Skepthical review: *Self-Supervised Feature Learning for Robust and Interpretable Step Event Detection in Multi-Fidelity Wearable Data*

Summary

The manuscript proposes a two-stage pipeline for step-event detection from triaxial wearable accelerometer data across multiple sensor "fidelity" conditions (hip vs. wrist; **100 Hz** vs. **25 Hz**). Stage 1 pre-trains a 1D-CNN encoder with contrastive self-supervised learning (NT-Xent) on unlabeled **2.56 s** windows using standard time-series augmentations (Secs. 2.1–2.2). Stage 2 reuses the encoder inside a 1D U-Net-like dense predictor fine-tuned with Focal Loss to produce per-sample step probabilities, then converts probabilities to discrete events via peak finding and evaluates with event matching within a $\pm$**0.2 s** tolerance and step-count error (MAE/MAPE) under 5-fold group-stratified cross-validation (Secs. 2.3–2.4). The paper also includes a UMAP-based representation analysis intended to support interpretability and to compare structure across sensor conditions (Sec. 2.5, early Sec. 3).

The research question (reducing annotation burden via SSL; robustness to placement and sampling rate) is timely and relevant. However, the current manuscript is not scientifically valid as an empirical study because Sec. 3 explicitly states that a processing error prevented completion of supervised fine-tuning/evaluation and that all quantitative results are "hypothesized", yet the Abstract/Results/Discussion/Conclusions report concrete numbers and statistical claims as if measured (Abstract; Sec. 3; Secs. 4.3–4.4). In addition, several core methodological details needed for reproducibility and correct interpretation are missing or ambiguous (dataset/label definition, split integrity for SSL, event matching and peak-finding parameters, hyperparameters/architectures, and figure placeholders). A substantial revision is required: either execute the full evaluation and report real results with clear protocols and uncertainty, or reframe as a methods/protocol paper without numerical performance claims.

Strengths

- Timely and practically important problem formulation: event-level step detection with reduced labeling needs and attention to deployment realities (placement and sampling-rate variability) (Introduction).
- Overall pipeline structure is sensible and aligned with current practice in time-series SSL: contrastive pretraining with augmentations followed by downstream fine-tuning for dense prediction (Secs. 2.2–2.3).
- Use of group-based cross-validation at the participant level reflects good awareness of subject leakage risks in wearable sensing (Secs. 2.1.2, 2.3.1, 2.4).
- Evaluation intent covers both event-detection quality (Precision/Recall/F1 with temporal tolerance) and counting error (MAE/MAPE), which is appropriate for step-event systems (Sec. 2.4).

- UMAP-based representation visualization is a potentially useful diagnostic to compare learned features across sensor conditions (Sec. 2.5, early Sec. 3).
- Basic metric definitions for Precision/Recall/F1 are correct and window length conversions ($2.56\,\mathrm{s} \rightarrow 256$ samples at $100\,\mathrm{Hz}$; 64 at $25\,\mathrm{Hz}$) are internally consistent.

Major issues

1. **Empirical results are not executed: Sec. 3 states supervised fine-tuning/evaluation could not be completed due to a data processing error and that quantitative outcomes are "hypothesized", yet the manuscript reports specific performance numbers, standard deviations, and statistical significance claims (Tables 1–2; Abstract; Sec. 3 narrative; Secs. 4.3–4.4) as if they were measured.** This is not merely "overclaiming": it presents fabricated metrics and invalidates the paper as an experimental contribution.

   Recommendation: Choose one of the following and revise consistently end-to-end (Abstract → Sec. 4): (a) Fix the processing/training pipeline, rerun all experiments, and replace all hypothesized numbers with real outputs (per-fold + aggregated), including uncertainty (e.g., bootstrap CIs over participants) and properly specified statistical tests; or (b) remove Tables 1–2 and all numeric/statistical claims, and reframe the paper explicitly as a methodological/protocol + qualitative analysis paper (UMAP only), with claims limited to what is actually demonstrated. In either case, perform a line-by-line audit of Abstract, Sec. 3, and Secs. 4.3–4.4 to ensure there is no language implying validated quantitative superiority when none is presented.

2. **Dataset and labeling protocol are under-specified, preventing interpretation and reproducibility. The manuscript does not clearly define what a "step event" label corresponds to (heel strike vs toe-off vs another gait event), how labels were obtained (e.g., motion capture, pressure insoles, manual annotation), expected timing accuracy, and the recording protocol (activities, environment, duration, walking/non-walking distribution). This is critical because the evaluation uses a $\pm 0.2\,\mathrm{s}$ tolerance and the model predicts single-sample impulses (Secs. 2.1.1–2.1.2, 2.4.1).**

   Recommendation: Expand Sec. 2.1.1–2.1.2 with: dataset name/source (and whether it is used as-is or subset/merged); sensor setup and synchronization assumptions (hip vs wrist recorded simultaneously or treated independently); activity protocol and duration per participant; definition of step event (exact biomechanical event); label acquisition method and timing precision; and class composition (approximate walking vs non-walking share). If labels come from another paper/dataset, explicitly cite and summarize the annotation procedure and known error bounds.

3. **Potential split/leakage ambiguity for SSL pretraining: the paper uses participant-grouped CV for supervised fine-tuning, but it appears SSL pretraining may be done using "all participants within each sensor condition"**

(Secs. 2.2, 2.3.1). If unlabeled data from validation participants is used for representation learning, downstream evaluation becomes transductive with respect to the validation set, which can inflate performance and changes the claim being made.

Recommendation: Make the SSL pretraining/evaluation setting explicit in Sec. 2.2 and Sec. 2.3.1 and align the pipeline accordingly: (i) Preferred for clean claims: within each CV fold, pretrain SSL using only the fold's training participants (unlabeled), then fine-tune/evaluate on held-out participants; or (ii) If pretraining globally on all participants, clearly state this is transductive w.r.t. unlabeled validation data, justify why it matches the intended deployment, and (if possible) add a comparison to fold-restricted pretraining to quantify the effect.

4. **Evaluation protocol is ambiguous in ways that can materially change reported F1/MAE/MAPE and any fairness/robustness conclusions: (a) peak-finding parameters and thresholding are underspecified (Sec. 2.4.1: threshold given only as an example "e.g., 0.5"); (b) the one-to-one matching policy within the $\pm 0.2\,\mathrm{s}$ tolerance is not defined (multiple predicted peaks near one true step and vice versa); (c) aggregation over windows/participants/folds (macro vs micro) is unclear; (d) the Wilcoxon testing plan lacks unit of analysis and multiplicity handling (Secs. 2.4.1–2.4.3).**

   Recommendation: Rewrite Sec. 2.4.1–2.4.3 to be fully operational: specify the exact peak detection implementation (library/function), threshold selection procedure (fixed vs tuned; where tuned), minimum peak distance (in seconds/samples), and any smoothing. Define the matching algorithm (e.g., one-to-one greedy nearest-neighbor with exclusivity, or Hungarian) so TP/FP/FN are well-defined. In Sec. 2.4.2 specify precisely how MAE/MAPE and F1 are computed per participant and then aggregated across participants and folds (macro vs micro). In Sec. 2.4.3 define the paired samples for Wilcoxon (e.g., participant-level metric averaged over that participant's validation appearances), α level, and multiple-comparison correction across sensor conditions/metrics.

5. **Core model/training details are insufficient for reproducibility and for judging whether the design is appropriate for extremely sparse targets (especially at $25\,\mathrm{Hz}$). Missing: architecture specifics (channels/kernels/strides/padding, U-Net depth, skip connections), optimizer and schedule, batch size, epochs, NT-Xent temperature, projection head details, normalization, weight decay/dropout, and Focal Loss parameters (γ and α/class weighting) (Secs. 2.2.2, 2.3.2–2.3.3).**

   Recommendation: Add a compact but complete implementation specification (Secs. 2.2.2 and 2.3.2–2.3.3 or an Appendix): layer-by-layer encoder and decoder tables; projection head definition (if any) and embedding dimension; training hyperparameters

for SSL and fine-tuning (optimizer, LR, schedule, epochs, batch size, temperature, augmentations and their magnitudes/probabilities); Focal Loss equation and the exact γ/α used; early stopping criteria; and compute budget (hardware/runtime). This is especially important given label sparsity (single-sample positives at $25\,\mathrm{Hz}$).

6. **Windowing/label construction choices are not fully specified and may create boundary artifacts and optimization instability: SSL windows are stated as non-overlapping $2.56\,\mathrm{s}$ (Sec. 2.2.1), but supervised window stride/overlap is unclear; handling of steps at window boundaries is unspecified; and it is unclear whether training uses any label dilation/smoothing (a single 1 at the exact sample index) which is extremely sparse at $25\,\mathrm{Hz}$ (Secs. 2.3.2, 2.4).**

   Recommendation: In Sec. 2.3.2 explicitly state supervised training/inference window stride/overlap and how boundary events are handled. Specify how timestamps map to sample indices (rounding policy) and how multiple steps within a window are encoded. Consider (and report) a sensitivity/ablation: overlap vs non-overlap; label dilation (mark $\pm k$ samples positive) vs single-sample impulses; and/or training directly on event times (if applicable). At minimum, justify the current choices and discuss expected boundary effects.

7. **Interpretability claims from UMAP risk being overstated and potentially confounded. Coloring embeddings by "contains a step" can show correlation, but separation may be driven by participant identity, activity type, or amplitude/orientation differences (especially wrist), and UMAP is sensitive to hyperparameters. UMAP settings and sampling details are not reported; several figure references are placeholders (Sec. 2.5; Sec. 3: "Figure ??").**

   Recommendation: Tighten Sec. 2.5 and related Sec. 3 text: report UMAP hyperparameters ($n_{\text{neighbors}}$, min_dist, metric, random_state), feature preprocessing, and sampling scheme (how many windows per participant/condition). Add at least one quantitative complement such as linear-probe performance (e.g., logistic regression to classify step/non-step from frozen features), k-NN accuracy, or silhouette score. To address confounds, add plots/analyses colored by participant and (if available) activity type. Ensure all referenced figures exist and are numbered consistently (remove "Figure ??" and "not shown" references).

8. **Demographic subgroup definitions/counts are inconsistent across the manuscript, undermining the stated stratification and any fairness/robustness narrative: Sec. 2.1.2 uses different age bins than Table 2, and sex counts differ by 1 between Sec. 2.1.2 and Table 2 (Secs. 2.1.2, 2.3.1, Sec. 3 Table 2). With $n = 39$, subgroup comparisons are also likely underpowered without uncertainty reporting.**

Recommendation: Standardize age bins and sex counts across Sec. 2.1.2, the stratification description (Sec. 2.3.1), and Table 2 so totals match $n = 39$ exactly and definitions are consistent. Once real results exist, report subgroup metrics with uncertainty (e.g., bootstrap CIs) and explicitly note limited power; avoid strong fairness claims unless supported by statistically and practically meaningful evidence.

9. **Related work and novelty positioning are not yet cohesive. The manuscript references SSL and step detection in a scattered way but lacks a focused comparison to (i) classical step-event/step-count pipelines (filtering + peak detection, heuristic thresholds), (ii) supervised deep sequence models for gait/step event detection, and (iii) prior SSL-for-HAR/time-series work. This makes it difficult to evaluate what is genuinely new beyond applying a standard SSL + downstream model template (Secs. 1–2).**

   Recommendation: Add a dedicated Related Work subsection (e.g., Sec. 1.1) that separately covers: step event detection/counting methods (signal-processing and deep learning), SSL for time-series/HAR (contrastive and non-contrastive), and multi-domain/multi-sensor robustness approaches (domain adaptation, multi-rate learning). Then state clearly what this paper contributes (e.g., dense event formulation + multi-fidelity comparison + representation visualization) and calibrate novelty claims accordingly.

Minor issues

1. Training objectives are referenced but not written explicitly: NT-Xent and Focal Loss are named but not given as equations with defined symbols/variants (Secs. 2.2–2.3).

   Recommendation: Add explicit mathematical definitions for NT-Xent (including similarity, temperature, number of views/positives) and Focal Loss (probabilities vs logits, γ and α/class weights), and specify the exact variants used.

2. Signal preprocessing is not described in enough detail, yet it strongly affects cross-participant and cross-placement generalization (Secs. 2.1–2.3). It is unclear whether signals are standardized per window/participant, whether gravity is removed, whether axes are reoriented/aligned, or whether magnitude is used.

   Recommendation: In Sec. 2.1 or Sec. 2.2.1, specify preprocessing steps: filtering (if any), detrending/gravity handling, coordinate frame/orientation treatment, normalization (per participant/window/global), and whether raw axes or magnitude are used. Justify choices especially for wrist where orientation variability is large.

3. "Multi-fidelity robustness" is emphasized, but the methods train separate models per condition (hip/wrist $\times$ $\mathbf{100/25\,Hz}$) and (as written) do not evaluate cross-condition transfer or a unified model (Secs. 2.2–2.3; Secs. 4.3–4.4).

Recommendation: Either soften claims to "within-condition performance across four settings" or add experiments once the pipeline runs: joint pretraining across all conditions, multi-condition fine-tuning, and/or transfer (pretrain on one condition, fine-tune/evaluate on another) to directly test robustness to placement/frequency changes.

4. MAPE is mentioned but not defined as a formula, and handling of potential zero-denominator cases is unspecified (Sec. 2.4.2).

   Recommendation: Provide an explicit MAPE definition (per participant vs pooled; normalization by true step count) and state how any edge cases are handled (e.g., windows/participants with zero true steps if applicable).

5. The manuscript would benefit from clearer separation of "executed analyses" vs "planned analyses," even after fixing the main results issue (Sec. 3).

   Recommendation: In Sec. 3, structure subsections into (i) executed qualitative analyses (e.g., UMAP) and (ii) quantitative results (only once actually run). If some analyses remain future work, label them explicitly as such and avoid presenting placeholder numbers.

6. Ethics/privacy considerations are absent despite using human-subject wearable data (Secs. 1–2).

   Recommendation: Add a brief statement (Sec. 2.1 or Sec. 4) referencing the original dataset's consent/IRB (if applicable) and note privacy and bias considerations relevant to deployment.

Very minor issues

1. Numerous presentation issues reduce readability and actionability: placeholders like "Figure ??" and references to panels "not shown"; inconsistent naming of sensor conditions ("Hip 100Hz", "Hip_100Hz", "Hip 100 Hz"); typographical/LaTeX issues (e.g., broken line breaks); malformed citations with trailing "?"; and spaced TP/FP/FN notation that can be misread (Secs. 1–4).

   Recommendation: Do a full editorial pass: resolve all figure numbers/captions; standardize condition naming; fix LaTeX/typos; clean up all citations; and standardize notation (TP/FP/FN). Ensure every in-text reference points to an existing figure/table and correct panel.

2. Keyword list includes peripheral terms (e.g., "Distributed computing") and misses key terms central to the manuscript (Abstract).

   Recommendation: Revise keywords to reflect the core contribution (e.g., self-supervised learning, contrastive learning, wearable accelerometers, step event detection, human activity recognition, U-Net, interpretability).

3. Some citations appear generic or weakly tied to specific claims (Sec. 2).

Recommendation: Tighten citations to foundational and domain-relevant sources (original NT-Xent/SimCLR-style contrastive learning; focal loss; step detection/gait event detection; SSL-for-HAR). Ensure each citation supports a specific statement or design choice.

Key statements and references

- ✖ **The dataset used in this study comprised triaxial accelerometer data and corresponding step annotations collected from 39 participants, organized by hip and wrist placement at 100 Hz and 25 Hz, derived from a multi-sensor human gait dataset with associated demographic metadata provided in a separate metadata_csv file.**
- *Reference(s):* Zhang et al., 2024, Santos et al., 2021, Bayat et al., 2022
- *Justification:* None of the attached papers describe a dataset with **39** participants organized by hip and wrist placement sampled at both **100 Hz** and **25 Hz** with step annotations and a separate metadata_csv. Santos et al., 2021 has **25** subjects with IMU and optical data at **100 Hz** attached to the leg (not hip/wrist) and no explicit step annotations; Bayat et al., 2022 uses smartphone accelerometers at the left waist and right thigh from **93** subjects at **100 Hz** without step annotations; Zhang et al., 2024 uses foot-worn IMUs with step labels for **10** subjects, not hip/wrist nor **25 Hz**. Thus the statement is not supported by the papers collectively.

- ✖ **For self-supervised learning, the continuous triaxial accelerometer time series for each sensor condition was segmented into non-overlapping 2.56-second windows (256 samples at 100 Hz and 64 samples at 25 Hz), following prior work on self-supervised human activity recognition using windowed wearable data.**
- *Reference(s):* Taghanaki et al., 2021, Sridhar and Myers, 2021, Yuan et al., 2024
- *Justification:* None of the cited papers use non-overlapping **2.56**-second windows with **256** samples at **100 Hz** or **64** at **25 Hz** for self-supervised learning. Taghanaki et al., 2021 uses **2.56 s** segments but with **50%** overlap and data sampled at **50 Hz** (not **100/25 Hz**) and a special $2.08\,\mathrm{s} + 0.48\,\mathrm{s}$ z-axis setup. Sridhar and Myers, 2021 resample to **30 Hz** and use non-overlapping **10**-second windows. Yuan et al., 2024 also resample to **30 Hz** and use **10**-second windows. Thus the specific segmentation claimed is not supported.

- ✖ **The self-supervised pre-training phase employed a contrastive learning framework based on the NT-Xent loss, in which two augmented views of each 2.56-second window (generated via jitter, scaling, and time-warping) formed positive pairs and all other windows in the mini-batch served as negatives, following established contrastive representation learning methods.**

- *Reference(s):* Chen et al., 2020, Le-Khac et al., 2020, Taghanaki et al., 2021
- *Justification:* Taghanaki et al., 2021 pre-trains via cross-dimensional motion prediction (regressing z from past z and past/present x, y) using MSE, not a contrastive NT-Xent framework; it uses $2.56\,\mathrm{s}$ inputs for x/y and $2.08\,\mathrm{s}$ for z with a $0.48\,\mathrm{s}$ prediction, without forming positive/negative pairs or using jitter/scaling/time-warping augmentations. Chen et al., 2020 (SimCLR) does use NT-Xent with two augmented views (crop/color/blur) and in-batch negatives, but in images, not the stated HAR setup. Le-Khac et al., 2020 reviews contrastive learning generally and does not provide evidence for the specific protocol claimed. Hence the statement is not supported.

- ✖ **To address the severe class imbalance in continuous step detection, the supervised fine-tuning phase used Focal Loss within a 5-fold group cross-validation scheme where participants were the grouping variable and folds were stratified by sex and age, consistent with recommended practices for cross-validation and imbalanced classification in sensor-based activity recognition.**

- *Reference(s):* Khan and Abedi, 2022, Li et al., 2025, Sedaghati et al., 2024
- *Justification:* None of the attached papers report using Focal Loss or a 5-fold group cross-validation stratified by sex and age for step detection. Khan and Abedi, 2022 formulate step counting as regression with MAE loss and use five-fold CV on WDSC, leave-one-person-out on WeAllWalk, and leave-two-person-out on Pedometer—no focal loss or sex/age stratification. Li et al., 2025 develops a CV method for panel models, unrelated to sensor-based step detection or focal loss. Sedaghati et al., 2024 uses a simple train/test split for HAR with CNN, not group CV nor focal loss.

- ⚠ **Discrete step events were obtained from the model's continuous probability output by applying a peak-finding algorithm that identifies local maxima exceeding a fixed threshold (e.g., 0.5) and higher than their immediate neighbors, in line with established peak search methods for uniformly sampled time series.**

- *Reference(s):* Guidorzi, 2015, Lee et al., 2024
- *Justification:* Lee et al., 2024 convert continuous anomaly scores to discrete predicted locations using a local maxima-finding algorithm (SciPy find_peaks), which aligns with using peak-finding to discretize continuous outputs. Guidorzi, 2015 presents established peak-search methods for uniformly sampled time series (MEPSA, LFA) based on comparisons to neighboring bins. However, neither paper supports using a fixed threshold like 0.5: MEPSA/LFA use sigma-based criteria, and Lee et al., 2024 do not specify a fixed threshold, instead using local maxima (and quantile thresholds only for counterfactual sampling). Thus, the peak-finding aspect is supported but the fixed-threshold detail is not.

- ✖ **Feature interpretability was assessed by extracting high-dimensional feature vectors from the global average pooling layer of the self-supervised encoders and projecting them into 2D using UMAP, a non-linear dimensionality reduction technique designed to preserve local and global structure, enabling visualization of clusters corresponding to stepping versus non-stepping windows.**
- *Reference(s):* Khaertdinov and Asteriadis, 2023, Han et al., 2024, Ren et al., 2025
- *Justification:* None of the attached papers describe extracting features from the global average pooling layer of self-supervised encoders and visualizing them with UMAP to separate stepping vs non-stepping windows. Khaertdinov and Asteriadis, 2023 analyze SSL representations using occlusion, Guided Grad-CAM, and probing, without UMAP, GAP features, or a stepping/non-stepping task. Han et al., 2025 focus on backdoor mitigation for SSL encoders and do not use UMAP for interpretability. Ren et al., 2025 discusses a general embedding visualization tool that can use UMAP, but not in the context of HAR or stepping windows nor GAP features from SSL encoders.

Mathematical consistency audit

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

Maths relevance: light

The paper contains relatively little explicit mathematics: it mainly states sampling/window conversions and defines standard classification/counting metrics (Precision/Recall/F1; MAE/MAPE described verbally). Core optimization objectives (NT-Xent and Focal Loss) and model equations are not written explicitly, so central derivations/notation cannot be audited beyond high-level consistency checks.

Checked items

1. ✔ **Window duration to sample count conversion** (Sec. 2.2.1, p.3)
    - **Claim:** A 2.56-second window corresponds to **256** samples at **100 Hz** and **64** samples at **25 Hz**.
    - **Checks:** algebra, unit/dimensional consistency
    - **Verdict:** PASS; confidence: high; impact: minor
    - **Assumptions/inputs:** Sampling frequency is exactly **100** samples/s or **25** samples/s, Window length is exactly **2.56 s**
    - **Notes:** $2.56\,\mathrm{s} \times 100\,\mathrm{Hz} = 256$ samples; $2.56\,\mathrm{s} \times 25\,\mathrm{Hz} = 64$ samples. Units ($\mathrm{s} \times \mathrm{samples/s}$) are consistent.
2. ⚠ **Binary target vector definition for event labeling** (Sec. 2.3.2 (Data Labeling), p.5)

- **Claim:** For each window, create a target vector of equal length with **1.0** at indices of annotated step events and **0.0** otherwise.
- **Checks:** definition consistency, well-posedness
- **Verdict:** UNCERTAIN; confidence: medium; impact: moderate
- **Assumptions/inputs:** Step annotations are point events aligned to sample indices (or discretized to nearest index), A window can contain zero, one, or multiple steps
- **Notes:** The definition is plausible, but the paper does not specify how timestamp annotations are discretized to indices (round/floor/nearest) and what happens if multiple steps map to the same index or fall on a window boundary. These choices affect the precise mathematical target.

3. ✔ **Precision formula** (Sec. 2.4.2 (Event Detection Accuracy), p.6)
   - **Claim:** Precision is $TP/(TP + FP)$.
   - **Checks:** algebra, definition consistency
   - **Verdict:** PASS; confidence: high; impact: minor
   - **Assumptions/inputs:** TP, FP are nonnegative integers, $TP + FP > 0$ when precision is computed
   - **Notes:** Standard and algebraically correct definition.
4. ✔ **Recall formula** (Sec. 2.4.2 (Event Detection Accuracy), p.6)
   - **Claim:** Recall is $TP/(TP + FN)$.
   - **Checks:** algebra, definition consistency
   - **Verdict:** PASS; confidence: high; impact: minor
   - **Assumptions/inputs:** TP, FN are nonnegative integers, $TP + FN > 0$ when recall is computed
   - **Notes:** Standard and algebraically correct definition.
5. ✔ **F1-score formula** (Sec. 2.4.2 (Event Detection Accuracy), p.6)
   - **Claim:** F_1 is $2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$.
   - **Checks:** algebra, definition consistency
   - **Verdict:** PASS; confidence: high; impact: minor
   - **Assumptions/inputs:** $\text{Precision} + \text{Recall} > 0$ when F_1 is computed
   - **Notes:** Algebraically correct harmonic-mean form.
6. ⚠ **Event matching tolerance definition** (Sec. 2.4.2 (Event Detection Accuracy), p.6)
   - **Claim:** A predicted step is a TP if within $\pm 0.2\,\text{s}$ of a true step; otherwise it is FP; unmatched true steps are FN.
   - **Checks:** definition consistency, well-posedness
   - **Verdict:** UNCERTAIN; confidence: medium; impact: moderate

- **Assumptions/inputs:** Predicted steps and true steps are timestamped point events, A unique matching policy is used
- **Notes:** The tolerance window itself is clear, but without specifying a one-to-one matching rule, $TP/FP/FN$ counts are not uniquely determined in cases of multiple predictions near one true step (or vice versa).

7. ✔ **Cross-validation fold arithmetic** (Sec. 2.3.1, p.4)
   - **Claim:** With 39 participants and 5 folds, each fold uses ~ 31 for training and 8 for validation.
   - **Checks:** arithmetic consistency, definition consistency
   - **Verdict:** PASS; confidence: high; impact: minor
   - **Assumptions/inputs:** Participants are the grouping unit and are not split across train/val within a fold
   - **Notes:** $31 + 8 = 39$, consistent with total participant count; approximate language is acceptable given uneven fold sizes may occur.

8. ⚠ **MAPE definition completeness** (Sec. 2.4.2 (Counting Accuracy), p.5–6)
   - **Claim:** MAPE is the average absolute percentage difference between predicted and true step counts.
   - **Checks:** definition completeness, symbol/notation clarity
   - **Verdict:** UNCERTAIN; confidence: high; impact: moderate
   - **Assumptions/inputs:** Per-participant true step counts are positive
   - **Notes:** No explicit formula is provided (e.g., whether percentage is $|p - t|/t$, whether averaging is over participants vs pooled totals, and how any zero-true cases are handled). This is a definitional ambiguity.

9. ✖ **Demographic subgroup internal consistency** (Sec. 2.1.2, p.3 vs Table 2, p.9)
   - **Claim:** Participant sex and age-group counts are consistent across the paper.
   - **Checks:** definition consistency, internal consistency
   - **Verdict:** FAIL; confidence: high; impact: moderate
   - **Assumptions/inputs:** All counts refer to the same cohort of 39 participants
   - **Notes:** Sec. 2.1.2 states 18 females/21 males and age bins 18–$25/26$–$40/41$–65 with counts $12/14/13$, while Table 2 uses Female $n = 19$/Male $n = 20$ and different age bins 19–$30/31$–$44/45$–81 (each $n = 13$). These cannot all simultaneously hold for the same dataset without explanation (e.g., re-binning, exclusions, or corrected counts).

10. ⚠ **Loss function specification (NT-Xent and Focal Loss)** (Sec. 2.2.2 (NT-Xent), p.3–4; Sec. 2.3.3 (Focal Loss), p.5)

- **Claim:** The paper's training uses NT-Xent for SSL and Focal Loss for supervised fine-tuning as defined objectives.
- **Checks:** missing derivation/definition, notation completeness
- **Verdict:** UNCERTAIN; confidence: high; impact: critical
- **Assumptions/inputs:** A specific variant of NT-Xent and Focal Loss is implemented
- **Notes:** The objectives are only described conceptually; no equations, symbol definitions, or exact variants are provided. This prevents symbolic auditing of the central mathematical optimization setup.

Limitations

- The provided PDF text contains almost no explicit equations beyond metric formulas; key losses/objectives and model mappings are described verbally only.
- Figures and their captions reference placeholders (e.g., Figure ??), and the audit cannot verify any mathematical content that would have been in those missing/unnumbered figures.
- Because the paper reports "hypothesized" quantitative outcomes, this audit intentionally does not assess numerical plausibility, only internal definitional/algebraic consistency.

Numerical results audit

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

16 numerical checks were executed: **14** PASS and **2** FAIL. Passes include multiple exact parts-to-total count validations (participant totals by sex and age bins, and Table 2 age subgroup totals), unit-consistent window-size-to-sample conversions, and multiple F_1 identity cross-checks against reported mean precision/recall (within stated tolerances). Failures relate to cross-section consistency of subgroup counts (sex) and a structural age-binning consistency check.

Checked items

1. ✔ **C1** (p.3, Sec. 2.1.2 (EDA baseline characteristics))
    - **Claim:** Sex distribution: **18** females and **21** males out of **39** participants.
    - **Checks:** parts_vs_total
    - **Verdict:** PASS
    - **Notes:** $18 + 21 = 39$.
2. ✔ **C2** (p.3, Sec. 2.1.2 (EDA baseline characteristics))
    - **Claim:** Age ranges: **12** participants aged **18–25**, **14** aged **26–40**, and **13** aged **41–65** (total **39**).
    - **Checks:** parts_vs_total
    - **Verdict:** PASS

- **Notes:** $12 + 14 + 13 = 39$.

3. ✔ **C3** (p.4, Sec. 2.2.1 (Windowing details for SSL))
   - **Claim:** A fixed window size of 2.56 seconds translated to 256 samples at $100\,\mathrm{Hz}$ and 64 samples at $25\,\mathrm{Hz}$.
   - **Checks:** unit_consistency_recompute
   - **Verdict:** PASS
   - **Notes:** $2.56 \times 100 = 256$ and $2.56 \times 25 = 64$.
4. ✔ **C4** (p.4, Sec. 2.3.1 (Cross-validation setup))
   - **Claim:** In each of five folds, approximately 31 participants were allocated for training and 8 for validation (total 39).
   - **Checks:** parts_vs_total
   - **Verdict:** PASS
   - **Notes:** $31 + 8 = 39$.
5. ✔ **C5** (p.3, Sec. 2.1.2 (EDA baseline characteristics))
   - **Claim:** Low step annotation ratio: approximately one step per 1.67 seconds of data.
   - **Checks:** rate_recompute_from_stated_ratio
   - **Verdict:** PASS
   - **Notes:** Implied rates computed from the stated ratio: ~ 0.5988 steps/s and ~ 35.9281 steps/min; no explicit target rate was provided to assert against.
6. ✔ **C6** (p.8, Table 1 (Hip $100\,\mathrm{Hz}$, SSL-Pretrained row))
   - **Claim:** Hip $100\,\mathrm{Hz}$ SSL-Pretrained: F_1-Score 0.96 ± 0.03, Precision 0.97 ± 0.02, Recall 0.95 ± 0.04. F_1 should equal $2PR/(P+R)$ from the reported mean P and R (approx).
   - **Checks:** metric_identity_recompute
   - **Verdict:** PASS
   - **Notes:** Implied F_1 from mean $P, R = 0.9598958333$ vs reported 0.96.
7. ✔ **C7** (p.8, Table 1 (Hip $100\,\mathrm{Hz}$, Baseline row))
   - **Claim:** Hip $100\,\mathrm{Hz}$ Baseline: F_1-Score 0.92, Precision 0.93, Recall 0.91. Check $F_1 \approx 2PR/(P+R)$.
   - **Checks:** metric_identity_recompute
   - **Verdict:** PASS
   - **Notes:** Implied F_1 from mean $P, R = 0.9198913043$ vs reported 0.92.
8. ✔ **C8** (p.8, Table 1 (Hip $25\,\mathrm{Hz}$, SSL-Pretrained row))
   - **Claim:** Hip $25\,\mathrm{Hz}$ SSL-Pretrained: F_1 0.94, Precision 0.95, Recall 0.93. Check $F_1 \approx 2PR/(P+R)$.

- **Checks:** metric_identity_recompute
- **Verdict:** PASS
- **Notes:** Implied F_1 from mean $P, R = 0.9398936170$ vs reported 0.94.

9. ✔ **C9** (p.8, Table 1 (Hip $25\,\mathrm{Hz}$, Baseline row))
   - **Claim:** Hip $25\,\mathrm{Hz}$ Baseline: F_1 0.89, Precision 0.90, Recall 0.88. Check $F_1 \approx 2PR/(P+R)$.
   - **Checks:** metric_identity_recompute
   - **Verdict:** PASS
   - **Notes:** Implied F_1 from mean $P, R = 0.8898876404$ vs reported 0.89.
10. ✔ **C10** (p.8, Table 1 (Wrist $100\,\mathrm{Hz}$, SSL-Pretrained row))
    - **Claim:** Wrist $100\,\mathrm{Hz}$ SSL-Pretrained: F_1 0.88, Precision 0.89, Recall 0.87. Check $F_1 \approx 2PR/(P+R)$.
    - **Checks:** metric_identity_recompute
    - **Verdict:** PASS
    - **Notes:** Implied F_1 from mean $P, R = 0.8798863636$ vs reported 0.88.
11. ✔ **C11** (p.8, Table 1 (Wrist $100\,\mathrm{Hz}$, Baseline row))
    - **Claim:** Wrist $100\,\mathrm{Hz}$ Baseline: F_1 0.81, Precision 0.83, Recall 0.79. Check $F_1 \approx 2PR/(P+R)$.
    - **Checks:** metric_identity_recompute
    - **Verdict:** PASS
    - **Notes:** Implied F_1 from mean $P, R = 0.8095061728$ vs reported 0.81.
12. ✔ **C12** (p.8, Table 1 (Wrist $25\,\mathrm{Hz}$, SSL-Pretrained row))
    - **Claim:** Wrist $25\,\mathrm{Hz}$ SSL-Pretrained: F_1 0.85, Precision 0.86, Recall 0.84. Check $F_1 \approx 2PR/(P+R)$.
    - **Checks:** metric_identity_recompute
    - **Verdict:** PASS
    - **Notes:** Implied F_1 from mean $P, R = 0.8498823529$ vs reported 0.85.
13. ✔ **C13** (p.8, Table 1 (Wrist $25\,\mathrm{Hz}$, Baseline row))
    - **Claim:** Wrist $25\,\mathrm{Hz}$ Baseline: F_1 0.76, Precision 0.79, Recall 0.74. Check $F_1 \approx 2PR/(P+R)$.
    - **Checks:** metric_identity_recompute
    - **Verdict:** PASS
    - **Notes:** Implied F_1 from mean $P, R = 0.7641830065$ vs reported 0.76 (within allowed tolerance).
14. ✖ **C14** (p.9, Table 2 vs p.3, Sec. 2.1.2 (sex counts inconsistency check))

- **Claim:** Table 2 reports Sex subgroup counts Female ($n = 19$), Male ($n = 20$), while earlier EDA reports 18 females and 21 males.
- **Checks:** repeated_constant_consistency
- **Verdict:** FAIL
- **Notes:** Sex counts differ between EDA (18F/21M) and Table 2 (19F/20M), though both sums equal 39.

15. ✔ **C15** (p.9, Table 2 (Age subgroup counts))
    - **Claim:** Table 2 age range subgroup sizes: 19–30 ($n = 13$), 31–44 ($n = 13$), 45–81 ($n = 13$) total 39.
    - **Checks:** parts_vs_total
    - **Verdict:** PASS
    - **Notes:** $13 + 13 + 13 = 39$.
16. ✖ **C16** (p.3 (EDA age bins) vs p.9, Table 2 (different age bins))
    - **Claim:** EDA uses age bins (18–25, 26–40, 41–65) with counts (12, 14, 13) while Table 2 uses (19–30, 31–44, 45–81) with counts (13, 13, 13).
    - **Checks:** repeated_constant_consistency
    - **Verdict:** FAIL
    - **Notes:** Age bin edges differ; only sums should be checked without raw ages, but the executed check output shows sums (55 and 62) inconsistent with the stated counts in the extracted text.

Limitations

- Audit performed from provided PDF text only; no access to underlying datasets, code, logs, or supplementary materials.
- No checks proposed that require extracting numeric values from plotted figures/images; figure captions without explicit numeric endpoints cannot be validated.
- Several reported statistics (means/SDs, class imbalance ratio as observed property, significance claims) cannot be recomputed without participant-level data; these are listed as unverified.
- F_1 consistency checks using reported mean Precision/Recall are approximate because the mean of F_1 over folds/participants is not necessarily equal to F_1 computed from mean Precision and mean Recall.