Skepthical review: *An Investigation into Deep Generative Reconstruction for Low-Frequency Step Counting: Unveiling Data Integrity and Workflow Challenges*

Summary

The manuscript proposes a conditional variational autoencoder (CVAE) pipeline to mitigate loss of fine-grained gait information when triaxial accelerometer data are downsampled from **100 Hz** to **25 Hz** for step counting. The intended workflow (Sec. 1–2.5) uses participant-level splits on OxWalk, extracts **2 s** step-centered windows (**200** samples at **100 Hz**; **50** samples at **25 Hz**), trains separate hip/wrist CVAEs to reconstruct **100 Hz** windows from **25 Hz** inputs, stitches reconstructed windows via overlap-add, and evaluates step counts using SVM + scipy.signal.find_peaks with MAE/MAPE and demographic subgroup analyses.

In the dataset distribution used, ground-truth step annotations required for step-centered segmentation and absolute step-count evaluation are absent (Sec. 2.1, Sec. 3.2). Additionally, the computational environment contained legacy artifacts (e.g., pre-existing .npz and step_count_evaluation_results.csv) that were silently loaded by scripts, producing plausible but invalid tables/plots/metrics (Sec. 3.3–3.4). The authors acknowledge that all reported quantitative results are invalid and partly reframe the submission as a cautionary case study on data verification and workflow isolation (Sec. 3.5, Sec. 4). As written, the paper contains a useful proposed methodology and a transparent failure analysis, but it does not provide empirical evidence for the central claim that CVAE-based reconstruction improves step counting at **25 Hz**, and several key methodological and provenance details remain underspecified.

Strengths

- Motivates the practical importance of accurate step counting under low sampling rates and provides an intuitive rationale for generative reconstruction as a way to recover step-relevant high-frequency structure (Sec. 1, Sec. 2.3).
- Presents a coherent end-to-end *planned* pipeline (splits, windowing, reconstruction, overlap-add, peak-based counting, subgroup analysis) that is easy to follow and would be useful if completed with valid data/provenance (Sec. 2.1–2.5).
- Is unusually transparent about critical experimental failures (missing labels; artifact contamination) and attempts to analyze them rather than obscuring them (Sec. 3.2–3.5).
- Highlights valuable meta-scientific lessons on early dataset validation, strict experiment isolation, and reproducibility practices (Sec. 3.5, Sec. 4).
- Uses standard, interpretable signal constructs (e.g., SVM) and clearly explains the intended **25 Hz** vs **100 Hz** window-size relationship (Sec. 2.1–2.2), even though the implementation later appears inconsistent.

Major issues

1. **The central scientific claim (CVAE reconstruction improves step counting from 25 Hz) is not supported by valid evidence: the OxWalk distribution used lacks ground-truth step annotations, so step-centered segmentation and absolute MAE/MAPE step-count evaluation cannot be performed, and all reported quantitative results are acknowledged to be invalid (Sec. 2.1–2.2, Sec. 2.4, Sec. 3.2–3.4, Sec. 4).**

   Recommendation: Pick and execute one consistent scope: (a) **Methods + empirical study**: obtain/create step annotations (OxWalk alternative release if it exists; manual labeling on a subset; synchronized reference sensor; or another dataset with step labels) and rerun the entire pipeline in a clean environment; or (b) **Case study**: fully reframe the paper so the primary contribution is the data/provenance/workflow postmortem, and explicitly state (Abstract, Sec. 1, Sec. 4) that the step-counting efficacy hypothesis remains untested.

If staying with OxWalk *without* step labels, also consider adding an intermediate, label-free evaluation (Major Issue 2) so the CVAE reconstruction component is empirically assessed even if absolute step-count accuracy is not.

2. **The manuscript treats missing step labels as blocking *all* model training, but step annotations are not strictly necessary for training a 25 Hz → 100 Hz reconstruction model if aligned paired signals exist; labels are primarily required for step-centered window sampling and evaluation against ground-truth step counts (Sec. 2.2–2.4, Sec. 3.2). This conceptual coupling weakens the methodology and the paper's usefulness even as a proposal.**

   Recommendation: Decouple (i) **reconstruction training** from (ii) **step-count evaluation**: - In Sec. 2.2–2.3, define an alternative windowing scheme that does not require step events (e.g., uniformly sampled aligned windows from walking bouts, or sliding windows across the full stream with stratification by participant and activity if available). - Add reconstruction metrics that do not require step labels (e.g., time-domain MAE/RMSE on acceleration/SVM, correlation, spectral/PSD similarity, coherence) and report them in Sec. 3 if feasible. - For step counting, if no ground truth exists, explicitly label any comparison as *proxy* (e.g., comparing peak counts on reconstructed 100 Hz against peak counts on the true 100 Hz using the same detector) and keep it separate from claims about real-world step-count accuracy. This yields an executable study even under limited labels, while clearly stating what remains unvalidated.

3. **Workflow contamination by legacy artifacts (pre-existing .npz feature tensors and step_count_evaluation_results.csv) undermines the credibility of the entire results section, yet the current description is not sufficiently**

forensic/reproducible for readers to learn from it or verify the claims (Sec. 3.3–3.4).

Recommendation: Expand Sec. 3.3–3.4 into an auditable post-mortem: - Enumerate exact artifact filenames/paths loaded by each script (e.g., train_cvae.py, synthesis/evaluation scripts), including how paths are resolved and any fallbacks. - Provide concrete evidence (timestamps, directory listings, hashes/checksums, tensor shapes, dataset version identifiers) demonstrating mismatch between intended inputs and loaded artifacts. - Add explicit "failure mode" and "fix" steps (e.g., require-empty output dirs; fail-fast if expected raw-data files/labels are missing; embed dataset version + commit hash into every artifact; containerize; pin dependencies; write-protect artifact directories). - Consider moving invalid plots/tables to an appendix labeled as *artifact outputs from contaminated runs*, and add a short checklist readers can reuse (Sec. 3.5 or Sec. 4).

4. **Dataset provenance and validation are insufficiently specified: the paper does not clearly document the exact OxWalk version/source/date, what files/columns were expected vs actually present, and what checks were performed to rule out labels stored under different names or in separate files (Sec. 2.1, Sec. 3.2).**

   Recommendation: In Sec. 2.1, add a concise dataset audit table: - dataset release identifier (URL/DOI), date accessed, checksum (if available); - directory/file listing (at least for one participant) and example headers/columns; - explicit statement of what was searched for (e.g., 'step' column; separate annotation files; alternative column names) and results of that search. Then add a short, explicit *data validation protocol* (must-pass checks) that precedes any modeling (e.g., non-empty label fields; plausible step rates; alignment between **25 Hz** and **100 Hz** streams).

5. **Methods are described in a way that blends intended protocol with what was actually executed, which can mislead readers given that the reported outputs are invalid (Sec. 2.1–2.5 vs Sec. 3).**

   Recommendation: Systematically label procedures as **Planned** vs **Executed**: - Either split Sec. 2 into "Planned Methods" and add a new "Actual Execution / Deviations" subsection before Sec. 3, or annotate each subsection (Sec. 2.2–2.5) with an explicit status note. - Ensure Sec. 3 contains only verified observations (e.g., missing labels, artifact loading behavior) and does not read like a completed performance evaluation.

6. **Key technical specifications are underspecified, limiting replicability even as a proposal: CVAE conditioning definition, architecture details, loss terms (KL), training hyperparameters, alignment between 25 Hz and 100 Hz windows, and overlap-add stitching are not implementation-ready (Sec. 2.3–2.4).**

Recommendation: Augment Sec. 2.3–2.4 with a concrete specification: - Define the conditional model explicitly (e.g., $q(z|X_{\text{high}}, X_{\text{low}})$, $p(X_{\text{high}}|z, X_{\text{low}})$, and whether $p(z|X_{\text{low}})$ is used). - Provide layer-by-layer architecture (Conv1D channels, kernel/stride, activations, latent dim), optimizer and schedule, epochs, batch size, normalization/dropout, and any β-VAE weighting. - State the exact KL formula and aggregation/normalization. - Clarify synchronization: whether **25 Hz** is a native stream aligned by timestamps to **100 Hz** or is derived by downsampling; document the exact resampling/downsampling method. - Specify overlap-add details: hop length, overlap fraction, any tapering window, and boundary handling.

7. **Related work is not systematically covered, making it hard to judge novelty and to place the CVAE reconstruction idea within existing step-counting, low-frequency sensing, and time-series super-resolution literature (Sec. 1–2).**

   Recommendation: Add a Related Work section (between Sec. 1 and Sec. 2): - classical step counting (thresholding/peak methods) and learning-based alternatives; - impacts of sampling rate reduction on gait/step detection; - time-series super-resolution / reconstruction for wearable sensors; - VAEs/CVAEs (and diffusion/other generative models if relevant) in gait or activity analysis. Clearly state what is novel here (e.g., placement-specific reconstruction + downstream counting pipeline) while qualifying that step-count accuracy gains are not yet validated.

Minor issues

1. Invalid figures/tables (Figures 1–8 and associated result summaries) remain in the main narrative and can be misread despite caption caveats, especially given the paper's original framing as a performance study (Sec. 3.3–3.4).

   Recommendation: Move invalid performance figures/tables to a clearly labeled appendix ("Outputs from contaminated runs; not scientifically valid"), or remove them entirely from the main Results. If retained anywhere, add prominent in-figure watermarks and include provenance annotations (dataset version, script name, commit hash).

2. Internal inconsistency in low-resolution window length: protocol says **2 s** at **25 Hz** → **50** samples, but an artifact tensor is reported as $X_{\text{low}} = (38071, 51, 3)$, which also breaks the implied **4 : 1** correspondence with **200**-sample **100 Hz** windows (Sec. 2.2; shape mention in Sec. 3.3–3.4).

   Recommendation: Resolve and document the exact indexing convention: - If **51** is correct (inclusive endpoints / center sample handling / padding), update Sec. 2.2 and the **4 : 1** alignment description accordingly. - If **50** is correct, fix preprocessing and ensure scripts fail-fast when shapes deviate from spec. Also clarify window-centering arithmetic (e.g., whether the center sample is included) to avoid off-by-one ambiguity.

3. Peak-detection parameterization risks unfair comparisons across sampling rates: find_peaks 'distance' is in samples but described as fixed across **25 Hz** vs **100 Hz** conditions, implying different physiological constraints in time (Sec. 2.4).

   Recommendation: Define peak constraints in physical units (seconds) and convert to samples per sampling rate/condition. Report tuned parameters separately for **25 Hz**, **100 Hz**, and reconstructed signals, and describe how tuning avoids participant/test leakage (e.g., tuning only on training/validation participants).

4. Title/Abstract framing suggests an evaluated generative method for low-frequency step counting, but the main body reveals the evaluation is invalid/unexecuted; this mismatch affects reader expectations (Title, Abstract, Sec. 1, Sec. 4).

   Recommendation: Revise Title/Abstract to align with the actual contribution. If keeping the case-study angle, signal it early (e.g., "case study", "lessons learned", "post-mortem"). If aiming for a methods paper, separate: (i) proposed pipeline, (ii) what failed and why, (iii) what remains to be evaluated.

5. Demographic subgroup analysis is presented as part of the evaluation plan but lacks design details and (given $N \approx \mathbf{39}$) is likely underpowered; the manuscript should better qualify feasibility (Sec. 2.5, Sec. 4).

   Recommendation: Label subgroup analysis explicitly as planned/unexecuted; add expected subgroup sizes, handling of small-N groups, and planned statistical tests/effect-size reporting. Consider simplifying to an exploratory analysis plan or moving it to Future Work.

6. VAE math is incomplete in places: encoder outputs $\log(\sigma^2)$ but the transformation to σ is not stated; KL term is described but not given as an explicit formula with aggregation/weighting (Sec. 2.3).

   Recommendation: State $\sigma = \exp(0.5 \cdot \log \sigma^2)$ explicitly, and provide the exact KL divergence expression used (sign conventions, sum over latent dims, mean over batch/time) and any β coefficient.

7. Ethics/data-governance considerations are not addressed, despite reconstructing higher-resolution human motion signals potentially increasing identifiability or sensitivity (Sec. 2.1, Sec. 4).

   Recommendation: Add a brief statement on dataset ethics (public/de-identified; no new data collection), and note any privacy implications of reconstructing higher-resolution gait signatures (even if only conceptual).

Very minor issues

1. The manuscript contains unresolved cross-references (e.g., "Section ??") and polishing issues that hinder navigation and signal incompleteness (Sec. 3.2–3.4 and elsewhere).

Recommendation: Replace all placeholders with proper cross-references using the typesetting system's referencing tools; do a final consistency pass on section numbering, capitalization, and terminology.

2. Figure readability/accessibility issues (small fonts, low DPI, color palettes not robust to color-vision deficiency) and repetitive captions reduce clarity (Figures 1–8).

   Recommendation: Export high-DPI or vector figures; increase font/line sizes; use colorblind-safe palettes with redundant encodings; streamline captions and add concise provenance/setting notes per panel.

3. The term "Conditional VAE" is used but the conditioning mechanism is not fully specified (e.g., whether conditioning is on X_{low} at encoder, decoder, and/or prior) (Sec. 2.3).

   Recommendation: Explicitly define what variables condition each distribution (q, p decoder, and optional conditional prior) and reflect this consistently in diagrams/text.

Key statements and references

- • **Wearable accelerometers sampled at high frequencies around 100 Hz provide the detailed signal morphology needed for robust and precise step detection, but continuous collection at this rate in free-living settings is constrained by battery life, storage, and transmission limitations, motivating the use of lower sampling rates such as 25 Hz for prolonged monitoring [11].**
- *Reference(s):* [11]
- • **The OxWalk dataset is a publicly available cohort for physical activity monitoring that provides paired triaxial accelerometer recordings from hip and wrist at 100 Hz and 25 Hz, along with demographic metadata (sex and age range) for 39 unique participants [11].**
- *Reference(s):* [11]
- • **Step counting from accelerometer data in activity monitoring commonly relies on peak detection applied to the Signal Vector Magnitude (SVM), for example using the scipy.signal.find_peaks function with parameters such as peak prominence and minimum inter-peak distance tuned to human gait characteristics [11].**
- *Reference(s):* [11]
- • **Typical step-counting accuracy reported in the literature does not exhibit Mean Absolute Error (MAE) values on the order of thousands of steps or Mean Absolute Percentage Error (MAPE) values approaching 100%, so results of that magnitude (e.g., MAE $>$ 3000 steps and MAPE $\approx$**

100% for CVAE-based methods) are inconsistent with established step-counting performance and indicate invalid or contaminated evaluation data [11].

- *Reference(s):* [11]

Mathematical consistency audit

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

Maths relevance: light

The paper contains limited formal mathematics; most content is methodological description. The main analytic elements are the SVM definition, sampling-rate/window-length relationships, the VAE reparameterization expression, and qualitative loss-function components (MSE + KL). The primary internal inconsistency arises in stated vs reported tensor/window shapes (50 vs 51 samples at 25 **Hz**), along with unit-consistency concerns for peak-detection parameters across sampling rates and missing explicit formulas for KL and latent-variance transformation.

Checked items

1. ✔ **SVM definition** (Sec. 2.1, p.2 (signal characteristics paragraph))
    - **Claim:** Signal Vector Magnitude is computed as $SVM = \sqrt{x^2 + y^2 + z^2}$.
    - **Checks:** dimensional/units, notation consistency
    - **Verdict:** PASS; confidence: high; impact: minor
    - **Assumptions/inputs:** x, y, z are synchronous triaxial accelerometer axes in the same units
    - **Notes:** The expression is well-formed and preserves units (same as axis units, stated as g).
2. ✔ **100 Hz 2-second window length** (Sec. 2.2, p.3 (High-Resolution Target bullet))
    - **Claim:** A 2-second window at 100 **Hz** yields 200 samples (200×3).
    - **Checks:** dimensional/units, counting/sanity check
    - **Verdict:** PASS; confidence: high; impact: minor
    - **Assumptions/inputs:** Sampling rate is exactly 100 samples/second, Window duration is exactly 2 seconds
    - **Notes:** 2 s $\times$ 100 Hz $= 200$ samples is consistent.
3. ✔ **25 Hz 2-second window length** (Sec. 2.2, p.3 (Low-Resolution Input bullet))
    - **Claim:** A 2-second window at 25 **Hz** yields 50 samples (50×3).
    - **Checks:** dimensional/units, counting/sanity check
    - **Verdict:** PASS; confidence: high; impact: minor

- **Assumptions/inputs:** Sampling rate is exactly **25** samples/second, Window duration is exactly **2** seconds
- **Notes:** **2 s** $\times$ **25 Hz** $=$ **50** samples is consistent.

4. ✔ **Claimed 4:1 alignment between 100 Hz and 25 Hz windows** (Sec. 2.2, p.3 (alignment sentence))
    - **Claim:** Temporal alignment is maintained leveraging a $4:1$ sample ratio between **100 Hz** and **25 Hz** data.
    - **Checks:** dimensional/units, consistency across definitions
    - **Verdict:** PASS; confidence: medium; impact: moderate
    - **Assumptions/inputs:** The **25 Hz** stream is perfectly downsampled from the **100 Hz** stream with no offset, Window lengths are **200** samples at **100 Hz** and **50** samples at **25 Hz**
    - **Notes:** Given **200** vs **50** samples, the ratio is $4:1$. This becomes inconsistent later when X_{low} is reported with length **51** (see separate item).

5. ⚠ **Window centering sample counts** (Sec. 2.2, p.3 (High-Resolution Target bullet: "100 before and 99 after"))
    - **Claim:** A **200**-sample window is centered on a step annotation using **100** samples before and **99** samples after the event.
    - **Checks:** counting/sanity check, definition clarity
    - **Verdict:** UNCERTAIN; confidence: medium; impact: minor
    - **Assumptions/inputs:** The annotated event corresponds to a specific sample index within the **200**-sample window
    - **Notes:** **100** (before) $+$**99** (after) $=$ **199** samples; the claim is consistent only if the annotated event sample itself is included as the $\mathbf{200}^{\text{th}}$ sample. This inclusion is not explicitly stated.

6. ✖ **Reported loaded training tensor shapes vs defined window shapes** (Sec. 3.3, p.5 (quoted log: $X_{\text{low}} = (38071, 51, 3)$, $X_{\text{high}} = (38071, 200, 3)$) vs Sec. 2.2–2.3, pp.3–4 (X_{low} 50×3, X_{high} 200×3))
    - **Claim:** The model uses **2**-second **25 Hz** windows (**50** samples) paired with **2** -second **100 Hz** windows (**200** samples).
    - **Checks:** definition consistency, dimensional/shape consistency
    - **Verdict:** FAIL; confidence: high; impact: critical
    - **Assumptions/inputs:** The stated windowing protocol is the one actually used to produce training arrays, Sampling rates are fixed at **25 Hz** and **100 Hz**
    - **Notes:** The paper defines X_{low} as 50×3 for a **2**-second **25 Hz** window, but later reports X_{low} length **51** in loaded data. This contradicts the window-duration/sampling-rate relationship and undermines the stated $4:1$ alignment with $X_{\text{high}} = 200$.

7. ✔ **Reparameterization trick equation** (Sec. 2.3, p.3 (Latent Space Sampling paragraph))
    - **Claim:** Sampling is done via $z = \mu + \sigma \cdot \epsilon$ with $\epsilon \sim \mathcal{N}(0, I)$.
    - **Checks:** algebraic form, notation consistency
    - **Verdict:** PASS; confidence: high; impact: minor
    - **Assumptions/inputs:** μ and σ are vectors of the same dimension as z, ϵ is standard normal of matching dimension
    - **Notes:** The expression is algebraically consistent for Gaussian latent sampling.
8. ⚠ **Consistency of log-variance output with σ used in sampling** (Sec. 2.3, p.3 (Encoder output: $\log(\sigma^2)$ vs sampling uses σ))
    - **Claim:** Encoder outputs μ and $\log(\sigma^2)$, and sampling uses σ in $z = \mu + \sigma \cdot \epsilon$.
    - **Checks:** missing derivation step, notation/definition consistency
    - **Verdict:** UNCERTAIN; confidence: medium; impact: moderate
    - **Assumptions/inputs:** σ is obtained from the reported log-variance vector
    - **Notes:** The necessary transformation from $\log(\sigma^2)$ to σ is not specified. Without this, the sampling step is under-defined symbolically (though likely intended).
9. ✔ **Reconstruction loss definition** (Sec. 2.3, p.4 (Reconstruction Loss bullet))
    - **Claim:** Reconstruction loss is MSE between $X'{\rm high}andX$.
    - **Checks:** dimensional/shape consistency, definition sanity check
    - **Verdict:** PASS; confidence: medium; impact: minor
    - **Assumptions/inputs:** $X'{\rm high}andXhaveidenticalshapes(intended 200\times 3$ per window)
    - **Notes:** Given matching target/output dimensions, MSE is well-defined. Exact averaging convention (per-sample, per-axis, per-window) is not specified but not internally contradictory.
10. ⚠ **KL divergence loss term (explicit formula missing)** (Sec. 2.3, p.4 (KL Divergence Loss bullet))
    - **Claim:** A KL divergence regularizer is used between the learned latent Gaussian and a standard normal prior.
    - **Checks:** missing derivation step, normalization/definition completeness
    - **Verdict:** UNCERTAIN; confidence: high; impact: moderate
    - **Assumptions/inputs:** Latent distribution is Gaussian parameterized by μ and σ^2, Prior is $\mathcal{N}(0, I)$

- **Notes:** No explicit KL expression, sign convention, or normalization is provided; cannot verify algebraic correctness or whether the implemented objective matches the described one.

11. ✖ **Peak-detection distance parameter across sampling rates** (Sec. 2.4, p.4 (peak detection description: fixed 'distance' in samples applied to **100 Hz**, **25 Hz**, and reconstructed **100 Hz**))

    - **Claim:** Using the same peak detection with fixed 'distance' and 'prominence' parameters yields a fair comparison reflecting gait physiology.
    - **Checks:** dimensional/units, methodological consistency
    - **Verdict:** FAIL; confidence: medium; impact: moderate
    - **Assumptions/inputs:** The 'distance' parameter is interpreted as a number of samples (as stated), The same numeric value is used at **25 Hz** and **100 Hz**
    - **Notes:** A fixed minimum peak separation in samples corresponds to different minimum times between peaks at **25 Hz** vs **100 Hz**, so the physiological constraint is not preserved across conditions unless 'distance' is scaled by sampling rate (not described).

Limitations

- The audit is based on the provided **9**-page PDF text/images; the paper contains few explicit equations and omits explicit formulas for the ELBO/KL term and total loss, limiting symbolic verification.
- Several internal references are placeholders (e.g., "Section ??"), so precise cross-referencing of definitions across sections is partially impeded.
- Figures are present but do not add derivation steps; no additional mathematical content is extractable beyond the surrounding captions/text.

Numerical results audit

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

12 numeric checks were run: **11** PASS and **1** FAIL. The only detected internal arithmetic/shape inconsistency is a **51**-sample low-resolution window reported by the training script versus a **50**-sample low-resolution window implied by **2 s** at **25 Hz** and stated elsewhere.

Checked items

1. ✔ **C1** (Page 2, Section 2.2 (Data Splitting and Segmentation Protocol))

    - **Claim:** **39** participants were intended to be divided into training (**31** participants, approximately **80%**) and test (**8** participants, approximately **20%**).
    - **Checks:** parts_vs_total_and_percentage
    - **Verdict:** PASS

- **Notes:** Counts sum exactly ($31 + 8 = 39$); computed fractions $31/39 \approx 0.7949$ and $8/39 \approx 0.2051$ match the approximate 80%/20% claims within tolerance; fractions sum to 1.0.

2. ✔ **C2** (Page 3, Section 2.2 (Segmentation windows))
   - **Claim:** High-resolution target is a 2-second window (200 samples) at 100 Hz; low-resolution input is a 2-second window (50 samples) at 25 Hz.
   - **Checks:** unit_consistent_sample_count
   - **Verdict:** PASS
   - **Notes:** Exact arithmetic matches: $2\ \text{s} \times 100\ \text{Hz} = 200$ samples; $2\ \text{s} \times 25\ \text{Hz} = 50$ samples.
3. ✔ **C3** (Page 3, Section 2.2 (High-resolution centering description))
   - **Claim:** The 200-sample 100 Hz window is centered on the step, with 100 samples before and 99 samples after the annotated event.
   - **Checks:** window_length_consistency
   - **Verdict:** PASS
   - **Notes:** Exact centering relationship holds: 100 (before) +1 (center) +99 (after) = 200.
4. ✔ **C4** (Page 3, Section 2.2 (Sample ratio statement))
   - **Claim:** Alignment leveraged the inherent $4:1$ sample ratio between the 100 Hz and 25 Hz datasets.
   - **Checks:** ratio_consistency
   - **Verdict:** PASS
   - **Notes:** Exact ratio check: $100/25 = 4$.
5. ✔ **C5** (Page 3, Section 2.3 (Encoder input dimensions))
   - **Claim:** Encoder takes low-resolution input window of dimensions 50 samples $\times 3$ axes.
   - **Checks:** dimension_product_check
   - **Verdict:** PASS
   - **Notes:** Implied per-window scalar count: $50 \times 3 = 150$.
6. ✔ **C6** (Page 3, Section 2.3 (Decoder output dimensions))
   - **Claim:** Decoder reconstructs high-resolution window of dimensions 200 samples $\times 3$ axes.
   - **Checks:** dimension_product_check
   - **Verdict:** PASS
   - **Notes:** Implied per-window scalar count: $200 \times 3 = 600$.
7. ✔ **C7** (Page 2 (Section 2.1) and Page 5 (Table 1))

- **Claim:** Mean/Std SVM statistics are given for four sources: Hip_100Hz ($1.012, 0.435$), Hip_25Hz ($1.012, 0.431$), Wrist_100Hz ($0.998, 0.612$), Wrist_25Hz ($0.998, 0.609$).
- **Checks:** repeated_constants_match_across_mentions
- **Verdict:** PASS
- **Notes:** The values match exactly across the two mentions in the provided payload.

8. ✔ **C8** (Page 5, Section 3.2 (Data integrity verification))
   - **Claim:** Reported total count of ground-truth steps for the entire dataset was "Total annotated ground-truth steps: 0".
   - **Checks:** nonnegativity_and_integer_sanity
   - **Verdict:** PASS
   - **Notes:** Sanity check only: value is an integer and nonnegative.
9. ✖ **C9** (Page 5, Section 3.3 (Training script loaded shapes))
   - **Claim:** Training script reported "Loaded data shapes: $X_{\text{low}} = (38071, 51, 3)$, $X_{\text{high}} = (38071, 200, 3)$".
   - **Checks:** shape_consistency_and_dimensional_logic
   - **Verdict:** FAIL
   - **Notes:** Window counts match (38071) and axes=3, but X_{low} time length is 51, differing from the earlier stated 50 for a 2-second 25 Hz window.
10. ✔ **C10** (Page 6, Figure 1 caption (Hip loss convergence))
    - **Claim:** Hip CVAE loss converging to approximately 0.994 over 25 epochs.
    - **Checks:** approximate_value_format_sanity
    - **Verdict:** PASS
    - **Notes:** Sanity checks only: epochs is a positive integer; loss is finite and non-negative.
11. ✔ **C11** (Page 6, Figure 3 caption (Wrist epochs))
    - **Claim:** Wrist CVAE model trained over 30 epochs.
    - **Checks:** integer_sanity
    - **Verdict:** PASS
    - **Notes:** Sanity check only: epochs is a positive integer.
12. ✔ **C12** (Page 7, Table 2 (Overall performance metrics))
    - **Claim:** Table 2 lists MAE (steps) and MAPE (%) for 6 rows (Hip/Wrist × 100Hz/25Hz/CVAE): Hip 100Hz MAE 213.12 MAPE 11.87%; Hip 25Hz MAE 83.50 MAPE 5.71%; Hip CVAE MAE 3102.25 MAPE 99.92%; Wrist 100Hz MAE 1588.12 MAPE 141.80%; Wrist 25Hz MAE 1169.38 MAPE 107.43%; Wrist CVAE MAE 3099.38 MAPE 99.69%.

- **Checks:** percentage_format_and_nonnegativity
- **Verdict:** PASS
- **Notes:** Sanity checks only: all MAE and MAPE values are finite and non-negative; MAPE exceeds 100% for wrist_100 and wrist_25 (not treated as an error).

Limitations

- Only parsed PDF text/images were available; no access to underlying dataset files, scripts, logs, or generated artifacts referenced in the paper.
- Plot-based numeric verification (reading values from Figures 1-8) is out of scope; only caption-embedded numbers can be sanity-checked.
- Many claims are qualitative or procedural (e.g., contamination, missing columns) and cannot be independently validated without external run artifacts.
- FAST checks here are limited to internal arithmetic/ratio/dimension consistency using explicit numbers present in the PDF.