Appendix

A More Dataset Details

Figure A.1 presents the distribution of UPDRS-gait scores in the four labeled datasets. Score 0 (normal) is most common across cohorts, while score 3 (severe) is rare—especially in PD-GaM and 3DGait, highlighting class imbalance challenges. Figure A.2 visualizes the distribution of (a) medication states and (b) diagnostic labels. BMCLab offers a balanced ON/OFF medication split, while E-LC is skewed toward ON-medication. DNE includes healthy, Parkinsonian, and other disease groups for broader contrastive training. Figure A.3 shows label distributions for FoG-related cohorts. BMCLab and KUL-DT-T distinguish freezers vs. non-freezers, while E-LC includes subtypes such as PD with FoG, PD without FoG, and non-PD with FoG symptoms.

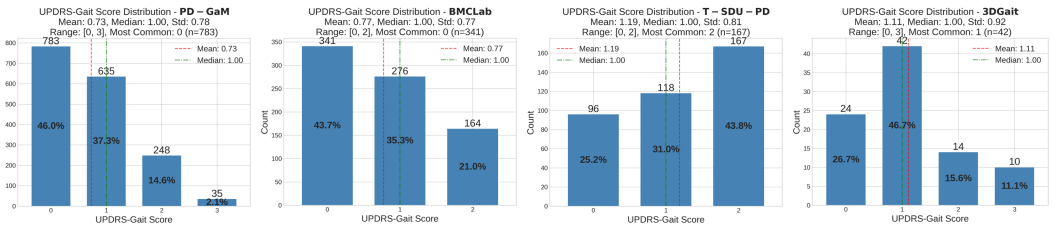

Figure A.1: Class distributions for the four datasets with UPDRS-gait labels.

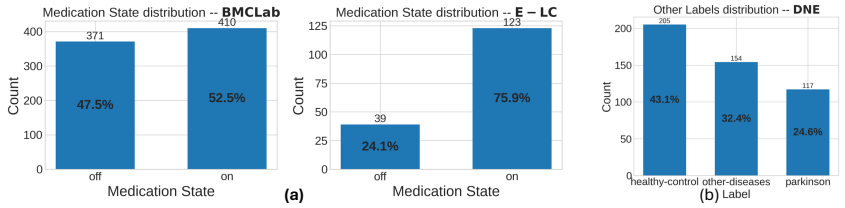

Figure A.2: (a) Medication state breakdown for BMCLab and E-LC datasets. (b) Diagnostic categories in DNE dataset.

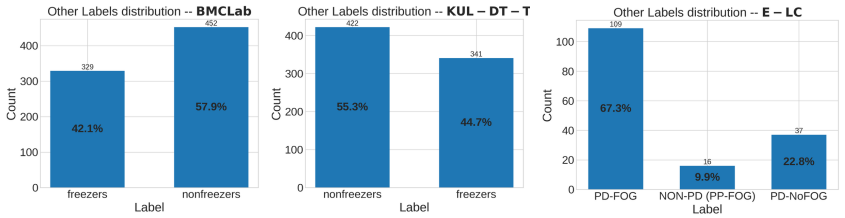

Figure A.3: Label distributions for freezing status for BMCLab, KUL-DT-T and E-LC datasets.

A.1 Slope Correction

In the CARE-PD datasets recorded from ceiling-mounted cameras (T-SDU, T-LTC, and T-SDU-PD), we observed that sometimes subjects appeared to walk along a sloped or curved plane, rather than a flat floor. This artifact likely stems from the unusual top-down perspective—different from the front-facing or side views seen in WHAM's training data [1]. While motion encoder-based models may be robust to such distortions, feature-based gait classifiers rely on precise kinematic measurements and thus require carefully corrected input data. To correct this slope artifact, we perform a frame-wise rigid alignment of the reconstructed SMPL skeleton using the Kabsch algorithm [2]. The goal is to rotate each frame so that anatomical directions align with canonical coordinate axes (up, forward), while preserving natural gait structure. Let the SMPL skeleton at time t be a set of 3D joint positions: $\mathbf{J}^t \in \mathbb{R}^{22\times 3}$. We define three key anatomical vectors per frame:

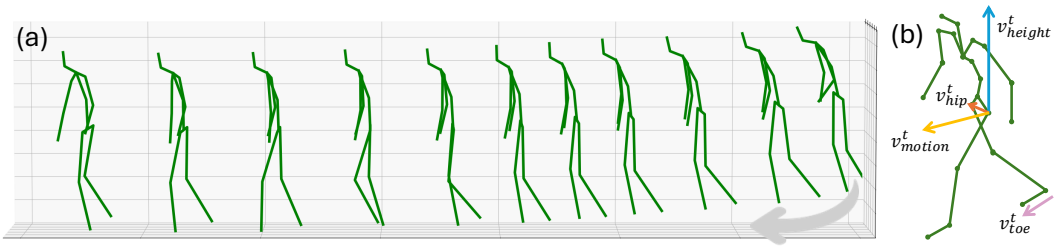

Figure A.4: a) Illustration of the slope artifcat b) An example of the vertical height vector (blue), the direction of movement vector (yellow), and the hip vector (orange).

(1) Height Vector (posture): defined as the offset between the sacrum and the average of the ankle and knee joint positions.

$$\mathbf{v}^t_{\text{height}} = \mathbf{j}^t_{\text{sacrum}} - \frac{1}{4}\left(\mathbf{j}^t_{\text{left ankle}} + \mathbf{j}^t_{\text{right ankle}} + \mathbf{j}^t_{\text{left knee}} + \mathbf{j}^t_{\text{right knee}}\right)$$

This approximates the vertical posture and should align with the global y-axis: $\hat{\mathbf{y}} = [0, 1, 0]^T$. Misalignment suggests the subject appears tilted in 3D space.

(2) Motion Vector (walking direction): To estimate walking direction, we compute the offset between the sacrum at frame t and frame $t + 15$, representing ~ 0.5 seconds ahead. This motion vector is then projected onto the ground plane (xz-plane) and used as the walking axis.

$$\mathbf{v}^t_{\text{motion}} = \text{Proj}_{xz}\left(\mathbf{j}^{t+15}_{\text{sacrum}} - \mathbf{j}^t_{\text{sacrum}}\right)$$

where $\text{Proj}_{xz}(\cdot)$ zeroes out the y-component. In frames where the sacrum displacement is less than 4mm—indicating near-stationary posture—we fall back on a proxy direction: the cross product of the hip vector (left hip to right hip) and the vertical vector. This gives a third perpendicular vector — ideally pointing forward along the walking direction.

$$\mathbf{v}^t_{\text{motion}} = \mathbf{v}^t_{\text{hip}} \times \mathbf{v}^t_{\text{height}}, \qquad \text{If } \|\mathbf{v}^t_{\text{motion}}\| < 4\text{mm}$$

This proxy is adjusted to ensure consistency with foot orientation (by checking the sign of its dot product with toe direction and flipping the fallback direction when). We ensure alignment by flipping the fallback direction when

$$\text{sign}\left((\mathbf{v}^t_{\text{motion}})^\top \cdot \mathbf{v}^t_{\text{toe}}\right) < 0, \qquad \mathbf{v}^t_{\text{toe}} = \mathbf{j}^t_{\text{toe}} - \mathbf{j}^t_{\text{heel}}$$

We normalize and smooth $\mathbf{v}^t_{\text{motion}}$ over time using a Savitzky–Golay filter [3] (window=90, order=4) to ensure temporal coherence.

(3) Hip Vector (rotation anchor): We assigned different importance to different pairs of vectors that should be aligned in the Kabsch algorithm. We set the weight for the alignment of the hip vector to infinity while the other two alignments were given a weight of 1. Thereby, we forced the hip vector ($\mathbf{v}^t_{\text{hip}} = \mathbf{j}^t_{\text{right hip}} - \mathbf{j}^t_{\text{left hip}}$) to stay aligned perfectly with itself while the other two vectors were allowed to deviate slightly from their targets. This prevents the correction from introducing unnatural body twisting to the subject's gait. Let

$$\mathcal{S} = \{(\mathbf{v}_i, \hat{\mathbf{v}}_i, w_i)\}$$

where $\mathbf{v}_i \in \{\mathbf{v}^t_{\text{height}}, \mathbf{v}^t_{\text{motion}}, \mathbf{v}^t_{\text{hip}}\}$, target $\hat{\mathbf{v}}_i \in \{\hat{\mathbf{y}}, \hat{\mathbf{z}}, \mathbf{v}^t_{\text{hip}}\}$, and weights $w_i \in \{1, 1, \infty\}$. We solve the weighted orthogonal Procrustes problem:

$$\mathbf{R}^t = \arg\min_{\mathbf{R} \in SO(3)} \sum_i w_i \|\mathbf{R}\mathbf{v}_i - \hat{\mathbf{v}}_i\|^2$$

The solution $\mathbf{R}^t$ is the optimal rotation aligning anatomical directions. We then apply this rotation to the entire skeleton around the root joint (sacrum) and translate the rotated skeleton vertically so that the lowest foot joint rests at $y = 0$, ensuring ground contact consistency. This method corrects the slope artifacts while preserving the gait dynamics and anatomical validity of each sequence. An illustration of the process and vector definitions is shown in Fig. A.4.

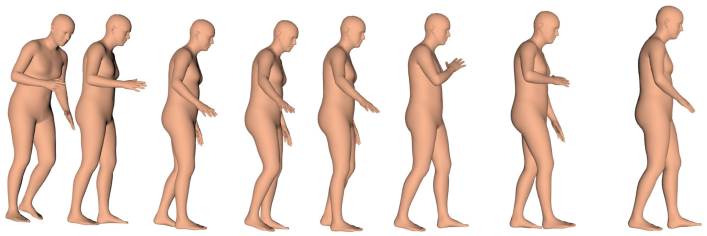

Figure A.5: Example of the 6890 vertices SMPL mesh at different frames of the gait sequence.

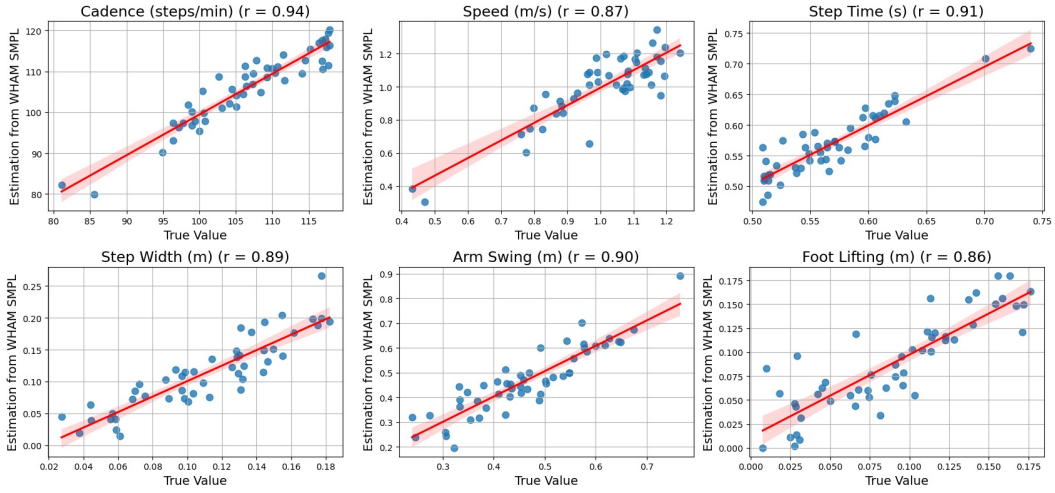

Figure A.6: Correlation between WHAM estimations and IMU ground-truth gait features. Pearson correlations ranged from $r = 0.86$ to $r = 0.94$.

A.2 Clinical Validity of WHAM

We validated the WHAM SMPL estimations against a publicly available video + IMU benchmark, the Toronto Older Adults Gait Archive (TOAGA) [4]. For each walk (from 14 participants) we extracted cadence, walking speed, step time, step width, arm-swing amplitude and foot-lifting height from the WHAM meshes and compared them to the features from their synchronised Xsens MVN Analyze 3D IMU recordings, which incorporate a reliable biomechanical model. Pearson correlations ranged from $r = 0.86$ to $r = 0.94$ (Fig. A.6), closely matching the high correlation originally observed between 2D pose-tracking and IMU measures in the TOAGA paper, supporting the biomechanical and clinical reliability of estimations. Furthermore, to quantify geometric accuracy we computed root-relative MPJPE between WHAM key-points and synchronised Xsens ground truth in TOAGA. The mean error was 39 mm, comfortably within the 35–65 mm range reported for multi-view pose estimation systems [5, 6, 7] and the TULIP dataset for PD gait task [8]. This level of agreement, together with high spatiotemporal feature correlations (Fig. A.6), supports the use of WHAM meshes as a reliable surrogate for markerless gait analysis in our study.

A.3 Baseline Models and Baseline-specific Data Preprocessing

To ensure that every backbone receives input in the format and frame-rate it was trained on, we applied a unified preprocessing pipeline along with baseline-specific preprocessing steps. All motion sequences were converted to 30 FPS to match the expected input frequency of the pretrained encoders. All preprocessing steps and motion representation generation procedures are available in our public code repository.

A.3.1 Motion Formats

To facilitate rigorous evaluation of motion encoder performance in clinical gait settings, we selected state-of-the-art models that operate on two broad classes of motion representations: skeleton-based (joint locations) and mesh-based (SMPL parameters). The SMPL-based models use either raw pose parameters or a redundant representation optimized for motion generation tasks. All formats are derived from SMPL as described below.

SMPL The SMPL model [9] represents human body shape β and pose θ using 24 joints and a 10-dimensional shape vector. The pose is expressed as a set of joint rotations (e.g., axis-angle), and can be rendered as a mesh with 6890 vertices, an example of which can be seen in Fig. A.5. For each time step t, the SMPL input sequence $\mathbf{M}^{1:T}$ has shape $\mathbb{R}^{T\times 24\times D}$, where D is the dimension of the rotation representation. SMPL serves as the base representation for generating other formats.

Human3.6M Joints Many encoders in our study were originally trained on the Human3.6M dataset [10], which uses a 17-joint skeleton. We project SMPL mesh vertices to this joint format using a linear regressor matrix $\mathbf{R} \in \mathbb{R}^{17\times 6890}$, as done in MotionBERT [11]. For each frame, the 3D Human3.6M joint coordinates are computed by multiplying the mesh with this regressor. The resulting motion sequence has shape $\mathbb{R}^{T\times 17\times 3}$.

HumanML3D The HumanML3D representation [12], originally introduced for text-to-motion generation, encodes each frame $\mathbf{m}^t \in \mathbb{R}^{263}$ as a tuple of interpretable features $\mathbf{m}^t = \{\dot{r}_a, \dot{r}_x, \dot{r}_z, r_y, \mathbf{j}_p, \mathbf{j}_v, \mathbf{j}_r, \mathbf{c}_f\}$, where $\dot{r}_a \in \mathbb{R}$, is the root joint's angular velocity along the y-axis; $\dot{r}_x, \dot{r}_z \in \mathbb{R}$, are the root's linear velocities in the xz-plane; and $r_y \in \mathbb{R}$, is the vertical height of the root joint. Joint-level features include $\mathbf{j}_p \in \mathbb{R}^{3(N_j-1)}$, the 3D positions of all joints except the root (21 joints); $\mathbf{j}_v \in \mathbb{R}^{3N_j}$, the linear joint velocities; and $\mathbf{j}_r \in \mathbb{R}^{6(N_j-1)}$, the 6D joint rotations relative to parent joints in the skeletal hierarchy. Finally, $\mathbf{c}_f \in \mathbb{R}^4$ encodes four binary foot contact indicators derived from heel and toe velocities. This representation was computed from SMPL joints using the procedure introduced in [12], yielding input tensors of shape $\mathbb{R}^{T\times 263}$.

A.3.2 Models

Our benchmark intentionally spans *diverse pre-training objectives, input formats, and architectural choices* so that conclusions about clinical transfer do not depend on a single modelling paradigm. To assess the clinical utility of pretrained motion representations, we evaluate seven state-of-the-art encoders spanning a range of architectures, training objectives, and input formats. These models were selected for their strong performance on benchmark motion tasks such as 2D to 3D lifting, motion reconstruction, prediction, and generation. They cover both skeleton-based and mesh-based representations and include both discriminative and generative paradigms. All models are used as fixed backbones; we extract their latent representations from last layer before final head (pooled over temporal dimension) and train lightweight classifiers on top for UPDRS-gait severity prediction.

POTR [13] is a transformer-based model originally developed for non-autoregressive human pose forecasting. Although designed for forecasting, its encoder learns strong spatiotemporal representations of input motion sequences shown to be useful for clinical gait assessment task [14]. We use the encoder's temporally pooled token embeddings as input features for our downstream clinical classifier. Input: 3D Human3.6M joints.

MixSTE [15] is a 2D to 3D joint lifting model that factorizes spatial and temporal dependencies using stacked blocks of transformer encoders. Each block in its stacked architecture consists of a spatial transformer that captures joint-to-joint relationships within a single frame, followed by a temporal transformer that models how each joint evolves across time. Input: 2D projected (prespective) Human3.6M joints.

PoseFormerV2 [16] is a transformer-based model for 2D to 3D lifting that addresses two key challenges: computational efficiency and robustness to noisy 2D inputs. It applies a Discrete Cosine Transform (DCT) to each joint trajectory to obtain a compact, frequency-domain representation of global motion. Only a subset of low-frequency DCT coefficients are retained, effectively reducing noise from 2D pose estimators and shortening the sequence length. A spatial transformer encodes relations among joints using a fixed number of central frames, while the frequency features are linearly projected and concatenated with the spatial output. This combined representation is processed by a

temporal transformer to model motion dynamics, and finally decoded back to the time domain. This architecture allows the model to capture long-range dependencies with reduced computational cost. Input: 2D projected (prespective) Human3.6M joints.

MotionBERT [11] is a dual-stream spatiotemporal transformer designed for 2D to 3D pose lifting. It takes 2D joint sequences as input and learns representations that capture both spatial relations among joints and temporal dynamics across frames. The model consists of stacked transformer blocks, each with two parallel branches: one applies multi-head self-attention in a spatial-first order (joint-wise attention followed by temporal), and the other in a temporal-first order. This design allows MotionBERT to learn complementary patterns in human motion while retaining frame-wise features useful for action recognition. The outputs from both streams are merged using a learned weighted average. In our setting, the final representation is obtained by averaging the output tokens across time. Input: 2D projected (orthographic) Human3.6M joints.

MotionAGFormer [17] extends the dual-stream transformer design of MotionBERT by integrating Graph Convolutional Networks (GCNs) into one of the branches. One stream uses MHSA to capture long-range dependencies, while the other applies spatial and temporal GCNs to model local joint interactions. The spatial GCN encodes the human body structure, while the temporal GCN builds connections based on feature similarity across time. This hybrid attention–graph architecture enhances robustness to localized variations in movement. Final features are obtained by temporally averaging the outputs across frames. Input: 2D projected (orthographic) Human3.6M joints.

MotionCLIP [18] is a transformer-based motion autoencoder trained for text-to-motion generation. During training, its latent space is aligned with the CLIP embedding space, enabling it to bridge motion and language domains; yet its motion encoder is a strong semantic aggregator. Including it tests whether language-aligned features, which never saw clinical labels, can be transferred to severity scoring. The model encodes SMPL pose sequences using stacked transformer layers and reconstructs them from the latent representation. For our experiments, we use its motion encoder as a frozen backbone and extract frame-level representations by averaging token outputs. MotionCLIP requires SMPL input in 6D rotation format, which avoids discontinuities associated with axis-angle representations and improves learning stability [19]. Input: SMPL (6D rotation).

MoMask [20] is a Vector Quantized VAE (VQ-VAE) based framework for text-conditioned 3D motion generation. It comprises a Residual VQ-VAE encoder–decoder for motion reconstruction and representation learning plus two transformers: a masked transformer for predicting base-layer motion tokens, and a residual transformer for refining higher-layer tokens. Unlike standard VQ-VAEs, MoMask uses multiple codebooks to iteratively quantize the residuals, enabling finer motion detail. We use the pretrained RVQ-VAE encoder as a feature extractor and obtain motion representations by summing tokens across all residual layers and averaging over time. MoMask operates on the HumanML3D representation and requires normalized features; normalization statistics are computed per dataset or per LODO training split. Input: HumanML3D.

A.3.3 2D Projection Pipeline

To evaluate motion encoders trained on 2D joint data, we converted every 3D sequence into the appropriate 2D format via projection. To ensure a fair comparison between 2D and 3D models, given that projection discards depth information, we defined a multi-view setup for 2D encoders, using both back views, which minimize limb occlusion, and side views, which better preserve stride length. Using ground-truth 2D skeletons isolates an encoder's *representation capacity* from the performance of upstream key-point detectors and avoids confounds introduced by varying video quality across the eight sites.

The projection pipeline involves several steps: 1) Canonicalizing orientation of each regressed SMPL pose so that the initial walking direction faces $+z$. 2) *Perspective projection*. We render "perfect" skeletons using two virtual pinhole camera models, viewing the walk from side and back[1]. *Orthographic projection* for MotionBERT and MotionAGformer by removing the z axis in camera coordinate. Views from the side and back were chosen to reflect common clinical perspectives. 3) Pruning out-of-frame projections. Frames in which any joint projects outside the image plane are discarded. Also, sequences shorter than 30 frames after clipping are excluded.

[1]For additional implementation details of the projection setup, including camera configuration and rendering, refer to `data/preprocessing/smpl2h36m.py` in our codebase.

dataset file (.pkl) → subject_id → walk_id → [data]

	fields	data type	description
[data]	pose	array	SMPL pose parameters
	trans	array	translation parameters
	beta	array	pose blend shapes[†]
	fps	integer	frames per second
	UPDRS_GAIT	integer	UPDRS gait score (0-3)[*]
	medication	string	medication status[*]
	other	string	additional labels (e.g., FoG)[*]

Figure A.7: Each dataset within CARE-PD is provided as a single .pkl data file, structured as illustrated. [†]Pose blend shapes are set to zero to preserve anonymity. [*]Label information varies by dataset and is explicitly set as None if unavailable.

To test whether complementary viewpoints help severity scoring, we build a "Side & Back" variant: a Side (lateral) probe and a Back (posterior) probe are trained independently on their respective projections and their softmax outputs are averaged at inference time. All 2D encoder results reported in the manuscript use this multi-view fusion setup, as it consistently outperforms either Side or Back views alone.

A.3.4 Input Normalizations

For each model we followed its original preprocessing scheme. For MixSTE and PoseFormerV2, input 2D joint coordinates were re-scaled to $[-1, 1]$ in image space. For MotionAGFormer and MotionBERT cropping and rescaling normalization is used. Specifically, valid joint coordinates in the 2D image plane are tightly cropped to the bounding box of the motion, then linearly rescaled to the $[-1, 1]$ range. The scaling is performed independently per clip using the larger of the height or width of the bounding box to preserve aspect ratio. POTR, which operates on 3D joint coordinates, centers each pose (i.e., per-frame joint set) on the pelvis and applies z-score normalization from the training set. MotionCLIP expects SMPL rotations in continuous 6D form; we therefore convert every axis-angle in the walk to 6D. For MoMask, we computed per-dataset mean/std (or, in LODO, mean/std on the pooled training sets) and divided the std of four root-velocity channels by a factor of 5, as recommended by the authors to emphasize global trajectory [20]. For all the encoders, if a motion clip is shorter than the required input length, zero-padding is applied and a binary mask is used to track valid (non-padded) frames. For PoseFormerV2, which processes the central frames through a spatial transformer, we apply symmetric padding to preserve the alignment of meaningful motion content with the model's receptive field.

A.3.5 Generality of CARE-PD.

While our benchmarks focus on widely used motion formats and pretrained encoders, CARE-PD is not restricted to these configurations. Its unified SMPL representation enables future work to explore other input types as well as specialized model architectures tailored to clinical gait analysis. We therefore view the present baselines as a starting point: future work can freely experiment with new motion formats and model classes that may prove even better suited to clinical gait analysis.

A.4 Data Access and Preparation

The CARE-PD database is publicly accessible via the University of Toronto Dataverse. It is hosted by the University of Toronto Libraries, with data storage provided by the Ontario Library Research Cloud, a secure and geographically distributed cloud storage network developed in collaboration with partner universities across Ontario, Canada. The database is released under a CC-BY-NC license, allowing for open but non-commercial use with appropriate attribution. Detailed instructions for accessing the database can be found directly on the Dataverse project page (Data) and the GitHub code base (CARE-PD). The structure of the CARE-PD database's metadata and SMPL data

is visualized in Fig. A.7. In addition to the SMPL data, CARE-PD includes three derived assets to facilitate ease of use: Human3.6M, HumanML3D, and SMPL6D formats. For more information on these derived assets, we refer users to supplementary documentation in Sec. A.3.1 and our GitHub code base.

B Gait Feature Extraction Details

To build an interpretable baseline for UPDRS-gait classification, we extract a set of clinically meaningful gait features from 3D joint trajectories in Human3.6M format. These features, inspired by established clinical guidelines and prior work [21, 22, 23], span spatiotemporal, stability, and posture-related dimensions relevant to parkinsonian gait.

Heel Strike Detection. Accurate detection of heel strike events is necessary for estimating step-level features. We compute the Euclidean distance between the left and right ankle joints over time identifying local maxima that are at least 8 frames apart and have a prominence of at least 0.02. These peaks approximate the alternating steps and define the heel strike timestamps.

Extracted Gait Features. Following [24], we compute the following gait features, using the detected heel strikes:

- *Cadence*: steps per minute, based on the total number of detected heel strikes.
- *Step Length / Width / Time*: computed between consecutive heel strikes. Step length is the distance measured along the walking (z) axis, step width along the mediolateral (x) axis at the time of each detected heel strike, and step time as the duration between strikes. Both the mean and standard deviation of these values are calculated.
- *Walking Speed*: total sacrum displacement between first and last heel strike, divided by total time.
- *Estimated Margin of Stability (eMoS)*: computed as the minimum distance between the extrapolated center of mass (XCoM) and base of support (feet) along the mediolateral direction. The hip vector approximates this axis. We calculated both the minimum (capturing the most unstable moment) and the standard deviation across steps.
- *Foot Lifting*: the vertical range of ankle movement.
- *Stoop Posture*: defined as the forward-lean distance is the vertical displacement between neck and sacrum, projected onto the direction of walk.
- *Arm Swing*: horizontal displacement of the hand joints along the forward axis, after translating the sacrum to the origin to remove global motion.

To ensure consistency, all sequences are pre-aligned to a canonical coordinate system (z-forward, y-up, x-lateral). This alignment is critical for ensuring geometric consistency when computing direction-sensitive features such as step length, step width, and stoop posture. Previous studies [25, 26, 27] have demonstrated the relevance of these gait features to the severity of PD symptoms. A low cadence and short step length are characteristic of slowness of movement, one of the hallmark symptoms of PD. While narrower step width and lower eMoS values reflect stability issues [4]. PD may also manifest as patients taking shorter steps, resulting in elevated cadence [26]. Moreover, a stooped posture is commonly seen in PD and is directly associated with postural instability [27].

We use a Random Forest classifier to map the extracted gait features to UPDRS-gait score classes. The model is trained and evaluated using the same data splits, evaluation metrics, and hyperparameter tuning strategy as the encoder-based models (detailed in Appendix C), ensuring a consistent comparison across representation-learning and handcrafted approaches.

C Reproducibility

The experiments in this work can be reproduced using our Github repository, available at this link: https://github.com/TaatiTeam/CARE-PD/. Steps for how to reproduce evaluation experiments are available in our code README.md and dataset.md.

Compute resources. All clinical score estimation task experiments were conducted on one NVIDIA A40 GPU hosted on a HPC cluster and pretext task experiments were conducted on a single RTX6000

GPU. In pretext experiments, training MotionAGFormer for 50 epochs took approximately 15 hours, while MoMask required around 2 hours for 30 epochs. All code are implemented in PyTorch. More information on dependencies can be found on the Github page, installation guideline. Hyperparameter tuning was performed using Optuna [28] with 50 trials per model-dataset pair. In all the experiments best set of hyperparameters were found in the first $\sim$30 trials.

Hyperparameter Tuning Details During classifier training, all encoder backbones were kept frozen. We trained only the classifier head and tuned its hyperparameters using 6-fold stratified cross-validation on the BMClab dataset. BMClab was chosen due to its large size, clean motion capture quality, and pre-extracted walking segments. Hyperparameter search was conducted using the Optuna framework [28] to explore a wide range of options, including learning rate $\{1e^{-2}, 1e^{-3}, 1e^{-4}, 1e^{-5}\}$, batch size $\{64, 128, 256\}$, number of training epochs $\{10, 20, 30, 50, 70\}$, weight decay $\{0, 0.001, 0.01\}$, and loss type (weighted cross-entropy or focal loss). For focal loss, the $\gamma \in \{1, 2\}$ parameter were included in the search and α was set to one.

The best-performing hyperparameters discovered on BMClab were reused across all datasets (see Fig. 1 in the paper). However, the optimal number of epochs was selected individually per dataset to account for differences in dataset size. Json file for the best set of hyperparameters used for each experiment is available in our GitHub page (Link) in `configs/best_configs_augmented` folder. All splits used in cross-validation were subject-disjoint and stratified by label to prevent data leakage and ensure robust estimates. The exact fold splits used for each dataset and evaluation protocol are provided in the `folds` directory of our data repository.

For the LODO and MIDA experiments, the classifier head hyperparameters were again tuned on the combined training set (train set of the target dataset plus all the other datasets excluding the target's test set), using the same Optuna-based approach. The same hyperparameter tuning procedure was applied to the pretext task experiments.

The Random Forest classifier used in the engineered-feature baseline was also tuned using 6-fold subject-stratified cross-validation on the BMClab dataset. The optimal configuration was then applied uniformly across all experiments.

D More Experimental Results

D.1 Cross-site robustness vs. in-site accuracy

The two scatter plots in Fig. D.8 summarize every pair of ⟨encoder × source-cohort⟩ probe by plotting its LOSO (within-dataset) macro-F_1 on the horizontal axis and the mean of its three off-site scores on the vertical axis; the grey diagonal marks perfect transfer. The left panel uses the 3-class metric (labels 0-2) whereas the right panel includes the rare severe class 3. The circled region highlights MoMask models that consistently combine strong within-dataset accuracy with robust cross-dataset generalization, with PD-GaM-trained variants showing the most prominent and reliable transferability, confirming that (i) breadth and heterogeneity of the source data are critical and (ii) this backbone make best use of that breadth.

Adding class 3 shifts every points trained on BMCLab and T-SDU-PD (the two dataset without label 3) downward, often by 5–10pp on the y-axis, but the relative ordering is unchanged; models that were robust in the 3-class setting remain the most robust once the challenging severe cases are re-introduced. This pattern reinforces the earlier conclusion that scarcity of severe samples, is a major failure mode on cross-site tests.

D.2 Multi-dataset in-domain adaptation (MIDA) vs. baseline accuracy

Figure D.9 contrasts standard LOSO evaluation (x-axis), where each model is trained solely on the target dataset, with MIDA (y-axis), where training includes both the target dataset and additional cohorts. Most points rise above the diagonal, showing that supplementing a site's own data with external cohorts usually helps, even though the test split is unchanged.

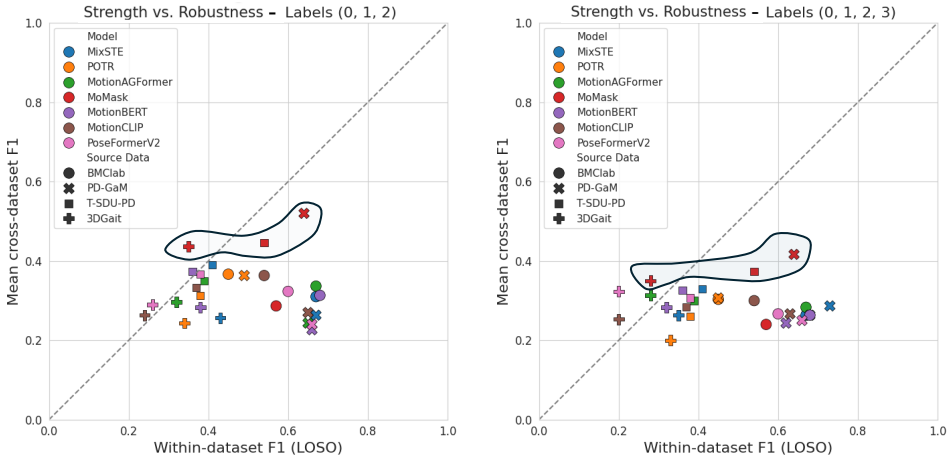

Figure D.8: Accuracy vs. robustness analysis. Each marker represents an encoder plus linear prob trained on one dataset (marker shape) and evaluated on that dataset (x-axis) and, on average, on the other three (y-axis). Colours distinguish encoder backbones; The left plot reports macro-F1$_{0-2}$, the right F1$_{0-3}$. The enclosed region highlights the most robust backbone and probes.

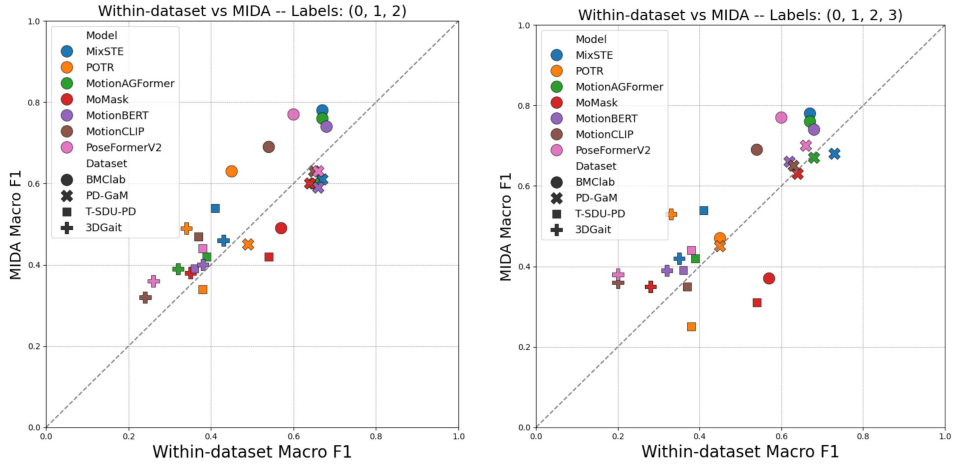

Figure D.9: Within-dataset vs. MIDA evaluation (effect of in-domain adaptation and external data for training). Each point compares macro-F$_1$ scores of an encoder trained with (y-axis) and without (x-axis) access to additional out-of-domain datasets. LOSO uses training data only from the evaluation cohort; MIDA adds other datasets to training while still testing on the same held-out subjects. Colors indicate encoder backbone and shapes indicate target dataset. The left panel uses macro-F$_1$ over labels 0–2, the right panel over 0–3. Markers above the diagonal indicate improvement; colours denote backbones, shapes the source cohort.

D.3 View-point results for 2D encoders

Table D.1 reports the average macro-F$_1$ scores across datasets for four 2D models under the within- and cross-dataset, LODO, and MIDA evaluation protocols. We separately evaluated performance using posterior and lateral projections and their combination to assess the effect of viewpoint on model performance and robustness. When only posterior or lateral projections are available, accuracy varies with the backbone. Fusing both views ("Combined") reliably boosts performance, suggesting that the two projections supply complementary depth cues.

D.4 Variability Reporting

We report mean macro-F$_1$ scores alongside their standard deviation to quantify variability and support statistical interpretation. *(i) LOSO:* inside each cohort we perform leave-one-subject-out; the n held-out subjects yield n scores. We report mean$\pm$SD across these n folds. *(ii) Cross-dataset:*

Table D.1: Average macro-F_1 (%) of the 2D encoders across all datasets, grouped by evaluation protocol and viewpoint. "Posterior" and "Lateral" use single-view projections, while "Combined" averages a posterior and a lateral probe at score level. The upper half evaluates all four UPDRS classes, the lower half excludes the rare score 3. Means (last column) are taken across the four backbones.

Protocol	View	MixSTE	MotionAGFormer	MotionBERT	PoseFormerV2	Mean
			Included Labels: {0,1,2,3}			
Within/Cross	Posterior	**35.19**	**35.69**	25.81	28.87	31.39
	Lateral	34.31	32.38	**33.38**	**34.06**	**33.53**
	Combined	34.94	34.38	**33.31**	33.00	**33.91**
LODO	Posterior	32.25	34.00	25.75	30.00	30.50
	Lateral	33.01	33.75	28.00	30.50	31.31
	Combined	**35.75**	**38.75**	**32.25**	**36.75**	**35.88**
MIDA	Posterior	55.75	39.75	45.51	54.25	48.81
	Lateral	39.75	52.25	40.00	46.75	44.69
	Combined	**60.51**	**55.67**	**54.51**	**57.25**	**56.99**
			Included Labels: {0,1,2}			
Within/Cross	Posterior	37.06	**37.50**	28.88	30.88	33.58
	Lateral	**37.38**	34.19	35.25	**37.12**	**35.99**
	Combined	36.51	35.69	**35.44**	34.75	**35.60**
LODO	Posterior	33.75	**39.01**	29.01	34.75	34.13
	Lateral	**37.25**	33.25	28.75	33.00	33.06
	Combined	35.75	36.05	**32.25**	**35.75**	**34.94**
MIDA	Posterior	55.75	43.75	45.51	52.25	49.31
	Lateral	45.25	48.51	44.02	46.52	46.07
	Combined	**59.75**	**54.25**	**53.08**	**55.07**	**55.54**

Table D.2: **Between-subject and between-site variability.** Mean±SD macro-F_1 (%), labels 0–3 over the seven encoders.

	Target dataset			
Protocol	BMClab	PD-GaM	T-SDU-PD	3DGait
	LOSO (within-site train and test)			
Mean F_1	55.9±13.6	62.0±5.6	41.7±5.2	27.1±8.2
	Cross-dataset (train on source dataset test on target)			
Mean F_1	27.6±12.3	28.9±11.2	29.0±12.0	28.7±10.8
	MIDA (LOSO: train on target train split + auxiliary datasets, test on target test split)			
Mean F_1	61.5±12.2	65.2±4.4	43.6±4.3	37.2±8.3

training on one cohort and testing on the other three gives $n = 3$ off-site scores; the same formula provides mean ± SD.*(iii) MIDA:* we re-run LOSO after adding external data to the training split, so n and the computation are identical to (i). These statistics quantify, respectively, *between-subject* and *between-dataset* heterogeneity.

Table D.2 reports the resulting mean±SD over all models. Within-site LOSO yields the highest and most stable scores when the cohort itself is large and diverse (PD-GaM 62.0 ± 5.6 pp), but collapses on the small 3DGait set (27.1 ± 8.2 pp). Cross-site transfer is markedly harder: mean macro-F_1 drops by ~25 pp on average, with wider confidence intervals, confirming that domain shift, is a major source of error. Adding auxiliary cohorts during training improves the accuracy in all the datasets. The persistent spread, however, shows that even with extra data the smaller or more idiosyncratic sites (T-SDU-PD, 3DGait) remain challenging, underscoring the importance of both scale and diversity in future clinical gait datasets.

E Ethics and Documentation

CARE-PD includes nine datasets, six of which are existing retrospective datasets that did not require new participant instructions. Ethical approval for use of these retrospective datasets was obtained from the Social Sciences, Humanities & Education Research Ethics Board of the University of Toronto (REB #47891). For the three newly collected datasets ethical approval was provided by the University Health Network Research Ethics Board (CAPCR ID 24-5835). Participants were informed clearly about the data acquisition process and provided informed consent. All data were anonymized to protect participant identity and personal health information. The dataset is distributed under a CC-BY-NC research-only license to prevent misuse and ensure alignment with clinical and ethical standards. Detailed documentation supports transparency and reproducibility, and we expect CARE-PD to drive clinically meaningful, generalizable machine learning research in PD assessment. Full ethical and procedural details can be found in the original publications for each dataset.

F Limitations and Broader Impact

While CARE-PD represents a major step toward clinically grounded gait modeling, several limitations remain.

First, despite its scale and diversity, the dataset remains imbalanced with respect to severe gait impairment (UPDRS-gait score 3), which is both clinically rare and difficult to capture due to mobility constraints. Future work may explore data augmentation or synthetic generation to address this gap. *Second*, while the dataset covers diverse clinical environments and capture modalities, RGB recordings can introduce additional noise that may impact reconstruction quality. Although SMPL fitting and WHAM recovery have shown clinical utility, validated via TOAGA (A.2), monocular errors in depth and distal-joint estimation may still affect downstream tasks. Future releases could extend support from MoCap and RGB to wearable sensor modalities like IMUs to broaden compatibility and enable multimodal learning. *Third*, some datasets use the original UPDRS rubric, while others follow the revised MDS-UPDRS. While the two scales are largely compatible and map onto the same four severity levels, small wording and scoring adjustments, together with per-subject or per-session (rather than per-walk) annotations in several datasets, introduce additional label variability. Moreover, the UPDRS-III gait score was also found to have the highest inter-rater variability among all UPDRS-III scores, with an intraclass correlation coefficient of 0.746‘[29]. *Fourth*, all data are recorded in clinical corridors or labs; outdoor and in-home walking are absent. *Fifth*, our clinical evaluation focuses on gait severity classification; more fine-grained symptom estimation (e.g., stride irregularity, freezing episodes) is left for future work. *Finally*, while CARE-PD provides a strong foundation for representation learning, clinical decision-making often requires temporal context across multiple visits or activities. Most datasets in CARE-PD consist of single-task, short-segment gait walks; however, three of the cohorts (i.e., T-SDU, T-LTC, T-SDU-PD) include logitudinal recordings and could be explored in future work for temporal modeling.

Future releases will target richer labels (e.g. stride-level events, patient-reported outcomes), additional capture modalities, and semi-synthetic augmentation pipelines to balance class 3. As a future direction, we aim to release an identity-preserving, photorealistic video synthesis layer, turning the real videos into paired synthetic clips, so researchers can benchmark the entire video to clinical downstream pipeline end-to-end. Despite these limitations, we believe CARE-PD is a crucial step toward scalable, clinically meaningful motion AI. We encourage future work to build on its protocols and extend the dataset to even richer and more representative clinical populations.

Broader Impact Misuse of CARE-PD is limited due to strict anonymization protocols detailed in Sec. 3.2. Nonetheless, improper training practices represent a potential misuse, particularly training models selectively on subsets biased towards certain demographics. For instance, there is an underrepresentation of women in the severe FoG PD datasets such as BMCLab, KUL-DT-T, and E-LC, each having more than 75% male participants. Given this imbalance, caution should be exercised when extrapolating results. This underrepresentation of women in clinical FoG datasets is, however, a widely recognized phenomenon [30]. More broadly, there is a risk that clinical decision-making could become overly reliant on automated predictions, which may fail to generalize to underrepresented subgroups if not carefully validated.

Despite these potential issues, the contributions of CARE-PD toward advancing AI-driven gait analysis significantly outweigh the risks associated with its misuse, as long as clinical applications developed from CARE-PD undergo thorough and independent validation. CARE-PD has strong potential for positive societal impact: it enables scalable and objective assessments of Parkinsonian gait, encourages reproducibility through public release, and fosters standardization in a fragmented research area. To maximize impact and minimize harm, models developed using CARE-PD should be rigorously validated in diverse clinical contexts.

References

[1] Soyong Shin, Juyong Kim, Eni Halilaj, and Michael J. Black. WHAM: Reconstructing world-grounded humans with accurate 3D motion. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2024. 1

[2] Jim Lawrence, Javier Bernal, and Christoph Witzgall. A purely algebraic justification of the Kabsch-Umeyama algorithm. *Journal of research of the National Institute of Standards and Technology*, 124:1, 2019. 1

[3] Abraham Savitzky and Marcel JE Golay. Smoothing and differentiation of data by simplified least squares procedures. *Analytical chemistry*, 36(8):1627–1639, 1964. 2

[4] Sina Mehdizadeh, Hoda Nabavi, Andrea Sabo, Twinkle Arora, Andrea Iaboni, and Babak Taati. The toronto older adults gait archive: video and 3d inertial motion capture data of older adults' walking. *Scientific data*, 9(1):398, 2022. 3, 7

[5] Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social motion capture. In *Proceedings of the IEEE international conference on computer vision*, pages 3334–3342, 2015. 3

[6] Andrea Avogaro, Federico Cunico, Bodo Rosenhahn, and Francesco Setti. Markerless human pose estimation for biomedical applications: a survey. *Frontiers in Computer Science*, 5:1153160, 2023. 3

[7] José Carrasco-Plaza and Mauricio Cerda. Evaluation of human pose estimation in 3d with monocular camera for clinical application. In *International Symposium on Intelligent Computing Systems*, pages 121–134. Springer, 2022. 3

[8] Kyungdo Kim, Sihan Lyu, Sneha Mantri, and Timothy W Dunn. TULIP: Multi-camera 3d precision assessment of parkinson's disease. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22551–22562, 2024. 3

[9] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, October 2015. 4

[10] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6M: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1325–1339, 2013. 4

[11] Wentao Zhu, Xiaoxuan Ma, Zhaoyang Liu, Libin Liu, Wayne Wu, and Yizhou Wang. MotionBERT: A unified perspective on learning human motion representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15085–15099, 2023. 4, 5

[12] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5152–5161, 2022. 4

[13] Angel Martínez-González, Michael Villamizar, and Jean-Marc Odobez. Pose transformers (POTR): Human motion prediction with non-autoregressive transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2276–2284, 2021. 4

[14] 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 *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 130–139. Springer, 2022. 4

[15] Jinlu Zhang, Zhigang Tu, Jianyu Yang, Yujin Chen, and Junsong Yuan. MixSTE: Seq2seq mixed spatio-temporal encoder for 3d human pose estimation in video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13232–13242, 2022. 4

[16] Qitao Zhao, Ce Zheng, Mengyuan Liu, Pichao Wang, and Chen Chen. PoseFormerV2: Exploring frequency domain for efficient and robust 3d human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8877–8886, 2023. 4

[17] Soroush Mehraban, Vida Adeli, and Babak Taati. MotionAGFormer: Enhancing 3d human pose estimation with a transformer-gcnformer network. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 6920–6930, 2024. 5

[18] Guy Tevet, Brian Gordon, Amir Hertz, Amit H Bermano, and Daniel Cohen-Or. MotionCLIP: Exposing human motion generation to clip space. In *European Conference on Computer Vision*, pages 358–374. Springer, 2022. 5

[19] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5745–5753, 2019. 5

[20] Chuan Guo, Yuxuan Mu, Muhammad Gohar Javed, Sen Wang, and Li Cheng. MoMask: Generative masked modeling of 3d human motions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1900–1910, 2024. 5, 6

[21] Anat Mirelman, Hagar Bernad-Elazari, Avner Thaler, Eytan Giladi-Yacobi, Tanya Gurevich, Mali Gana-Weisz, Rachel Saunders-Pullman, Deborah Raymond, Nancy Doan, Susan B Bressman, et al. Arm swing as a potential new prodromal marker of parkinson's disease. *Movement Disorders*, 31(10):1527–1534, 2016. 7

[22] Fraje Watson, Peter C Fino, Matthew Thornton, Constantinos Heracleous, Rui Loureiro, and Julian JH Leong. Use of the margin of stability to quantify stability in pathologic gait–a qualitative systematic review. *BMC musculoskeletal disorders*, 22:1–29, 2021. 7

[23] Andrea Sabo, Sina Mehdizadeh, Kimberley-Dale Ng, Andrea Iaboni, and Babak Taati. Assessment of parkinsonian gait in older adults with dementia via human pose tracking in video data. *Journal of neuroengineering and rehabilitation*, 17:1–10, 2020. 7

[24] Kimberley-Dale Ng, Sina Mehdizadeh, Andrea Iaboni, Avril Mansfield, Alastair Flint, and Babak Taati. Measuring gait variables using computer vision to assess mobility and fall risk in older adults with dementia. *IEEE journal of translational engineering in health and medicine*, 8:1–9, 2020. 7

[25] Seung Min Kim, Dae Hyun Kim, YoungSoon Yang, Sang Won Ha, and Jeong Ho Han. Gait patterns in parkinson's disease with or without cognitive impairment. *Dementia and neurocognitive disorders*, 17(2):57, 2018. 7

[26] Ana Paula Janner Zanardi, Edson Soares da Silva, Rochelle Rocha Costa, Elren Passos-Monteiro, Ivan Oliveira Dos Santos, Luiz Fernando Martins Kruel, and Leonardo Alexandre Peyré-Tartaruga. Gait parameters of parkinson's disease compared with healthy controls: a systematic review and meta-analysis. *Scientific reports*, 11(1):752, 2021. 7

[27] Ji-yeon Yoon, Sun-shil Shin, Jin-se Park, and Won-gyu Yoo. The effects of stooped posture on gait and postural sway in korean patients with parkinson's disease. *Neurology Asia*, 24(3), 2019. 7

[28] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2623–2631, 2019. 8

[29] T de Deus Fonticoba, D Santos García, and M Macías Arribí. Inter-rater variability in motor function assessment in parkinson's disease between experts in movement disorders and nurses specialising in pd management. *Neurología (English Edition)*, 34(8):520–526, 2019. 11

[30] Anouk Tosserams, Masood Mazaheri, Priya Vart, Bastiaan R Bloem, and Jorik Nonnekes. Sex and freezing of gait in parkinson's disease: a systematic review and meta-analysis. *Journal of Neurology*, 268:125–132, 2021. 11