GAITREADER: FROM LARGE-SCALE HEALTHY GAIT PRETRAINING TO CROSS-CENTER DISEASE RECOGNITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Six-degree-of-freedom (6-DoF) knee kinematics provide a non-invasive means of assessing lower-limb function and recognizing musculoskeletal disorders. However, existing learning-based approaches largely depend on disease-specific annotations, which are difficult to obtain at scale across diverse patient populations and clinical centers. Although learning from healthy movement has shown promise for gait analysis, whether healthy-only pretraining can support label-efficient knee disease recognition across centers remains insufficiently explored. To investigate this question, we introduce KineGait, a multicenter dataset containing 4,168 healthy participants for self-supervised training and validation, alongside Healthy, anterior cruciate ligament deficiency (ACLD), and knee osteoarthritis (KOA) cohorts for downstream evaluation. Building on this resource, we propose GaitReader, which organizes recordings into cycle-level gait units and learns a gait vocabulary comprising six DoF-specific codebooks. The frozen tokenizer provides discrete targets for contextual pretraining, combining masked code prediction with attribute prediction and waveform reconstruction without disease annotations. Following supervised fine-tuning, GaitReader achieves mean accuracies of 93.04% and 89.87% on fixed internal and external test sets, respectively, across five training–validation folds. Pretraining improves accuracy over the same architecture trained from scratch at every evaluated label fraction; using half of the training labels, it exceeds scratch training with all labels on both test sets. These findings support healthy gait as a source of transferable representations for label-efficient disease recognition across the evaluated centers.

1 INTRODUCTION

Gait analysis provides a non-invasive means of assessing lower-limb function during walking. Compared with analyses based on a single joint angle or static structural information, six-degree-of-freedom (6-DoF) knee kinematics simultaneously capture flexion-extension, varus-valgus, and internal-external rotation, together with anterior-posterior, medial-lateral, and superior-inferior translations, providing a comprehensive description of knee motion during walking (Lafortune et al., 1992; Li et al., 1996; Zhang et al., 2015). Such kinematic measurements have revealed systematic alterations associated with musculoskeletal disorders and have been used for functional assessment and automated disease recognition, including knee osteoarthritis (KOA) and anterior cruciate ligament deficiency (ACLD) (Zeng et al., 2017; Wu et al., 2018). For broader clinical application, gait-based models need to capture information relevant to disease recognition while generalizing across individuals and clinical centers. This motivates the development of transferable representations of knee kinematics that can support recognition across different disease populations.

However, existing learning-based approaches to 6-DoF knee kinematics remain predominantly supervised and depend on disease-specific annotations (Wu et al., 2018; Wang et al., 2022; Zeng et al., 2022). Their representations are therefore shaped by the disease categories and patient cohorts available during training, which may not adequately capture the variability encountered across individuals and clinical centers. Expanding this coverage requires large, diverse patient cohorts with reliable clinical annotations, which are difficult to assemble at scale. In contrast, healthy gait data, which are more readily obtainable in large quantities, provide a more scalable resource for learning recur-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

ring patterns of human walking. Self-supervised learning offers a way to exploit these recordings by deriving training signals from the motion itself, without requiring disease annotations. Recent studies illustrate the potential of learning from normal or healthy movement: FSGait learns normal gait patterns through reconstruction and prediction for abnormality detection (Duan et al., 2024), while GaitDynamics learns a generative model of healthy gait through diffusion-based denoising and demonstrates generalization to biomechanical prediction in individuals with osteoarthritis (Tan et al., 2026). However, these studies focus on abnormality detection or biomechanical prediction, leaving insufficiently explored whether healthy-only pretraining can yield representations for distinguishing knee disorders with limited disease annotations across clinical centers.

Addressing this question requires considering both how knee kinematics are represented and what supervisory targets are used during pretraining. Contrastive methods learn representations by contrasting augmented views of time series (Eldele et al., 2021), whereas reconstruction-based methods recover masked signal values (Goswami et al., 2024). For disease recognition, augmentation-induced invariances require careful consideration, as waveform variations may carry clinically relevant information; direct reconstruction, meanwhile, does not explicitly group recurring patterns into shared units. Vector quantization provides a complementary basis for representation learning by mapping continuous features to discrete codes, as established by VQ-VAE (van den Oord et al., 2017). Subsequent vector-quantized approaches have explored discrete representations for temporal shapes and physiological signals, including VQShape, VQ-MTM, and HeartLang (Wen et al., 2024; Gui et al., 2024; Jin et al., 2025). This formulation is particularly appealing for gait: repeated walking cycles provide multiple realizations of similar motion patterns, which can be represented by shared prototypes despite differences between individual recordings. Such prototypes offer discrete targets for learning relationships among recurring patterns, motivating the investigation of vocabulary-based pretraining for transferable knee kinematic representations.

To investigate this direction, we construct KineGait, a multicenter dataset of bilateral 6-DoF knee kinematics comprising a large healthy cohort for self-supervised pretraining and Healthy, ACLD, and KOA cohorts for downstream disease recognition and cross-center evaluation. Building on this resource, we propose GaitReader, a framework that learns reusable gait representations from healthy gait before introducing disease supervision. GaitParser first organizes continuous recordings into cycle-level gait units. VQ-Gait then learns a gait vocabulary comprising six DoF-specific codebooks, which represent recurring waveform shapes as discrete codes. With the tokenizer frozen, GaitFormer learns contextual representations through masked code prediction, attribute prediction, and waveform reconstruction, capturing relationships across cycles, DoFs, and sides. The pretrained encoder is subsequently fine-tuned for disease classification. Through these approaches, our method can reduce reliance on patient annotations while supporting disease recognition across clinical centers. The main contributions of this work are summarized below:

- We introduce KineGait, a large-scale, multicenter dataset of 6-DoF knee kinematics containing 4,168 healthy participants for self-supervised pretraining, alongside healthy, ACLD, and KOA cohorts for downstream and cross-center evaluation. The code and KineGait dataset will be publicly released upon acceptance to support reproducible research.
- We propose GaitReader, a self-supervised framework that organizes knee kinematics into cycle-level gait units and learns a gait vocabulary comprising six DoF-specific codebooks. It combines discrete code prediction with continuous attribute prediction and waveform reconstruction to learn contextual gait representations from healthy subjects.
- We systematically evaluate healthy-to-disease representation transfer through comparisons with supervised and self-supervised baselines, controlled ablations, and experiments with varying amounts of labeled training data.

2 Related Work

2.1 Six-Degree-of-Freedom Knee Kinematics Analysis

Six-degree-of-freedom (6-DoF) knee kinematics have been widely used to characterize three-dimensional knee motion during walking. Early studies established normative rotational and translational motion patterns in healthy subjects (Lafortune et al., 1992; Li et al., 1996; Zhang et al., 2015), while subsequent work examined disease-related alterations, including those associated with

KOA (Zeng et al., 2017). Building on these observations, Wu et al. (2018) and Wang et al. (2022) developed neural-network and support-vector-machine models for ACLD detection, respectively, and Zeng et al. (2022) constructed KOA diagnostic and treatment-decision models using 1,201 participants. Complementing these supervised approaches, Li et al. (2024) established normative gait references from 500 healthy participants for functional assessment of ACL-injured knees.

2.2 Representation Learning for Gait Analysis

Representation learning has been explored as a means of reducing reliance on task-specific clinical gait data. GaitForeMer demonstrated that pretraining on general human motion can support Parkinsonian gait severity estimation with limited clinical examples (Endo et al., 2022). A related direction learns from normal or healthy movement: FSGait uses self-supervised reconstruction and prediction to identify gait abnormalities (Duan et al., 2024), while GaitDynamics learns healthy gait dynamics that generalize to biomechanical prediction in individuals with osteoarthritis (Tan et al., 2026). Masked reconstruction of joint angles has similarly improved data efficiency in joint moment estimation (Li et al., 2025). Complementing these approaches, studies incorporating pathological gait seek representations that support multiple clinical endpoints. PathoFM learns from patient gait cycles for classification and regression (Dey & Paez-Granados, 2025), while GaitEncoder uses mixed healthy and pathological populations to learn representations transferable across clinical tasks and previously unseen disorders (Magruder et al., 2026).

2.3 Self-Supervised Representation Learning for Time-Series Data

Self-supervised learning has emerged as an effective strategy for learning time-series representations from unlabeled data. Contrastive methods such as TS-TCC and TS2Vec learn invariant representations across temporal views and scales (Eldele et al., 2021; Yue et al., 2022), while masked and large-scale pretraining methods such as MOMENT learn reusable representations through reconstruction-based objectives (Goswami et al., 2024). VQ-VAE provides a foundation for learning discrete latent representations (van den Oord et al., 2017). Subsequent vector-quantized approaches include VQShape for shape-and-attribute representations (Wen et al., 2024), VQ-MTM for masked EEG modeling with random-projection quantization (Gui et al., 2024), and HeartLang for heartbeat vocabulary learning and contextual ECG pretraining (Jin et al., 2025). In clinical gait analysis, GAIT-Gen combines residual vector quantization and masked transformers for pathology-conditioned motion generation (Adeli et al., 2026).

3 Method

In this section, we describe the structure of the GaitReader framework. We denote a bilateral recording by $\mathcal{X} \in \mathbb{R}^{S\times D\times T}$, where $S = 2$ denotes the two sides, $D = 6$ denotes the number of kinematic degrees of freedom, and $T = 600$ is the number of time steps in each recording. The overview of the framework is shown in Figure 1. The use of the framework can be divided into four steps. First, GaitParser constructs cycle-level gait units (Section 3.1). Second, VQ-Gait learns a gait vocabulary from healthy subjects (Section 3.3). Third, GaitFormer is pretrained through masked code prediction, attribute prediction, and waveform reconstruction (Section 3.4). Finally, the pretrained GaitFormer is fine-tuned with labeled data for downstream disease classification.

3.1 GaitParser: Constructing Cycle-Level Gait Units

GaitParser defines a gait unit as one DoF waveform within a detected cycle. For each side, we estimate the cycle period from the flexion–extension trajectory and locate extension minima following swing-flexion peaks as kinematic anchors. Consecutive anchors delimit candidate cycles, which are filtered by absolute and period-relative duration limits. The retained boundaries are shared across all six DoFs, preserving their within-cycle correspondence.

Let $[a_{s,n}, b_{s,n})$ denote the sample interval of retained cycle n on side s. We resample each original waveform to L points and organize the resulting units chronologically:

$$\mathbf{x}_{s,n,d} = \frac{\mathcal{R}_L\left(\mathcal{X}_{s,d,\,a_{s,n}:b_{s,n}}\right) - m_d}{v_d}, \tag{1}$$

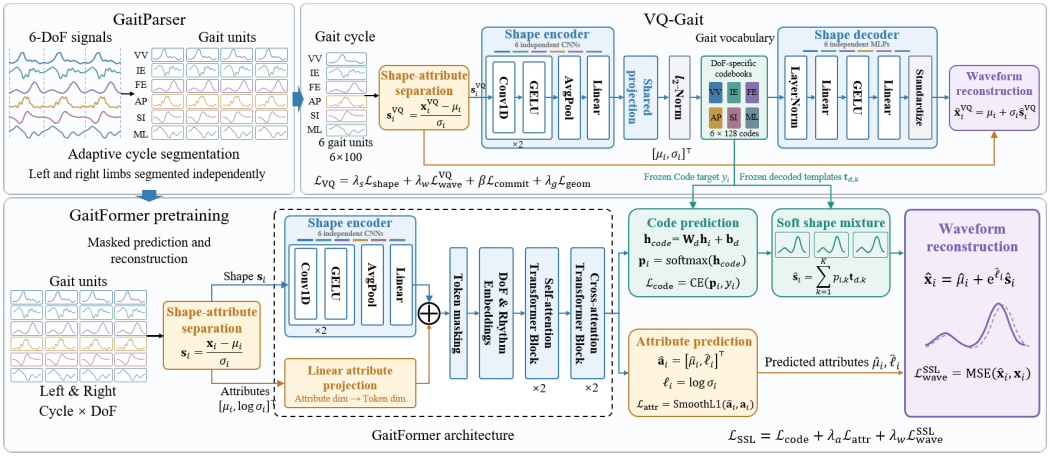

Figure 1: **Overview of GaitReader.** GaitParser constructs cycle-level gait units, and VQ-Gait learns a gait vocabulary comprising six DoF-specific codebooks. GaitFormer is pretrained through masked code prediction, attribute prediction, and waveform reconstruction.

where $\mathcal{R}_L$ is linear interpolation, N_s is the retained cycle count, and m_d, v_d are the channel mean and standard deviation fitted on accepted healthy training cycles and subsequently fixed. Units are ordered into $\mathcal{G}_s \in \mathbb{R}^{N_s \times 6 \times L}$ for each side independently. Padding masks identify valid cycles.

To preserve gait rhythm information after resampling, we retain cycle duration and the interval between consecutive retained cycle centers:

$$\tau_{s,n} = \frac{b_{s,n} - a_{s,n}}{f_{\text{samp}}}, \qquad \delta_{s,n} = \frac{c_{s,n} - c_{s,n-1}}{f_{\text{samp}}} \quad (n > 1), \tag{2}$$

where $c_{s,n} = (a_{s,n} + b_{s,n})/2$ and f_{samp} is the sampling frequency. These descriptors accompany the waveform units, allowing subsequent modeling of both cycle morphology and temporal context. Further details of GaitParser are provided in Appendix A.1.

3.2 GaitFormer: Encoding Multi-DoF Gait Context

Shape and attribute embeddings Inspired by the separation of shape and attributes in VQShape (Wen et al., 2024), we represent each gait unit by its normalized shape and amplitude attributes. For a unit $\mathbf{x}$ from DoF d, we compute its mean μ and population standard deviation σ, floored for numerical stability. The token embedding is

$$\mathbf{e} = \phi_d\left(\frac{\mathbf{x} - \mu}{\sigma}\right) + \psi_d\left([\mu, \log\sigma]^\top\right), \tag{3}$$

where ϕ_d is a 1-D convolutional encoder with temporal pooling and projection, and ψ_d is a linear attribute projection. This retains offset and amplitude alongside waveform shape. Both mappings are DoF-specific and shared between sides.

DoF and rhythm embeddings For each side, we flatten the token grid into a sequence $\mathbf{E} \in \mathbb{R}^{N_s D \times m}$ by concatenating all D DoF tokens of each cycle in cycle order. Here, m is the embedding dimension. We then add

$$\mathbf{E}' = \mathbf{E} + \mathbf{E}_{\text{DoF}} + \mathbf{E}_{\text{rhythm}}. \tag{4}$$

Here, $\mathbf{E}_{\text{DoF}}$ identifies each token's DoF, while $\mathbf{E}_{\text{rhythm}}$ is obtained by projecting cycle duration and inter-cycle interval through a shared MLP. The same rhythm embedding is added to all six tokens of a cycle. A learnable summary token is prepended to each side sequence.

Transformer encoder We adopt multi-head attention (Vaswani et al., 2017), whose per-head operation is

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V}, \tag{5}$$

where $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ are projected queries, keys, and values, and d_{k} is the key dimension. Within-side self-attention derives $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ from the same side sequence, allowing each gait-unit token to attend to all valid tokens across its cycles and DoFs. Each token aggregates within-cycle and cross-cycle information using waveform content, DoF identity, and timing descriptors. Bilateral cross-attention then uses queries from one side and keys and values from the other. Both directions share parameters and are updated simultaneously. Attention and feed-forward sublayers use pre-layer normalization (Xiong et al., 2020). Further architectural details of GaitFormer are provided in Appendix A.2.

3.3 VQ-GAIT: LEARNING A DOF-SPECIFIC GAIT VOCABULARY

VQ-Gait learns reusable discrete gait codes from healthy gait units through vector-quantized reconstruction, following VQ-VAE (van den Oord et al., 2017). We maintain a separate codebook for each DoF to capture its characteristic waveform patterns. The resulting gait vocabulary is shared across subjects and sides.

Shape encoding and vector quantization We index gait units by i, with $d = d(i)$ denoting the associated DoF. Using the shape-attribute separation in Section 3.2, we obtain the normalized shape $\mathbf{s}_i^{\mathrm{VQ}} = \left(\mathbf{x}_i^{\mathrm{VQ}} - \mu_i\right)/\sigma_i$. Each kinematic DoF is encoded by an independent convolutional encoder, followed by a shared projection head and ℓ_2-normalized feature $\mathbf{z}_i$. The encoder processes each gait unit independently, without incorporating information from adjacent cycles or the contralateral side. For the corresponding codebook $\mathcal{C}_d = \{\mathbf{c}_{d,k}\}_{k=1}^{K}$, both features and code vectors have unit norm, and quantization selects

$$y_i = \arg\min_k \|\mathbf{z}_i - \mathbf{c}_{d,k}\|_2^2, \qquad \mathbf{z}_{\mathrm{q},i} = \mathbf{c}_{d,y_i}. \tag{6}$$

Here, K is the codebook size and y_i represents waveform shape, while the continuous attributes μ_i and σ_i bypass quantization.

Shape and waveform reconstruction A DoF-specific decoder g_d reconstructs the normalized shape; the retained attributes restore its offset and scale:

$$\widehat{\mathbf{s}}_i^{\mathrm{VQ}} = \mathrm{Standardize}\left(g_d\left(\mathbf{z}_{\mathrm{q},i}\right)\right), \qquad \widehat{\mathbf{x}}_i^{\mathrm{VQ}} = \mu_i + \sigma_i \widehat{\mathbf{s}}_i^{\mathrm{VQ}}. \tag{7}$$

Here, Standardize normalizes each decoded waveform $g_d\left(\mathbf{z}_{\mathrm{q},i}\right)$ using its own mean and floored population standard deviation. This gives decoded shapes a common normalization before the input attributes are reapplied. The learning objective is

$$\mathcal{L}_{\mathrm{VQ}} = \lambda_s \mathcal{L}_{\mathrm{shape}} + \lambda_w \mathcal{L}_{\mathrm{wave}}^{\mathrm{VQ}} + \beta \mathcal{L}_{\mathrm{commit}} + \lambda_g \mathcal{L}_{\mathrm{geom}}. \tag{8}$$

The first two terms use smooth-L_1 reconstruction losses against $\mathbf{s}_i^{\mathrm{VQ}}$ and $\mathbf{x}_i^{\mathrm{VQ}}$, respectively. Commitment loss anchors encoder features to assigned codes; geometry regularization matches within-DoF shape neighborhoods in the continuous latent space. Codebooks are initialized by k-means and updated using exponential moving averages (EMA). All loss components and further details of VQ-Gait are provided in Appendix A.3.

3.4 MASKED PREDICTION AND RECONSTRUCTION FOR GAITFORMER PRETRAINING

After vocabulary learning, we freeze the entire VQ-Gait tokenizer and train a separately initialized GaitFormer. The tokenizer provides code targets and decoded shape templates; GaitFormer takes gait units as input and embeds their shapes and attributes. We sample valid gait units to form a mask set $\mathcal{M}$ and replace their combined shape-attribute embeddings $\mathbf{e}_i$ with a learnable mask token, retaining DoF and rhythm embeddings. All three objectives below use the same masked positions and contextual features $\mathbf{h}_i$ from a single backbone pass.

Masked code and attribute prediction The frozen tokenizer assigns a target code y_i to each original gait unit. GaitFormer uses visible units within and across sides to contextualize each masked

position. DoF-specific linear heads predict its code distribution and continuous attributes, represented by $\mathbf{a}_i = [\mu_i, \ell_i]^\mathsf{T}$, where $\ell_i = \log \sigma_i$. We denote the predicted attributes by $\widehat{\mathbf{a}}_i = [\widehat{\mu}_i, \widehat{\ell}_i]^\mathsf{T}$ and optimize code cross-entropy and a smooth-L_1 loss summed over the two attribute components:

$$\mathbf{p}_i = \mathrm{softmax}\left(\mathbf{W}_d \mathbf{h}_i + \mathbf{b}_d\right), \qquad \mathcal{L}_{\mathrm{code}} = -\frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} \log p_{i,y_i}, \tag{9}$$

$$\widehat{\mathbf{a}}_i = \mathbf{A}_d \mathbf{h}_i + \mathbf{b}_d^{\mathrm{attr}}, \qquad \mathcal{L}_{\mathrm{attr}} = \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} \mathrm{SmoothL1}(\widehat{\mathbf{a}}_i, \mathbf{a}_i). \tag{10}$$

Here, p_{i,y_i} denotes the probability assigned to the target code.

Masked waveform reconstruction Let $\mathbf{t}_{d,k} = \mathrm{Standardize}(g_d(\mathbf{c}_{d,k}))$ denote the frozen decoded shape template for code k. The predicted code probabilities form a shape mixture, $\widehat{\mathbf{s}}_i = \sum_{k=1}^{K} p_{i,k} \mathbf{t}_{d,k}$. We recover the waveform using the predicted attributes and minimize the reconstruction error averaged over masked units and their L samples:

$$\widehat{\mathbf{x}}_i = \widehat{\mu}_i + \exp(\widehat{\ell}_i) \widehat{\mathbf{s}}_i, \qquad \mathcal{L}_{\mathrm{wave}}^{\mathrm{SSL}} = \frac{1}{|\mathcal{M}| L} \sum_{i \in \mathcal{M}} \|\widehat{\mathbf{x}}_i - \mathbf{x}_i\|_2^2 . \tag{11}$$

The total target loss function for this stage is:

$$\mathcal{L}_{\mathrm{SSL}} = \mathcal{L}_{\mathrm{code}} + \lambda_a \mathcal{L}_{\mathrm{attr}} + \lambda_w \mathcal{L}_{\mathrm{wave}}^{\mathrm{SSL}}. \tag{12}$$

4 Experiments

4.1 Experimental Setup

Dataset To train and evaluate GaitReader, we construct KineGait, a multicenter dataset of bilateral 6-DoF knee kinematics collected using the OptiKnee optical motion capture system (Shanghai Innomotion Co., Ltd., China). The healthy pretraining cohort comprises 4,168 subjects, divided into 3,751 training and 417 validation subjects; after quality control, 3,747 and 417 subjects remain, respectively. For downstream Healthy/ACLD/KOA classification, KineGait includes a development cohort of 689 subjects and an external test cohort of 318 subjects. Patients are recruited from two hospitals for development and a third hospital for external testing; healthy controls come from separate community sources. We reserve 20% of the development cohort for internal testing. After quality control, 549 subjects remain available for downstream training and validation, while the fixed internal and external test sets contain 138 and 316 subjects, respectively. Further dataset details are provided in Appendix B. The study received institutional ethics approval, and written informed consent was obtained; further details are provided in the Ethics Statement.

Evaluation protocol We partition the 549 downstream training and validation subjects into five folds using subject-level stratification by disease class. Each model is trained on four folds, with the remaining fold used for checkpoint selection based on validation macro-F1. The selected model is evaluated on the same fixed internal and external test sets in every fold, and we report mean performance across the five fold-specific models. Pretrained methods reuse their respective pretrained checkpoints across folds, with downstream training performed separately for each fold. Neither test set is used for model selection.

Compared methods We compare GaitReader with three supervised baselines: TimesNet (Wu et al., 2023), PatchTST (Nie et al., 2023), and iTransformer (Liu et al., 2024); and five self-supervised baselines: TS-TCC (Eldele et al., 2021), TS2Vec (Yue et al., 2022), T-Rep (Fraikin et al., 2024), VQShape (Wen et al., 2024), and HeartLang (Jin et al., 2025). The supervised baselines are trained from scratch on labeled downstream data, whereas the self-supervised baselines are pretrained on the same healthy cohort before downstream adaptation. All methods share the same subject eligibility criteria, training–validation folds, and fixed test sets. GaitReader and the gait-adapted HeartLang operate on cycle-level gait units; the remaining methods process full-length recordings from each side. Method-specific adaptations are detailed in Appendix C.3.

Table 1: Healthy/ACLD/KOA classification performance on the internal and external test sets. Accuracy and macro-F1 (%) are averaged across five fold-specific models evaluated on fixed test sets. Best results are shown in bold.

Method	Internal		External	
	Acc.	Macro-F1	Acc.	Macro-F1
TimesNet (Wu et al., 2023)	79.42	74.32	48.23	47.02
PatchTST (Nie et al., 2023)	76.38	74.03	63.92	62.74
iTransformer (Liu et al., 2024)	74.06	68.59	67.22	67.15
TS-TCC (Eldele et al., 2021)	74.35	70.77	53.86	52.70
TS2Vec (Yue et al., 2022)	87.25	84.23	85.76	85.54
T-Rep (Fraikin et al., 2024)	87.83	85.73	83.61	83.37
VQShape (Wen et al., 2024)	84.64	80.06	73.61	73.90
HeartLang (Jin et al., 2025)	91.45	89.88	74.56	75.23
GaitReader (ours)	**93.04**	**92.09**	**89.87**	**89.42**

Implementation GaitParser resamples each cycle to 100 points. VQ-Gait uses six DoF-specific codebooks, each containing 128 codes with a dimensionality of 128. During GaitFormer pretraining, we randomly mask 15% of valid gait units. We optimize GaitReader using AdamW with a learning rate of 3×10^{-4} and a batch size of 32. Vocabulary learning and self-supervised pretraining each run for up to 200 epochs, while downstream training runs for up to 200 epochs per fold. Early stopping uses a patience of 10 epochs, with checkpoints selected by healthy validation loss during pretraining and validation macro-F1 during downstream training. Full fine-tuning updates both the backbone and classifier. Experiments are conducted on an NVIDIA GeForce RTX 5080 GPU, with seed 42 used for data partitioning and training. Additional implementation details are provided in the appendix.

4.2 Disease Recognition and Cross-Center Generalization

Table 1 evaluates whether representations learned from healthy gait can support disease recognition within and across centers. GaitReader achieves the highest mean accuracy and macro-F1 among the compared methods on both test sets, with accuracies of 93.04% internally and 89.87% externally. Following supervised adaptation, the model distinguishes Healthy, ACLD, and KOA subjects despite using only healthy subjects during self-supervised pretraining. This finding supports the use of healthy gait as a source of transferable representations for downstream disease recognition.

The external evaluation further examines whether this transfer extends to subjects from a different center. While HeartLang is the strongest baseline internally, TS2Vec performs best among the baselines externally, indicating that strong internal performance does not necessarily translate into equally strong performance across centers. GaitReader maintains its leading performance under this center shift, although its accuracy decreases relative to the internal evaluation. Together, these results support the applicability of the learned representations across the evaluated cohorts.

4.3 Label-Efficient Transfer

To examine whether healthy-only pretraining reduces the need for labeled downstream training data, we compare pretrained GaitReader with the same architecture trained from scratch using 10%, 25%, 50%, and 100% of the training labels. Within each fold, both models use identical class-stratified labeled subsets, while the validation and test sets remain unchanged. As shown in Table 2, pretraining yields higher mean accuracy at every label fraction on both test sets, indicating that its benefits extend to settings with limited disease annotations.

With only half of the training labels, pretrained GaitReader surpasses the scratch model trained with all labels, achieving 91.88% versus 88.41% internal accuracy and 87.72% versus 82.03% external accuracy. At the lowest label fraction of 10%, pretraining also improves internal and external accuracy by 4.78 and 9.18 percentage points, respectively. These findings suggest that representations learned from healthy gait provide a useful initialization for disease recognition, allowing supervised

Table 2: Label-efficient disease recognition with pretrained GaitReader and the same architecture trained from scratch. Accuracies (%) are averaged across five fold-specific models evaluated on fixed test sets. Label fractions refer to each fold's training subjects; improvements are reported in percentage points.

Method	Internal Acc.				External Acc.			
	10%	25%	50%	100%	10%	25%	50%	100%
Scratch	79.86	85.07	85.65	88.41	74.56	80.13	78.23	82.03
Full fine-tuning	84.64	88.70	91.88	93.04	83.73	85.82	87.72	89.87
Improvement	4.78	3.62	6.23	4.64	9.18	5.70	9.49	7.85

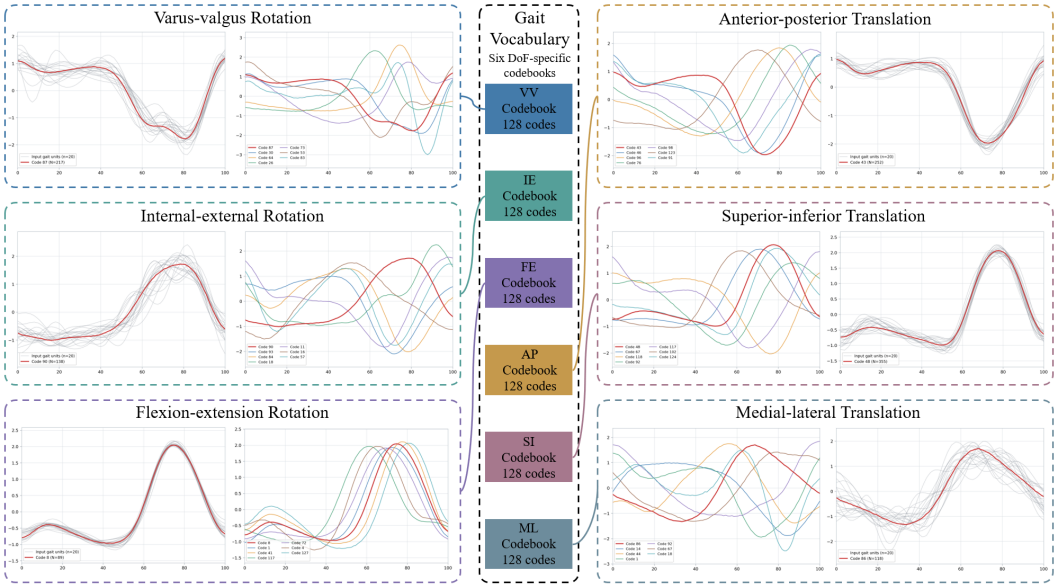

Figure 2: **Visualization of the gait vocabulary.** Inner panels show seven decoded prototypes from each 128-code DoF-specific codebook. Outer panels compare the highlighted prototypes with 20 randomly sampled assigned healthy validation gait units (gray). All curves show standardized waveform shapes.

adaptation with fewer labeled subjects. The benefit extends to the external cohort, supporting label-efficient transfer across the evaluated centers.

4.4 Gait Vocabulary Visualization

Figure 2 examines whether the learned gait vocabulary captures recurring waveform shapes. Within each DoF, the decoded prototypes differ in the relative timing and shape of peaks and troughs. For example, the displayed flexion–extension prototypes share a prominent peak whose position varies across codes, whereas varus–valgus and medial–lateral motion exhibit more varied profiles. Because all curves are standardized, these differences reflect waveform morphology beyond absolute amplitude and offset. The vocabulary thus provides a compact description of variation within each motion channel, making the learned discrete targets directly inspectable in waveform space.

The outer panels show that units assigned to the same highlighted code follow a shared profile while retaining variation across cycles and subjects. Each highlighted code receives units from multiple healthy validation subjects, supporting its interpretation as a reusable motion pattern. Correspondence is tighter for flexion–extension than for medial–lateral motion, a difference also observed in the full validation analysis in Appendix E. This variation indicates that the prototypes summarize common structure with differing fidelity across DoFs. These findings support using the vocabulary to define recurring shape targets for pretraining; continuous attributes and contextual relationships

Table 3: Ablation study of GaitReader. Accuracy and macro-F1 (%) are averaged across five fold-specific models evaluated on fixed internal and external test sets. Best results are shown in bold.

Variant	Internal		External	
	Acc.	Macro-F1	Acc.	Macro-F1
w/o gait vocabulary	90.29	89.35	88.10	87.80
w/o waveform loss	92.75	91.88	87.72	87.22
w/o attribute and waveform losses	91.88	90.54	88.61	88.16
w/o GaitParser	88.41	86.62	82.09	82.30
GaitReader (full)	**93.04**	**92.09**	**89.87**	**89.42**

provide additional information in GaitReader. Prototype selection, usage distributions, quantitative correspondence, and the complete vocabulary are detailed in Appendix E.

4.5 Ablation Studies

Table 3 examines the contributions of cycle-based segmentation, the gait vocabulary, and the pretraining objectives. Replacing GaitParser's cycle-based segmentation with fixed-length windows reduces performance on both test sets, with external accuracy decreasing from 89.87% to 82.09%. This result supports organizing recordings into complete gait cycles as meaningful units for representation learning.

To assess the vocabulary-based pretraining design, we replace it with direct masked waveform reconstruction optimized using mean squared error. This variant performs worse on both cohorts, suggesting that direct waveform reconstruction alone is less effective than the proposed combination of discrete motion targets and continuous reconstruction objectives. The comparison supports the vocabulary-based design as a whole, rather than isolating the effect of discrete supervision alone.

The objectives exhibit complementary but non-additive effects. Adding attribute supervision to code prediction improves internal accuracy and macro-F1 but lowers both external metrics. Adding waveform reconstruction to these two objectives improves all four metrics, giving the full model the best performance among the main ablations. Thus, attribute supervision alone does not consistently improve transfer; the results support jointly learning discrete codes and continuous attributes through waveform reconstruction. Additional module, embedding, and vocabulary-size analyses are provided in Appendix D.3 and Appendix D.4.

5 Conclusion and Limitations

Conclusion We introduce KineGait and GaitReader to investigate healthy-only pretraining for transferable disease recognition from 6-DoF knee kinematics. KineGait comprises 4,168 healthy participants for self-supervised training and validation, alongside Healthy, ACLD, and KOA cohorts for downstream evaluation across clinical centers. GaitReader first adaptively segments raw recordings into cycle-level gait units using GaitParser, then learns a gait vocabulary comprising six DoF-specific codebooks and contextual representations through masked code prediction, attribute prediction, and waveform reconstruction. After fine-tuning, these representations support disease recognition on internal and external cohorts and improve accuracy over training from scratch across all evaluated label fractions. These findings support healthy gait as a pretraining resource for reducing reliance on disease annotations across the evaluated centers. The code and dataset will be publicly released upon acceptance to support further research.

Limitations Our evaluation covers Healthy/ACLD/KOA classification using a single motion capture system; transfer across acquisition systems and disease severity assessment remain to be evaluated. Differences in demographic composition and recruitment settings warrant further validation in more closely matched cohorts. These considerations define the scope of the current cross-center findings.

AI USE STATEMENT

In this work, we used generative AI tools to assist with translation, writing portions of the implementation and analysis code, checking consistency between mathematical descriptions and their implementation. Additionally, we used these tools for literature searches, polishing and revising manuscript text, preparing LaTeX and references, and creating scientific plots and tables from recorded experimental outputs. All AI-assisted work was reviewed and verified by the authors. We take responsibility for the final content of this work, including text, claims, code, and other artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This study was conducted in accordance with the Declaration of Helsinki and approved by the relevant institutional ethics committee, with approval covering all participating centers and cohorts. Written informed consent was obtained from participants or their legal guardians before participation, with participant assent where applicable. The committee name and approval number are withheld for double-blind review. Direct personal identifiers were removed, and participants were assigned coded identifiers to protect their privacy. Permission was obtained for publication of participant photographs and public sharing of the de-identified data. The dataset will be publicly released upon acceptance, subject to the applicable ethics approval and participant consent. The code will also be publicly released upon acceptance.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we provide detailed descriptions of GaitParser, the model architecture, and training objectives in Appendix A; dataset composition, preprocessing, and subject partitions in Appendix B; and training configurations, baseline adaptations, and evaluation protocols in Appendix C. All methods use the same downstream training–validation folds and fixed test sets, with implementation settings and random seeds specified. The code and dataset will be publicly released upon acceptance.

REFERENCES

Vida Adeli, Soroush Mehraban, Majid Mirmehdi, Alan Whone, Benjamin Filtjens, Amirhossein Dadashzadeh, Alfonso Fasano, Andrea Iaboni, and Babak Taati. GAITGen: Disentangled motion-pathology impaired gait generative model – bringing motion generation to the clinical domain. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 3150–3161, 2026. doi: 10.1109/WACV61042.2026.00308. URL https://openaccess.thecvf.com/content/WACV2026/html/Adeli_GAITGen_Disentangled_Motion-Pathology_Impaired_Gait_Generative_Model_--_Bringing_Motion_WACV_2026_paper.html.

Sharmita Dey and Diego Paez-Granados. PathoFM: Toward a foundation model for pathological gait. In *NeurIPS Workshop on Learning from Time Series for Health*, 2025. URL https://openreview.net/forum?id=Hq9Afq7EMk.

Bingzhi Duan, Xiaoyue Wan, and Xu Zhao. FSGait: Fine grained self-supervised gait abnormality detection. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, pp. 2248–2264, 2024. URL https://openaccess.thecvf.com/content/ACCV2024/html/Duan_FSGait_Fine_Grained_Self-Supervised_Gait_Abnormality_Detection_ACCV_2024_paper.html.

Emadeldeen Eldele, Mohamed Ragab, Zhenghua Chen, Min Wu, Chee Keong Kwoh, Xiaoli Li, and Cuntai Guan. Time-series representation learning via temporal and contextual contrasting. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pp. 2352–2359, 2021. doi: 10.24963/ijcai.2021/324. URL https://www.ijcai.org/proceedings/2021/324.

Mark Endo, Kathleen L. Poston, Edith V. Sullivan, Li Fei-Fei, Kilian M. Pohl, and Ehsan Adeli. GaitForeMer: Self-supervised pre-training of transformers via human motion forecasting for few-shot gait impairment severity estimation. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pp. 130–139, 2022. URL https://arxiv.org/abs/2207.00106.

Archibald Fraikin, Adrien Bennetot, and Stéphanie Allassonnière. T-Rep: Representation learning for time series using time-embeddings. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/84b946c4c4162b29464fc7aa57e480e1-Abstract-Conference.html.

M. Goswami, K. Szafer, A. Choudhry, Y. Cai, S. Li, and A. Dubrawski. MOMENT: A family of open time-series foundation models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 16115–16152, 2024.

H. Gui, X. Li, and X. Chen. Vector quantization pretraining for EEG time series with random projection and phase alignment. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 16731–16750, 2024.

Jiarui Jin, Haoyu Wang, Hongyan Li, Jun Li, Jiahui Pan, and Shenda Hong. Reading your heart: Learning ECG words and sentences via pre-training ECG language model. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=6Hz1Ko087B.

M. A. Lafortune, P. R. Cavanagh, H. J. Sommer, III, and A. Kalenak. Three-dimensional kinematics of the human knee during walking. *Journal of Biomechanics*, 25(4):347–357, 1992. doi: 10.1016/0021-9290(92)90254-X.

Junqiao Li, Yunhe Mao, Tian Lan, Liang Huangfu, Yan Xiong, and Jian Li. A dynamic knee function scoring system for anterior cruciate ligament injuries based on normative six-degree-of-freedom gait pattern. *BMC Musculoskeletal Disorders*, 25:1083, 2024. doi: 10.1186/s12891-024-08208-0. URL https://doi.org/10.1186/s12891-024-08208-0.

X. M. Li, B. Liu, B. Deng, and S. M. Zhang. Normal six-degree-of-freedom motions of knee joint during level walking. *Journal of Biomechanical Engineering*, 118(2):258–261, 1996. doi: 10.1115/1.2795970.

Yifan Li, Jiayu He, Bernard Liew, David S. Hollinger, Qichang Mei, Behnam Gholami, Maria Fasli, Klaus McDonald-Maier, and Xiaojun Zhai. Self-supervised learning enhances accuracy and data efficiency in lower-limb joint moment estimation from gait kinematics. *Frontiers in Bioengineering and Biotechnology*, 13:1633513, 2025. doi: 10.3389/fbioe.2025.1633513.

Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. iTransformer: Inverted transformers are effective for time series forecasting. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/2ea18fdc667e0ef2ad82b2b4d65147ad-Abstract-Conference.html.

R. Daniel Magruder, Selim Gilon, Antoine Falisse, and Scott D. Uhlrich. GaitEncoder: A foundation model of gait kinematics for diverse clinical applications and pathologies. *medRxiv*, 2026. doi: 10.64898/2026.07.07.26357479. URL https://www.medrxiv.org/content/10.64898/2026.07.07.26357479v1.

Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2211.14730.

Tian Tan, Tom Van Wouwe, Keenon F. Werling, C. Karen Liu, Scott L. Delp, Jennifer L. Hicks, and Akshay S. Chaudhari. GaitDynamics: a generative foundation model for analyzing human walking and running. *Nature Biomedical Engineering*, 10:1659–1671, 2026. doi: 10.1038/s41551-025-01565-8. URL https://www.nature.com/articles/s41551-025-01565-8.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/7a98af17e63a0ac09ce2e96d03992fbc-Abstract.html.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

G. Wang, X. Zeng, G. Lai, G. Zhong, K. Ma, and Y. Zhang. Efficient subject-independent detection of anterior cruciate ligament deficiency based on marine predator algorithm and support vector machine. *IEEE Journal of Biomedical and Health Informatics*, 26(10):4936–4947, 2022. doi: 10.1109/JBHI.2022.3152846.

Yunshi Wen, Tengfei Ma, Tsui-Wei Weng, Lam M. Nguyen, and Anak Agung Julius. Abstracted shapes as tokens – a generalizable and interpretable model for time-series classification. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://papers.nips.cc/paper_files/paper/2024/hash/a7c57c8942b7d2adba1ac615ba0e60b9-Abstract-Conference.html.

Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. TimesNet: Temporal 2D-variation modeling for general time series analysis. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2210.02186.

Wenbao Wu, Wei Zeng, Limin Ma, Chengzhi Yuan, and Yu Zhang. Modeling and classification of gait patterns between anterior cruciate ligament deficient and intact knees based on phase space reconstruction, Euclidean distance and neural networks. *BioMedical Engineering OnLine*, 17:165, 2018. doi: 10.1186/s12938-018-0594-1. URL https://link.springer.com/article/10.1186/s12938-018-0594-1.

R. Xiong, Y. Yang, D. He, K. Zheng, S. Zheng, C. Xing, H. Zhang, Y. Lan, L. Wang, and T.-Y. Liu. On layer normalization in the transformer architecture. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 10524–10533, 2020.

Zhihan Yue, Yujing Wang, Juanyong Duan, Tianmeng Yang, Congrui Huang, Yunhai Tong, and Bixiong Xu. TS2Vec: Towards universal representation of time series. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(8):8980–8987, 2022. doi: 10.1609/aaai.v36i8.20881. URL https://ojs.aaai.org/index.php/AAAI/article/view/20881.

X. Zeng, Y. Zhu, Z. Xie, G. Zhong, W. Huang, L. Ma, Y. Zhang, and C. Mao. 3D knee kinematic parameters effectively diagnose knee osteoarthritis and assess its therapeutic strategy. *Advanced Intelligent Systems*, 4(6):2100161, 2022. doi: 10.1002/aisy.202100161.

Xiaolong Zeng, Limin Ma, Zefeng Lin, Wenhan Huang, Zhiqiang Huang, Yu Zhang, and Chuanbin Mao. Relationship between Kellgren-Lawrence score and 3D kinematic gait analysis of patients with medial knee osteoarthritis using a new gait system. *Scientific Reports*, 7:4080, 2017. doi: 10.1038/s41598-017-04390-5. URL https://www.nature.com/articles/s41598-017-04390-5.

Y. Zhang, Z. Yao, S. Wang, W. Huang, L. Ma, H. Huang, and H. Xia. Motion analysis of Chinese normal knees during gait based on a novel portable system. *Gait & Posture*, 41(3):763–768, 2015. doi: 10.1016/j.gaitpost.2015.01.020.

A Additional Method Details

We retain the notation of Section 3. All averages and losses below exclude padded units.

A.1 GaitParser and Input Normalization

Period estimation and boundary detection Each side is segmented independently from its flexion–extension trajectory at 60 Hz. The procedure is:

1. Smooth the reference trajectory with a third-order, nine-sample Savitzky–Golay filter.
2. Estimate the period from autocorrelation over 0.4–4.0 s. Select the earliest local peak reaching both 0.05 correlation and 85% of the largest peak; otherwise select the strongest local peak. If no local peak exists, use the interval maximum only when its correlation reaches 0.05.
3. Estimate a second period from median flexion-peak spacing, with prominence at least 15% of the waveform range. Prefer this estimate when the autocorrelation estimate is unavailable, near a search boundary, or their ratio lies in $[0.75, 1.33]$.
4. Given period P_s in samples, detect peaks at least $0.75P_s$ apart. Locate extension anchors at minima within $[0.05P_s, 0.40P_s]$ after each peak, with an additional backward search for a leading anchor.

Anchors are inferred from kinematics and do not represent measured foot-contact events.

Cycle selection and resampling Sorted, unique anchors define candidate intervals $[a, b)$. A candidate is retained if

$$\max(0.4f_{\text{samp}}, 0.55P_s) \le b - a \le \min(4.0f_{\text{samp}}, 1.60P_s). \tag{A.1}$$

The retained boundaries are applied to all six unsmoothed trajectories, and each segment is linearly interpolated to $L = 100$ points. Quality control requires at least two valid cycles per side and rejects non-finite recordings, nearly constant reference trajectories, or insufficient detectable cycles.

Timing conventions Duration and center-to-center intervals follow Eq. (2), with $\delta_{s,1} = \delta_{s,2}$. Intervals between retained centers preserve gaps from discarded cycles. Both descriptors enter the rhythm projection.

Channel and unit normalization Channel statistics m_d and v_d in Eq. (1) are pooled across accepted, resampled cycles from both sides of the healthy training subjects, after subject-level quality control. The population standard deviation v_d is floored at 10^{-6}. These statistics are fixed for validation and downstream data. Within these standardized channel coordinates, each unit is further decomposed into shape and continuous attributes:

$$\begin{aligned} \mu_i &= \frac{1}{L}\sum_{t=1}^{L} x_{i,t}, \qquad \sigma_i = \max\left\{ \sqrt{\frac{1}{L}\sum_{t=1}^{L}(x_{i,t} - \mu_i)^2}, \epsilon_s \right\}, \\ \mathbf{s}_i &= \frac{\mathbf{x}_i - \mu_i}{\sigma_i}, \qquad \ell_i = \log \sigma_i, \end{aligned} \tag{A.2}$$

where $\epsilon_s = 10^{-3}$. Channel normalization establishes common coordinates; unit normalization separates shape from offset and scale. The decoder's $\mathrm{Standardize}$ uses its own output's mean and standard deviation. Decoded templates have zero mean and unit variance unless the scale floor is active. Waveform losses operate in standardized channel coordinates.

A.2 GaitFormer Architecture

Local embeddings and sequence construction Table A.1 specifies the modules. Each DoF's shape and attribute embeddings are added, using the same modules for both sides.

Within a side, tokens follow the order $(1,1), \ldots, (1,D), (2,1), \ldots, (N_s, D)$, where each pair denotes cycle and DoF indices. Each token receives a DoF embedding and a rhythm embedding

shared across the six tokens of its cycle. A shared learnable summary token is prepended, producing $1 + N_s D$ valid tokens. Shorter sequences are padded within a mini-batch, and padded keys are excluded from attention. The encoder uses no causal mask: each valid token can attend to every valid cycle–DoF position on the same side.

Table A.1: Architecture used for vocabulary learning and contextual pretraining. Linear dimensions denote input and output widths; convolution settings are kernel size, stride, and padding. DoF-specific modules are shared between sides.

Component	Configuration
Shape encoder, per DoF	Conv1D $1 \to 64$ $(9, 4, 4)$; GELU; Conv1D $64 \to 64$ $(5, 2, 2)$; GELU; temporal average pooling; Linear $64 \to 128$
Attribute embedding, per DoF	Linear $2 \to 128$
Rhythm embedding	Linear $2 \to 128$; GELU; Linear $128 \to 128$
Structural embeddings	Six DoF embeddings of width 128
Within-side encoder	Two Transformer layers; width 128; four heads; feed-forward width 256; GELU; pre-layer normalization; final LayerNorm
Bilateral encoder	Two cross-attention blocks; width 128; four heads; feed-forward width 256; GELU; pre-layer normalization
Transformer dropout	0.1
VQ projection, shared	Linear $128 \to 128$; Tanh; Linear $128 \to 128$; ℓ_2 normalization
Codebooks	Six codebooks; 128 codes per DoF; code dimension 128
Shape decoder, per DoF	LayerNorm; Linear $128 \to 256$; GELU; Linear $256 \to 100$; waveform standardization
SSL prediction heads, per DoF	Linear $128 \to 128$ for codes; Linear $128 \to 2$ for attributes

Bilateral contextualization After within-side encoding, bilateral blocks exchange information between the complete side sequences, including their summary tokens. Let $\mathbf{U}_{\mathrm{L}}$ and $\mathbf{U}_{\mathrm{R}}$ be the inputs to one block. Its simultaneous updates are

$$\mathbf{U}'_{\mathrm{L}} = \mathcal{B}(\mathbf{U}_{\mathrm{L}}, \mathbf{U}_{\mathrm{R}}), \qquad \mathbf{U}'_{\mathrm{R}} = \mathcal{B}(\mathbf{U}_{\mathrm{R}}, \mathbf{U}_{\mathrm{L}}). \tag{A.3}$$

Both directions share parameters and use the pre-update sequences. Queries come from the layer-normalized first argument, and keys and values from the second. Residual cross-attention and feed-forward sublayers retain padding masks and token positions. Attention covers all valid contralateral units without requiring matched cycle indices.

A.3 VQ-Gait Objectives and Codebook Learning

Tokenizer architecture and reconstruction VQ-Gait uses the local encoders and shared projection in Table A.1, with parameters separate from GaitFormer and no contextual attention. Features and codes are ℓ_2-normalized, making maximum inner-product assignment equivalent to Eq. (6).

Let $\mathcal{V}$ be the set of valid gait units in a mini-batch and C the code dimension. The smooth-L_1 penalty is

$$h(r) = \begin{cases} \frac{1}{2} r^2, & |r| < 1, \\ |r| - \frac{1}{2}, & |r| \geq 1. \end{cases} \tag{A.4}$$

The reconstruction and commitment terms in Eq. (8) are

$$\begin{aligned} \mathcal{L}_{\mathrm{shape}} &= \frac{1}{|\mathcal{V}|L} \sum_{i \in \mathcal{V}} \sum_{t=1}^{L} h(\widehat{s}_{i,t}^{\mathrm{VQ}} - s_{i,t}^{\mathrm{VQ}}), \\ \mathcal{L}_{\mathrm{wave}}^{\mathrm{VQ}} &= \frac{1}{|\mathcal{V}|L} \sum_{i \in \mathcal{V}} \sum_{t=1}^{L} h(\widehat{x}_{i,t}^{\mathrm{VQ}} - x_{i,t}^{\mathrm{VQ}}), \\ \mathcal{L}_{\mathrm{commit}} &= \frac{1}{|\mathcal{V}|C} \sum_{i \in \mathcal{V}} \left\| \mathbf{z}_i - \mathrm{sg}(\mathbf{c}_{d(i),y_i}) \right\|_2^2, \end{aligned} \tag{A.5}$$

where sg denotes stop-gradient. During vocabulary learning, waveform reconstruction uses the observed input mean and scale in Eq. (7). The shape term measures normalized morphology, while the waveform term also reflects the effect of the input scale on reconstruction error.

Geometry regularization This term encourages nearby waveform shapes to retain similar neighborhoods in the continuous feature space before quantization. We select at most 256 valid side–cycle positions per mini-batch, using the same positions for every DoF. Selection is random during training and evenly spaced in batch order during validation. For DoF d, let $\mathcal{I}_d$ index the selected units. Excluding self-pairs, define

$$
\begin{aligned}
a_{ij} &= \frac{1}{L}\left\|\mathbf{s}_i^{\mathrm{VQ}} - \mathbf{s}_j^{\mathrm{VQ}}\right\|_2^2, \qquad \overline{a}_d = \max\left\{\frac{\sum_{\substack{i,j\in\mathcal{I}_d \\ i\neq j}} a_{ij}}{|\mathcal{I}_d|(|\mathcal{I}_d|-1)}, 10^{-8}\right\}, \\
P_{ij} &= \frac{\exp[-a_{ij}/(\overline{a}_d T_s)]}{\sum_{j'\in\mathcal{I}_d\setminus\{i\}}\exp[-a_{ij'}/(\overline{a}_d T_s)]}, \\
Q_{ij} &= \frac{\exp[-\|\mathbf{z}_i - \mathbf{z}_j\|_2^2/T_z]}{\sum_{j'\in\mathcal{I}_d\setminus\{i\}}\exp[-\|\mathbf{z}_i - \mathbf{z}_{j'}\|_2^2/T_z]}.
\end{aligned}
\tag{A.6}
$$

The resulting loss is

$$
\mathcal{L}_{\mathrm{geom}} = \frac{1}{D}\sum_{d=1}^{D}\frac{1}{|\mathcal{I}_d|}\sum_{i\in\mathcal{I}_d}\sum_{j\in\mathcal{I}_d\setminus\{i\}} P_{ij}\log\frac{P_{ij}}{Q_{ij}}. \tag{A.7}
$$

The target P is detached. Mean-distance normalization provides a comparable scale across DoFs, and $T_s = T_z = 0.2$. In Eq. (8), $\lambda_s = \lambda_w = 1$, $\beta = 0.25$, and $\lambda_g = 0.1$.

Straight-through estimation and EMA updates Following vector-quantized learning (van den Oord et al., 2017), reconstruction gradients reach the encoder through

$$
\mathbf{z}_i^{\mathrm{ST}} = \mathbf{z}_i + \mathrm{sg}(\mathbf{c}_{d(i),y_i} - \mathbf{z}_i). \tag{A.8}
$$

Its forward value equals the assigned code. Code vectors are updated by EMA rather than gradient descent. For training step u, let $\mathcal{A}_{d,k}^{(u)}$ contain valid units assigned to code k of DoF d. Running counts $n_{d,k}$ and feature sums $\mathbf{r}_{d,k}$ are updated as

$$
\begin{aligned}
n_{d,k}^{(u)} &= \gamma n_{d,k}^{(u-1)} + (1-\gamma)|\mathcal{A}_{d,k}^{(u)}|, \\
\mathbf{r}_{d,k}^{(u)} &= \gamma \mathbf{r}_{d,k}^{(u-1)} + (1-\gamma)\sum_{i\in\mathcal{A}_{d,k}^{(u)}}\mathbf{z}_i, \\
\widetilde{n}_{d,k}^{(u)} &= \frac{n_{d,k}^{(u)} + \epsilon_e}{\sum_{j=1}^{K} n_{d,j}^{(u)} + K\epsilon_e}\sum_{j=1}^{K} n_{d,j}^{(u)}, \\
\mathbf{c}_{d,k}^{(u)} &= \mathrm{Normalize}_2\left(\mathbf{r}_{d,k}^{(u)}/\widetilde{n}_{d,k}^{(u)}\right),
\end{aligned}
\tag{A.9}
$$

where $\gamma = 0.99$, $\epsilon_e = 10^{-5}$, and $\mathrm{Normalize}_2$ denotes ℓ_2 normalization. Initialization uses ten k-means iterations on each DoF's first-batch features, with normalized centers, unit counts, and feature sums equal to the centers. Codes whose EMA counts fall below one are replaced by sampled features of the same DoF, resetting counts to one and feature sums to the replacement vectors. Assignments precede each EMA update; codebooks remain fixed during validation.

A.4 Masked Pretraining Details

Masking and fixed targets VQ-Gait is frozen and kept in evaluation mode, and GaitFormer is initialized separately. Each valid side–cycle–DoF position is independently masked with probability 0.15. The complete shape-plus-attribute embedding is replaced before adding structural and rhythm embeddings; cycle duration and interval remain visible. Uncorrupted units supply code, mean, and log-scale targets. All three objectives use the same valid mask set, with DoF-specific prediction heads shared across sides.

Differentiable waveform reconstruction The shape mixture $\widehat{\mathbf{s}}_i$ defined in Section 3.4 uses softmax temperature one to combine frozen templates $\mathbf{t}_{d,k}$. The mixture is not standardized again; predicted mean and exponentiated log-scale recover the waveform in Eq. (11). Reconstruction is differentiable with respect to these probabilities and attributes. In Eq. (10), $\mathrm{SmoothL1}(\widehat{\mathbf{a}}_i, \mathbf{a}_i) = h(\widehat{\mu}_i - \mu_i) + h(\widehat{\ell}_i - \ell_i)$, with h defined in Eq. (A.4). Thus, the loss sums the two attribute penalties and averages over masked units. The squared norm in Eq. (11) sums sample-wise errors, with division by $|\mathcal{M}|L$ yielding waveform MSE. We use $\lambda_a = \lambda_w = 1$ in Eq. (12), with equal weights on mean and log-scale and no separate timing-recovery loss.

B KineGait Dataset

B.1 Acquisition and Reconstruction of Knee Kinematics

Optical tracking and anatomical calibration Knee kinematics were acquired with the OptiKnee optical tracking system (Shanghai, China). Stereo cameras tracked four-marker rigid-body clusters attached to the lateral thigh and shank during treadmill walking (Figure B.1). A tracked pointer digitized anatomical landmarks, including the greater trochanter, femoral epicondyles, and malleoli, to register each cluster to its anatomical coordinate system (Figure B.2).

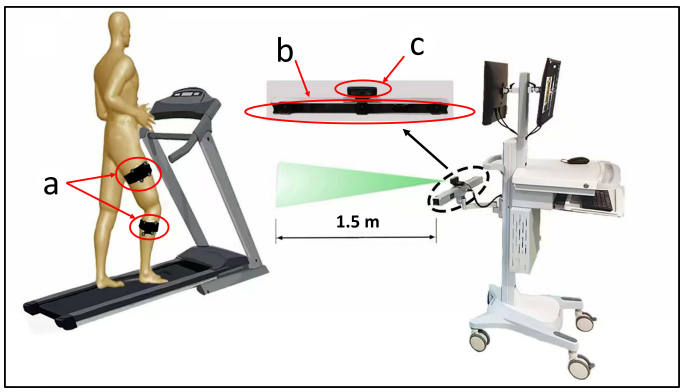

Figure B.1: OptiKnee acquisition setup: (a) thigh and shank marker clusters, (b) stereo infrared tracking camera, and (c) camera lens.

Kinematic reconstruction The knee center was the midpoint between the femoral epicondyles. Dynamic tracking reconstructed femoral and tibial coordinate systems; relative tibial motion yielded three rotations and three translations (Figure B.3). Superior–inferior translation is also termed proximal–distal translation in the acquisition exports.

Acquisition protocol The three hospitals used the same tracking-system and treadmill models, with common marker-placement, calibration, and preparation procedures. The clinical protocol specified walking at 3 km/h after 5 min of acclimatization, followed by a 10-s recording. Experimental recordings contain 600 samples per side at 60 Hz. Community recruitment sources are distinguished from hospital cohorts below.

B.2 Recruitment Sources and Clinical Labels

Community and hospital cohorts KineGait comprises a healthy pretraining cohort from 15 centers, downstream community controls, and patients from three hospitals. For anonymous review, Hospitals A and B denote the development patient sources and Hospital C the external patient source. Community controls are separate from these hospitals. Table B.1 reports source-specific counts before and after quality control.

Eligibility assessment Eligibility was assessed using participant-reported medical history, available clinical records, and clinical examination, supplemented by disease-specific imaging for patients. Participants had to walk independently and complete the test without substantial pain exacerbation or risk of injury. Healthy controls had no history of knee injury, surgery, or chronic knee pain, and no clinical evidence of degenerative knee disease. Available imaging findings were considered, but imaging was not performed solely to enroll healthy controls. Healthy-control exclusions included lower-limb fracture, ligament injury or joint surgery, neurological or musculoskeletal disorders affecting gait, and deformity or pain limiting walking.

Disease ascertainment ACLD was established by clinical examination and MRI. Patients had unilateral deficiency, were assessed before reconstruction, and were in a stable post-injury phase

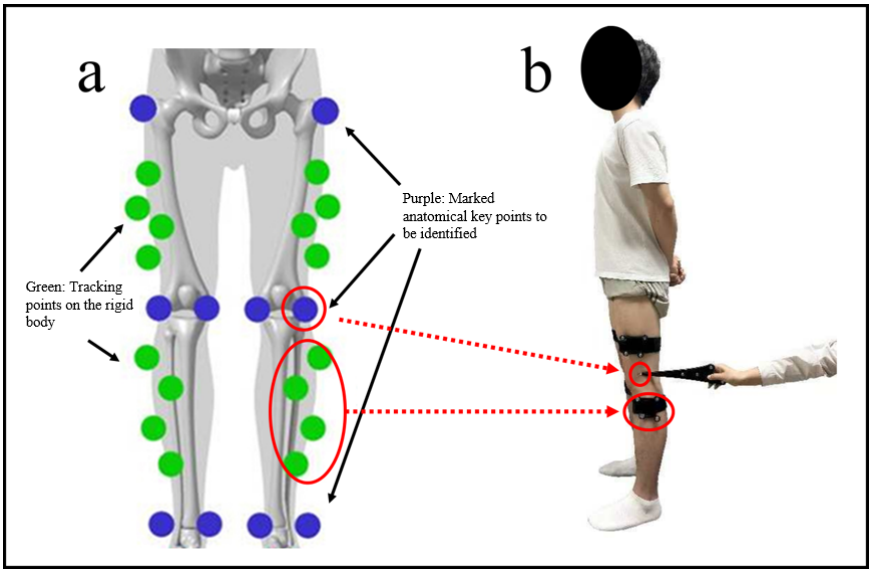

Figure B.2: Anatomical calibration: (a) anatomical landmarks and rigid-body tracking points; (b) marker fixation and landmark digitization using a tracked pointer.

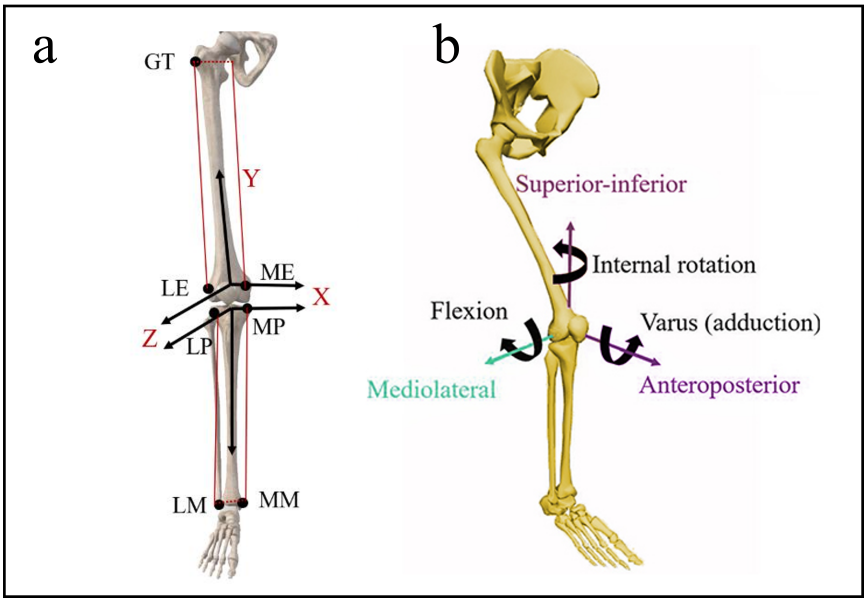

Figure B.3: Knee coordinate system and motion components: (a) anatomical reference landmarks; (b) three rotational and three translational DoFs. GT: greater trochanter; LE/ME: lateral/medial femoral epicondyle; LP/MP: lateral/medial tibial plateau; LM/MM: lateral/medial malleolus.

Table B.1: Recruitment sources and subject counts. Class counts precede GaitParser quality control; the last column gives the total retained across classes. Community controls are separate from the hospital cohorts.

Source	Role	Healthy	ACLD	KOA	Retained
15 pretraining centers	Pretraining	4168	0	0	4164
Community controls	Development	248	0	0	248
Hospital A	Development	0	211	63	272
Hospital B	Development	0	108	59	167
Community controls	External	95	0	0	95
Hospital C	External	0	103	120	221

permitting independent walking. Associated meniscal findings were limited to Stoller grades 0–2 and cartilage damage to International Cartilage Repair Society grades 0–2. Combined multi-ligament injuries, prior lower-limb surgery, and other gait-affecting disorders were excluded. KOA was established from clinical symptoms and imaging, with radiographic severity assessed using Kellgren–Lawrence grading. Patients had to complete walking without assistive devices; inflammatory arthritis, recent acute knee injury or surgery, and neurological disorders affecting gait were exclusion criteria. Disease severity grades are not available as structured fields in the experimental CSV files, so no grade-specific cohort counts are inferred.

Subject-level labels Each participant contributes two limb records and one disease label. An ACLD participant remains in the ACLD class even if the contralateral limb is labeled Healthy. No participant has both ACLD and KOA labels. Age, sex, and BMI are descriptive metadata and are not model inputs.

B.3 Quality Control and Dataset Partitions

Before waveform quality control, the experimental data contain 4,168 pretraining, 689 development, and 318 external subjects. The criteria in Appendix A.1 exclude four pretraining subjects and two KOA subjects each from Hospitals A and C. These counts describe exclusions from the experimental data, not earlier recruitment attrition.

Before quality control, the healthy cohort is randomly split 90/10 into training and validation sets, and a class-stratified 20% of the development cohort is reserved for internal testing. Table B.2 reports the resulting counts and class composition. Both limbs remain in one subject-level partition. The retained downstream pool supports five-fold training and validation, with internal and external tests held fixed (Appendix C.1).

Table B.2: Final experimental partitions. Before/after counts refer to GaitParser quality control; disease-class counts refer to retained subjects.

Partition	Before	After	Healthy	ACLD	KOA
Pretraining: training	3751	3747	3747	0	0
Pretraining: validation	417	417	417	0	0
Downstream: training/validation	551	549	198	255	96
Fixed internal test	138	138	50	64	24
Fixed external test	318	316	95	103	118

B.4 Participant Characteristics

Table B.3 summarizes retained participants, counted once each. Continuous variables are reported as mean and sample standard deviation. Recorded ages span 11–84 years in development and 14–81 years externally; groups were not age-matched.

For descriptive statistics, unverifiable ages outside $(0, 110]$ years and BMI values outside $[10, 60]$ kg/m^2 were treated as missing without imputation. Age is missing for 10 pretraining training sub-

Table B.3: Participant characteristics after waveform quality control. Age and BMI are mean ± sample standard deviation. M/F denotes male/female counts; BMI is in kg/m^2.

Partition	Group	n	M/F	Age (years)	BMI
Pretraining train	Healthy	3747	1990/1757	54.67 ± 13.43	24.54 ± 3.57
Pretraining validation	Healthy	417	236/181	55.49 ± 12.75	24.57 ± 3.70
Downstream pool	Healthy	198	103/95	58.52 ± 11.17	23.90 ± 3.22
	ACLD	255	182/73	30.60 ± 12.71	24.07 ± 3.61
	KOA	96	26/70	64.02 ± 8.13	24.85 ± 3.30
Internal test	Healthy	50	17/33	58.36 ± 9.49	24.53 ± 3.09
	ACLD	64	51/13	28.86 ± 11.36	24.34 ± 3.73
	KOA	24	9/15	63.21 ± 12.97	24.68 ± 3.77
External test	Healthy	95	49/46	51.86 ± 14.25	24.28 ± 3.30
	ACLD	103	83/20	30.36 ± 10.16	24.24 ± 3.43
	KOA	118	28/90	59.96 ± 10.93	23.89 ± 3.36

Available demographic sample sizes for pretraining are: training age $n = 3{,}737$, training BMI $n = 3{,}728$, validation age and BMI $n = 416$. Downstream demographic sample sizes equal the corresponding group counts. Unverified numerical outliers are treated as missing for these descriptive statistics only; subjects remain in the experiments.

jects and one validation subject; BMI is missing for 19 and one, respectively. Downstream records have no missing values under these rules. Metadata review does not change waveform eligibility or experimental inclusion.

Age and sex composition differ across groups, particularly for ACLD. Community recruitment of controls and hospital recruitment of patients also confound diagnosis with recruitment setting. Common clinical acquisition procedures do not remove these differences; external performance should be interpreted within the evaluated cohorts, without assuming independence from demographic or recruitment effects.

All recordings were collected with the OptiKnee system. Although using a common system limits acquisition variability, testing across centers does not establish transfer to other measurement systems or protocols. The downstream task evaluates only Healthy/ACLD/KOA labels; the reported results therefore provide no evidence of severity grading or longitudinal disease monitoring. Future evaluation should include other acquisition systems, cohorts more closely matched for demographic characteristics and recruitment settings, and severity or longitudinal outcomes.

B.5 Data Organization and Availability

Each CSV record contains a coded participant identifier, recruitment source, demographic metadata, limb side and label, and a 600×6 kinematic matrix. Rotations are expressed in degrees and translations in centimeters. Raw channels are ordered as varus–valgus, internal–external rotation, flexion–extension, anterior–posterior, superior–inferior, and medial–lateral translation. Model inputs reorder them to flexion–extension, varus–valgus, internal–external rotation, anterior–posterior, medial–lateral, and superior–inferior translation.

The code and dataset will be publicly released upon acceptance. Data sharing will follow the arrangements described in the Ethics Statement. The data release will include a data dictionary, preprocessing configuration, and subject partitions. Original clinical files containing personal identifiers are excluded from the release.

C Experimental and Implementation Details

C.1 Evaluation Protocol and Model Selection

We use the subject partitions in Table B.2. The downstream pool is divided into five class-stratified folds with seed 42, keeping both limbs and any repeated records of a subject together. Each run trains on four folds and validates on the remaining fold. All methods share these folds and the fixed internal and external test sets; channel normalization follows Appendix A.1.

Vocabulary learning and SSL select the lowest healthy-validation loss. Downstream training selects the highest validation macro-F1. Architecture and pretraining hyperparameter choices are based solely on training and validation results. For pretrained methods, each fold starts from the same pretrained checkpoint and a newly initialized classifier. Test data are used only after checkpoint selection. We report the arithmetic mean of the five model-specific test scores, without prediction ensembling. Where reported, sample standard deviation describes variation across training–validation partitions on the same test subjects, not across independent test cohorts.

C.2 Training Configuration and Downstream Adaptation

Table C.1 summarizes optimization. Learning rates remain constant, gradients are clipped at global norm 1.0, and all applicable stages have a 200-epoch budget with early stopping. Data splitting, label sampling, and training use seed 42. GaitReader's training masks vary by epoch, while validation uses a fixed mask seed. GaitReader uses CUDA automatic mixed precision; comparison adapters use FP32. Experiments run on an NVIDIA GeForce RTX 5080 GPU.

Table C.1: Training settings. Checkpoints follow the listed validation criterion, with early-stopping patience measured in epochs.

Setting	Vocabulary learning	SSL pretraining	Downstream training
Optimizer	AdamW	AdamW	AdamW
Learning rate	3×10^{-4}	3×10^{-4}	3×10^{-4}
Weight decay	10^{-4}	10^{-2}	10^{-2}
Epoch budget	200	200	200
Batch size (subjects)	32	32	32
Early-stopping patience	10	10	10
Selection criterion	Validation loss	Validation loss	Validation macro-F1

GaitReader uses 128 codes per DoF. It averages the two side summary embeddings and applies a linear three-class classifier. Full fine-tuning updates both encoder and classifier; scratch training uses the same architecture with random initialization. Linear probing freezes the pretrained encoder in evaluation mode. VQ-Gait and pretraining heads are not used for downstream inference.

All methods minimize class-weighted cross-entropy. Weights are inverse labeled-training class frequencies normalized to mean one and recomputed for each label subset. The mini-batch loss is normalized by the sum of target-class weights. Validation and test labels do not determine these weights.

C.3 Baseline Implementations and Gait Adaptations

Shared protocol The baselines use upstream model components; their repository revisions will be included in the code release upon acceptance. They share GaitReader's eligible subjects, normalization, and evaluation partitions. Except for HeartLang, inputs are complete 600×6 side recordings; HeartLang uses the same 100-point cycle units as GaitReader. Baselines encode both sides with shared parameters, concatenate their representations, and apply dropout 0.1 followed by a linear classifier. Full fine-tuning updates the downstream feature extractor and classifier, with method-specific frozen components described below. Table C.2 collects architecture and objective settings.

Supervised models TimesNet uses flattened temporal features from its classification backbone. PatchTST concatenates the last-patch features across channels and is trained without SSL in this

Table C.2: Baseline configurations. Widths and depths refer to the adapted encoders unless otherwise specified; FF denotes feed-forward width. Shared training settings are in Table C.1.

Method	Configuration
TimesNet (Wu et al., 2023)	Width 32, FF 64, depth 2; three selected periods and three convolution kernels.
PatchTST (Nie et al., 2023)	Width 128, FF 256, depth 2, four heads; patch length 20, stride 10; supervised training.
iTransformer (Liu et al., 2024)	Width 128, FF 256, depth 2, four heads; instance normalization.
TS2Vec (Yue et al., 2022)	Hidden width 32, output width 128, depth 6; hierarchical contrastive objective.
T-Rep (Fraikin et al., 2024)	Hidden width 32, output width 128, depth 6; sinusoidal time embedding width 16; four loss weights of 0.25.
TS-TCC (Eldele et al., 2021)	Output width 128; six future latent steps; scaling $\mathcal{N}(2, 1.1^2)$, 1–7 permutation segments, jitter SD 0.8; contextual loss weight 0.7, temperature 0.2.
VQShape (Wen et al., 2024)	Patch length 10, width 64, 20 tokens, decoded shape length 64; encoder/tokenizer/decoder depth 2 with four heads; 128 codes of dimension 8; mask ratio 0.25; waveform, shape, quantization, and disentanglement loss-group weights all 1.
HeartLang (Jin et al., 2025)	Width 128, FF 256, four heads; tokenizer encoder/decoder depths 2/1 and contextual depth 2; 128 EMA codes of dimension 64; mask ratio 0.5; unit commitment weight and EMA decay 0.99.

comparison. iTransformer retains instance normalization and concatenates channel-token features. All three are trained from scratch.

Contrastive models Full-recording SSL samples one side per subject at each step, preventing paired limbs from becoming separate instances in the same contrastive batch. TS2Vec and T-Rep use overlapping crops and their hierarchical objectives; averaged pretrained encoder weights initialize downstream adaptation, which uses temporal max pooling without input masking. TS-TCC combines weak scaling with strong segment-permutation and jitter views. A common temporal permutation preserves correspondence across all six DoFs; downstream features are flattened convolutional outputs.

Quantized models VQShape applies shared parameters to individual DoF waveforms. Classification concatenates quantized tokens with predicted start, length, mean, and scale attributes in DoF order; the healthy codebook remains frozen during fine-tuning. HeartLang maps ECG channel and temporal indices to DoF and gait-cycle indices. It trains its tokenizer with waveform MSE and commitment loss, then trains a separate contextual encoder to predict frozen code targets. Both sides contribute to pretraining. Sequences are grouped by valid length to avoid introducing padding into the upstream convolution and attention; side summary tokens support classification.

C.4 Label-Efficient Transfer Protocol

Within each fold, a class-wise random subject permutation with seed 42 defines nested 10%, 25%, 50%, and 100% labeled training subsets. Counts are rounded to the nearest integer, retaining at least one subject per class. Fine-tuning, scratch training, and linear probing use identical subsets and independently initialized fits; the healthy checkpoint remains fixed across fractions. The complete validation fold remains labeled for model selection, and both test sets remain unchanged. Label fractions therefore refer only to downstream training labels.

C.5 Ablation Protocols

Table C.3 defines the main ablations. Each uses all training labels, the shared folds and fixed tests, and the optimization settings above. SSL loss ablations reuse the full model's VQ-Gait tokenizer and train separate contextual encoders. For fixed windows, both vocabulary learning and contextual pretraining use the changed units. Thus, the vocabulary comparison tests the combined pretraining design, while the segmentation comparison also changes unit counts and timing descriptors.

Table C.3: Ablation definitions relative to GaitReader. Waveform and attribute loss removal refers to contextual pretraining.

Variant	Change	Retained components
w/o gait vocabulary	Direct 100-point waveform prediction with DoF-specific linear heads and MSE; no tokenizer, code targets, or attribute supervision.	GaitParser, GaitFormer, and shape, attribute, DoF, and rhythm input embeddings.
w/o waveform loss	Remove only the SSL waveform reconstruction term.	Code and attribute prediction; VQ-Gait training, including its waveform loss.
w/o attribute and waveform losses	Use code cross-entropy alone; omit attribute heads and waveform reconstruction.	Input attribute embeddings and the frozen VQ-Gait tokenizer.
w/o GaitParser	Use non-overlapping 100-sample windows: six units per DoF and side.	Adaptive-reference subject eligibility and channel standardization; model architecture and objectives.

D Additional Experimental Results

D.1 Adaptation Results and Variation Across Folds

Table D.1 extends Table 2 with macro-F1, linear probing, and sample standard deviations. Full fine-tuning achieves higher accuracy and macro-F1 than both scratch training and linear probing at every label fraction on both fixed test sets. Linear probing performs worse than scratch training internally throughout, but externally exceeds scratch training at 10%, 50%, and 100% labels and falls below it at 25%.

Healthy-only pretraining exposes the encoder to neither patient gait nor disease labels, so its representations are not explicitly optimized to distinguish the downstream disease categories. Linear probing adapts only the classifier, leaving the encoder unable to adjust its features to disease-related waveform variations or cohort differences. This restriction may help explain its lower internal performance relative to scratch training, which updates the entire network using disease labels. The consistent advantage of full fine-tuning over linear probing across both test sets suggests that adapting the encoder with disease labels is important for translating representations learned from healthy gait into effective disease recognition.

Table D.1: Downstream adaptation with different label fractions. Values are mean ± sample standard deviation (%) across five fold-trained models evaluated on the same fixed tests.

Variant	Internal		External	
	Acc.	Macro-F1	Acc.	Macro-F1
Scratch (10%)	79.86 ± 1.73	76.15 ± 1.53	74.56 ± 4.86	74.11 ± 4.48
Linear probe (10%)	72.90 ± 1.89	70.16 ± 2.78	76.58 ± 7.21	76.02 ± 6.64
Fine-tuning (10%)	84.64 ± 1.94	82.13 ± 2.95	83.73 ± 3.69	83.54 ± 3.36
Scratch (25%)	85.07 ± 2.44	82.71 ± 2.77	80.13 ± 2.04	79.51 ± 2.13
Linear probe (25%)	76.23 ± 4.57	74.16 ± 4.78	78.86 ± 4.32	78.19 ± 3.84
Fine-tuning (25%)	88.70 ± 2.74	86.86 ± 3.80	85.82 ± 2.45	85.36 ± 2.53
Scratch (50%)	85.65 ± 3.22	83.33 ± 3.15	78.23 ± 2.22	78.02 ± 2.04
Linear probe (50%)	78.70 ± 3.18	76.96 ± 3.38	82.03 ± 1.92	81.56 ± 1.80
Fine-tuning (50%)	91.88 ± 2.37	91.23 ± 2.76	87.72 ± 1.51	87.35 ± 1.36
Scratch (100%)	88.41 ± 1.15	87.49 ± 1.12	82.03 ± 2.25	81.51 ± 2.03
Linear probe (100%)	80.14 ± 1.41	79.04 ± 1.46	82.72 ± 2.71	82.51 ± 2.51
Fine-tuning (100%)	93.04 ± 1.50	92.09 ± 1.54	89.87 ± 0.78	89.42 ± 0.79

D.2 Class-Wise Performance and Confusion Matrices

We analyze GaitReader with all training labels on the fixed internal and external tests. Table D.2 reports precision, recall, and F1 separately for Healthy, ACLD, and KOA. Each metric is computed from the confusion matrix of an individual fold-trained model and then summarized as mean and sample standard deviation across the five models. Class supports count unique test subjects, not repeated evaluations. In particular, mean class F1 is computed by averaging fold-specific F1 scores, rather than taking the harmonic mean of the reported mean precision and recall.

Figure D.1 shows confusion matrices averaged across these models. Internally, the largest confusion by mean count is ACLD predicted as KOA (4.4 subjects per model, 6.88% of ACLD subjects); the reverse error affects 9.17% of the smaller KOA group. This asymmetry contributes to KOA's lower precision (83.31%) relative to its recall (90.83%). Externally, errors are concentrated between Healthy and ACLD: 13.68% of healthy subjects are classified as ACLD, while 8.54% of ACLD subjects are classified as healthy. KOA has the highest external recall (97.63%) and F1 (95.78%). The external reduction in overall performance therefore primarily reflects difficulty separating Healthy and ACLD in this cohort, rather than a uniform decline across all classes. These descriptive patterns do not isolate the effects of disease from demographic or recruitment differences.

Table D.2: Class-wise GaitReader performance using all training labels. Precision, recall, and F1 are mean ± sample standard deviation (%) across five fold-trained models on fixed test sets. n is the number of unique subjects per class in each test set.

Test set	Class	n	Precision	Recall	F1
Internal	Healthy	50	99.20 ± 1.10	94.80 ± 4.60	96.89 ± 2.16
Internal	ACLD	64	92.68 ± 3.82	92.50 ± 2.04	92.52 ± 1.43
Internal	KOA	24	83.31 ± 3.75	90.83 ± 3.49	86.86 ± 2.92
External	Healthy	95	90.35 ± 2.15	85.26 ± 6.86	87.56 ± 3.03
External	ACLD	103	85.17 ± 4.34	85.24 ± 6.78	84.93 ± 1.67
External	KOA	118	94.07 ± 3.73	97.63 ± 2.11	95.78 ± 2.44

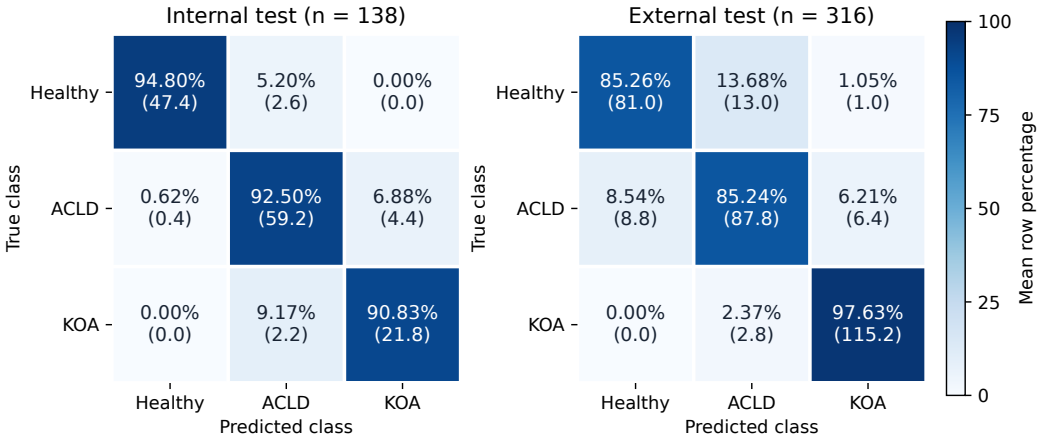

Figure D.1: GaitReader confusion matrices using all training labels. Rows denote true classes and columns predicted classes. Each cell shows the mean row-normalized percentage across five fold-trained models; parentheses give the mean number of subjects assigned to that cell per model. All models evaluate the same fixed test subjects; these are not ensemble predictions or counts of independent subjects pooled across folds.

D.3 Extended Module and Embedding Ablations

Vocabulary learning and contextual objectives Table D.3 removes individual vocabulary-learning components, SSL losses, or bilateral attention while retaining 128 codes per DoF. Removing VQ shape–attribute separation quantizes the channel-standardized waveform directly. The three VQ loss ablations set only the indicated weight to zero and relearn the tokenizer. In contrast, SSL loss ablations retain the full tokenizer; removing SSL code loss preserves vocabulary-based waveform reconstruction and is therefore different from removing the vocabulary. Removing bilateral attention sets its depth to zero while retaining the within-side encoders and downstream pooling.

The full model has higher mean accuracy and macro-F1 than each of these variants on both test sets. Geometry regularization contributes more visibly internally, whereas removing bilateral attention lowers external accuracy from 89.87% to 86.96%. These comparisons support the evaluated components within the full configuration without establishing that their effects are independent or additive.

Table D.3: Additional module ablations. All variants use 128 codes per DoF. Values are mean accuracy and macro-F1 (%) across five fold-trained models; best results are bold.

Variant	Internal		External	
	Acc.	Macro-F1	Acc.	Macro-F1
Full GaitReader	**93.04**	**92.09**	**89.87**	**89.42**
w/o VQ shape–attribute separation	92.03	91.01	86.65	86.45
w/o VQ geometry loss	91.59	90.53	89.37	88.95
w/o VQ shape loss	91.30	90.19	87.91	87.51
w/o VQ waveform loss	91.16	90.27	87.09	86.53
w/o SSL attribute loss	92.03	90.93	88.54	88.10
w/o SSL code loss	91.30	90.76	87.72	87.25
w/o bilateral attention	92.17	91.22	86.96	86.68

Input information Table D.4 removes specific inputs to GaitFormer. Attribute removal suppresses both mean and log-scale; their individual ablations suppress one component. Duration and interval ablations set the respective rhythm-projection input to zero; removing both disables rhythm embeddings. These changes apply during contextual pretraining and downstream adaptation, leaving VQ-Gait unchanged. The shape-only and shape-plus-attributes variants additionally omit DoF and rhythm embeddings.

Removing rhythm embeddings produces the largest external degradation among the individual embedding ablations, from 89.87% to 68.10% accuracy. Removing shape or scale information also reduces performance on both cohorts. However, removing duration slightly improves internal metrics while reducing external metrics, so the contribution of each input is not uniform across cohorts.

D.4 Vocabulary Size and Code Dimension

Table D.5 varies the number of codes per DoF at fixed code dimension 128, then varies code dimension at fixed vocabulary size 128. All configurations retain GaitFormer width 128 and use the same training and evaluation protocol. With code dimension 128, the three vocabulary sizes have identical mean internal accuracy, while external accuracy varies from 87.34% to 89.87%. The reported $K = 128$, $C = 128$ configuration has the highest external accuracy and macro-F1 among these settings, whereas $K = 256$ has slightly higher internal macro-F1. Increasing vocabulary capacity therefore does not uniformly improve transfer. Code-dimension effects are also non-monotonic, and these finite comparisons do not establish a universally optimal vocabulary size or dimension.

D.5 Masking Ratio Sensitivity

We vary the gait-unit masking probability during contextual pretraining over 10%, 15%, 30%, 50%, and 75%. All settings reuse the same frozen VQ-Gait tokenizer and retain the architecture, training

Table D.4: Input embedding ablations. Values are mean accuracy and macro-F1 (%) across five fold-trained models. The shape + attributes + DoF variant is identical to w/o rhythm embedding and is reported once. Best results are bold.

Variant	Internal		External	
	Acc.	Macro-F1	Acc.	Macro-F1
Full GaitReader	93.04	92.09	**89.87**	**89.42**
w/o shape embedding	86.96	85.68	83.10	82.71
w/o attribute embedding	91.59	91.03	84.18	83.47
w/o mean embedding	92.46	91.47	88.10	87.63
w/o scale embedding	90.00	88.75	84.37	83.66
w/o DoF embedding	90.58	89.59	87.59	87.15
w/o rhythm embedding	87.83	86.21	68.10	68.46
w/o duration input	**93.19**	**92.46**	87.78	87.18
w/o interval input	91.16	90.24	89.05	88.55
Shape only	87.10	85.50	68.54	69.10
Shape + attributes	87.83	86.16	70.95	71.47

Table D.5: Vocabulary size and code dimension. K is the number of codes per DoF and C is the code dimension; GaitFormer width remains 128. Values are mean accuracy and macro-F1 (%) across five fold-trained models. Best results are bold.

Variant	Internal		External	
	Acc.	Macro-F1	Acc.	Macro-F1
$K = 64,\ C = 128$	**93.04**	91.93	88.35	87.90
$K = 128,\ C = 128$	**93.04**	92.09	**89.87**	**89.42**
$K = 256,\ C = 128$	**93.04**	**92.17**	87.34	86.77
$K = 128,\ C = 32$	92.46	91.21	89.11	88.68
$K = 128,\ C = 64$	92.03	90.70	83.92	83.82

objectives, and downstream protocol. DoF and rhythm embeddings remain visible at masked positions. Each pretrained encoder is fully fine-tuned using all training labels and evaluated across the same five folds and fixed test sets.

Table D.6 reports the resulting mean performance and sample standard deviations. The 10% setting has the highest internal accuracy and macro-F1, whereas the 15% setting used in the main experiments has the highest external accuracy and macro-F1. Internal accuracy decreases from 93.62% at 10% masking to 92.17% at 75%; external performance varies non-monotonically. Masking controls how much waveform content remains available for contextual prediction, and these results show that increasing its difficulty does not consistently improve downstream transfer.

Table D.6: Sensitivity to the gait-unit masking probability during pretraining. Accuracy and macro-F1 are mean $\pm$ sample standard deviation (%) across five fold-trained models evaluated on the same fixed test sets. Downstream training uses all labels.

Variant	Internal		External	
	Acc.	Macro-F1	Acc.	Macro-F1
10%	93.62 ± 1.65	92.46 ± 1.85	87.85 ± 1.06	87.37 ± 1.03
15% (default)	93.04 ± 1.50	92.09 ± 1.54	89.87 ± 0.78	89.42 ± 0.79
30%	92.75 ± 2.46	91.73 ± 2.58	87.97 ± 0.92	87.53 ± 0.90
50%	92.17 ± 0.94	91.37 ± 0.91	88.99 ± 0.93	88.50 ± 0.88
75%	92.17 ± 1.57	91.19 ± 1.68	87.22 ± 1.02	86.94 ± 0.91

E GAIT VOCABULARY ANALYSIS

E.1 PROTOTYPE DECODING AND VISUALIZATION

We analyze the frozen VQ-Gait checkpoint used by the full model, with a gait vocabulary comprising six DoF-specific codebooks of 128 codes each. Each code is passed through its DoF-specific decoder and standardized using the same operation as during vocabulary learning (Appendix A.3). The resulting 100-point curve represents waveform shape, with its mean and scale removed. Its horizontal axis denotes relative position within a resampled gait cycle; its vertical axis is dimensionless, rather than an angle or translation in physical units. No additional smoothing, phase shifting, or sign reversal is applied for visualization.

To obtain a compact display, we select seven existing prototypes per DoF that represent both commonly used codes and waveform diversity. Specifically, we minimize the sum of squared waveform distances from eligible prototypes to their nearest selected prototype, weighted by the training assignment frequencies maintained by the tokenizer's exponential moving average (EMA). Codes with an EMA count of at least one are eligible; all 128 codes meet this criterion in each codebook. Selection uses deterministic greedy initialization followed by improving single-prototype swaps, yielding an approximate solution. The highest-usage code among the seven selected prototypes is highlighted. These training usage estimates guide display selection only; the validation frequencies reported below are computed from actual assignments.

Figure E.1 shows the resulting prototypes. Within each DoF, they differ in the relative timing and shape of peaks and troughs. For example, the FE prototypes share a prominent peak whose position varies across codes, whereas other DoFs exhibit a wider range of waveform profiles. Such differences illustrate the vocabulary's capacity to represent multiple recurring shapes within a channel. Code indices are arbitrary identifiers local to each DoF; equal indices in different codebooks do not indicate equivalent motion.

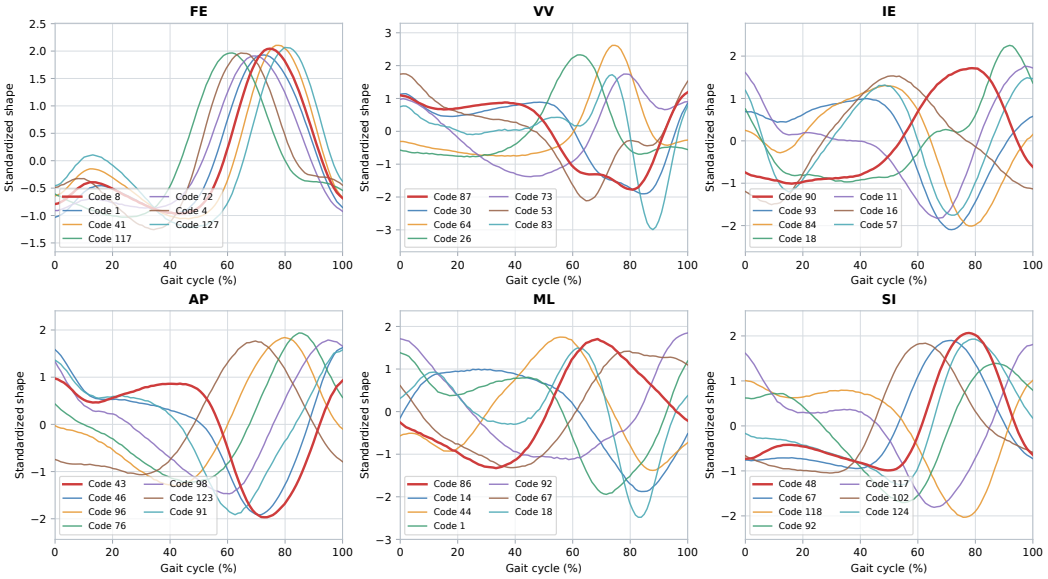

Figure E.1: Seven decoded prototypes per DoF from the current 128-code tokenizer. Red highlights the most frequently used code among the selected prototypes according to training EMA counts.

E.2 VOCABULARY USAGE ON HEALTHY VALIDATION SUBJECTS

We assign all retained cycles from the 417 healthy validation subjects using the frozen tokenizer and the preprocessing fitted on the healthy training partition. Pooling both sides yields 8,117 cycles and 48,702 DoF-specific gait units. Thus, each codebook receives 8,117 assignments. Padded cycles are excluded, and neither the encoder nor the codebooks are updated during this analysis.

Table E.1 and Figure E.2 summarize the assignment distribution. We report the number of codes receiving at least one assignment and the usage perplexity, $\exp(-\sum_{k:p_k>0} p_k \log p_k)$, where p_k is the fraction of validation gait units assigned to code k within a DoF. Perplexity equals 128 for a uniform distribution and decreases as assignments become more concentrated.

All 128 codes are used in every DoF, with perplexities between 99.15 and 113.11. The most frequently assigned code accounts for 1.79–4.37% of units, depending on the DoF. Thus, validation assignments are distributed across the vocabulary rather than concentrated in a few codes. This occupancy analysis does not by itself establish that all decoded prototypes are distinct or have separate biomechanical meanings.

Table E.1: Vocabulary usage and shape correspondence across all healthy validation gait units. Each DoF contains 8,117 units from 417 subjects. Pearson correlation and shape RMSE are averaged over units.

DoF	Used codes	Perplexity	Largest share (%)	Mean r	Mean RMSE
FE	128/128	110.12	1.79	0.992	0.113
VV	128/128	107.30	3.14	0.900	0.401
IE	128/128	99.58	3.46	0.890	0.425
AP	128/128	105.16	3.10	0.953	0.273
ML	128/128	113.11	1.95	0.850	0.506
SI	128/128	99.15	4.37	0.947	0.276

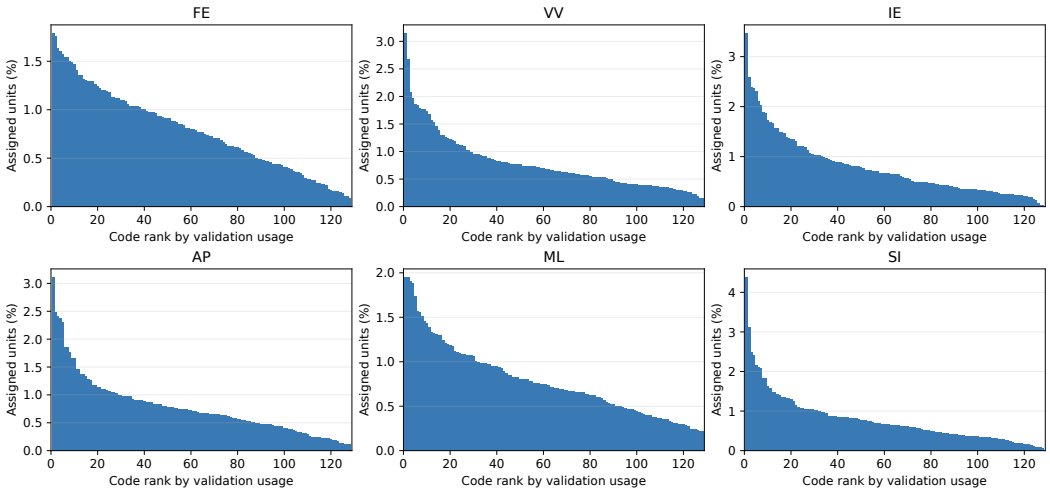

Figure E.2: Validation assignment frequencies within each DoF. Codes are ranked independently by frequency; ranks are not code indices. Each panel summarizes 8,117 gait units.

E.3 Correspondence to Recorded Gait Units

For each highlighted prototype, we randomly sample 20 assigned validation gait units without replacement, using seed $42 + d$ for zero-based model DoF index d in the order FE, VV, IE, AP, ML, and SI. Sampling is independent of waveform similarity or reconstruction error. Multiple cycles from one subject can be selected, so 20 curves do not necessarily represent 20 subjects. Figure E.3 overlays these units after the same shape normalization used by VQ-Gait. Table E.2 reports statistics for all units assigned to each highlighted code, not only the displayed sample.

The highlighted prototypes capture shared waveform profiles despite variation among individual cycles. Their assigned units come from 44–103 subjects per code, supporting reuse across subjects rather than correspondence to a single recording. The full validation analysis in Table E.1 also shows that correspondence varies by DoF: mean Pearson correlation is highest for FE (0.992) and lowest for ML (0.850), with larger shape reconstruction errors for ML. The highlighted examples

generally have closer correspondence than the all-code averages, underscoring the need to consider the complete assignment statistics alongside the visual examples.

For these statistics, we compute Pearson correlation and root mean squared error (RMSE) between each standardized input unit and the decoded prototype selected by the tokenizer, then average the resulting unit-level values. Subject coverage counts each subject once per code, pooling both sides and all retained cycles. The metrics characterize normalized shape fidelity; they do not measure recovery of physical amplitudes or clinical discrimination. Because this healthy validation partition is also used for checkpoint selection, the analysis is descriptive and is not an independent test of generalization. Repeated cycles are not treated as independent subjects for statistical inference.

Table E.2: Correspondence statistics for the six highlighted codes, using all their assigned validation units. Subject counts are unique within each code and can overlap across codes. Code indices are zero-based.

DoF	Code	Units	Subjects	Mean r	Mean RMSE
FE	8	89	54	0.996	0.087
VV	87	217	79	0.964	0.252
IE	90	138	44	0.960	0.268
AP	43	252	78	0.980	0.192
ML	86	118	44	0.916	0.394
SI	48	355	103	0.988	0.152

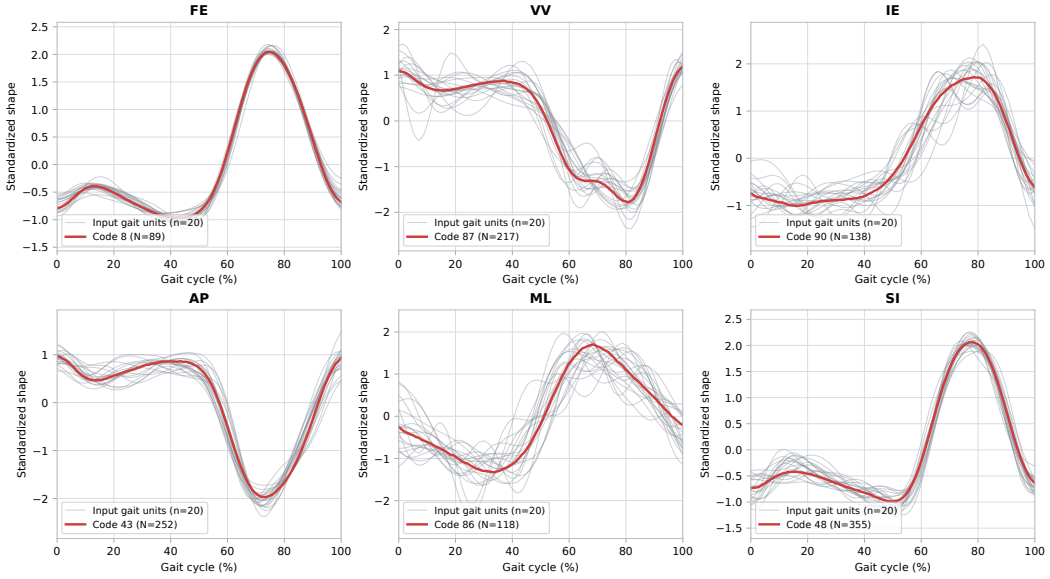

Figure E.3: Decoded prototypes (red) and 20 randomly sampled assigned validation gait units (gray). N denotes all validation units assigned to the highlighted code; n denotes the displayed sample.

E.4 Complete Gait Vocabulary

Figures E.4–E.9 display all 128 decoded prototypes for each DoF, ordered by code index with identical axis limits. These unselected views expose both shape variation and similarity across codes; distinct indices do not necessarily imply distinct biomechanical patterns. Each panel is labeled with its code index and the number of assigned healthy validation gait units, n, pooling both sides.

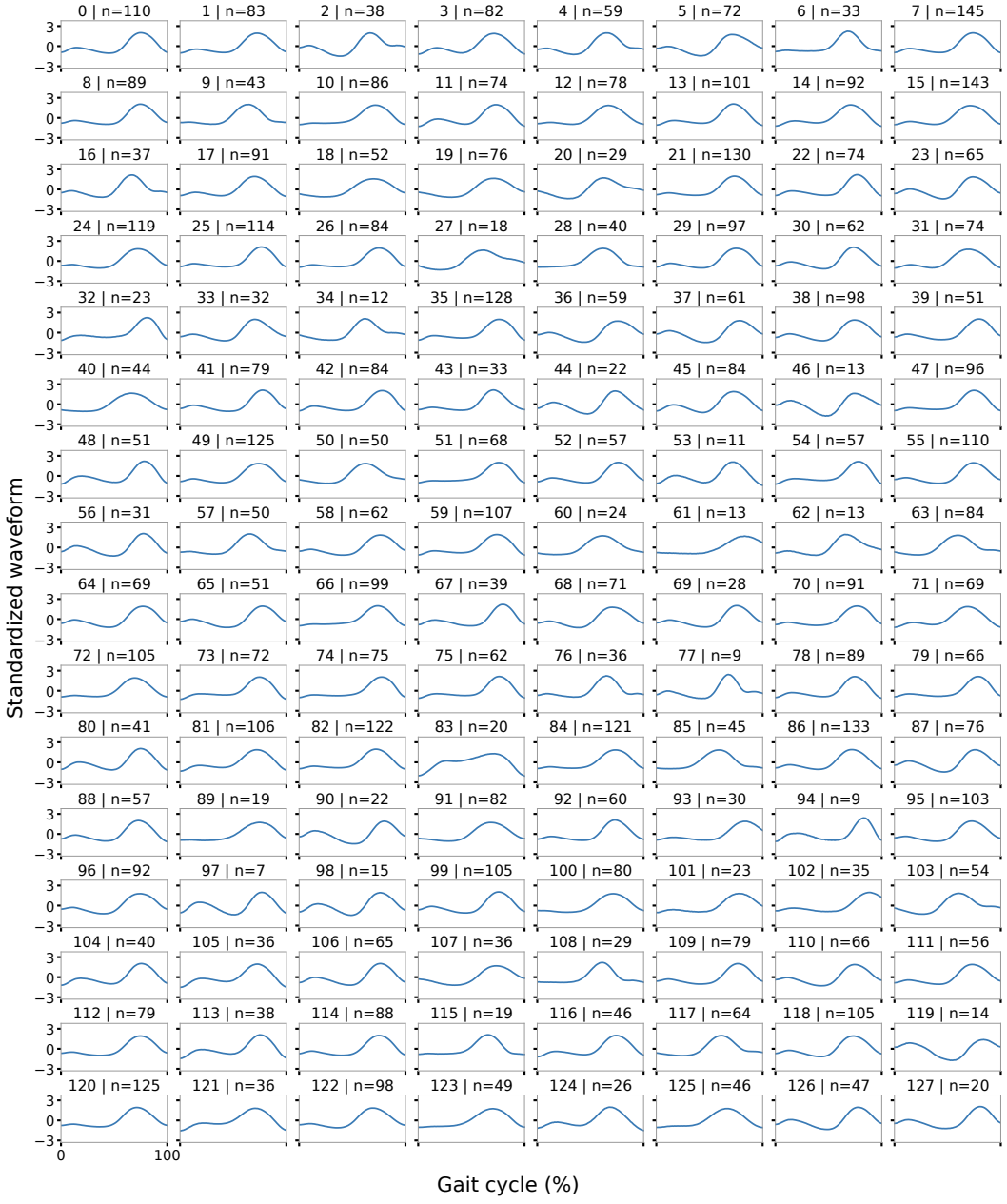

Figure E.4: Complete flexion–extension (FE) codebook. Curves show standardized decoded shapes; n counts assigned validation gait units, not subjects.

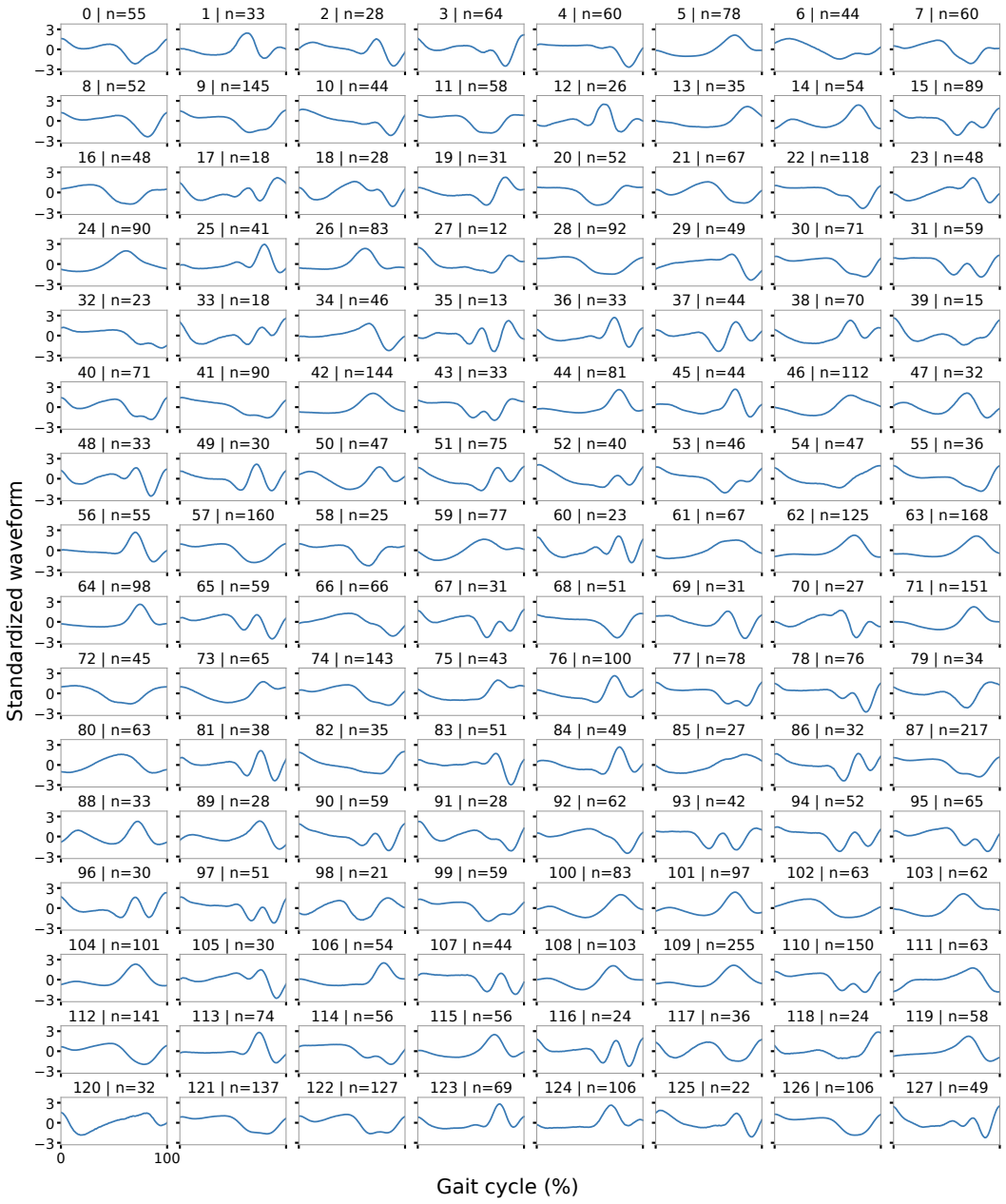

Figure E.5: Complete varus–valgus (VV) codebook. Curves show standardized decoded shapes; n counts assigned validation gait units, not subjects.

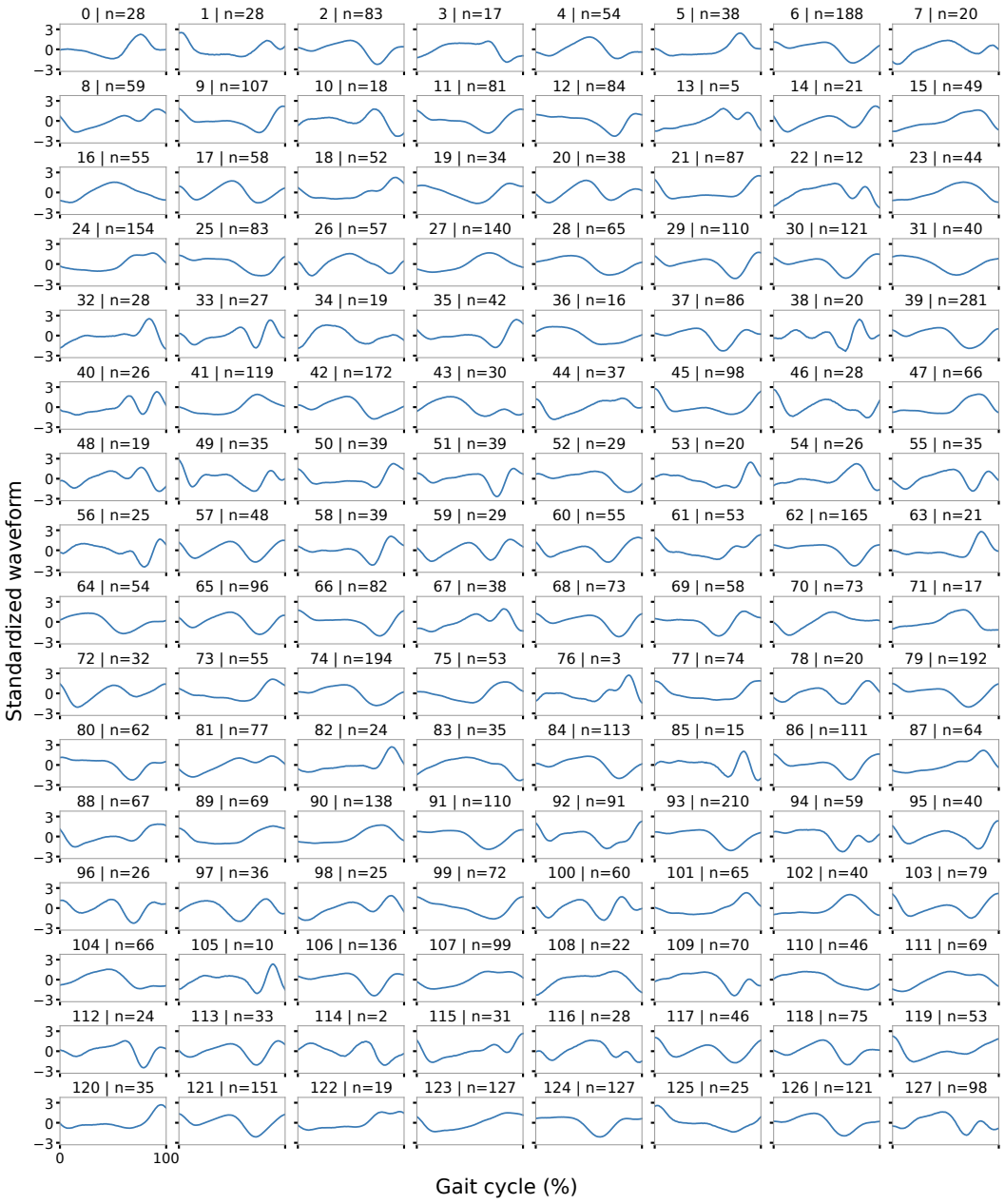

Figure E.6: Complete internal–external rotation (IE) codebook. Curves show standardized decoded shapes; n counts assigned validation gait units, not subjects.

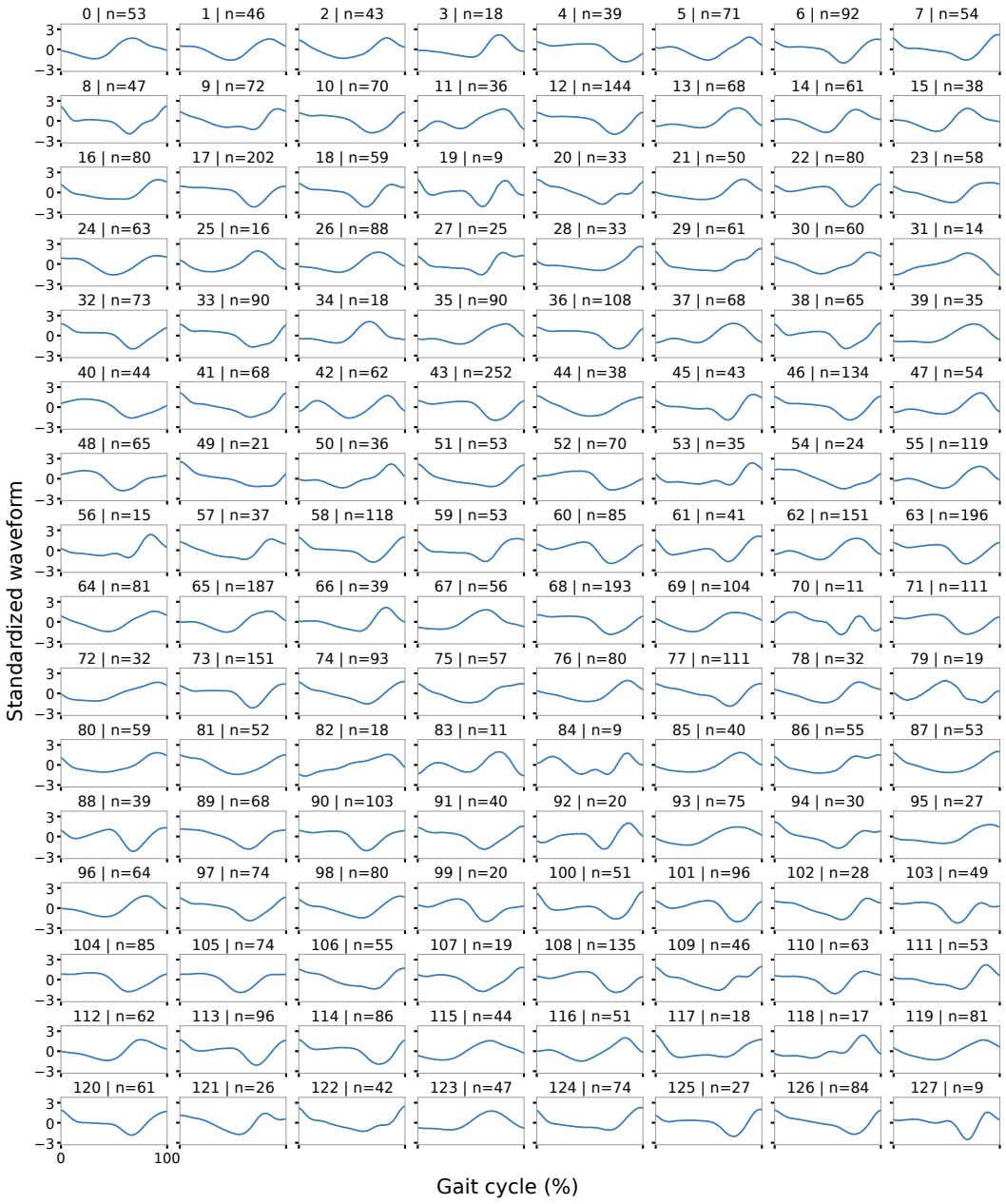

Figure E.7: Complete anterior–posterior translation (AP) codebook. Curves show standardized decoded shapes; n counts assigned validation gait units, not subjects.

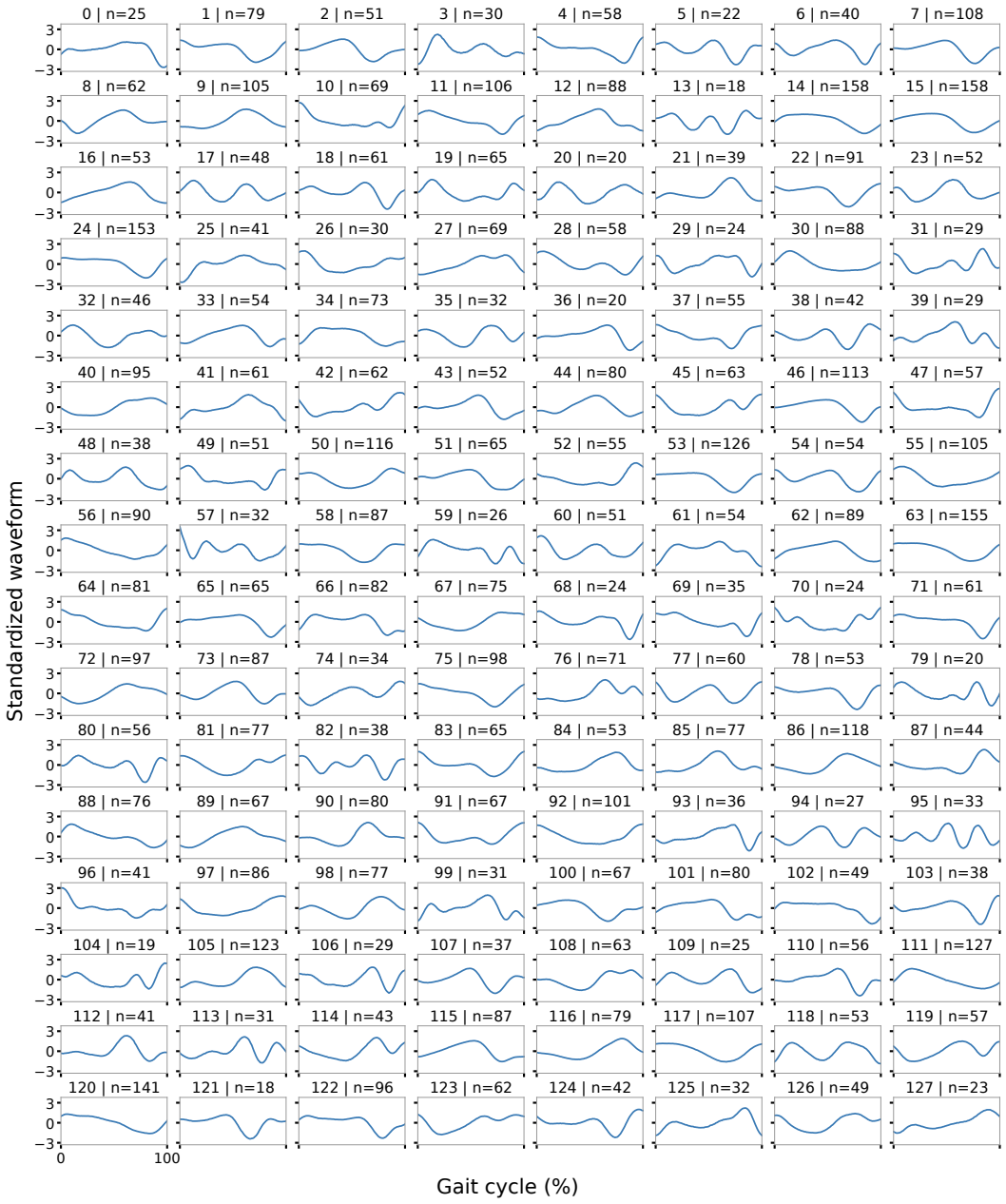

Figure E.8: Complete medial–lateral translation (ML) codebook. Curves show standardized decoded shapes; n counts assigned validation gait units, not subjects.

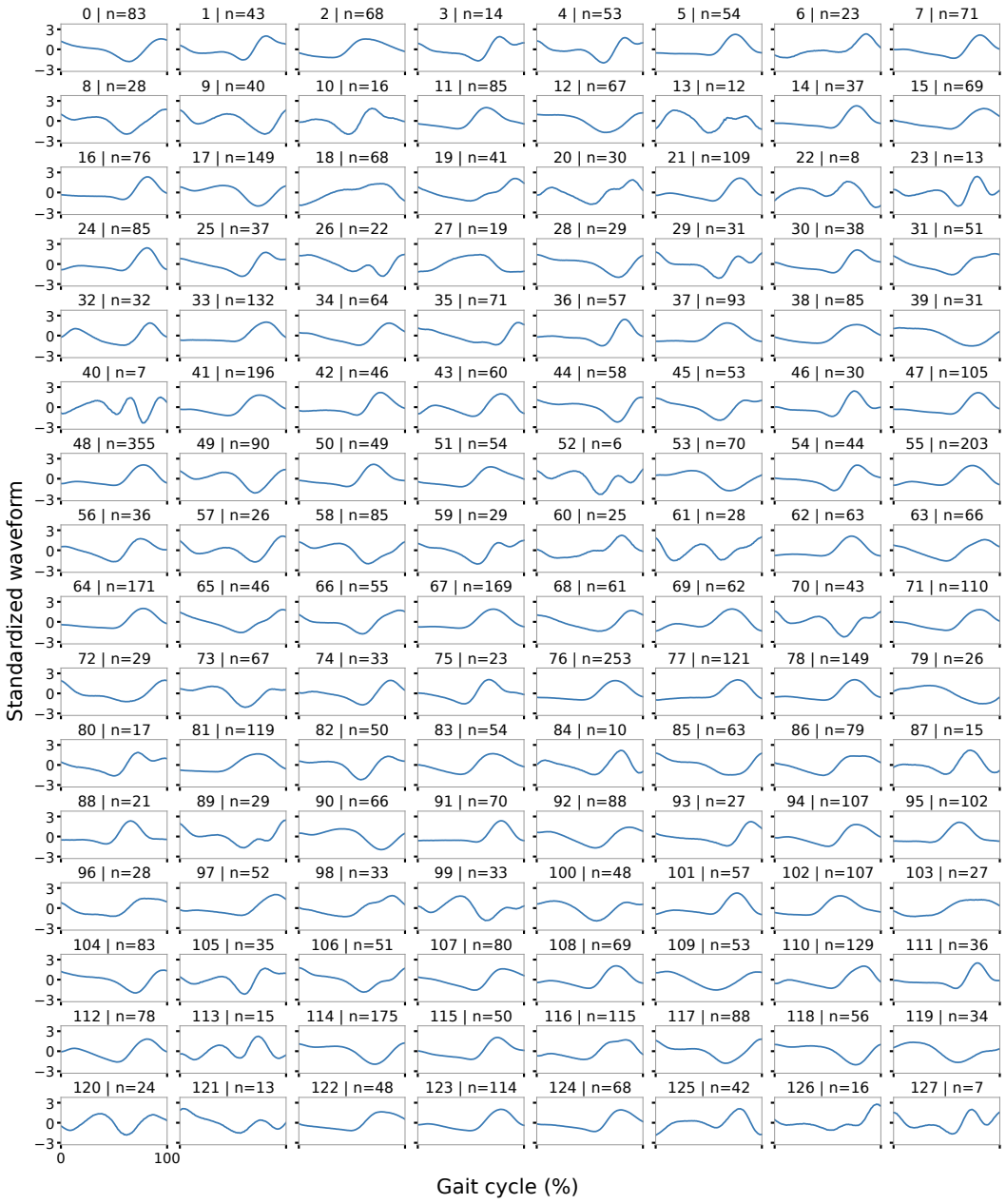

Figure E.9: Complete superior–inferior translation (SI) codebook. Curves show standardized decoded shapes; n counts assigned validation gait units, not subjects.