Skepthical review: *Challenges in Learning Universal Gait Fingerprints: Evaluating Adversarial Invariance and Demographic Bias for Wearable Step Counting*

Summary

This paper investigates whether "universal gait fingerprints" can be learned from wearable tri-axial accelerometer data—i.e., low-dimensional step representations that enable step detection/counting while being invariant to sensor placement (hip vs wrist) and sampling frequency (25 vs 100 Hz), and that behave equitably across demographics. The authors propose a 1D-CNN encoder trained with multi-task adversarial objectives using a Gradient Reversal Layer (Sec. 2.3) to suppress information about location and sampling frequency. Experiments on OxWalk (39 participants; four concurrent configurations) use 2-second windows with 50% overlap and an upsampling step to a common length (Sec. 2.2), and evaluate window-level step/non-step classification and trial-level step counting (Sec. 2.4, Sec. 3.3), alongside representation diagnostics via linear probes and t-SNE (Sec. 3.4) and subgroup analyses by sex/age (Sec. 3.5).

Empirically, step counting performance is relatively weak overall (reported test **MAPE** $\approx 44\%$), much better on hip than wrist data, and shows counterintuitive condition effects (e.g., wrist 25 Hz outperforming wrist 100 Hz). The invariance diagnostics indicate that embeddings still strongly encode sensor location (linear probe $\approx$ **96%** accuracy) while frequency information is only partially attenuated ($\approx$ **59%** accuracy). Demographic stratification suggests large disparities, especially for older adults (**MAPE** up to $\sim$ **75%**) and differences by sex, but these claims are currently fragile due to the very small test cohort. The paper's main value is as a careful, largely negative/diagnostic study showing that a single monolithic adversarially trained encoder does not achieve robust location invariance or equitable performance in this setting. However, key gaps in reproducibility (architecture/training details), the step-count aggregation protocol, missing baselines/ablations, and limited statistical robustness (single split; small test set) weaken the strength and generality of the conclusions. Addressing these would substantially strengthen the manuscript and better support the "universal fingerprint" claims and their limitations.

Strengths

- Well-motivated problem framing around cross-device generalization (placement and sampling rate) and demographic equity in wearable step counting (Sec. 1).
- Coherent use of adversarial representation learning (GRL) to target nuisance invariance, paired with a diagnostic evaluation via post hoc linear probes on frozen embeddings (Sec. 2.3, Sec. 3.4).

- Appropriate use of OxWalk's concurrent multi-condition recordings (hip/wrist × 25/100 Hz) for studying invariance questions more directly than many datasets allow (Sec. 2.1).
- Evaluation goes beyond a single metric: per-condition performance, representation probes, and demographic stratification provide a richer picture than aggregate accuracy alone (Sec. 3.3–3.5).
- Honest reporting and discussion of negative findings; the manuscript does not oversell success and includes thoughtful future directions (Sec. 3.6–3.7, Conclusions).
- Figures generally aim to connect training dynamics, downstream performance, and embedding behavior, which helps interpret what the model is and is not learning.

Major issues

1. **Step counting from overlapping window predictions is under-specified and may be ill-posed as described, which undermines interpretation of MAE/MAPE (Sec. 2.4.2, Sec. 3.3–3.3.2).** The paper states that window-level step/non-step outputs are "aggregated" into trial-level step counts, but does not define the aggregation algorithm. With 50% overlap and a "positive if annotation is in the central 25% of the window" labeling rule (Sec. 2.2), naive summation of positive windows will not generally equal the number of steps; different debouncing/peak-picking rules can change **MAE**/**MAPE** substantially.

   Recommendation: In Sec. 2.4.2, provide an explicit, reproducible algorithm (or mathematical definition) mapping window probabilities/labels to step-event counts per trial: threshold choice (fixed vs tuned on validation), how contiguous positives are merged, any refractory period, peak-finding/non-maximum suppression, and how window indices map to time. Define clearly how **MAE** and **MAPE** are computed/averaged (per-trial then averaged vs global sums; macro vs micro averaging; handling of small denominators). Consider adding an event-based evaluation (step-event precision/recall with a tolerance window) that more directly matches the "central 25%" event-labeling scheme, and/or a regression formulation (steps-per-window / step-rate) as a sensitivity analysis.

2. **Insufficient architectural and training details prevent replication and make it hard to interpret the negative results (Sec. 2.3.1–2.3.2).** Critical missing information includes the full 1D-CNN layout (layers, kernel sizes/strides/pooling, activations, normalization, dropout), embedding dimensionality, adversarial head architectures/capacities, optimizer and hyperparameters (learning rate, schedule, batch size, weight decay), number of epochs/steps, early stopping/checkpoint selection (epoch 20 is mentioned), and the exact loss formulation/weighting (task vs location/frequency adversaries; GRL λ and any α/β weights).

Recommendation: Expand Sec. 2.3 with a precise specification of (i) encoder and head architectures (ideally a table of layers/parameters) and (ii) training protocol. In Sec. 2.3.2, write the full objective explicitly (e.g., $L = L_{\text{task}} + \alpha\, L_{\text{loc}} + \beta\, L_{\text{freq}}$ with GRL λ and whether λ is shared or per-adversary), and state all hyperparameters and selection criteria. Add an Appendix with a full config and/or pseudocode sufficient for exact reproduction.

3. **No baselines and limited ablations make it unclear whether the observed failures are due to adversarial training, model capacity, optimization choices, or dataset difficulty (Sec. 3.3–3.4, Sec. 3.6).** The manuscript does not report: (i) a task-only version of the same CNN, (ii) single-adversary variants (location-only, frequency-only), (iii) sensor-specific models (hip-only, wrist-only), or (iv) simple classical/heuristic baselines. Without these, it is hard to attribute performance/invariance outcomes to the GRL strategy or to judge whether 44% **MAPE** is expected/competitive.

   Recommendation: Add at least: (1) task-only CNN baseline, (2) frequency-adversary-only and location-adversary-only ablations, and (3) a simple baseline (e.g., a classical peak-based step counter or shallow classifier on handcrafted features), plus optionally hip-only/wrist-only specialized models. Report both downstream metrics (Sec. 3.3) and probe metrics (Sec. 3.4) for all baselines. If compute is a constraint, prioritize task-only and single-adversary ablations and clearly qualify conclusions about adversarial learning.

4. **Demographic/fairness conclusions are fragile given the very small test set and lack of uncertainty quantification (Sec. 3.1, Sec. 3.5–3.5.2).** The paper reports large disparities (e.g., older adults much worse), but with only 6 test participants, subgroup metrics can be dominated by one participant/trial. The manuscript does not report subgroup sample sizes in the test set (participants and trials), participant-level distributions, or confidence intervals.

   Recommendation: In Sec. 3.1 and Sec. 3.5, report exact counts: number of test participants (and trials/windows) per sex and per age bin. Add participant-level plots or summaries (e.g., per-participant **MAE**/**MAPE**) and bootstrap confidence intervals at the participant level. If feasible, repeat experiments over multiple participant-level random splits (e.g., 10–50 repeats) or use leave-one-subject-out / group k-fold CV, reporting variability for overall, invariance, and subgroup metrics. Temper fairness claims in Sec. 3.6–Conclusions to reflect uncertainty and the current cohort size.

5. **Claims about sampling-frequency invariance are not statistically supported and lack appropriate baselines/metrics (Sec. 3.4, Sec. 3.6, Conclusions).** A frequency probe accuracy of $\sim$ **59%** vs "chance 50\%" may or may not be meaningful depending on class balance and variance, and without a non-adversarial comparison it is unclear how much the adversary reduced frequency information.

Recommendation: In Sec. 3.4, report class proportions for the probe datasets and use balanced accuracy and/or AUC in addition to raw accuracy; define the chance baseline explicitly (e.g., majority-class accuracy). Provide uncertainty (bootstrap CIs). Add probe results for a task-only model to quantify how much frequency information is removed by the adversarial objective. Rephrase claims in Sec. 3.6/Conclusions to "partial attenuation" unless the reduction is clearly demonstrated and statistically meaningful.

6. **Resampling/upsampling design may confound the "frequency invariance" question (Sec. 2.2, Sec. 3.4).** Upsampling 25 Hz windows to 200 samples via interpolation enforces a common input size but also changes signal smoothness/bandwidth and may create artifacts that either hide or introduce frequency cues. This makes "invariance to sampling frequency" hard to disentangle from "invariance to interpolation-induced differences."

   Recommendation: Add a sensitivity analysis comparing alternative standardizations: (i) downsample 100 Hz to 25 Hz, (ii) resample both to a common intermediate rate, and/or (iii) architectures that avoid fixed-length resampling (variable-length models, time encodings). Report how downstream performance and frequency-probe separability change across these preprocessing choices (Sec. 3.3–3.4).

7. **Probe methodology details are incomplete, and potential leakage/overfitting in probe training is not ruled out (Sec. 2.4.2, Sec. 3.4).** The paper says probes are trained on frozen embeddings, but does not fully specify which splits are used for probe training/validation/testing, whether hyperparameters are tuned, and whether evaluation is strictly on held-out participants. Given a large number of overlapping windows, probes can also be sensitive to subtle artifacts; reporting only accuracy can be misleading under imbalance.

   Recommendation: In Sec. 2.4.2, specify the probe protocol precisely: which participant splits are used, whether probes are trained on train (or train+val) embeddings and evaluated on test embeddings, any cross-validation, regularization, and hyperparameter tuning procedure. Ensure the probe is not trained and evaluated on the same embeddings. In Sec. 3.4, report balanced accuracy/AUC and class proportions, and consider reducing dependence among overlapping windows by subsampling windows per trial or aggregating embeddings per step/trial for the probe as a robustness check.

8. **Table 3 is misformatted and appears to mix step-counting metrics with probe-related columns, making core results difficult to read and potentially error-prone (Sec. 3.3.2–3.4).** This is a high-impact presentation issue because it affects the interpretability of the main per-condition results.

   Recommendation: Rebuild the tables so that per-condition step-count metrics (hip/wrist × 25/100 Hz; **MAE/MAPE**) are presented cleanly in Table 3 (Sec. 3.3.2), and move probe tasks/accuracies/interpretations into a separate Table 4 (Sec. 3.4).

Cross-check that all values quoted in the text match the corrected tables and remove any LaTeX/OCR artifacts causing column mixing.

Minor issues

1. Class imbalance (reported $\sim$ **23%** positive windows) and decision-thresholding are not fully specified, complicating interpretation of the conservative bias/high FN rate (Sec. 2.2–2.3.2, Sec. 3.3).

   Recommendation: State explicitly whether class weighting, resampling, focal loss, or threshold tuning was used; report the probability threshold used to binarize window outputs and whether it was selected on validation data. Consider reporting PR-AUC and/or calibrating thresholds per condition as an additional analysis.

2. Terminology sometimes conflates window-level step/non-step classification with step counting, which can obscure what the model is actually trained to do versus what is derived post hoc (Sec. 1, Sec. 2.2–2.4, Sec. 3.3, Conclusions).

   Recommendation: Standardize language: clearly state that the primary task is window-level classification and that trial-level counts are derived via a specified aggregation method (Sec. 2.4.2). Align Conclusions accordingly.

3. Age bin definitions are inconsistent across the manuscript (e.g., Table 1 vs Sec. 3.1/Table 5), which makes subgroup comparisons hard to reconcile (Sec. 2.1.2, Sec. 3.1, Sec. 3.5.2).

   Recommendation: Harmonize age bins across Table 1, Sec. 3.1, and Table 5 (including endpoint conventions). If re-binning was done for analysis, explicitly describe the mapping and rationale and update all counts accordingly.

4. The relationship between sex-group **MAE** and **MAPE** appears internally inconsistent (large **MAE** gap but modest **MAPE** gap), suggesting scale differences in true step counts or outliers (Sec. 3.5.1).

   Recommendation: Report group-wise distributions of true steps per trial and include robust summaries (median absolute error) alongside means. Add scatter/box plots at participant level (true vs predicted; error vs true) to diagnose whether results are driven by a few long trials or outliers.

5. The better wrist-25 Hz vs wrist-100 Hz performance explanation is presented somewhat speculatively without supporting analysis (Sec. 3.3.2).

   Recommendation: Label this explanation explicitly as a hypothesis and, if feasible, add a brief supporting check (e.g., PSD comparisons or qualitative signal examples). Otherwise, soften the claim.

6. Figure 1–2 presentation makes it harder to interpret convergence and error distributions (loss scale, lack of uncertainty, limited stratification by condition) (Sec. 3.2–3.3; figures).

   Recommendation: Plot mean loss per sample/batch with clearer scaling; add adversarial classifier accuracy with a chance baseline; annotate the selected checkpoint/early stopping. For error distributions, add condition-stratified views (hip/wrist × 25/100) and include uncertainty (bootstrap CIs) and/or ECDF/CCDF plots.

7. Related work and positioning could be expanded to better contextualize novelty and the negative result (Sec. 1–2).

   Recommendation: Add a focused Related Work subsection covering (i) placement-robust step counting, (ii) domain-adversarial learning (e.g., DANN) for sensor time series, and (iii) fairness in wearable/digital health models; explicitly position this work as a diagnostic/negative result under the tested setup.

8. Ethics and dataset governance are not explicitly discussed despite using human-subject wearable data (Sec. 2.1).

   Recommendation: Add a short ethics/data statement summarizing OxWalk's consent/IRB status as documented by the dataset, anonymization, and how this study complies with the dataset's terms.

Very minor issues

1. Front-matter and formatting contain template/OCR artifacts (e.g., unrelated keywords like "Astronomy data analysis," inconsistent headings/LaTeX formatting, minor notation inconsistencies such as C_{freq} vs C_{freq}) (Abstract, Sec. 2–3).

   Recommendation: Clean keywords and standardize formatting/headings and notation (e.g., use C_{loc} and C_{freq} consistently). Proofread to remove LaTeX/OCR artifacts in units and percentages.

2. Multi-panel figures are referenced as "left/middle/right" without explicit (a)/(b)/(c) labels, and some captions omit sample sizes/binning choices (Figures 1–5).

   Recommendation: Add panel labels (a,b,c) directly on figures and reference them in text; include key caption details (N windows/trials, bin widths, kernel/bandwidth if KDE is used).

3. The statement that 50% overlap ensures "comprehensive coverage" is somewhat in tension with the central-25% positive labeling rule, which can create systematic gaps depending on stride timing (Sec. 2.2).

   Recommendation: Clarify what "coverage" means under the chosen labeling rule and briefly justify the design trade-off (e.g., precision vs recall of step-event labeling).

Key statements and references

- • **Adversarial training with a Gradient Reversal Layer achieved only partial invariance to sampling frequency, as evidenced by a linear probe trained on the learned embeddings attaining an accuracy of 59.20% for discriminating 25 Hz vs. 100 Hz data, which is only moderately above the 50% chance level for binary classification (Table 4).**
- *Reference(s):* Table 4
- • **The same adversarial training framework critically failed to achieve invariance to sensor location, with a linear probe trained on the learned embeddings classifying hip vs. wrist data with an accuracy of 96.47%, indicating that the purported universal gait fingerprints still strongly encode sensor placement information (Table 4).**
- *Reference(s):* Table 4
- • **Step-counting performance of the proposed universal model was highly dependent on sensor configuration, with hip-worn data at 100 Hz yielding an MAE of 164.33 steps and a MAPE of 38.96%, while wrist-worn data at 100 Hz produced the worst performance with an MAE of 617.83 steps and a MAPE of 51.66%, demonstrating substantial degradation for wrist signals (Table 3).**
- *Reference(s):* Table 3
- • **The model exhibited pronounced demographic bias by age, with Mean Absolute Percentage Error increasing from 21.24% for participants aged 19–30 years to 75.04% for participants aged 45–81 years, indicating that the learned gait representation generalizes poorly to older adults (Table 5).**
- *Reference(s):* Table 5
- • **Sex-stratified evaluation revealed substantial disparity in absolute error, with female participants experiencing a Mean Absolute Error of 576.25 steps compared to 89.67 steps for male participants, despite relatively similar MAPE values (46.02% vs. 42.70%), suggesting more severe misestimations for women (Figure 5).**
- *Reference(s):* Figure 5

Mathematical consistency audit

This section audits **symbolic/analytic** mathematical consistency (algebra, derivations, dimensional/unit checks, definition consistency).

Maths relevance: light

The paper contains few explicit equations and largely describes an ML pipeline in prose (windowing, upsampling, GRL-based adversarial training, and evaluation metrics). The main analytic audit points are internal consistency of dimensional reasoning, correctness of stated GRL gradient behavior, clarity/definitions of evaluation metrics and baselines, and consistency of subgroup definitions used for stratified reporting.

Checked items

1. ✔ **Window sample counts from sampling frequency** (Sec. 2.2 (Data Windowing), p.3)
    - **Claim:** A 2-second window contains 200 samples at 100 Hz and 50 samples at 25 Hz.
    - **Checks:** algebra/arithmetic consistency, dimensional reasoning
    - **Verdict:** PASS; confidence: high; impact: minor
    - **Assumptions/inputs:** Sampling frequency in Hz equals samples per second., Window duration is exactly 2 seconds.
    - **Notes:** $\mathbf{2\ s \times 100\ samples/s = 200}$ samples; $\mathbf{2\ s \times 25\ samples/s = 50}$ samples. Units and arithmetic are consistent.
2. ✔ **Uniform input tensor shape after upsampling** (Sec. 2.2 (Handling Mismatched Frequencies), p.3)
    - **Claim:** Upsampling 25 Hz windows from 50 to 200 points yields a consistent input shape of (200, 3) for all windows.
    - **Checks:** shape/dimension consistency, definition consistency
    - **Verdict:** PASS; confidence: high; impact: minor
    - **Assumptions/inputs:** Three accelerometer axes are used (x,y,z)., Upsampling changes only the time dimension.
    - **Notes:** Given 3 axes, standardizing time points to 200 makes (time, axes) = (200, 3) consistent across frequencies.
3. ⚠ **50% overlap vs central-25% labeling coverage** (Sec. 2.2 (Windowing + Window Labeling), p.3)
    - **Claim:** 50% overlap mitigates boundary effects and ensures comprehensive coverage of events, while labeling a window positive only if a step lies in the central 25% reduces ambiguity.
    - **Checks:** logical consistency, sanity/edge-case reasoning
    - **Verdict:** UNCERTAIN; confidence: medium; impact: moderate
    - **Assumptions/inputs:** Window length is 2 seconds., Stride/step event times are not synchronized to the window grid.
    - **Notes:** With 2 s windows and 50% overlap, window starts are spaced by 1 s. The central 25% spans only 0.5 s per window. Central regions of consecutive windows are disjoint (leaving 0.5 s gaps per 1 s step). Steps falling in those

gaps would be labeled non-step despite occurring within a window, which contradicts an unqualified claim of "comprehensive coverage." This may be intended as a trade-off, but it is not analytically clarified.

4. ✔ **GRL backward-pass sign behavior** (Sec. 2.3.2, p.4)
    - **Claim:** The GRL multiplies gradients from adversarial losses by $-\lambda$ during backpropagation, causing the encoder to update in a direction that maximizes adversarial classifier loss.
    - **Checks:** calculus/optimization sign logic, definition consistency
    - **Verdict:** PASS; confidence: medium; impact: moderate
    - **Assumptions/inputs:** Adversarial classifiers minimize their own cross-entropy losses., Encoder receives reversed gradient contribution from those losses.
    - **Notes:** Reversing (multiplying by $-\lambda$) the gradient contribution from adversarial losses to the encoder is consistent with encouraging the encoder to increase those losses while the adversaries themselves minimize them. However, the combined objective is not explicitly written, limiting deeper verification.
5. ⚠ **Missing explicit total loss/objective definition** (Sec. 2.3.2, p.4)
    - **Claim:** Training optimizes the main task while enforcing invariance via adversarial objectives.
    - **Checks:** derivation completeness, symbol/definition completeness
    - **Verdict:** UNCERTAIN; confidence: high; impact: moderate
    - **Assumptions/inputs:** There exists an implicit multi-term loss combining task and adversarial losses with weights (including λ).
    - **Notes:** No explicit formula is provided for the overall optimization objective (e.g., how task loss and multiple adversarial losses are combined and weighted). Without it, one cannot audit algebraic consistency of training dynamics beyond the qualitative GRL description.
6. ⚠ **Step-count aggregation from window predictions** (Sec. 2.4.2 (Step-Counting Performance), p.5)
    - **Claim:** Window-level binary predictions are aggregated to produce a total predicted step count per trial.
    - **Checks:** definition completeness, estimator consistency under overlap
    - **Verdict:** UNCERTAIN; confidence: high; impact: critical
    - **Assumptions/inputs:** Windows overlap by 50%., Each window produces a step/non-step label or probability.
    - **Notes:** The aggregation rule is not specified. Because windows overlap, different reasonable aggregations (sum of positives, debiasing by overlap, event detection, clustering contiguous positives, etc.) produce different step-count

estimators. Without a defined mapping, **MAE/MAPE** on "step counts" is not analytically well-defined.

7. ✔ **Confusion matrix internal consistency** (Sec. 3.3.1 + Fig. 3 caption, p.7)
    - **Claim:** Confusion matrix reports $\mathbf{TN = 56,543}$; $\mathbf{FN = 9,592}$; $\mathbf{FP = 2,251}$; $\mathbf{TP = 17,056}$ for test windows.
    - **Checks:** basic consistency (partition of outcomes)
    - **Verdict:** PASS; confidence: high; impact: minor
    - **Assumptions/inputs:** Entries are nonnegative integer counts of the same evaluated set.
    - **Notes:** Counts are mutually compatible as a complete partition: $\mathbf{TN + FP + FN + TP = 56,543 + 2,251 + 9,592 + 17,056 = 85,442}$ total evaluated windows.
8. ⚠ **Chance baseline for invariance probes** (Sec. 2.4.2 (Representation Invariance Analysis), p.5; Sec. 3.4, p.8)
    - **Claim:** For binary nuisance-variable probes, chance level is 50% accuracy; thus 59.20% indicates partial invariance.
    - **Checks:** probability/baseline consistency, assumption checking
    - **Verdict:** UNCERTAIN; confidence: medium; impact: moderate
    - **Assumptions/inputs:** Probe dataset is balanced or chance is defined as 50-50 random guessing.
    - **Notes:** Chance accuracy depends on class priors unless balanced accuracy is used. The paper does not state the class balance for frequency/location in the probe training/evaluation data, so interpreting 50% as chance is not fully justified from the text alone.
9. ✖ **Age-range bin consistency across sections** (Table 1, p.3; Sec. 3.1, p.6; Table 5, p.9)
    - **Claim:** Age ranges used for demographic summaries and fairness analysis are consistent throughout the paper.
    - **Checks:** definition consistency
    - **Verdict:** FAIL; confidence: high; impact: moderate
    - **Assumptions/inputs:** Age bins define subgroup membership for stratified evaluation.
    - **Notes:** Table 1 uses bins 18–29, 30–39, 40–49, 50+. Sec. 3.1 claims three groups 19–30, 31–44, 45–81 (equal counts). Table 5 uses 19–30, 31–44, 45–81. These are different partitions; endpoints shift and grouping changes (including collapsing/expanding ranges). This undermines internal consistency of demographic stratification.

10. ⚠ **Metric definition completeness (MAE/MAPE)** (Sec. 2.4.2 and Sec. 3.3, pp.5–7)
    - **Claim:** MAE and MAPE used for step-count evaluation are well-defined from the text.
    - **Checks:** definition completeness, aggregation convention clarity
    - **Verdict:** UNCERTAIN; confidence: high; impact: moderate
    - **Assumptions/inputs:** Errors are computed on trial-level step counts., MAPE uses a denominator based on ground-truth steps.
    - **Notes:** The paper names MAE and MAPE but provides no formulas and does not specify averaging conventions (per trial vs per participant, macro vs micro), nor any handling of potential small denominators. This prevents a strict analytic audit of metric usage.
11. ✔ **Adversary notation consistency (frequency classifier)** (Sec. 2.3.1–2.3.2, p.4)
    - **Claim:** The sampling-frequency adversary is consistently denoted.
    - **Checks:** notation consistency
    - **Verdict:** PASS; confidence: medium; impact: minor
    - **Assumptions/inputs:** C_{freq} and C_{freq} refer to the same object as C_{freq}.
    - **Notes:** This appears to be a typesetting/spacing inconsistency rather than a mathematical one, and the intended meaning remains clear in context.

Limitations

- The provided paper text contains essentially no explicit equations (loss functions, metric formulas, or aggregation rules), limiting the audit to consistency of stated definitions, shapes, and optimization-sign logic.
- Several central evaluative quantities (trial-level step count estimator; MAE/MAPE formulas; probe baselines) are described only qualitatively, so their mathematical correctness cannot be fully verified without additional specification in the document.

Numerical results audit

This section audits **numerical/empirical** consistency: reported metrics, experimental design, baseline comparisons, statistical evidence, leakage risks, and reproducibility.

Across 16 checks, internal arithmetic and basic consistency validations all passed. Verified items include demographic totals, dataset split totals and implied percentages, mean-to-total consistency for annotated steps, window-label percentage, epoch inequality consistency, bounds checks for overall vs condition metrics, confusion-matrix cell summation, probe arithmetic vs chance, and windowing/upsampling point calculations. One cross-section comparability concern remains: inconsistent age-bin definitions across sections cannot be reconciled without underlying ages.

Checked items

1. ✔ **C1_parts_vs_total_sex_table1** (Page 3, Table 1 (Participant Demographic Summary, $N = 39$))
    - **Claim:** Sex counts are Male 19 and Female 20 for $N = 39$ participants.
    - **Checks:** parts_vs_total
    - **Verdict:** PASS
    - **Notes:** $19 + 20 = 39$.
2. ✔ **C2_parts_vs_total_age_table1** (Page 3, Table 1 (Participant Demographic Summary, $N = 39$))
    - **Claim:** Age range counts are 18-29: 15, 30-39: 12, 40-49: 8, 50+: 4 for $N = 39$ participants.
    - **Checks:** parts_vs_total
    - **Verdict:** PASS
    - **Notes:** $15 + 12 + 8 + 4 = 39$.
3. ✔ **C3_table2_total_steps_from_mean** (Page 3, Table 2 (Descriptive Statistics of Annotated Steps per Participant) and Table 1 ($N = 39$))
    - **Claim:** Table 2 reports Mean Step Count 2150.4 and Total Annotated Steps 83,865 across $N = 39$ participants.
    - **Checks:** total_from_mean_times_n
    - **Verdict:** PASS
    - **Notes:** $2150.4 \times 39 = 83,865.6$ vs reported 83,865 (difference consistent with mean rounding).
4. ✔ **C4_window_count_step_percent** (Page 6, Section 3.1 (Data and Cohort Summary))
    - **Claim:** Out of 545,350 windows, 126,127 windows (23.1%) were labeled as step windows.
    - **Checks:** percentage_from_ratio
    - **Verdict:** PASS
    - **Notes:** Computed $100 \times 126,127/545,350 = 23.1277\%$, which rounds to 23.1%.
5. ✔ **C5_split_counts_from_percentages** (Page 5, Section 2.4.1 (Data Splitting))
    - **Claim:** Split is 70% (27), 15% (6), 15% (6) of 39 participants.
    - **Checks:** percent_to_count_and_total
    - **Verdict:** PASS
    - **Notes:** Counts sum: $27 + 6 + 6 = 39$. Achieved percentages are 69.23%, 15.38%, 15.38% (within stated rounding tolerance).

6. ✔ **C6_epoch_numbers_consistency** (Page 6, Section 3.2 (Model Training and Convergence))
    - **Claim:** Trained maximum 50 epochs; concluded after 30 epochs; optimal weights restored from epoch 20; lowest validation loss 0.2382 at epoch 20.
    - **Checks:** inequality_and_identity_checks
    - **Verdict:** PASS
    - **Notes:** Inequalities hold: $\mathbf{20 \leq 30 \leq 50}$.
7. ✔ **C7_overall_mape_vs_group_mapes_weighted_by_trials** (Page 6, Section 3.3.1 (Overall Accuracy) and Page 8, Table 3 (by sensor condition))
    - **Claim:** Overall **MAPE** is 44.36%, while condition **MAPE**s are 47.31%, 38.96%, 39.51%, 51.66%.
    - **Checks:** bounds_check_weighted_average_possible
    - **Verdict:** PASS
    - **Notes:** Overall 44.36% lies within [38.96%, 51.66%], as required for a weighted average over disjoint subsets.
8. ✔ **C8_overall_mae_vs_group_maes_bounds** (Page 6, Section 3.3.1 (Overall Accuracy) and Page 8, Table 3 (by sensor condition))
    - **Claim:** Overall **MAE** is 332.96 steps, while condition **MAE**s are 198.50, 164.33, 351.17, 617.83 steps.
    - **Checks:** bounds_check_weighted_average_possible
    - **Verdict:** PASS
    - **Notes:** Overall 332.96 lies within [164.33, 617.83], as required for a weighted average over disjoint subsets.
9. ✔ **C9_confusion_matrix_sum_check** (Page 7, Figure 3 text (confusion matrix counts))
    - **Claim:** Confusion matrix shows True Negatives 56,543 and False Negatives 9,592; matrix cells (from figure) appear to be 56,543; 2,251; 9,592; 17,056.
    - **Checks:** confusion_matrix_totals_and_consistency
    - **Verdict:** PASS
    - **Notes:** All cells are nonnegative integers; total $= \mathbf{56,543 + 2,251 + 9,592 + 17,056 = 85,442}$.
10. ✔ **C10_table4_frequency_probe_above_chance_delta** (Page 8, Table 4 (Accuracy of Linear Probes on Learned Embeddings))
    - **Claim:** Sampling Frequency probe accuracy is 59.20% and described as only moderately above chance level 50%.
    - **Checks:** difference_from_chance
    - **Verdict:** PASS

- **Notes:** $59.20 - 50 = 9.20$ percentage points.

11. ✔ **C11_table4_location_probe_extreme_value_range** (Page 8, Table 4 (Accuracy of Linear Probes on Learned Embeddings) and Abstract/Page 1)
    - **Claim:** Sensor Location probe accuracy is 96.47% (also stated in abstract) and is for a binary classification task (chance 50%).
    - **Checks:** range_and_repetition_consistency
    - **Verdict:** PASS
    - **Notes:** The 96.47% value matches in both places and lies within [0,100]; it is 46.47 percentage points above 50% chance.
12. ✔ **C12_table5_age_group_counts_sum** (Page 6, Section 3.1 (Data and Cohort Summary))
    - **Claim:** Age distribution stated as 13 participants aged 19-30, 13 aged 31-44, and 13 aged 45-81 (total 39).
    - **Checks:** parts_vs_total
    - **Verdict:** PASS
    - **Notes:** $13 + 13 + 13 = 39$.
13. ✔ **C13_age_binning_inconsistency_between_table1_and_results** (Page 3, Table 1 vs Page 6, Section 3.1 (Data and Cohort Summary) vs Page 9, Table 5)
    - **Claim:** Table 1 uses age bins 18-29/30-39/40-49/50+ with counts 15/12/8/4, while Results uses three bins 19-30/31-44/45-81 each with 13 participants; Table 5 uses 19-30, 31-44, 45-81.
    - **Checks:** cross_section_count_consistency
    - **Verdict:** PASS
    - **Notes:** Both schemes independently sum to 39, but differing bin edges prevent further reconciliation without participant-level ages.
14. ✔ **C14_window_size_points_100hz** (Page 3, Section 2.2 (Data Windowing and Structuring), item 1)
    - **Claim:** For 100 Hz data, 2-second windows comprise 200 time points (2 seconds × 100 Hz).
    - **Checks:** unit_consistent_recomputation
    - **Verdict:** PASS
    - **Notes:** $2\,\mathrm{s} \times 100\,\mathrm{Hz} = 200$ points.
15. ✔ **C15_window_size_points_25hz** (Page 3, Section 2.2 (Data Windowing and Structuring), item 1)
    - **Claim:** For 25 Hz data, 2-second windows contain 50 time points (2 seconds × 25 Hz).
    - **Checks:** unit_consistent_recomputation

- **Verdict:** PASS
- **Notes:** $2\,\mathrm{s} \times 25\,\mathrm{Hz} = 50$ points.

16. ✔ **C16__upsampling__ratio__25__to__100** (Page 3, Section 2.2 (Data Windowing and Structuring), item 2)
    - **Claim:** 25 Hz windows (50 points) were upsampled to 200 points to match the 100 Hz window length.
    - **Checks:** ratio_check
    - **Verdict:** PASS
    - **Notes:** $200/50 = 4\times$ upsampling.

Limitations

- Audit is based only on the provided parsed PDF text; no access to underlying CSVs, participant-level step counts, embeddings, or model outputs, so many reported metrics cannot be recomputed.
- Figures are not used for extracting numeric values except where the paper text explicitly states the numbers; no plot-pixel or image-based extraction is performed.
- Some cross-section comparisons (e.g., differing age-bin definitions between Table 1 and Results/Table 5) can be flagged but not fully reconciled without participant-level age data.