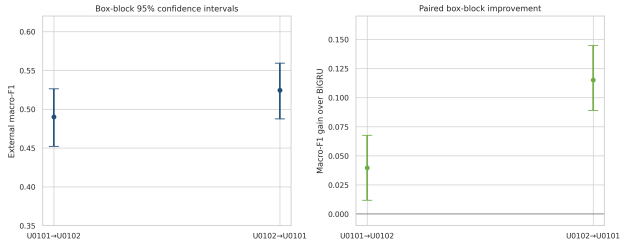

Figure 6 Target Box-Block Macro-F1 Intervals And Paired Gains Over Bigru.

Transfer-Component Ablation

The forward ablation in Table 6 isolates each transformation under the same Conformer-lite classifier. Masked reconstruction raised external macro-F1 from 0.385 to 0.443 and reduced ECE from 0.326 to 0.256. Adding the language gate increased source-validation F1 from 0.803 to 0.827 but reduced external F1 to 0.422. The gate therefore captured source-discriminative body-location semantics without consistently matching the new worker's execution style. The direction-sensitive result is consistent with the broader caution from subject-independent BCI, cross-cloud transfer, cold-start forecasting, and masked log modeling that a useful source representation need not remain optimal under every target shift [6], [67], [15], [16].

Table 6 Forward U0101→U0102 Transfer-Component Ablation Under The Conformer-Lite Classifier.

Variant	Source-val F1	Internal F1	External F1	External acc.	External ECE	External AURC
Raw Conformer-lite	0.768	0.818	0.385	0.425	0.326	0.432
+ masked reconstruction	0.803	0.793	0.443	0.488	0.256	0.358
+ reconstruction + language gate	0.827	0.796	0.422	0.451	0.235	0.422
+ CORAL only	0.795	0.811	0.383	0.383	0.383	0.456
+ reconstruction + CORAL	0.823	0.817	0.423	0.452	0.303	0.400
+ reconstruction + gate + CORAL	0.816	0.768	0.402	0.431	0.316	0.442
LG-SSL-CORAL ensemble	0.887	0.864	0.490	0.546	0.071	0.300

CORAL reduced the adaptation-pool covariance Frobenius distance from 38.234 to 0.254 forward and from 61.332 to 0.297 in reverse (Fig. 7). This exact statistical alignment did not guarantee class alignment. Forward CORAL-only F1 was 0.383, reconstruction plus CORAL was 0.423, and the fully sequential gate/CORAL branch was 0.402. In the reverse direction, reconstruction plus CORAL reached 0.496, demonstrating a strong directional interaction. Matching global covariance can merge or displace class-conditional structure, especially when brief labeling motions and large box movements share feature ranges. This limitation parallels cross-dataset wearable findings [5], approximately shared-feature analysis [7], and topology-sensitive cross-cloud transfer [67], all of which distinguish global alignment from task-conditional invariance.

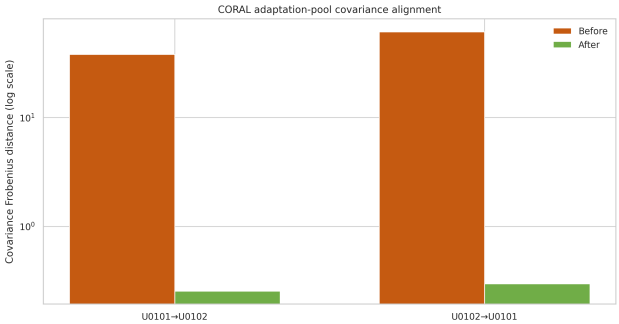

Figure 7 CORAL Adaptation-Pool Covariance Distance Before And After Target-To-Source Alignment.

The ensemble was more robust because validation selected nonzero contributions from raw recurrence, semantic gating, and covariance alignment. It reached forward F1 0.490 even though no standalone transformed branch exceeded 0.443. This is probability-level complementarity, not evidence that every component improves in isolation. The selected weights also retained at least 0.1 for each branch by design, so the method tests a language-and-adaptation ensemble rather than silently dropping a difficult component.

Sensor Selection

Extra-Trees sensor ablation averaged over both transfer directions is reported in Table 7 and Fig. 8. The four ATR devices achieved the best mean macro-F1, 0.523, slightly above all wearables at 0.519. Both wrists reached 0.494, and the right wrist alone reached 0.480. The left wrist, upper arms, and E4-only inputs reached 0.414, 0.396, and 0.390. Adding E4 physiology to all inertial features therefore did not improve this two-person cross-subject mean.

Table 7 Extra-Trees Sensor Ablation Averaged Over Both Transfer Directions.

Sensor set	Features/s	Extra-Trees inputs	Mean accuracy	Mean macro-F1
All ATR	128	640	0.575	0.523
All wearables	184	920	0.564	0.519
Both wrists	120	600	0.541	0.494
Right wrist (ATR+E4)	60	300	0.542	0.480
Left wrist (ATR+E4)	60	300	0.445	0.414
Upper arms	64	320	0.465	0.396
All E4	56	280	0.445	0.390

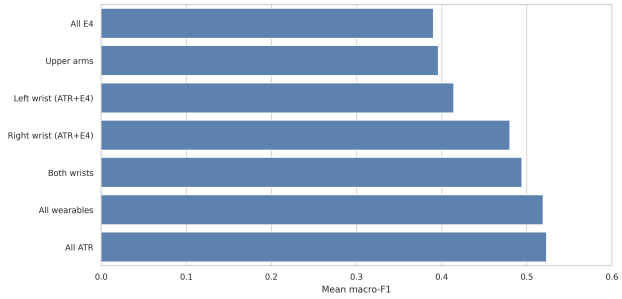

Figure 8 Mean Cross-Subject Extra-Trees Macro-F1 By Sensor Set.

The language matrix in Fig. 3 selected plausible locations: Fill out Order emphasized the right-wrist ATR, Insert Items emphasized the left wrist, Assemble Box emphasized the left upper arm, and Put on Back Table emphasized the right upper arm. Multipliers remained modest, approximately 0.92-1.11, which prevented a text mismatch from removing a sensor. These patterns support the gate's interpretability, while the ablation establishes that semantic plausibility and predictive transfer are different properties. The same distinction appears in calibrated multisensor fusion [28], language-guided network-feature selection [22], and LiDAR-camera tracking [44]: an interpretable or confidence-weighted fusion rule still requires task-specific validation.

D. Calibration and Selective Prediction

Before temperature scaling, the ensemble's target ECE was already low at 0.071 and 0.058 (Table 8). Source-fitted temperatures of 0.728 and 0.692 sharpened the distributions and worsened target ECE to 0.152 and 0.124. Mean negative log likelihood likewise increased from 1.401 to 1.585. Source-only temperature scaling therefore failed to transfer across these participants, even though it optimized source validation. Comparable calibration and abstention pipelines in late fusion, unanswerable QA, recruiter screening, credit risk, and resource allocation [28], [29], [30]-[33] make the same operating-point question explicit, but the present result shows that a source-fitted correction cannot be presumed portable.

Table 8 Transferred Calibration And Selective Prediction For The Final Ensemble.

Direction	ECE raw	Temperature	ECE scaled	Accuracy @ 50%	Risk @ 50 %	Source-rule coverage	Source-rule target risk
U0101→U0102	0.071	0.728	0.152	0.669	0.331	0.969	0.445
U0102→U0101	0.058	0.692	0.124	0.677	0.323	0.775	0.372

Confidence ranking remained useful. At 50% coverage, selective accuracy was 0.669 forward and 0.677 reverse, compared with full-coverage accuracies of 0.546 and 0.548. Figure 9 contrasts raw and temperature-scaled ECE and the 50%/100% coverage accuracies. Temperature scaling is monotonic, so confidence ordering changed only through numerical ties. The source-validation rule designed for at most 10% accepted error yielded target risks of 0.445 at 96.9% coverage and 0.372 at 77.5% coverage. It therefore served as a fixed empirical operating point, not a cross-subject risk guarantee. Conformal alert-control work similarly distinguishes a calibrated alert rule from universal safety [34]. A practical coach should expose coverage and observed target risk rather than label the threshold safe.

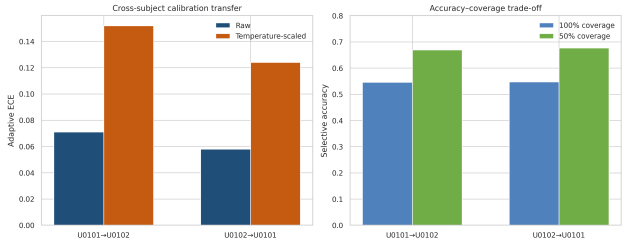

Figure 9 Cross-Subject ECE Transfer And Selective Accuracy At 50% Versus 100% Coverage.

Class Behavior and Coaching Cues

Table 9 aggregates class F1 across directions. Scan Label was easiest (mean 0.759), followed by Fill out Order (0.640) and Picking (0.607). Attach Box Label was hardest (0.185), consistent with its shortest median duration and smallest support. Relocate Item Label and Assemble Box changed markedly by direction: their directional F1 pairs were 0.638/0.309 and 0.338/0.589. These results favor operation-specific feedback over a single global confidence statement.

Table 9 Per-Operation External F1 And Support For Both Transfer Directions.

ID	Operation	Support U0102	F1 U0101→U0102	Support U0101	F1 U0102→U0101	Mean F1
100	Picking	94	0.599	110	0.615	0.607

ID	Operation	Support U0102	F1 U0101→U0102	Support U0101	F1 U0102→U0101	Mean F1
200	Relocate Item Label	180	0.638	157	0.309	0.474
300	Assemble Box	212	0.338	151	0.589	0.464
400	Insert Items	88	0.462	90	0.644	0.553
500	Close Box	141	0.503	92	0.368	0.435
600	Attach Box Label	51	0.265	42	0.105	0.185
700	Scan Label	185	0.821	169	0.697	0.759
800	Attach Shipping Label	76	0.382	75	0.548	0.465
900	Put on Back Table	65	0.364	71	0.619	0.491
1000	Fill out Order	104	0.530	84	0.750	0.640

The language branch generated directly inspectable cues. A correct Fill out Order prediction in U0102 S04 had probability 0.996 and emphasized the right-wrist ATR (1.11x) and right-wrist E4 (1.03x), leading to the instruction to review wrist timing and smoothness. A correct Assemble Box prediction had probability 0.879 and emphasized the left and right upper-arm ATR devices. The same mechanism also exposed a serious failure: one Fill out Order sequence was predicted as Picking with probability 0.980 and emphasized both upper arms. Because the cue follows the predicted class, a high-confidence recognition error produces a confidently wrong coaching focus. The failure aligns with the transferred calibration result and demonstrates why abstention must be validated on the deployment population. Evidence-linked incident, scientific-search, medical, counseling, e-commerce, and capacity-dashboard cards [50]-[57], [58], [59] provide useful interface precedents precisely because they keep uncertainty and evidence visible instead of allowing polished text to conceal a mistaken upstream decision.

Edge-Oriented Profile

The forward TensorFlow.js measurements are shown in Table 10 and Fig. 10. Raw Conformer-lite used 55.6 KiB and 1.09 ms per five-second sequence; masked reconstruction retained the same model size and used 1.10 ms. The language branch used 55.6 KiB, 1.13 ms, and achieved macro-F1 0.422. Raw BiGRU was more accurate at 0.450 but used 90.1 KiB and 6.94 ms. The feature transformations and coaching template are outside the neural weight file, so a production implementation must budget their preprocessing cost separately. This component-wise accounting is consistent with edge proxy work in quality prediction, progressive rendering, IoT autoencoding, and compact intrusion detection [41], [42], [43], [45].

TABLE 10.

Table 10 Tensorflow.Js Forward Host-CPU Proxy; FP32 Batch-One Inference After 20 Warm-Ups And Over 200 Repetitions.

Model	Variant	Parameters	Weights (KiB)	Latency (ms)	Macro-F1
Conformer-lite	Raw	14242	55.6	1.09	0.385
Conformer-lite	SSL	14242	55.6	1.10	0.443
Conformer-lite	SSL + language	14242	55.6	1.13	0.422
Conformer-lite	SSL + language + CORAL	14242	55.6	1.20	0.402
Transformer-lite	Raw	9466	37.0	1.35	0.291
Conformer-lite	SSL + CORAL	14242	55.6	1.41	0.423
Conformer-lite	CORAL	14242	55.6	1.82	0.383
BiGRU	Raw	23062	90.1	6.94	0.450

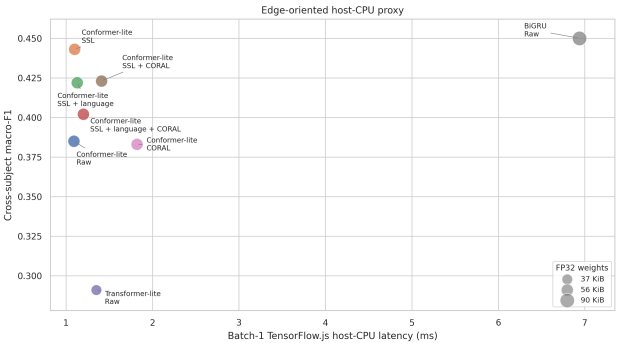

Figure 10 Forward Tensorflow.Js Host-CPU Latency And External Macro-F1; Marker Area Encodes FP32 Weight Size.

The three ensemble branches total 51,546 parameters and 201.3 KiB of FP32 weights. Their combined latency was not benchmarked as one deployed graph, so the individual measurements should not be summed and presented as a device result. The standalone language branch is the clearest edge-oriented coaching configuration: it produces a class, semantic device cue, and confidence with a small classifier, while the ensemble supplies the higher-accuracy server or workstation option.

Human-Facing Reliability and Transferability

The coaching cue should be treated as an evidence display, not an explanation of causality. Evidence-chain and evidence-grounded systems make the same separation between retrieval or provenance and verbalization [46], [47], [48], [49], while incident-visualization, prompt-injection oversight, and scientific-search cards organize source links and uncertainty for inspection [50], [51], [52]. The present template follows this principle by deriving text only from the language branch's predicted class, probability, and fixed semantic device weights.

The high-confidence error suggests that visual design must expose uncertainty and escalation rather than only render a polished instruction. Patient-facing safety cards, medical-image explanation cards, ultrasound cards, counseling recommendation cards, and risk-aware operations

dashboards all make uncertainty or evidence hierarchy visible [53], [54], [55], [56], [57], [59]. E-commerce recommendation cards provide a lower-stakes analog in which trust is calibrated through explanation structure rather than confidence alone [58]. These works are cross-domain design references; the present study did not run a human-subject UI evaluation.

A coach-facing implementation should therefore keep the cue editable, show the accepted or rejected state, preserve the underlying sensor evidence, and allow a human to override or defer. Text-grounded design rationale, design-critic evaluation, structured visual briefs, volleyball-app interface analysis, and rubric-grounded formative feedback provide complementary precedents for such editable oversight [23], [25], [60], [63], [62]. Reproducible vision-language prototyping further suggests that structural and visual consistency should be evaluated separately [61].

Scope and Limitations

Inference is limited to the two evaluated participants and five sessions per participant; the results do not estimate population-level OpenPack performance. The adaptation pool was annotation-filtered before class identities were discarded, so it represents label-blind covariance adaptation within known work intervals rather than continuous unsupervised field adaptation. Preprocessing used bidirectional within-session filling and is not a causal sensor pipeline. Each neural configuration was trained once, and the box bootstrap does not measure optimization variance. Runtime was measured on one host CPU under changing shared load, not on a named wearable device. More broadly, probabilistic forecasting under changing weather and seasonal regimes illustrates why operating characteristics can shift across external regimes [72]; here, that analogy reinforces rather than replaces the need for target-population wearable validation.

Directional results, fixed split counts, box-level intervals, and the negative calibration and CORAL findings characterize this bounded two-participant experiment. Exact archive checks, source-only model selection, and held-out target sessions delimit the conclusions to U0101 and U0102. A preregistered multi-participant leave-one-subject-out evaluation with causal preprocessing and target-free window formation is the appropriate next test.

D. CONCLUSIONS

Language-guided sensor selection, masked reconstruction, and covariance alignment addressed different parts of the cross-subject wearable problem, but none was uniformly beneficial alone. The LG-SSL-CORAL ensemble improved over BiGRU in both directions, reaching macro-F1 0.490 and 0.524 with positive box-block confidence intervals. Extra Trees remained slightly stronger on the two-direction mean, masked reconstruction produced the clearest standalone neural gain, and all four inertial devices outperformed the complete wearable set. Confidence ranking improved accepted accuracy at reduced coverage, whereas source-fitted temperature scaling and a source-derived 10% risk threshold did not transfer. The 55.6-KiB language branch provided fast host-CPU inference and readable sensor cues, but its high-confidence error confirmed that coaching text is only as reliable as the recognized action. These findings support compact, selective, language-grounded wearable coaching while defining the calibration, adaptation, population-validation, and evidence-interface work required before deployment. Future systems should combine target-population risk validation with editable, evidence-linked feedback rather than treating generated coaching prose as a standalone guarantee [51], [55], [57], [63], [62].

E. REFERENCES

[1] F. J. Ordóñez and D. Roggen, “Deep convolutional and LSTM recurrent neural networks for multimodal wearable activity recognition,” Sensors, vol. 16, no. 1, Art. no. 115, 2016, doi: 10.3390/s16010115.

[2] N. Yoshimura, J. Morales, T. Maekawa, and T. Hara, “OpenPack: A large-scale dataset for recognizing packaging works in IoT-enabled logistic environments,” in Proc. 2024 IEEE Int. Conf. Pervasive Computing and Communications (PerCom), Biarritz, France, 2024, pp. 90–97, doi: 10.1109/PerCom59722.2024.10494448.

[3] T. Maekawa, N. Yoshimura, J. Morales, and T. Hara, “Brief introduction of the OpenPack dataset and lessons learned from organizing activity recognition challenge using the dataset,” in Companion Proc. 2024 ACM Int. Joint Conf. Pervasive and Ubiquitous Computing, 2024, pp. 116–120, doi: 10.1145/3675094.3677597.

[4] J. Morales, N. Yoshimura, Q. Xia, A. Wada, Y. Namioka, and T. Maekawa, “MGA-Net+: Acceleration-based packaging work recognition using motif-guided attention networks,” Pervasive Mobile Comput., vol. 88, Art. no. 101735, 2023, doi: 10.1016/j.pmcj.2022.101735.

[5] J. Zhang, "From General Human Activity Recognition to Volleyball-Oriented Wearable Transfer Learning: Cross-Dataset Evidence from UCI HAR and WISDM for Domain Adaptation and Edge Deployment," J. Technol. Informatics Eng., vol. 4, no. 1, pp. 263-283, Apr. 2025, doi: 10.51903/jtie.v4i1.524.

[6] Q. Wu, G. Mi, and D. Wood, "Calibration-Light Subject-Independent Motor Imagery BCI via Self-Supervised Pretraining and Conformer," J. Technol. Informatics Eng., vol. 4, no. 1, pp. 239-262, Apr. 2025, doi: 10.51903/jtie.v5i1.493.

[7] Z. S. Zhong, X. Pan, and Q. Lei, "Bridging Domains with Approximately Shared Features," in Proc. 28th Int. Conf. Artificial Intelligence and Statistics (AISTATS), PMLR, vol. 258, pp. 559-567, 2025.

[8] E. Eldele, M. Ragab, Z. Chen, M. Wu, C. K. Kwoh, X. Li, and C. Guan, “Time-series representation learning via temporal and contextual contrasting,” in Proc. IJCAI, 2021, pp. 2352–2359, doi: 10.24963/ijcai.2021/324.

[9] S. Deldari, H. Xue, A. Saeed, D. V. Smith, and F. D. Salim, “COCOA: Cross modality contrastive learning for sensor data,” Proc. ACM Interact. Mob. Wearable Ubiquitous Technol., vol. 6, no. 3, Art. no. 108, 2022, doi: 10.1145/3550316.

[10] H. Yuan et al., “Self-supervised learning for human activity recognition using 700,000 person-days of wearable data,” npj Digit. Med., vol. 7, Art. no. 91, 2024, doi: 10.1038/s41746-024-01062-3.

[11] H. Haresamudram, I. Essa, and T. Plötz, “Assessing the state of self-supervised human activity recognition using wearables,” Proc. ACM Interact. Mob. Wearable Ubiquitous Technol., vol. 6, no. 3, Art. no. 116, 2022, doi: 10.1145/3550299.

[12] Z. Hong et al., “CrossHAR: Generalizing cross-dataset human activity recognition via hierarchical self-supervised pretraining,” Proc. ACM Interact. Mob. Wearable Ubiquitous Technol., vol. 8, no. 2, Art. no. 64, 2024, doi: 10.1145/3659597.

[13] Q. Xia, T. Maekawa, J. Morales, T. Hara, H. Oshima, M. Fukuda, and Y. Namioka, “Preliminary investigation of SSL for complex work activity recognition in industrial domain via MoIL,” in Proc. 2024 IEEE PerCom Workshops, 2024, pp. 465–468, doi: 10.1109/PerComWorkshops59983.2024.10503195.

[14] G. Mi, T. Ye, and D. Wood, "A Lightweight Medical Foundation Model for Cross-Modal Multi-Task Pretraining and Parameter-Efficient Few-Shot Transfer on MedMNIST," J. Technol. Informatics Eng., vol. 4, no. 3, pp. 572-589, Dec. 2025, doi: 10.51903/jtie.v4i3.492.

[15] S. He, C. Li, and H. Rao, "Few-Shot Cold-Start Workload Forecasting for New AI Inference Tenants with Time-Series Foundation Models," J. Technol. Informatics Eng., vol. 4, no. 1, pp. 306-324, Apr. 2025, doi: 10.51903/jtie.v4i1.546.

[16] Q. Xin, "Self-Supervised Log Anomaly Detection with LogBERT-Style Transformers: Full Empirical Evaluation on a Reproducible SynHDFS Benchmark," J. Electr. Eng. Comput. Sci., vol. 11, no. 1, pp. 23-35, 2026, doi: 10.54732/jeecs.v11i1.3.

[17] S. Miao and L. Chen, “GOAT: A generalized cross-dataset activity recognition framework with natural language supervision,” Proc. ACM Interact. Mob. Wearable Ubiquitous Technol., vol. 8, no. 4, Art. no. 184, 2024, doi: 10.1145/3699736.

[18] Z. Leng, A. Bhattacharjee, H. Rajasekhar, L. Zhang, E. Bruda, H. Kwon, and T. Plötz, “IMUGPT 2.0: Language-based cross modality transfer for sensor-based human activity recognition,” Proc. ACM Interact. Mob. Wearable Ubiquitous Technol., vol. 8, no. 3, Art. no. 112, 2024, doi: 10.1145/3678545.

[19] H. Yan, H. Tan, Y. Ding, P. Zhou, V. Namboodiri, and Y. Yang, “Large language model-guided semantic alignment for human activity recognition,” arXiv:2410.00003v2, Oct. 2024, doi: 10.48550/arXiv.2410.00003.

[20] C. Wang et al., “UbiPhysio: Support daily functioning, fitness, and rehabilitation with action understanding and feedback in natural language,” Proc. ACM Interact. Mob. Wearable Ubiquitous Technol., vol. 8, no. 1, Art. no. 20, 2024, doi: 10.1145/3643552.

[21] W. Chen, J. Cheng, L. Wang, W. Zhao, and W. Matusik, “Sensor2Text: Enabling natural language interactions for daily activity tracking using wearable sensors,” Proc. ACM Interact. Mob. Wearable Ubiquitous Technol., vol. 8, no. 4, Art. no. 192, 2024, doi: 10.1145/3699747.

[22] Y. Li and S. Lu, "Language-Guided Feature Selection for DDoS and Intrusion Detection on CICIDS2017," J. Technol. Informatics Eng., vol. 4, no. 1, pp. 284-305, Apr. 2025, doi: 10.51903/jtie.v4i1.531.

[23] Q. Wu, S. Meng, and J. Zhao, "Text-Grounded LLM-Assisted Design Rationale Interfaces: Turning Advertising Layout Metadata into Explainable UI/UX Decision Cards," Int. J. Graph. Des., vol. 3, no. 1, pp. 216-240, May 2025, doi: 10.51903/ijgd.v3i1.3713.

[24] H. Xu, Y. Chen, and A. Med, "Automatic Detection and Explanation of Dark Patterns from Interface Microcopy: Empirical Comparison of BERT-Style Encoders, RoBERTa-Style Encoders, and LLM-Style Decoders on the ec-darkpattern Dataset," J. Technol. Informatics Eng., vol. 4, no. 3, pp. 590-612, Dec. 2025, doi: 10.51903/jtie.v4i3.491.

[25] Y. Li, S. Lu, and L. Zhao, "LLM-as-Design-Critic: Aligning AI-Generated UI Feedback with Human Graphic Design Judgment," Int. J. Graph. Des., vol. 3, no. 1, pp. 196-215, May 2025, doi: 10.51903/ijgd.v3i1.3661.

[26] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, “On calibration of modern neural networks,” in Proc. 34th Int. Conf. Machine Learning, vol. 70, 2017, pp. 1321–1330.

[27] Y. Geifman and R. El-Yaniv, “SelectiveNet: A deep neural network with an integrated reject option,” in Proc. 36th Int. Conf. Machine Learning, vol. 97, 2019, pp. 2151–2159.

[28] Q. Xin, "Uncertainty-Aware Late Fusion for 3D Perception (Confidence Calibration + Fusion Rule Learning)," J. Technol. Informatics Eng., vol. 4, no. 1, pp. 215-238, Feb. 2025, doi: 10.51903/jtie.v4i1.485.

[29] Z. S. Zhong, J. Chen, E. Zhong, and X. Sun, "Evidence-Calibrated RAG for Unanswerable Question Answering: Retrieval Coverage, Abstention Calibration, and Hallucination-Proxy Analysis on SQuAD 2.0," J. Technol. Informatics Eng., vol. 4, no. 2, pp. 502-520, Aug. 2025, doi: 10.51903/jtie.v4i2.536.

[30] J. Jin, "Calibrated Resume-Job Matching for Trustworthy LLM-Assisted Recruiter Screening: Pairwise Matching, Probability Calibration, and Selective Refusal on Two Public Recruitment Datasets," J. Technol. Informatics Eng., vol. 4, no.

3, pp. 625-648, Dec. 2025, doi: 10.51903/jtie.v4i3.529.

[31] Y. Chen, Y. Zhang, D. Chau, and M. Sherman, "Credit Card Default Risk Tiering with Probability Calibration and Uncertainty-Driven Rejection: A Reproducible Study on the UCI Credit Card Clients Dataset," J. Adv. Comput. Syst., vol. 3, no. 4, pp. 31-47, Apr. 2023, doi: 10.69987/JACS.2023.30403.

[32] J. Jin, T. Huang, and S. Lu, "A Model-Risk-Friendly Probability of Default Workflow: Calibration, Distribution-Free Uncertainty Quantification, and SHAP Explanations on the UCI Credit Card Default Dataset," J. Adv. Comput. Syst., vol. 4, no. 6, pp. 74-85, Jun. 2024, doi: 10.69987/JACS.2024.40606.

[33] S. Chen, S. He, and E. Sun, "Risk-Bounded GPU Resource Oversubscription via Conformal Demand Envelopes in Production AI Clusters," J. Adv. Comput. Syst., vol. 4, no. 5, pp. 119-134, May 2024, doi: 10.69987/JACS.2024.40509.

[34] Q. Xin, "Log Anomaly Detection with Conformal Alert Control and Evidence-Grounded Incident Ticket Generation," AVITEC, vol. 8, no. 2, May 2026, doi: 10.28989/avitec.v8i2.3974.

[35] Y. Zhou, H. Zhao, Y. Huang, T. Riedel, M. Hefenbrock, and M. Beigl, "TinyHAR: A lightweight deep learning model designed for human activity recognition," in Proc. ACM ISWC, 2022, pp. 89–93, doi: 10.1145/3544794.3558467.

[36] D. Smilkov et al., "TensorFlow.js: Machine learning for the web and beyond," in Proc. 2nd MLSys Conf., 2019, pp. 309–321.

[37] T. Kautz, B. H. Groh, J. Hannink, U. Jensen, H. Strubberg, and B. M. Eskofier, "Activity recognition in beach volleyball using a deep convolutional neural network: Leveraging the potential of deep learning in sports," Data Min. Knowl. Discov., vol. 31, no. 6, pp. 1678–1705, 2017, doi: 10.1007/s10618-017-0495-0.

[38] Y. Wang, Y. Zhao, R. H. M. Chan, and W. J. Li, "Volleyball skill assessment using a single wearable micro inertial measurement unit at wrist," IEEE Access, vol. 6, pp. 13758–13765, 2018, doi: 10.1109/ACCESS.2018.2792220.

[39] Z. Wen, R. Zhang, and C. Wang, "Optimization of Bi-Directional Gated Loop Cell Based on Multi-Head Attention Mechanism for SSD Health State Classification Model," in Proc. 2025 6th Int. Conf. Electronic Communication and Artificial Intelligence (ICECAI), Chengdu, China, 2025, doi: 10.1109/ICECAI66283.2025.11171441.

[40] C. Wang, Z. Wen, R. Zhang, P. Xu, and Y. Jiang, "GPU Memory Requirement Prediction for Deep Learning Task Based on Bidirectional Gated Recurrent Unit Optimization Transformer," in Proc. 2025 5th Int. Conf. Artificial Intelligence, Virtual Reality and Visualization (AIVRV), Chengdu, China, pp. 31-35, 2025, doi: 10.1109/AIVRV67401.2025.11350369.

[41] B. Zhou, H. Wang, and X. Chang, "Distilling VMAF into an Edge-Deployable Quality Predictor: A Pilot Shot-Level Proxy with LLM-Ready Quality Tokens," J. Technol. Informatics Eng., vol. 4, no. 2, pp. 447-463, Aug. 2025, doi: 10.51903/jtie.v4i2.522.

[42] H. Wang, Y. Ren, and X. Chang, "Layout-Aware Progressive PDF Rendering: AI Prioritization of PDF Slices to Reduce Time-to-Functional-First-Frame on FUNSD," J. Technol. Informatics Eng., vol. 4, no. 2, pp. 425-446, Aug. 2025, doi: 10.51903/jtie.v4i2.523.

[43] Q. Xin, Z. Xu, L. Guo, F. Zhao, and B. Wu, "IoT Traffic Classification and Anomaly Detection Method Based on Deep Autoencoders," Appl. Comput. Eng., vol. 69, pp. 64-70, 2024, doi: 10.54254/2755-2721/69/20241511.

[44] Q. Xin, "LiDAR-Camera Object-Level Fusion for Multi-Target Tracking Using JPDA and EKF: A Reproducible Empirical Study on a PandaSet-Parameterised Five-Sequence Dataset," J. Technol. Informatics Eng., vol. 5, no. 1, pp. 54-76, Apr. 2026, doi: 10.51903/jtie.v5i1.486.

[45] S. Lu and D. Zhou, "TinyLLM-Assisted Intrusion Detection for Real-Time IoT Networks," J. Adv. Comput. Syst., vol. 4, no. 8, pp. 72-87, Aug. 2024, doi: 10.69987/JACS.2024.40809.

[46] J. Jin, "Evidence-Chain Reliable RAG: Hallucination Detection, Source Attribution, and Deterministic Provenance Explanations," J. Technol. Informatics Eng., vol. 4, no. 2, pp. 520-533, Aug. 2025, doi: 10.51903/jtie.v4i2.535.

[47] B. Zhang, H. Rao, and D. Zhao, "Evidence-Grounded RAG for Cloud-Native DevOps: Hallucination-Resistant AIOps Question Answering over Private Operations Documents," J. Adv. Comput. Syst., vol. 4, no. 3, pp. 109-125, Mar. 2024, doi: 10.69987/JACS.2024.40308.

[48] G. Liu, C. Li, and E. Zhang, "OpsLLM for Cloud Incident Triage: Bilingual RAG-Based Root Cause Analysis and Alert Summarization for AI Infrastructure Operations," J. Adv. Comput. Syst., vol. 4, no. 4, pp. 97-111, Apr. 2024, doi: 10.69987/JACS.2024.40408.

[49] C. Li, J. Bai, and S. Wang, "Evidence-Chain Reliable RAG: Word-Level Hallucination Detection, Source Attribution, and Provenance Explanation for LLM Applications," J. Adv. Comput. Syst., vol. 4, no. 2, pp. 76-92, Feb. 2024, doi: 10.69987/JACS.2024.40207.

[50] J. Nie, G. Liu, C. Li, and T. Zou, "Evidence-Constrained Incident Visualization Cards for Distributed Cloud Logs: A UI/UX Framework for Turning Hadoop, OpenStack, and ZooKeeper Logs into Actionable SRE Design Interfaces," Int. J. Graph. Des., vol. 4, no. 1, pp. 179-185, Apr. 2026, doi: 10.51903/ijgd.v4i1.3703.

[51] W. Su, H. Rao, and E. Ma, "Privacy and Data-Integrity Risk Cards for LLM Agents: A UI/UX Design Framework for Secure Human Oversight

under Prompt-Injection Attacks," Int. J. Graph. Des., vol. 4, no. 1, pp. 186-191, Apr. 2026, doi: 10.51903/ijgd.v4i1.3699.

[52] J. Jin, "LLM-Style Evidence Cards for Scientific Search Interfaces: A UI/UX Design Framework for Retrieval Transparency, Ranking Trust, and Visual Evidence Hierarchy," Int. J. Graph. Des., vol. 3, no. 2, pp. 397-414, Oct. 2025, doi: 10.51903/ijgd.v3i2.3698.

[53] B. Zhou, C. Li, and L. Liu, "Risk-Calibrated Patient-Facing AI Safety Cards: A UI/UX Design Framework for Rubric-Based Medical Risk Communication," Int. J. Graph. Des., vol. 3, no. 2, pp. 365-380, Oct. 2025, doi: 10.51903/ijgd.v3i2.3696.

[54] C. Li, B. Zhou, and K. Gao, "Risk-Calibrated Patient-Facing AI Safety Cards: A UI/UX Benchmark for Explainable Medical AI Response Interfaces," Int. J. Graph. Des., vol. 3, no. 2, pp. 381-394, Oct. 2025, doi: 10.51903/ijgd.v3i2.3709.

[55] Z. S. Zhong, Q. Wu, and G. Mi, "Uncertainty-Aware Medical Image Explanation Cards: LLM-Generated Visual Explanations for AI-Assisted Radiology Interfaces," Int. J. Graph. Des., vol. 3, no. 2, pp. 415-436, Oct. 2025, doi: 10.51903/ijgd.v3i2.3616.

[56] T. Ye, X. Chang, and E. Zhong, "Uncertainty-Aware Breast Ultrasound Explanation Cards: A Visual Communication Framework for Image-Based AI Diagnostic Support Using BreastMNIST_224," Int. J. Graph. Des., vol. 3, no. 2, pp. 365-380, Oct. 2025, doi: 10.51903/ijgd.v3i2.3701.

[57] Y. Zhang and H. Zhang, "Visualizing the Right Counseling Support: Evidence-Linked Recommendation Cards for Explainable Mental Health Intake Interfaces," Int. J. Graph. Des., vol. 3, no. 1, pp. 214-229, May 2025, doi: 10.51903/ijgd.v3i1.3722.

[58] B. Zhang, Y. Ren, and J. Zou, "LLM-Style Explainable E-Commerce Recommendation Cards: A UI/UX Design Framework for Trust-Calibrated Product Recommendation," Int. J. Graph. Des., vol. 3, no. 2, pp. 381-396, Oct. 2025, doi: 10.51903/ijgd.v3i2.3697.

[59] G. Liu, S. He, and H. Wong, "LLM-Compatible Visual Brief Cards for AI Infrastructure Capacity Dashboards: A UI/UX Framework for Turning Forecast Risk into Graphic Design Decisions," Int. J. Graph. Des., vol. 3, no. 1, pp. 196-213, May 2025, doi: 10.51903/ijgd.v3i1.3723.

[60] J. Mu, Y. Lu, and E. Hwang, "Structured Visual Brief Interfaces for Advertising Design: A UI/UX Framework for Turning Creative Intentions into Designer-Editable Graphic Design Cards," Int. J. Graph. Des., vol. 4, no. 1, pp. 192-208, Apr. 2026, doi: 10.51903/ijgd.v4i1.3702.

[61] Y. Chen and M. Li, "From Hand-Drawn Sketches to Interactive Web Prototypes: A Reproducible Vision-Language Approach with Structural and Visual Consistency Evaluation," J. Technol. Informatics Eng., vol. 4, no. 2, pp. 364-384, Aug. 2025, doi: 10.51903/jtie.v4i2.490.

[62] Q. Xin, "Auditable Automated Essay Scoring and Formative Feedback: A Rubric-Grounded Pipeline for Secondary and Higher Education," J. Appl. Artif. Intell. Educ., vol. 2, no. 1, pp. 1-19, Jul. 2026, doi: 10.66053/jaaie.v2i1.348.

[63] J. Zhang, "Adaptive User Interface Design for Volleyball Learning Apps: Empirical Evidence from Google Play Reviews and Mobile Screen Analysis," Int. J. Graph. Des., vol. 3, no. 1, pp. 175-195, May 2025, doi: 10.51903/ijgd.v3i1.3618.

[64] N. Yoshimura, J. Morales, and T. Maekawa, “OpenPack: Public multi-modal dataset for packaging work recognition in logistics domain,” Zenodo, ver. 1.1.0, Apr. 24, 2024, doi: 10.5281/zenodo.11059235.

[65] B. Sun and K. Saenko, “Deep CORAL: Correlation alignment for deep domain adaptation,” in Computer Vision–ECCV 2016 Workshops, 2016, pp. 443–450, doi: 10.1007/978-3-319-49409-8_35.

[66] X. Qin, Y. Chen, J. Wang, and C. Yu, “Cross-dataset activity recognition via adaptive spatial-temporal transfer learning,” Proc. ACM Interact. Mob. Wearable Ubiquitous Technol., vol. 3, no. 4, Art. no. 148, 2019, doi: 10.1145/3369818.

[67] S. He, X. Chang, and E. Sun, "Cross-Cloud Transfer Learning for AI Training Capacity Forecasting under Workload and Topology Distribution Shift," J. Adv. Comput. Syst., vol. 4, no. 1, pp. 100-120, Jan. 2024, doi: 10.69987/JACS.2024.40108.

[68] P. Geurts, D. Ernst, and L. Wehenkel, “Extremely randomized trees,” Mach. Learn., vol. 63, no. 1, pp. 3–42, 2006, doi: 10.1007/s10994-006-6226-1.

[69] K. Cho et al., “Learning phrase representations using RNN encoder–decoder for statistical machine translation,” in Proc. EMNLP, Doha, Qatar, 2014, pp. 1724–1734, doi: 10.3115/v1/D14-1179.

[70] A. Vaswani et al., “Attention is all you need,” in Advances in Neural Information Processing Systems 30, 2017, pp. 5998–6008.

[71] A. Gulati et al., “Conformer: Convolution-augmented Transformer for speech recognition,” in Proc. Interspeech, 2020, pp. 5036–5040, doi: 10.21437/Interspeech.2020-3015.

[72] Q. Xin, "Probabilistic Bike-Sharing Demand Forecasting under Changing Weather and Seasonal Regimes with Transformer-Based Models," Findings, Mar. 2026, doi: 10.32866/001c.157499.