CARE-PD: A Multi-Site Anonymized Clinical Dataset for Parkinson's Disease Gait Assessment

Vida Adeli[1,2,3†] **Ivan Klabučar**[3] **Javad Rajabi**[1,2] **Benjamin Filtjens**[2,3,4] **Soroush Mehraban**[1,2,3] **Diwei Wang**[5] **Hyewon Seo**[5] **Trung-Hieu Hoang**[7] **Minh N. Do**[7,8] **Candice Muller**[6] **Claudia Neves de Oliveira**[9] **Daniel Boari Coelho**[9] **Pieter Ginis**[10] **Moran Gilat**[10] **Alice Nieuwboer**[10] **Joke Spildooren**[11] **J. Lucas Mckay**[12,13] **Hyeokhyen Kwon**[12,13] **Gari Clifford**[12,13] **Christine Esper**[12] **Stewart A. Factor**[12] **Imari Genias**[12] **Amirhossein Dadashzadeh**[14] **Leia Shum**[3] **Alan Whone**[14] **Majid Mirmehdi**[14] **Andrea Iaboni**[2,3] **Babak Taati**[1,2,3†]

[1]University of Toronto [2]Vector Institute [3]KITE Research Institute-UHN
[4]Delft University of Technology [5]CNRS — University of Strasbourg
[6]University Hospitals of Strasbourg [7]University of Illinois Urbana-Champaign
[8]VinUniversity [9]Federal University of ABC [10]KU Leuven [11]Hasselt University
[12]Emory University [13]Georgia Institute of Technology [14]University of Bristol

Abstract

Objective gait assessment in Parkinson's Disease (PD) is limited by the absence of large, diverse, and clinically annotated motion datasets. We introduce CARE-PD, the largest publicly available archive of 3D mesh gait data for PD, and the first multi-site collection spanning 9 cohorts from 8 clinical centers. All recordings (RGB video or motion capture) are converted into anonymized SMPL meshes via a harmonized preprocessing pipeline. CARE-PD supports two key benchmarks: supervised clinical score prediction (estimating Unified Parkinson's Disease Rating Scale, UPDRS, gait scores) and unsupervised motion pretext tasks (2D-to-3D keypoint lifting and full-body 3D reconstruction). Clinical prediction is evaluated under four generalization protocols: within-dataset, cross-dataset, leave-one-dataset-out, and multi-dataset in-domain adaptation. To assess clinical relevance, we compare state-of-the-art motion encoders with a traditional gait-feature baseline, finding that encoders consistently outperform handcrafted features. Pretraining on CARE-PD reduces MPJPE (from 60.8 mm to 7.5 mm) and boosts PD severity macro-F1 by 17 percentage points, underscoring the value of clinically curated, diverse training data. CARE-PD and all benchmark code are released for non-commercial research at https://neurips2025.care-pd.ca.

1 Introduction

Accurate gait assessment is essential for PD diagnosis, monitoring, and treatment planning; but current clinical evaluations remain subjective and hard to scale. While automated motion analysis offers objective, reproducible metrics, progress is hindered by small, single-site datasets lacking standardization [1]. There is a critical need for large, diverse, and *publicly available* motion datasets with clinical labels to enable generalizable machine learning models for real-world use.

We introduce CARE-PD, a multi-institutional dataset aggregating nine datasets from eight clinical sites, encompassing optical motion-capture and RGB video data. All sequences are converted to a

[†]Corresponding authors: {vadeli, taati}@cs.toronto.edu

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

unified 3D parametric representation [2] using a reproducible preprocessing pipeline that includes data cleaning, temporal segmentation, sensor harmonization, and privacy-preserving SMPL mesh conversion. The anonymization process ensures compliance with institutional ethics and suitability for public research use. CARE-PD is the first clinically annotated motion dataset of this scale focused on parkinsonian gait. Over one-third of the walks include clinician-rated UPDRS-gait scores, referring to the gait-specific item of the Unified Parkinson's Disease Rating Scale [3] motor examination. The remaining data include other clinical attributes—e.g., medication status, Freezing-of-Gait (FoG) presence, PD diagnosis—or are unlabeled for pretraining and self-supervised learning.

Alongside the dataset, we release benchmark protocols for two tasks: (1) clinical severity estimation from gait sequences, and (2) motion pretext tasks, including 2D-to-3D lifting and 3D reconstruction. These tasks evaluate both the clinical utility of pretrained motion encoders and the benefit of incorporating clinically grounded data into self-supervised training. Our experiments show that while state-of-the-art encoders pretrained on generic motion retain *some* latent structure useful for clinical prediction, exposure to pathological gait in CARE-PD substantially improves their reconstruction quality and downstream severity estimation.

CARE-PD is publicly released under a research-only license, following privacy-preserving protocols and institutional approvals. It serves as a valuable resource for developing machine learning methods that bridge general-purpose motion modeling and clinical utility in PD care, and offers a testbed for evaluating representation learning, domain adaptation, and clinical motion understanding tasks.

2 Related Work

Gait Assessment. Recent machine learning approaches for Parkinsonism gait assessment span objectives such as diagnostic classification [4, 5, 6, 7], clinical severity scoring [8, 9, 10, 11, 1, 12, 13, 14], FoG detection [15, 16, 17, 18], and symptom measurement like bradykinesia [19] or tremor [20]. While methods using wearables or video show high accuracy and strong correlations with clinical ratings [21], they are typically limited by small, single-site datasets and perform poorly under real-world variability [22, 23]. Similarly, FoG models [15, 16, 17, 18] and motor symptom assessments [24] are largely validated in lab settings with task-specific protocols, limiting their generalizability [25, 26]. These challenges underscore the need for large, diverse datasets to support robust and generalizable Parkinsonism assessment.

General Motion Benchmarks. Large-scale datasets like Human3.6M [27], AMASS [28], and NTU RGB+D [29] offer extensive motion, RGB, and depth data from healthy individuals performing general activities. Although they include walking, they are not designed for clinical tasks and lack pathological gait patterns, limiting their usefulness for applications such as disease severity estimation or gait abnormality detection.

Gait Datasets. Parkinsonism gait has been studied using a wide range of modalities, including inertial sensors [30, 31, 32, 33, 34], pressure platforms [35, 32, 36], optical motion capture [37, 38, 33], RGB video [39, 40, 41, 34, 24], multimodal physiological sensors [42, 43], and radio-frequency (RF) methods [44]. Each modality has distinct strengths and trade-offs: IMUs offer portability but lack anatomical context [45]; pressure platforms provide spatiotemporal data but are lab-bound and anatomically limited; optical systems offer high biomechanical fidelity but lack scalability [45, 46]; video suffers from occlusions and limited accuracy; physiological sensors require complex setups [42]; and RF methods, while unobtrusive, lack anatomical resolution.To address the scalability challenges of prior work, this study merges RGB video and optical motion capture using the SMPL-based representation, which unifies their differing content and marker conventions within a common anatomical framework.

3 CARE-PD Dataset

CARE-PD is a clinically grounded dataset for automated PD gait assessment, unifying heterogeneous gait recordings from multiple clinical sites into a standardized 3D mesh format using SMPL. As the largest publicly available collection of 3D body meshes focused on PD gait, it supports model development and evaluation for PD severity estimation and related motion analysis tasks.

Table 1: CARE-PD dataset overview. "RGB" = monocular video; "MoCap" = optical-marker motion capture. "FPS" is the original frame or marker rate. "Duration" is the total time (in minutes) retained after gait-segment extraction. "Med" = medication state (on/off), "PD/HC" = Parkinson's vs. healthy control, "FoG" = subject-level freezing of gait (freezing/Non-freezing) label. "±" indicates (mean ± std).

Sub-dataset	Orig. Modality	Orig. FPS	#Subjects	#Walks	Duration (min:sec)	Age (years)	Sex (%male)	Annotation
PD-GaM [47]	RGB	25	30	1701	186:22	54.1 ± 8.1	56.7	UPDRS-gait
BMClab [37]	MoCap	150	23	781	46:57	65.6 ± 8.3	78.3	UPDRS-gait + FoG + Med
T-SDU-PD	RGB	30	14	381	49:39	76.2 ± 8.7	57.1	UPDRS-gait
3DGait [48]	RGB	30	43	90	14:59	78.43 ± 9.3	30.2	UPDRS-gait
KUL-DT-T [49]	MoCap	100	29	763	64:45	65.2 ± 6.8	79.3	FoG
DNE [50]	RGB	60	97	476	21:27	64.1 ± 14.3	48.7	PD/HC
E-LC [51]	MoCap	120	59	162	202:32	68.2 ± 8.3	76.3	FoG + Med
T-SDU	RGB	30	53	2799	341:09	77.1 ± 8.0	57.4	-
T-LTC	RGB	30	14	1324	196:04	81.4 ± 6.7	28.6	-
Total	-	-	362	8477	1123:54	70.0 ± 8.7	56.9	-

3.1 Participating Sites and Cohorts

CARE-PD aggregates gait recordings from 9 studies across 8 clinical centres in 6 countries, all collected under local institutional review board approval with written informed consent. Retrospective analysis of these existing datasets was approved by the Social Sciences, Humanities & Education Research Ethics Board of the University of Toronto (REB #47891). The cohorts differ in population, environment, and capture setup, offering diversity for training and evaluating generalizable models. Table 1 summarizes key metadata, including modality, frame rate, subject count, duration, and clinical annotations. Four datasets include UPDRS-gait scores (0–3), though score 3—reflecting severe impairment and often requiring assistive devices—is rare and appears only in PD-GaM and 3DGait. Further score distribution details are provided in the Appendix A.

The T-SDU and T-LTC datasets (*RGB*) were collected in prospective observational studies on gait changes and fall risk [13, 52] from inpatient participants at a specialized dementia unit and a long-term care facility in Toronto, Canada. These datasets are being publicly released for the first time in this work, following written informed consent and ethical approval from the University Health Network Research Ethics Board (CAPCR ID 24-5835). A 14-subject subset, T-SDU-PD, was selected to capture diverse parkinsonian gait patterns and annotated with per walk UPDRS-gait by expert clinicians. Gait data were recorded via a ceiling-mounted camera triggered by an RFID every time participants walked naturally through a hallway. For inclusion in CARE-PD, recordings were curated to retain only clean walking segments, excluding turns, stops, and other non-walking behaviors.

PD-GaM (*RGB*) [53] is a 3D mesh gait dataset derived from PD4T (University of Bristol, UK) [47], comprising per walk UPDRS-scored trials from four PD motor-function tasks. CARE-PD includes the gait task subset, where each trial captures a participant walking back and forth resulting in four walking segments per recording.

The DNE dataset (*RGB*) [50, 54] was collected across multiple clinical sites in the United States, including OSF HealthCare (Illinois), Bradley Physical Therapy (Washington, Pennsylvania), and Bon Secours St. Francis Inpatient Rehabilitation Center (South Carolina), as part of a neurological assessment study using smartphone video recordings. It includes recordings of participants performing up to five standardized tasks, including fine-motor, facial, and gait assessments. For CARE-PD, we include the stand-up and walk task subset, where participants walked back and forth with labels assigned per walk. Clean walking segments were extracted for inclusion, similar to PD-GaM.

The 3DGait dataset (*RGB*) [48] was collected at the University of Strasbourg (France) to support gait analysis in neurodegenerative diseases. It includes gait videos from individuals with Alzheimer's disease, dementia with Lewy bodies, and healthy controls, all assessed using the per walk MDS-UPDRS-gait criterion. Recordings were made during clinical exams using an RGB camera as patients walked across an 8-meter GAITRite walkway, with alternating front, back, or side views. CARE-PD includes the subset of trials featuring UPDRS-labeled straight walks.

The BMClab dataset (*MoCap*) [37] was collected in the Laboratory of Biomechanics and Motor Control, Federal University of ABC (Brazil), using a Raptor-4 optical motion-capture system (Motion Analysis Corp.) with 44 reflective markers placed following a clinical full-body protocol. It includes walking trials from PD participants with and without FoG, in both ON- and OFF-medication, with MDS-UPDRS-gait labels per participant in each medication state by expert clinicians.

The KUL-DT-T dataset (*MoCap*) [49, 18] was recorded in the Movement Disorders Clinic of the University Hospital Leuven (Belgium), using a 8-camera 3D optical motion capture system with 34 markers (Vicon Motion Systems), under dual-task and turning conditions designed to provoke FoG. It includes participants with and without FoG.

The E-LC dataset (*MoCap*) [51, 55] was acquired at the Emory Movement Disorders Clinic (Atlanta, USA) using a 14-camera 3D optical motion capture system (Motion Analysis Corp.) with 60 reflective markers, capturing high-resolution data from PD participants across medication states (ON/OFF) per walk and FoG subtypes per participant under standardized protocols.

3.2 Data Harmonization

CARE-PD harmonizes heterogeneous gait data using a preprocessing pipeline that converts all recordings into 30 Hz 3D SMPL mesh sequences, incorporating modality-specific processing, artifact correction, anonymization, and subject-level stratified evaluation splits.

MoCap Processing. MoCap data underwent: 1) Quality control to fix joint errors from dropped or noisy markers; 2) Joint standardization by mapping each system's layout to 22 SMPL joints; 3) SMPL fitting using SparseFusion optimization [56] to estimate full-body parameters from sparse 3D joints; and 4) Downsampling to 30 Hz and removal of corrupted or very short segments to ensure consistent, anatomically valid outputs.

Video Processing. For RGB datasets: 1) SMPL meshes were extracted using WHAM [57], a monocular mesh recovery method (see Appendix A for clinical validity experiments); 2) The extracted mesh was visually verified to correspond to the intended patient when multiple people appeared in a frame; 3) Only clean walking segments were retained by removing non-walking behaviors; and 4) A slope correction using the Kabsch algorithm [58] was applied to counteract distortions from ceiling-mounted cameras, aligning recovered 3D walks to a canonical ground plane and preserving gait dynamics while correcting the camera-induced skew (see Appendix A).

Anonymization. To ensure privacy, only textureless SMPL mesh parameters are released—excluding video frames, identifiable 3D point clouds, and potentially identity-revealing SMPL shape parameters. For longitudinal datasets (T-SDU, L-SDU, and T-SDU-PD), start dates are standardized to anonymize timelines, and all subject IDs and filenames are anonymized.

The resulting dataset contains 18.66 hours of anonymized 3D gait-mesh sequences, totaling 8,477 walking segments.We resample every recording to a common $\sim$30 FPS and provide subject-level train/val/test splits stratified by key UPDRS-gait scores, with multiple configurations depending on the evaluation protocol. Full label distributions, per-split statistics, and additional preprocessing details are included in Appendix A.

4 Benchmarks & Experiments

4.1 Clinical Score Estimation Task

A walk is represented as a sequence of body pose frames $\mathbf{M}^{1:T} = \{\mathbf{m}^1, \mathbf{m}^2, \ldots, \mathbf{m}^T\}$, where each frame $\mathbf{m}^t \in \mathbb{R}^F$ encodes the pose at time t using F parameters. Given a walk $\mathbf{M}^{1:T}$, the objective is to estimate its associated UPDRS-gait severity score $S \in \{0, 1, 2, 3\}$ reflecting the degree of gait impairment. We evaluate two baseline approaches for this task: *Representation-learning*, using deep encoders trained on motion sequences, and *Engineered gait features*, using traditional handcrafted features with classical classifiers. These baselines help assess the trade-offs between learned and interpretable features, and analyze their sensitivity to clinical labels across different data sources.

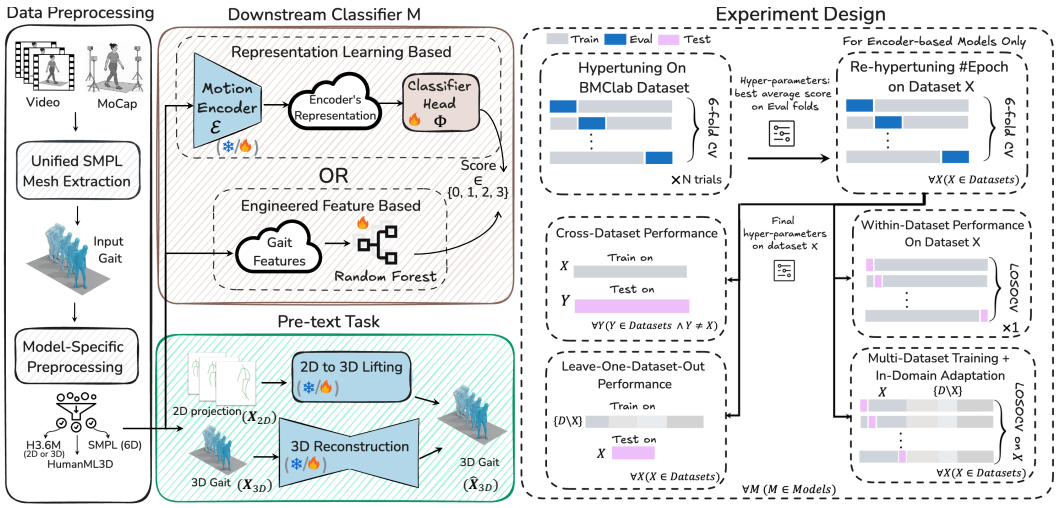

Figure 1: Overview of CARE-PD preprocessing and experimental design. Left: Unified pipeline for extracting SMPL gait meshes from MoCap and video data, followed by model-specific formatting. Right: Benchmarking setup across two pipelines (representation learning vs. gait features), pretext tasks, and four evaluation protocols.

Representation-Learning Baselines. To predict a person's UPDRS-gait score S from their walk $\mathbf{M}^{1:T}$ using a motion encoder $\mathcal{E}$, we first apply a series of preprocessing steps $P_{\mathcal{E}}$ to convert variable-length input into several non-overlapping motion clips of fixed length N, i.e., $\{\mathbf{p}_1^{1:N}, \ldots, \mathbf{p}_{\frac{T}{N}}^{1:N}\} = P_{\mathcal{E}}(\mathbf{M}^{1:T})$. The preprocessing steps are encoder-specific and include operations such as windowing, normalization, and format conversion, e.g., from SMPL to different joint coordinate format. Full details of these steps for each encoder are provided in Appendix A. Each motion clip $\mathbf{p}_i$ is fed into the encoder $\mathcal{E}$ to obtain a latent feature vector $\mathbf{e}_i$, which is then passed to a lightweight classifier head Φ to produce a predicted score $s_i = \Phi(\mathcal{E}(\mathbf{p}_i))$. The final predicted score S for the entire sequence $\mathbf{M}^{1:T}$ is obtained via majority voting across all predictions $\{s_i\}$. This approach enables score prediction even on long recordings by aggregating local clip-level predictions.

We deliberately keep each motion encoder trained on its original task frozen and evaluate it using two lightweight probes (a linear classifier and a k-nearest neighbors (k-NN) classifier) to examine whether state-of-the-art motion encoders, pre-trained on generic human motion data, already capture clinically relevant gait features. This setup allows us to investigate three key aspects: 1) Clinical usefulness: We compare pretrained encoders with traditional gait features to determine if their representations, when paired with simple probes, can match or outperform traditional approaches, quantifying their clinical utility. 2) Generalizability and out-of-distribution robustness: Subjects in Care-PD exhibit characteristics not typically seen in the encoders' pretraining on healthy human motion datasets (e.g., parkinsonian shuffling, tremors, or subtle asymmetries). We investigate whether these clinical patterns are effectively represented in the latent space or instead disregarded as noise by the pretrained encoders. 3) Real-world deployability: Many clinical settings cannot support full model fine-tuning. Frozen probes simulate "plug-and-play" usage, where public models are applied directly to clinical tasks with minimal adaptation.

We evaluate seven state-of-the-art validated motion encoders: POTR [59], MixSTE [60], PoseFormerV2 [61], MotionBERT [62], MotionAGformer [63], MotionCLIP [64], and MoMask [65], spanning a diverse range of human motion tasks including 2D-to-3D lifting, 3D pose reconstruction, action recognition, and motion generation. Further details on the encoders, encoder-specific preprocessings and motion representations used are provided in Appendix A.

Engineered-Feature Baseline. As a classical baseline, we train a Random Forest classifier on a set of interpretable gait features derived from the reconstructed body joints. These features include spatiotemporal descriptors (e.g., cadence, step length, step width, step time, walking speed) [66, 13], stability-related measures (e.g., estimated margin of stability) [67], and motor/postural indicators (e.g., foot lifting, arm swing, stoop posture) [68, 69]. Many of these directly correspond to criteria used in the UPDRS-gait scoring rubric [3]. Random Forest models built on such handcrafted variables

have shown to outperform several other machine-learning algorithms in differentiating PD from control gait [70]. We also tested kernel SVM and XGBoost, but Random Forest consistently yielded comparable or better results. For simplicity, we report only the Random Forest baseline. Full details on feature extraction are provided in Appendix B.

4.2 Evaluation Protocols

To understand how well different representations support clinical score estimation, we design four evaluation protocols that reflect realistic deployment scenarios (See Fig. 1).

Within and Cross Dataset Evaluation. Each dataset in CARE-PD is unique in terms of recording setup, capture geometry, and population demographics and movement instructions. To evaluate model performance under these conditions, we assess both within-dataset and cross-dataset generalization. *Within-dataset evaluation* uses a Leave-One-Subject-Out (LOSO) cross-validation strategy to ensure no subject-specific motion patterns leak between splits. *Cross-dataset evaluation* tests how well models trained on *one* dataset transfer to others. We train a classifier on a single dataset and evaluate it on the remaining ones. This simulates deployment in a new clinical environment without access to site-specific labeled data and highlights the model's robustness to changes in protocol, hardware, and patient characteristics. Together, these two protocols reveal the extent to which learned representations are sensitive to dataset-specific biases and whether they generalize across both subjects and sites.

Leave One Dataset Out (LODO). Clinical sets are small and often carry cohort-specific biases, which can cause classifiers to overfit to a single dataset. To test whether diversity can dilute these biases, we train on the union of $D-1$ cohorts and evaluate on the held-out cohort. A gain over the single-cohort cross-test of the previous section indicates that combining heterogeneous data lets the probe focus on pathology-related variation rather than spurious site cues. Conversely, a persistent gap reveals that the held-out cohort contains systematic differences that even large, diverse training data cannot bridge, highlighting where additional harmonization or domain adaptation is warranted.

Multi-dataset In-domain Adaptation (MIDA). While LODO tests generalization to unseen domains, many real-world deployments allow limited access to target-domain data. MIDA explores whether using a small amount of in-domain data can significantly boost performance. Starting from the LODO checkpoint, we fine-tune the probe (but keep the encoder frozen) on the target cohort's training split—again under LOSO—and test on its held-out subjects. Comparing MIDA to LODO quantifies how much performance can be recovered by a modest amount of in-domain supervision.

4.3 Training & Metrics

We tuned classifier head hyperparameters using 6-fold stratified cross-validation on the BMClab dataset, selected for its size and clean motion quality. The best combination was applied across all datasets, adjusting only the number of training epochs per dataset. Full tuning details, search space, and evaluation strategy are in Appendix C. UPDRS-gait score 3 (indicating severe impairment) is rare and is present in only two datasets (PD-GaM and 3DGait). In cross-dataset evaluations, it may be entirely absent from the training set but present in the test set. In such cases, the classifier is unable to predict label 3, resulting in an F1 score of 0 for that class. Therefore, including it in macro averaging artificially deflates the overall metric and make comparisons unfair. To address this, we report macro F1 scores in two setups: including all UPDRS-gait labels 0, 1, 2, 3 ($F1_{0-3}$), and excluding label 3 ($F1_{0-2}$). Importantly, all score-3 samples remain in the training folds; the exclusion applies only to the metric. We adopt macro (unweighted) averaging so that each class contributes equally despite imbalanced class distributions, and we choose F1 to balance precision and recall without favoring either. We report results using the linear probe, as it showed no substantial difference from the k-NN.

4.4 Motion Pre-text Tasks

To evaluate the utility of CARE-PD in improving motion representation learning and 3D pose estimation, we conducted experiments on two common pretext tasks: 2D-to-3D lifting and 3D reconstruction. We used two top-performing models: MotionAGFormer [63] for 2D lifting and MoMask [65] for motion reconstruction and generation. Both models were pre-trained on generic, able-bodied motion datasets. We then fine-tuned them or trained them from scratch on CARE-PD to

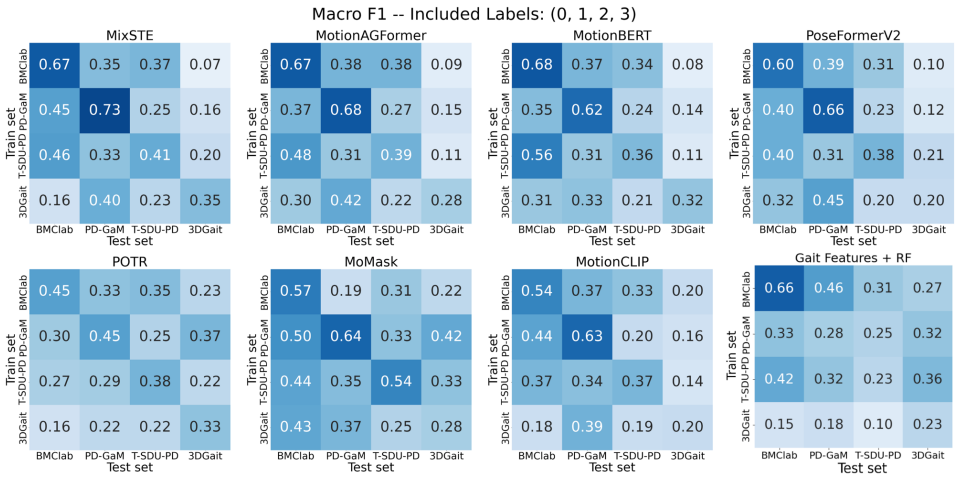

Figure 2: Within-dataset and cross-dataset macro-$F1_{0-3}$ scores for encoder and gait features-based models.

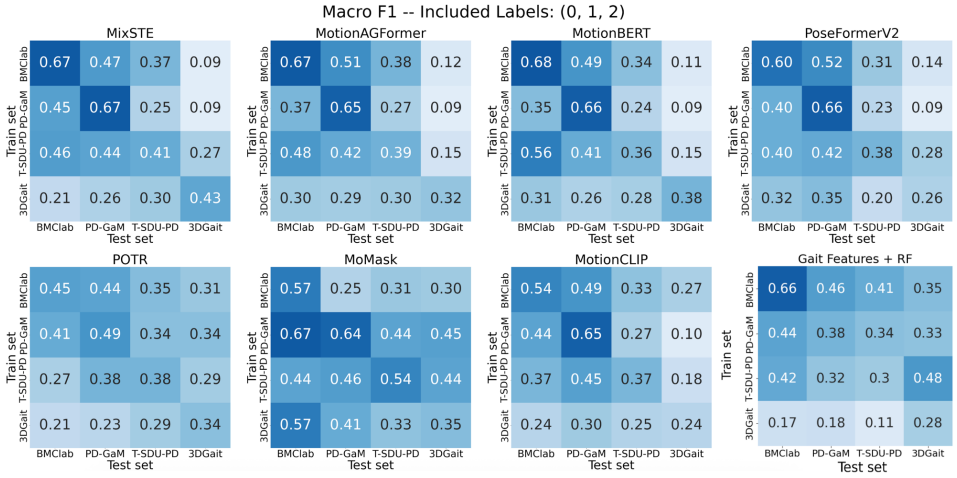

Figure 3: Cross-dataset and within-dataset macro-$F1_{0-2}$ scores using encoder-based models.

assess: (1) Clinical adaptation impact on proxy accuracy: whether exposure to ~19 hours of diverse pathological gait improves 3D estimation error measured via mean per joint position error (MPJPE), Procrustes-aligned MPJPE (PA-MPJPE), and acceleration error (Acc) common for the task [71]. (2) Downstream impact: whether such improvements translate to better downstream UPDRS-score prediction, when coupled with the same lightweight probe used in the clinical task. This experiment tells us whether injecting clinically rich motion into SOTA encoders (i) improves 3D pose accuracy on pathological gait, and (ii) yields clinical-task benefits *without* extra model capacity.

5 Results & Analysis

5.1 Severity Estimation Benchmarks

Within and Cross Dataset. We evaluate model performance using within-dataset (LOSOCV; diagonal) and cross-dataset (off-diagonal) protocols, reporting $F1_{0-3}$ and $F1_{0-2}$ in Fig. 2 and Fig. 3, respectively. Most encoders achieve strong in-site performance on large cohorts—up to 0.73 on PD-GaM and 0.68 on BMClab—demonstrating that frozen representations retain some clinically relevant information. Performance drops for the smaller T-SDU-PD set and falls further for 3DGait. Transferring to unseen datasets typically reduces F1 by 0.2 to 0.4, exposing domain gaps driven by data distribution shifts. Models trained on PD-GaM generalize best, likely due to its scale and diversity. Among all the backbones, the VQ-VAE MoMask prove the most robust: when it is trained

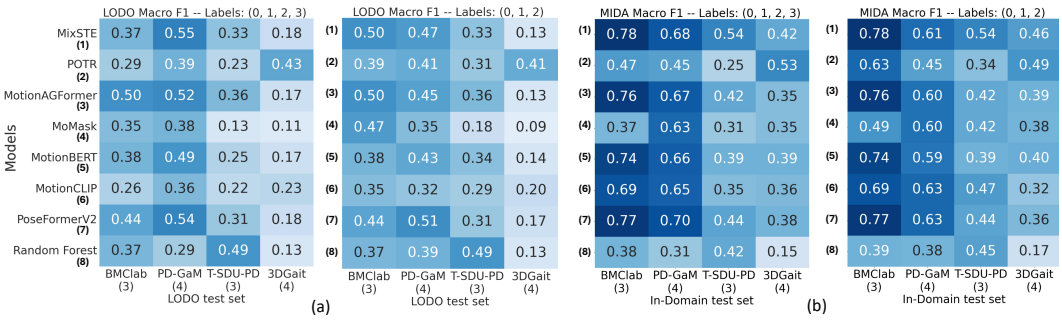

Figure 4: Macro-F1$_{0-3}$ and Macro-F1$_{0-2}$ scores for LODO (left two blocks) and MIDA (right two blocks) evaluations, comparing severity estimation across models and datasets.

Table 2: Impact of CARE-PD on motion pretext tasks and downstream severity prediction.

Model	Task	Train Data	Finetune Data	MPJPE ↓(mm)	PA-MPJPE ↓(mm)	Acc ↓(mm/s^2)	F1-score ↑
MotionAGFormer [63]	2D-3D lifting	H3.6M	-	60.7	21.4	99.8	48.1
		H3.6M	Healty Gait	29.8	7.3	35.4	50.1
		H3.6M	**CARE-PD**	**7.5**	**2.6**	**11.6**	**65.1**
		CARE-PD	-	9.0	3.2	13.8	62.3
MoMask [65]	3D reconst.	HumanML3D	-	22.5	17.8	4.3	41.4
		HumanML3D	Healty Gait	22.3	13.7	4.4	40.6
		HumanML3D	**CARE-PD**	**8.7**	**6.3**	**2.2**	**62.7**
		CARE-PD	-	9.6	7.3	3.3	59.8

on PD-GaM, its average cross-site F1 remains above 0.40, in several cases matching or surpassing the within-site result of weaker encoders (see Appendix D). Removing class 3 from evaluation consistently boosts metrics, especially on small datasets, confirming its sparsity and ambiguity.

Compared to these, the handcrafted-feature baseline underperforms on most cross-dataset settings, highlighting the superior portability of learned motion representations. Overall, the results underscore the value of multi-site evaluation and reveal the need for models that generalize beyond their training domain. The results validate our multi-site evaluation protocol: single-cohort scores over-estimate readiness for deployment, while cross-site tests reveal both the promise of modern encoders and the persistent domain gaps that future work must bridge.

LODO Analysis. In the LODO protocol (Fig. 4-a) we observe a clear degradation in performance compared to within-dataset training, confirming that domain shift remains a major challenge for generalization. The two most diverse target sets (BMClab and PD-GaM) are now handled best: MixSTE and MotionAGFormer reach macro-F1 ≈ 0.50 on both label configurations, with PoseFormerV2 close behind. By contrast, the small 3DGait cohort stays difficult for every deep backbone (≤ 0.18). Across backbones, MotionBERT is the most sensitive to the choice of target site. It peaks at 0.49 on PD-GaM but slips to 0.25 on T-SDU-PD, whereas MotionAGFormer and PoseFormerV2 yield the most consistent scores. Including the rare class 3 generally lowers every entry by 2–5 percentage points, yet leaves the relative ordering intact. The low performance in 3DGait is likely due to its limited size—only 90 videos from 43 participants—causing the LOSO setup to leave just one or two test samples per fold, making evaluation unstable. The Random Forest baseline remains surprisingly competitive on T-SDU-PD, especially in the 3-class setup (0.49), but its performance deteriorates in PD-GaM and 3DGait, indicating poor robustness outside the training domain.

MIDA analysis. The MIDA protocol yields higher F1 scores across the board (Fig. 4-b), confirming that augmenting training with a mix of diverse source domains and further tuning on the target significantly boosts performance and adaptation across nearly all settings. On BMClab, all backbones now exceeds 0.69, with MixSTE, MotionAGFormer, MotionBERT, and PoseFormerV2 reaching 0.74 to 0.78, halving the LODO error. PD-GaM shows similar gains, with the same models achieving 0.63–0.70 despite the inclusion of the more ambiguous class 3. Smaller datasets like T-SDU-PD also benefit: MotionAGFormer, MixSTE, and PoseFormerV2 improve by ~0.25 absolute F1 over their LODO scores. Comparing MIDA (Fig. 4-b) to within-dataset LOSO (diagonal elements of Fig. 2 and Fig. 3) highlights the value of the CARE-PD dataset in boosting performance. For instance, MixSTE's macro-F1 on BMCLab improves from 0.67 to 0.78 (see Appendix D).

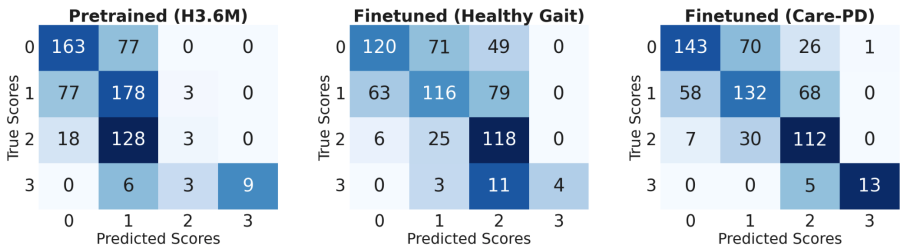

Figure 5: MotionAGFormer UPDRS-gait confusion matrices under different pretext training.

5.2 Motion Pre-text Results

For each of the two pretext tasks, we compare four training regimes: 1) Zero-shot: pre-trained on generic motion datasets (H3.6M for MotionAGFormer, HumanML3D for MoMask) and evaluated directly on CARE-PD. 2) Fine-tuning on CARE-PD. 3) Healthy-gait fine-tuning: fine-tuned instead on 7,971 healthy gait clips from [72, 73, 74, 75] datasets to isolate the impact of merely focusing on walking. 4) Training from scratch on CARE-PD without any external pre-training.

Testing is performed exclusively on CARE-PD. For each of the 9 CARE-PD datasets, we performed a separate 80/20 subject-stratified split (identical across all four training regimes) and report aggregate test error over all test splits. This pooled setting is considerably more challenging than the dataset-specific benchmarks used in previous sections, making the improvements especially meaningful.

Results in Tab. 2 show that fine-tuning on CARE-PD significantly boosts both reconstruction metrics across *all* metrics and downstream clinical score prediction. MPJPE dropped from 60.7 to 7.5 for MotionAGFormer and from 22.5 to 8.7 for MoMask, while the UPDRS-gait F1-score improved from 48.1 to 65.1 and from 41.4 to 62.7, respectively.

Notably, fine-tuning on healthy walks delivers far smaller gains, confirming that the improvement stems from exposure to pathological kinematics rather than from seeing more walking alone. Training from scratch on CARE-PD achieves comparable results, suggesting that the dataset is sufficiently rich to support effective learning of subtle biomechanical distinctions in pathological gait, yet external pre-training on diverse motions still leads to lower 3D reconstruction error.

Fig. 5 shows that the zero-shot probe with the model trained only on H3.6M, consistently over-predicts class 0 and fails on more severe classes. Fine-tuning on healthy gait helps separate classes 1 and 2 but still misses class 3. Only after fine-tuning on CARE-PD does the model begin to accurately identify all four classes, including class 3, reflecting better sensitivity to clinical severity.

Domain Gap and Transferability. While pretraining on generic motion datasets improves low-level reconstruction accuracy, transfer to clinical gait tasks remains limited due to a strong domain gap. These datasets feature young, healthy subjects performing broad actions with either action or text labels, whereas CARE-PD subjects are older adults with parkinsonian gait characterized by shorter steps, reduced arm swing, and temporal irregularity. As a result, pretrained representations capture general motion primitives but fail to encode subtle pathological cues essential for UPDRS scoring. Fine-tuning on CARE-PD improves severity-prediction, confirming that exposure to clinically relevant motion is necessary for effective adaptation. CARE-PD therefore provides a unique testbed for clinical domain adaptation, robust representation learning and motion foundation models that integrate both biomechanical and clinical priors.

5.3 Subgroup Sensitivity

We evaluated whether predicted UPDRS-gait scores from finetuned MotionAGFormer reflected clinically meaningful subgroup differences. Medication analysis used BMClab and E-LC, FoG comparison used BMClab, KUL-DT-T and E-LC, and PD vs. healthy control analysis used DNE.

Medication. Predicted UPDRS-gait scores were significantly lower when participants were on medication (median = 1.0, IQR = 1.0) compared to off-medication (median = 2.0, IQR = 2.0). This difference was statistically significant (Mann–Whitney U test, $p \leq 10^{-5}$), with a medium effect size

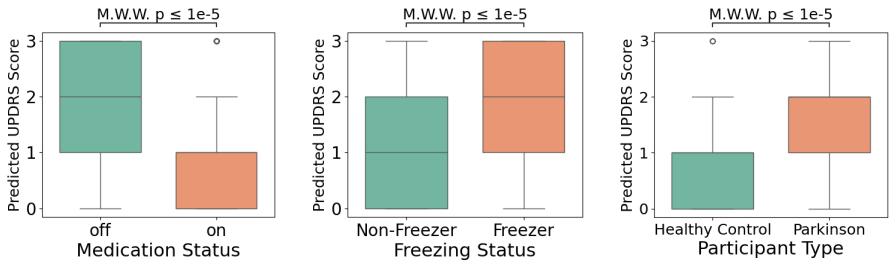

Figure 6: Predicted UPDRS-gait scores across three group comparisons: medication status (left), FoG status (middle), and diagnosis (right). M.W.W. refers to the Mann–Whitney–Wilcoxon test.

(Cliff's $\Delta = 0.42$) indicating that a random individual from group off-med has a 71% chance of having a higher predicted UPDRS-gait score than a random individual from group on-med. These findings demonstrate the model's sensitivity to medication status in assessing parkinsonian gait.

Freezing Status. Predicted UPDRS-gait scores were significantly higher among freezers (median = 2.0, IQR = 2.0) compared to non-freezers (median = 1.0, IQR = 2.0). This difference was statistically significant (Mann–Whitney U test, $p \leq 10^{-5}$), with a moderate effect size (Cliff's $\Delta = 0.25$), meaning a 62% likelihood that a freezer has a higher predicted UPDRS-gait score than a non-freezer. This indicates the model is sensitive to capture increased gait impairment associated with FoG.

Participant Type. Predicted UPDRS-gait scores were significantly higher among participants with PD (median = 2.0, IQR = 1.0) compared to healthy controls (median = 1.0, IQR = 1.0). This difference was statistically significant (Mann–Whitney U test, $p \leq 10^{-5}$), with a large effect size (Cliff's $\Delta = 0.50$), indicating that a randomly selected individual with PD has a 75% chance of receiving a higher predicted UPDRS-gait score than a healthy control. These results support the model's ability to distinguish between pathological and non-pathological gait patterns.

Figure 6 summarizes these results, showing that the model consistently captures clinically meaningful group differences in predicted UPDRS-gait scores across all three sensitivity analyses.

6 Conclusions

We introduced CARE-PD, a large multi-cohort 3D gait dataset for PD, enabling robust machine learning research through benchmark tasks in clinical score estimation and motion pretext learning. Results from seven backbone models and a handcrafted features baseline reveal four key lessons: First, pretrained encoders paired with classifier probe can capture some clinically relevant signals, but their accuracy collapses under distribution shifts—cross-dataset generalization is markedly harder, underscoring the need for multi-site evaluation. Second, dataset quality and size are as important as model architecture. Models trained or fine-tuned on larger cohorts generalize best, while those trained on smaller datasets underperform. Encoder choice also affects robustness to site variation. Third, classical gait features are competitive within domains but less generalizable than learned representations. Fourth, using CARE-PD for pretext tasks improves both 3D estimation and downstream clinical prediction, confirming its value for both supervised and self-supervised clinical learning. These findings motivate future work on clinical model development, domain-aware training, and test-time adaptation to capture pathological subtleties without overfitting site-specific biases. CARE-PD's scale and heterogeneity also give generative methods such as [53] a strong foundation for generating more diverse, clinically grounded motion samples, helping address the scarcity of severe cases. By providing data and evaluation protocols, CARE-PD aims to accelerate the translation of motion AI into objective, scalable support for PD care.

Acknowledgments and Disclosure of Funding

This work was supported by the Data Sciences Institute at the University of Toronto, the Walter and Maria Schroeder Institute for Brain Innovation and Recovery and the AGE-WELL Network of Centres of Excellence (AGE-WELL NCE). The authors sincerely thank these institutions for their support.

References

[1] Vida Adeli, Soroush Mehraban, Irene Ballester, Yasamin Zarghami, Andrea Sabo, Andrea Iaboni, and Babak Taati. Benchmarking skeleton-based motion encoder models for clinical applications: Estimating parkinson's disease severity in walking sequences. In *2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–10. IEEE, 2024. 1, 2

[2] 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. 2, 27

[3] Christopher G Goetz, Barbara C Tilley, Stephanie R Shaftman, Glenn T Stebbins, Stanley Fahn, Pablo Martinez-Martin, Werner Poewe, Cristina Sampaio, Matthew B Stern, Richard Dodel, et al. Movement disorder society-sponsored revision of the unified parkinson's disease rating scale (MDS-UPDRS): scale presentation and clinimetric testing results. *Movement disorders: official journal of the Movement Disorder Society*, 23(15):2129–2170, 2008. 2, 5

[4] Milla Juutinen, Cassia Wang, Justin Zhu, Juan Haladjian, Jari Ruokolainen, Juha Puustinen, and Antti Vehkaoja. Parkinson's disease detection from 20-step walking tests using inertial sensors of a smartphone: Machine learning approach based on an observational case-control study. *PLoS One*, 15(7):e0236258, 2020. 2

[5] Rana Zia Ur Rehman, Silvia Del Din, Yu Guan, Alison J Yarnall, Jian Qing Shi, and Lynn Rochester. Selecting clinically relevant gait characteristics for classification of early parkinson's disease: a comprehensive machine learning approach. *Scientific reports*, 9(1):17269, 2019. 2

[6] Lan Ma, Hua Huo, Wei Liu, Changwei Zhao, Jinxuan Wang, and Ningya Xu. Twin-tower transformer network for skeleton-based parkinson's disease early detection. *Complex & Intelligent Systems*, 10(5):6745–6765, 2024. 2

[7] Hua Huo, Chen Zhang, Wei Liu, Changwei Zhao, Lan Ma, Jinxuan Wang, and Ningya Xu. Early detection of parkinson's disease using a multi area graph convolutional network. *Scientific Reports*, 15(1):5561, 2025. 2

[8] Yi Han, Xiangzhi Liu, Ning Zhang, Xiufeng Zhang, Bin Zhang, Shuoyu Wang, Tao Liu, and Jingang Yi. Automatic assessments of parkinsonian gait with wearable sensors for human assistive systems. *Sensors*, 23(4):2104, 2023. 2

[9] Katsuki Eguchi, Ichigaku Takigawa, Shinichi Shirai, Ikuko Takahashi-Iwata, Masaaki Matsushima, Takahiro Kano, Hiroaki Yaguchi, and Ichiro Yabe. Gait video-based prediction of unified parkinson's disease rating scale score: a retrospective study. *BMC neurology*, 23(1):358, 2023. 2

[10] Navita, Pooja Mittal, Yogesh Kumar Sharma, Anjani Kumar Rai, Sarita Simaiya, Umesh Kumar Lilhore, and Vimal Kumar. Gait-based parkinson's disease diagnosis and severity classification using force sensors and machine learning. *Scientific Reports*, 15(1):328, 2025. 2

[11] Haoyu Tian, Haiyun Li, Wenjing Jiang, Xin Ma, Xiang Li, Hanbo Wu, and Yibin Li. Cross-spatiotemporal graph convolution networks for skeleton-based parkinsonian gait mds-updrs score estimation. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 32:412–421, 2024. 2

[12] 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. 2, 27

[13] 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. 2, 3, 5, 30

[14] Andrea Sabo, Sina Mehdizadeh, Andrea Iaboni, and Babak Taati. Estimating parkinsonism severity in natural gait videos of older adults with dementia. *IEEE journal of biomedical and health informatics*, 26(5):2288–2298, 2022. 2

[15] Luis Sigcha, Nélson Costa, Ignacio Pavón, Susana Costa, Pedro Arezes, Juan Manuel López, and Guillermo De Arcas. Deep learning approaches for detecting freezing of gait in parkinson's disease patients through on-body acceleration sensors. *Sensors*, 20(7):1895, 2020. 2

[16] Sara Abbasi and Khosro Rezaee. Deep learning–based prediction of freezing of gait in parkinson's disease with the ensemble channel selection approach. *Brain and Behavior*, 15(1):e70206, 2025. 2

[17] Thomas Bikias, Dimitrios Iakovakis, Stelios Hadjidimitriou, Vasileios Charisis, and Leontios J Hadjileontiadis. DeepFoG: an imu-based detection of freezing of gait episodes in parkinson's disease patients via deep learning. *Frontiers in Robotics and AI*, 8:537384, 2021. 2

[18] Benjamin Filtjens, Pieter Ginis, Alice Nieuwboer, Peter Slaets, and Bart Vanrumste. Automated freezing of gait assessment with marker-based motion capture and multi-stage spatial-temporal graph convolutional neural networks. *Journal of NeuroEngineering and Rehabilitation*, 19(1):48, 2022. 2, 4

[19] Albert Samà, Carlos Pérez-López, D Rodríguez-Martín, Andreu Català, Juan Manuel Moreno-Aróstegui, Joan Cabestany, Eva de Mingo, and Alejandro Rodríguez-Molinero. Estimating bradykinesia severity in parkinson's disease by analysing gait through a waist-worn sensor. *Computers in biology and medicine*, 84:114–123, 2017. 2

[20] Felipe Duque-Quiceno, Grzegorz Sarapata, Yuriy Dushin, Miles Allen, and Jonathan O'Keeffe. Deep learning for objective estimation of parkinsonian tremor severity. *arXiv preprint arXiv:2409.02011*, 2024. 2

[21] Kathrin Heye, Renjie Li, Quan Bai, Rebecca J St George, Kaylee Rudd, Guan Huang, Marjan J Meinders, Bastiaan R Bloem, and Jane E Alty. Validation of computer vision technology for analyzing bradykinesia in outpatient clinic videos of people with parkinson's disease. *Journal of the Neurological Sciences*, 466:123271, 2024. 2

[22] Lazzaro Di Biase, Alessandro Di Santo, Maria Letizia Caminiti, Alfredo De Liso, Syed Ahmar Shah, Lorenzo Ricci, and Vincenzo Di Lazzaro. Gait analysis in parkinson's disease: An overview of the most accurate markers for diagnosis and symptoms monitoring. *Sensors*, 20(12):3529, 2020. 2

[23] Andrea Sabo, Andrea Iaboni, Babak Taati, Alfonso Fasano, and Carolina Gorodetsky. Evaluating the ability of a predictive vision-based machine learning model to measure changes in gait in response to medication and dbs within individuals with parkinson's disease. *BioMedical Engineering OnLine*, 22(1):120, 2023. 2

[24] Daniel Deng, Jill L Ostrem, Vy Nguyen, Daniel D Cummins, Julia Sun, Anupam Pathak, Simon Little, and Reza Abbasi-Asl. Interpretable video-based tracking and quantification of parkinsonism clinical motor states. *npj Parkinson's Disease*, 10(1):122, 2024. 2

[25] Martina Mancini, Vrutangkumar V Shah, Samuel Stuart, Carolin Curtze, Fay B Horak, Delaram Safarpour, and John G Nutt. Measuring freezing of gait during daily-life: an open-source, wearable sensors approach. *Journal of neuroengineering and rehabilitation*, 18:1–13, 2021. 2

[26] Martina Mancini, J Lucas McKay, Helena Cockx, Nicholas D'Cruz, Christine D Esper, Benjamin Filtjens, Benedetta Heimler, Colum D MacKinnon, Luca Palmerini, Melvyn Roerdink, et al. Technology for measuring freezing of gait: Current state of the art and recommendations. *Journal of Parkinson's Disease*, page 1877718X241301065, 2025. 2

[27] 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. 2, 27

[28] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5442–5451, 2019. 2

[29] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. NTU RGB+D: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1010–1019, 2016. 2

[30] Daphnet freezing of gait. UCI Machine Learning Repository, https://archive.ics.uci.edu/dataset/245/daphnet+freezing+of+gait. Accessed: 2025-04-26. 2

[31] Brian M Bot, Christine Suver, Elias Chaibub Neto, Michael Kellen, Arno Klein, Christopher Bare, Megan Doerr, Abhishek Pratap, John Wilbanks, E Dorsey, et al. The mPower study, parkinson disease mobile data collected using researchkit. *Scientific data*, 3(1):1–9, 2016. 2

[32] Anthony J Anderson, David Eguren, Michael A Gonzalez, Naima Khan, Sophia Watkinson, Michael Caiola, Siegfried S Hirczy, Cyrus P Zabetian, Kelly Mills, Emile Moukheiber, et al. WearGait-PD: An open-access wearables dataset for gait in parkinson's disease and age-matched controls. *medRxiv*, pages 2024–09, 2024. 2

[33] Catherine Morgan, Emma L Tonkin, Alessandro Masullo, Ferdian Jovan, Arindam Sikdar, Pushpajit Khaire, Majid Mirmehdi, Ryan McConville, Gregory JL Tourte, Alan Whone, et al. A multimodal dataset of real world mobility activities in parkinson's disease. *Scientific data*, 10(1):918, 2023. 2

[34] Caroline Ribeiro De Souza, Runfeng Miao, Júlia Ávila De Oliveira, Andrea Cristina De Lima-Pardini, Débora Fragoso De Campos, Carla Silva-Batista, Luis Teixeira, Solaiman Shokur, Bouri Mohamed, and Daniel Boari Coelho. A public data set of videos, inertial measurement unit, and clinical scales of freezing of gait in individuals with parkinson's disease during a turning-in-place task. *Frontiers in Neuroscience*, 16:832463, 2022. 2

[35] Chariklia Chatzaki, Vasileios Skaramagkas, Nikolaos Tachos, Georgios Christodoulakis, Evangelia Maniadi, Zinovia Kefalopoulou, Dimitrios I Fotiadis, and Manolis Tsiknakis. The smart-insole dataset: Gait analysis using wearable sensors with a focus on elderly and parkinson's patients. *Sensors*, 21(8):2821, 2021. 2

[36] Jeffrey M Hausdorff, Apinya Lertratanakul, Merit E Cudkowicz, Amie L Peterson, David Kaliton, and Ary L Goldberger. Dynamic markers of altered gait rhythm in amyotrophic lateral sclerosis. *Journal of applied physiology*, 2000. 2

[37] Thiago Kenzo Fujioka Shida, Thaisy Moraes Costa, Claudia Eunice Neves de Oliveira, Renata de Castro Treza, Sandy Mikie Hondo, Emanuele Los Angeles, Claudionor Bernardo, Luana dos Santos de Oliveira, Margarete de Jesus Carvalho, and Daniel Boari Coelho. A public data set of walking full-body kinematics and kinetics in individuals with parkinson's disease. *Frontiers in Neuroscience*, 17:992585, 2023. 2, 3, 4

[38] Elke Warmerdam, Clint Hansen, Robbin Romijnders, Markus A Hobert, Julius Welzel, and Walter Maetzler. Full-body mobility data to validate inertial measurement unit algorithms in healthy and neurological cohorts. *Data*, 7(10):136, 2022. 2

[39] 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. 2, 26

[40] N Kour, S Gupta, and S Arora. Gait dataset for knee osteoarthritis and parkinson's disease analysis with severity levels. *Mendeley Data, V1*, 2020. 2

[41] Yefei He, Tao Yang, Cheng Yang, and Hong Zhou. Integrated equipment for parkinson's disease early detection using graph convolution network. *Electronics*, 11(7):1154, 2022. 2

[42] Wei Zhang, Zhuokun Yang, Hantao Li, Debin Huang, Lipeng Wang, Yanzhao Wei, Lei Zhang, Lin Ma, Huanhuan Feng, Jing Pan, et al. Multimodal data for the detection of freezing of gait in parkinson's disease. *Scientific data*, 9(1):606, 2022. 2

[43] Mohammadreza Abtahi, Seyed Bahram Borgheai, Roohollah Jafari, Nicholas Constant, Rassoul Diouf, Yalda Shahriari, and Kunal Mankodiya. Merging fNIRS-EEG brain monitoring and body motion capture to distinguish parkinsons disease. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 28(6):1246–1253, 2020. 2

[44] Wenhao Zhang, Haipeng Dai, Dongyu Xia, Yang Pan, Zeshui Li, Wei Wang, Zhen Li, Lei Wang, and Guihai Chen. mP-Gait: Fine-grained parkinson's disease gait impairment assessment with robust feature analysis. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(3):1–31, 2024. 2

[45] Yao Guo, Jianxin Yang, Yuxuan Liu, Xun Chen, and Guang-Zhong Yang. Detection and assessment of parkinson's disease based on gait analysis: A survey. *Frontiers in aging neuroscience*, 14:916971, 2022. 2

[46] Sevgi Z Gurbuz, Mohammad Mahbubur Rahman, Zahra Bassiri, and Dario Martelli. Overview of radar-based gait parameter estimation techniques for fall risk assessment. *IEEE Open Journal of Engineering in Medicine and Biology*, 2024. 2

[47] Amirhossein Dadashzadeh, Shuchao Duan, Alan Whone, and Majid Mirmehdi. PeCop: Parameter efficient continual pretraining for action quality assessment. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 42–52, 2024. 3

[48] Diwei Wang, Chaima Zouaoui, Jinhyeok Jang, Hassen Drira, and Hyewon Seo. Video-based gait analysis for assessing alzheimer's disease and dementia with lewy bodies. In *International Workshop on Applications of Medical AI*, pages 72–82. Springer, 2023. 3

[49] Joke Spildooren, Sarah Vercruysse, Kaat Desloovere, Wim Vandenberghe, Eric Kerckhofs, and Alice Nieuwboer. Freezing of gait in parkinson's disease: the impact of dual-tasking and turning. *Movement Disorders*, 25(15):2563–2570, 2010. 3, 4

[50] Trung-Hieu Hoang, Christopher Zallek, and Minh N Do. Smartphone-based digitized neurological examination toolbox for multi-test neurological abnormality detection and documentation. *IEEE Journal of Biomedical and Health Informatics*, 2024. 3

[51] J Lucas McKay, Felicia C Goldstein, Barbara Sommerfeld, Douglas Bernhard, Sahyli Perez Parra, and Stewart A Factor. Freezing of gait can persist after an acute levodopa challenge in parkinson's disease. *NPJ Parkinson's disease*, 5(1):25, 2019. 3, 4

[52] Sina Mehdizadeh, Elham Dolatabadi, Kimberley-Dale Ng, Avril Mansfield, Alastair Flint, Babak Taati, and Andrea Iaboni. Vision-based assessment of gait features associated with falls in people with dementia. *The Journals of Gerontology: Series A*, 75(6):1148–1153, 2020. 3

[53] Vida Adeli, Soroush Mehraban, Majid Mirmehdi, Alan Whone, Benjamin Filtjens, Amirhossein Dadashzadeh, Alfonso Fasano, and Andrea Iaboni Babak Taati. GAITGen: Disentangled motion-pathology impaired gait generative model–bringing motion generation to the clinical domain. *arXiv preprint arXiv:2503.22397*, 2025. 3, 10

[54] Trung-Hieu Hoang, Mona Zehni, Huaijin Xu, George Heintz, Christopher Zallek, and Minh N. Do. Towards a comprehensive solution for a vision-based digitized neurological examination. *IEEE Journal of Biomedical and Health Informatics*, 26(8):4020–4031, 2022. 3

[55] Hyeokhyen Kwon, Gari D Clifford, Imari Genias, Doug Bernhard, Christine D Esper, Stewart A Factor, and J Lucas McKay. An explainable spatial-temporal graphical convolutional network to score freezing of gait in parkinsonian patients. *Sensors*, 23(4):1766, 2023. 4

[56] Xinxin Zuo, Sen Wang, Jiangbin Zheng, Weiwei Yu, Minglun Gong, Ruigang Yang, and Li Cheng. SparseFusion: Dynamic human avatar modeling from sparse rgbd images. *IEEE Transactions on Multimedia*, 23:1617–1629, 2020. 4

[57] 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. 4, 24

[58] 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. 4, 24

[59] 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. 5, 27

[60] 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. 5, 27

[61] 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. 5, 27

[62] 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. 5, 27, 28

[63] 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, 6, 8, 28

[64] 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, 28

[65] 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, 8, 28, 29

[66] 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. 5, 30

[67] 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. 5, 30

[68] 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. 5, 30

[69] 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. 5, 30

[70] Beatriz Muñoz-Ospina, Daniela Alvarez-Garcia, Hugo Juan Camilo Clavijo-Moran, Jaime Andrés Valderrama-Chaparro, Melisa García-Peña, Carlos Alfonso Herrán, Christian Camilo Urcuqui, Andrés Navarro-Cadavid, and Jorge Orozco. Machine learning classifiers to evaluate data from gait analysis with depth cameras in patients with parkinson's disease. *Frontiers in Human Neuroscience*, 16:826376, 2022. 6

[71] Angjoo Kanazawa, Jason Y Zhang, Panna Felsen, and Jitendra Malik. Learning 3d human dynamics from video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5614–5623, 2019. 7

[72] Céline Schreiber and Florent Moissenet. A multimodal dataset of human gait at different walking speeds established on injury-free adult participants. *Scientific data*, 6(1):111, 2019. 9

[73] Geise Santos, Marcelo Wanderley, Tiago Tavares, and Anderson Rocha. A multi-sensor human gait dataset captured through an optical system and inertial measurement units. *Scientific Data*, 9(1):545, 2022. 9

[74] Aurélie Bertaux, Mathieu Gueugnon, Florent Moissenet, Baptiste Orliac, Pierre Martz, Jean-Francis Maillefert, Paul Ornetti, and Davy Laroche. Gait analysis dataset of healthy volunteers and patients before and 6 months after total hip arthroplasty. *Scientific Data*, 9(1):399, 2022. 9

[75] Gautier Grouvel, Lena Carcreff, Florent Moissenet, and Stéphane Armand. A dataset of asymptomatic human gait and movements obtained from markers, imus, insoles and force plates. *Scientific Data*, 10(1):180, 2023. 9

[76] Abraham Savitzky and Marcel JE Golay. Smoothing and differentiation of data by simplified least squares procedures. *Analytical chemistry*, 36(8):1627–1639, 1964. 25

[77] 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. 26, 30

[78] 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. 26

[79] 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. 26

[80] 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. 26

[81] 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. 27

[82] 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. 28

[83] 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. 30

[84] 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. 30

[85] 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. 31

[86] 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. 35

[87] 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. 36

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction accurately reflect the paper's contributions and scope by presenting the creation of the multi-site CARE-PD dataset, the benchmarking of supervised tasks under four generalization protocols, and the potential of pretraining tasks. The empirical results provide generalization insights, which may inform future directions that may be directly facilitated by the CARE-PD dataset.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We provided a dedicated limitation section in the Appendix.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: The paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Provided in the main paper Sec. 5 and the Appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: A data repository is created on the University of Toronto Dataverse (Data). Details and documentation are available in the Appendix, the Dataverse metadata, and in the GitHub codebase page (CARE-PD).

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Details are provided in the Appendix, data splits are provided in the Dataverse (Data), and training scripts in the GitHub page (CARE-PD).

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Error bars (mean ± SD) are reported for the subset of benchmarks that have led the paper's conclusions in the Appendix, with folds and seeds listed in the Dataverse and released code.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Compute resources are provided in the Appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: Our research adheres to the NeurIPS Code of Ethics by following institutional ethical review processes for retrospective use of existing clinical datasets (see Sec. 3.1), ensuring participant anonymity through harmonized preprocessing (see Sec. 3.2), explicitly addressing generalization considerations in our benchmarks, and transparently communicating dataset limitations and intended usage through structured data licenses and reproducibility measures (see Dataverse (Data)).

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: Details on the broader impacts of CARE-PD are provided in the Appendix.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [Yes]

Justification: CARE-PD has taken steps to safeguard against potential misuse as discussed in the main paper Sec. 3.2.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The original owners of the included datasets have been properly credited. The CARE-PD data repository includes the original owners, the name of the license and the terms of use, and the license of the re-packaged assets if relevant (`Data`).

Guidelines:

- The answer NA means that the paper does not use existing assets.

- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Clear documentation is provided both in our Dataverse (`Data`) and GitHub codebase page (`CARE-PD`)

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: CARE-PD utilizes nine datasets, six of which are existing retrospective datasets that did not require new participant instructions. For the three newly collected datasets, participants were informed clearly about the data acquisition process and provided informed consent; full details including the consent procedures, instructions provided to participants, and any applicable compensation are included in the Appendix.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [Yes]

Justification: CARE-PD has ethical approval for the retrospective use of the included datasets (see Sec. 3.1), as approved by the Social Sciences, Humanities & Education Research Ethics Board of the University of Toronto (REB #47891) and has ethical approval of the University Health Network Research Ethics Board (CAPCR ID 24-5835) for the release of the new T-SDU and T-LTC datasets.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: It does not impact the core methodology, scientific rigorousness, or originality of the research, therefore declaration is not required.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.

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.

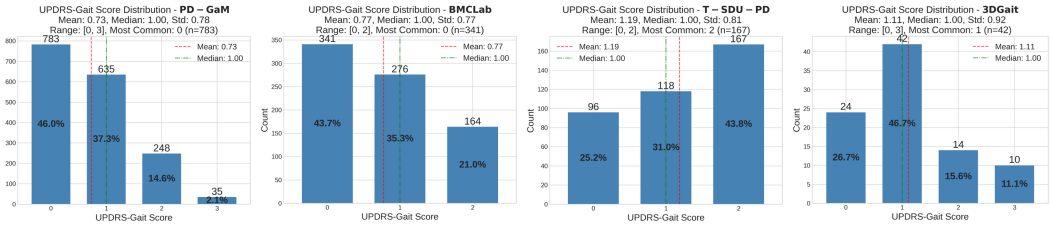

Figure A.1: Class distributions for the four datasets with UPDRS-gait labels.

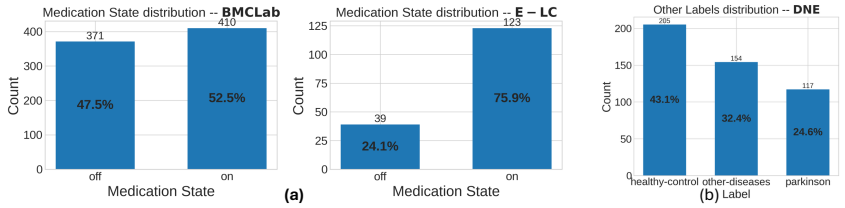

Figure A.2: (a) Medication state breakdown for BMClab and E-LC datasets. (b) Diagnostic categories in DNE dataset.

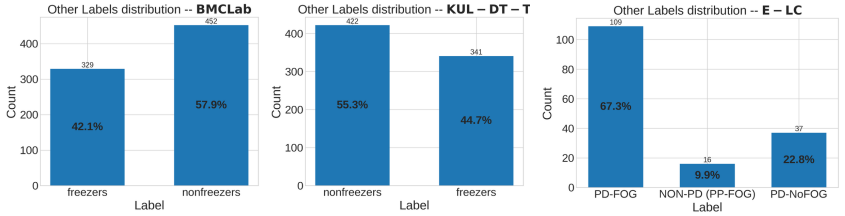

Figure A.3: Label distributions for freezing status for BMClab, KUL-DT-T and E-LC datasets.

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 [57]. 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 [58]. 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:

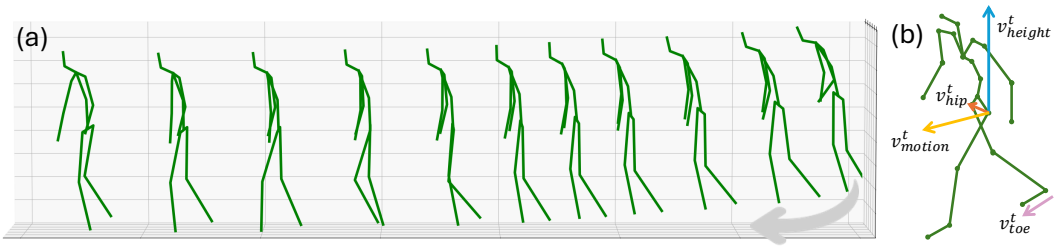

Figure A.4: a) Illustration of the slope artifcat b) An example of the vertical height vector (blue), the direction of movement vector (yellow), and the hip vector (orange).

(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 [76] (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.

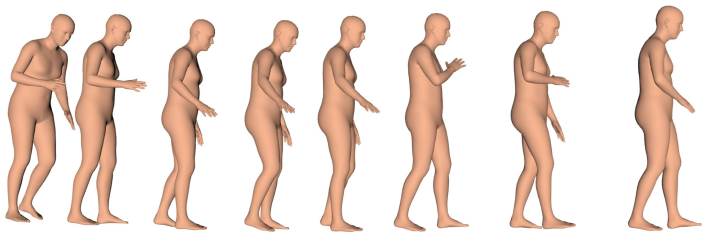

Figure A.5: Example of the 6890 vertices SMPL mesh at different frames of the gait sequence.

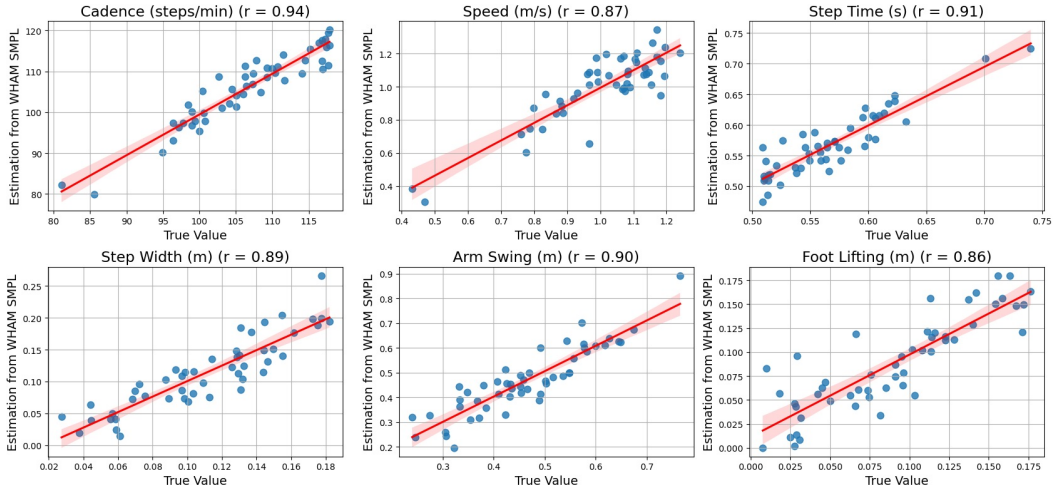

Figure A.6: Correlation between WHAM estimations and IMU ground-truth gait features. Pearson correlations ranged from $r = 0.86$ to $r = 0.94$.

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) [77]. 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 [78, 79, 80] and the TULIP dataset for PD gait task [39]. 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 [2] 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 [27], 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 [62]. 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 [81], 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 [81], 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 [59] 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 [12]. We use the encoder's temporally pooled token embeddings as input features for our downstream clinical classifier. Input: 3D Human3.6M joints.

MixSTE [60] 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 [61] 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 [62] 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 [63] 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 [64] 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 [82]. Input: SMPL (6D rotation).

MoMask [65] 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 projected 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 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.

We emphasize that for both MoCap and video-based sequences, 2D projections are generated directly from the same 3D SMPL meshes used by the 3D encoders. Consequently, both modalities share identical pose sources ensuring that any reconstruction noise or artifacts are consistently reflected in both 2D and 3D inputs. This setup provides a controlled comparison focused on encoder representation capacity rather than differences in keypoint detection or data quality.

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 [65]. 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 [68, 67, 13], 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 [83], 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 [84, 66, 69] 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 [77]. PD may also manifest as patients taking shorter steps, resulting in elevated cadence [66]. Moreover, a stooped posture is commonly seen in PD and is directly associated with postural instability [69].

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 [85] 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 [85] 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.

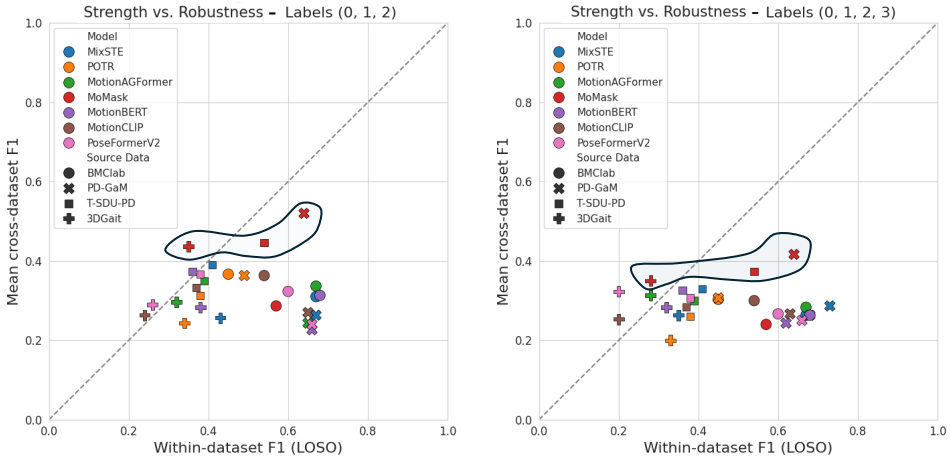

Figure D.8: Accuracy vs. robustness analysis. Each marker represents an encoder plus linear prob trained on one dataset (marker shape) and evaluated on that dataset (x-axis) and, on average, on the other three (y-axis). Colours distinguish encoder backbones; The left plot reports macro-F1$_{0-2}$, the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

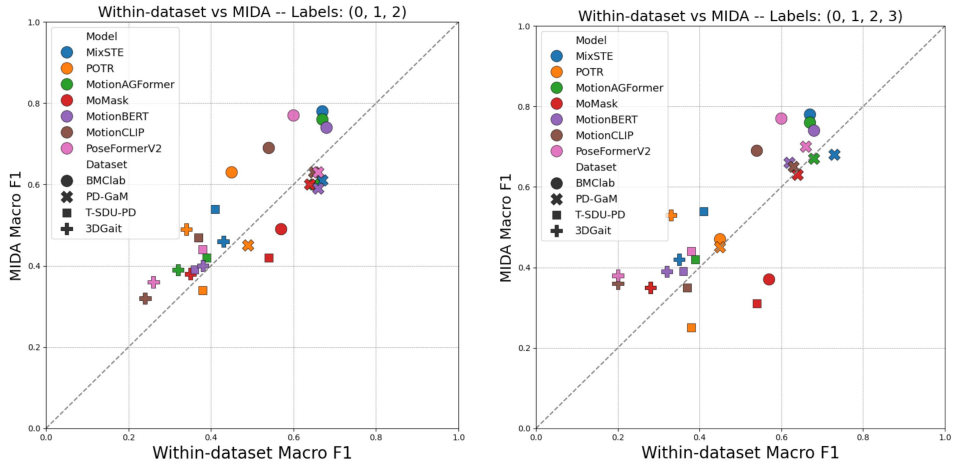

Figure D.9: Within-dataset vs. MIDA evaluation (effect of in-domain adaptation and external data for training). Each point compares macro-F$_1$ scores of an encoder trained with (y-axis) and without (x-axis) access to additional out-of-domain datasets. LOSO uses training data only from the evaluation 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.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.

D.3 Label Variability vs. Domain Shift

To disentangle the effects of annotation variability from domain shift, we conducted a focused analysis involving two annotators and two datasets (PD-GaM and T-SDU-PD).

1) Fixed dataset (PD-GaM), different annotators (A, B) to assess inter-rater variability: To quantify consistency between annotators (isolating annotation variability while holding the data distribution constant), we asked the clinician who labeled T-SDU-PD (Annotator A) to independently

re-score a subset of PD-GaM originally labeled by another clinician (Annotator B). The two clinicians showed strong agreement (Cohen's $\kappa = 0.92$, ICC = 0.92), indicating consistent scoring.

2) Fixed annotator (A), different datasets (T-SDU-PD, PD-GaM) to assess domain shift: We used both datasets with labels from the same clinician (A). A Momask encoder + classifier trained on T-SDU-PD (A-labeled) was evaluated on PD-GaM (also A-labeled). Table D.1 shows the results. With identical annotator and balanced class distributions, testing on PD-GaM reduces macro-F1 by 9 pp (0.54 $\rightarrow$ 0.45), while label harmonization changed macro-F1 by 1 pp (0.44 $\rightarrow$ 0.45). This confirms that the cross-site gap is driven primarily by dataset domain shift rather than inter-rater variability.

Table D.1: Macro-F1 for model trained on T-SDU-PD with Clinician A labels.

Test Label	**T-SDU-PD**	**PD-GaM**
Original (Clinician B)	0.54	0.44
Clinician A (harmonized)	same (0.54)	0.45

D.4 View-point results for 2D encoders

Table D.2 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.

Table D.2: 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**

D.5 Slope correction ablation

To quantify the effect of slope correction on model performance, we conducted an ablation study on the T-SDU-PD dataset, the only dataset in this experiment affected by ceiling-view slope artifacts. Two top-performing encoder models (MotionAGFormer and MoMask) were evaluated under multiple protocols, both with and without slope correction. Results are summarized in Table D.3.

In *within-dataset* experiment, slope correction had negligible effect in both models, confirming robustness when training and testing on the same (distorted) data. In *cross-dataset* evaluations, training on T-SDU-PD and testing on BMClab (no slope) showed minimal impact. Also, training on BMClab and testing on T-SDU-PD showed a similar minor effect. Under the *MIDA* protocol, training on multiple datasets and testing on T-SDU-PD or BMClab showed marginal gains from correction.

Overall, in all cases, the effect was less than 0.5 pp, supporting that encoder models are largely invariant to minor geometric distortions. In contrast, classical hand-crafted feature baseline, dropped 6–9 pp without correction. These results support our initial rationale that slope correction is essential for feature-based baselines that rely on accurate relative distances, while encoder models tolerate modest distortion due to learned representations.

Table D.3: Ablation on slope correction across evaluation protocols. Metrics are macro-F1 (%).

Protocol	Slope Corr.	Train	Test	MotionAGFormer	MoMask
Within	✗	T-SDU-PD	T-SDU-PD	39.2	54.3
	✓	T-SDU-PD	T-SDU-PD	38.9 (↓**0.3**)	54.4 (↑**0.1**)
Cross	✗	T-SDU-PD	BMClab	48.3	44.5
	✓	T-SDU-PD	BMClab	48.7 (↑**0.4**)	45.1 (↑**0.6**)
Cross	✗	BMClab	T-SDU-PD	38.2	31.3
	✓	BMClab	T-SDU-PD	38.6 (↑**0.4**)	31.5 (↑**0.2**)
MIDA	✗	T-SDU-PD (train) + All others	T-SDU-PD (test)	42.3	42.5
	✓	T-SDU-PD (train) + All others	T-SDU-PD (test)	42.8 (↑**0.5**)	42.9 (↑**0.4**)
MIDA	✗	All others + BMClab (train)	BMClab (test)	76.2	49.3
	✓	All others + BMClab (train)	BMClab (test)	76.1 (↓**0.1**)	49.5 (↑**0.2**)

D.6 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±SD across these n folds. *(ii) Cross-dataset:* 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.4 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

Table D.4: **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

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-ND 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‘[86]. *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.

Benchmark Scope. The benchmarking tasks in CARE-PD focus on UPDRS-gait score prediction and unsupervised motion representation learning, chosen to balance clinical grounding with general methodological relevance. These tasks provide baselines that connect motion representation learning to clinically validated outcomes. The dataset, however, was built with broader applicability in mind. Its preserved attributes, such as freezing episodes, step-level irregularities, variable segment lengths, and longitudinal recordings enable exploration of additional tasks.

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 [87]. 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.