
Politecnico di Torino


## Politecnico di Torino

Master's degree in Data Science and Engineering

A.A. 2025/2026

Graduation Session March 2026


# Exploiting skeleton-based Motion Encoder Networks to characterize Parkinsonian gait


Supervisors:

Gabriella Olmo

Gianluca Amprimo

Candidate:

Federica Mendoza

**Abstract**

Parkinson's disease is a neurodegenerative disorder, characterized by progressive motor impairments including tremor, rigidity, and bradykinesia. Clinical assessment relies primarily on the Movement Disorder Society-Unified Parkinson's Disease Rating Scale (MDS-UPDRS), which, despite being the gold standard, suffers from inherent subjectivity, inter-rater variability, and limited sensitivity to subtle motor changes—particularly for distinguishing between mild severity levels (scores 0-2). These limitations underscore the need for objective, quantitative assessment methods that can complement clinical evaluation. Recent advances in skeleton-based motion encoder networks have demonstrated remarkable capabilities in capturing complex human movement patterns from video or motion capture data. However, their application to clinical gait assessment in Parkinson's disease faces significant challenges: limited availability of labeled clinical data, underrepresentation of pathological movement patterns in training datasets, and substantial cross-dataset generalization gaps arising from variations in recording protocols and sensor modalities. This thesis investigates strategies for leveraging state-of-the-art motion encoder models to assess parkinsonian gait severity when training data is limited. We focus specifically on gait analysis, as walking sequences provide extended temporal information essential for characterizing diagnostically relevant features such as freezing of gait, stride length, posture, and balance. Our experiments employ two pre-trained motion encoders, PoseFormerV2 and MotionBERT, evaluated across three datasets with diverse acquisition modalities. We systematically compare multiple adaptation strategies, including full end-to-end fine-tuning and parameter-efficient Low-Rank Adaptation (LoRA). Moreover, we explore self-supervised pretraining using masked motion modeling with Taylor series expansion to enhance the models' understanding of gait dynamics. Finally, we benchmark our best-performing configuration against current state-of-the-art methods, identifying key factors that influence the clinical viability of motion encoder models as objective tools for parkinsonian gait assessment. Our results indeed demonstrate that models benefit from a more targeted finetuning strategy and self-supervised pretraining tasks. While this represents promising performance for cross-domain gait severity classification, our analysis reveals that substantial cross-dataset generalization gaps persist, attributable to heterogeneity in data acquisition protocols, skeletal representation formats, and population characteristics. We demonstrate that enhanced data harmonization procedures—including standardized skeleton normalization, temporal alignment, and domain-invariant feature learning—could further bridge these generalization gaps and improve clinical applicability.

# Acknowledgements

I would like to thank professor Gabriella Olmo and my supervisor Gianluca Amprimo for the constant guidance and support throughout this research. Thank you to my boyfriend, my friends, and my family for the never ending encouragement and love throughout these university years.

# Table of Contents

# List of Figures

# Chapter 1

# Introduction

## 1.1 Parkinson's Disease: Clinical Overview

Parkinson's disease (PD) is a neurodegenerative movement disorder, characterised by the loss of dopamine neurons in the *substantia nigra*, a region of the mid-brain. The dopamine deficiency leads to motor coordination impairments, the most evident characteristics being tremor, rigidity, instability, and bradykinesia (slowness of movement). PD patients manifest non-motor symptoms as well, such as autonomic dysfunction, hyposmia, and cognitive decline.
Although PD is the second most common neurodenerative disorder after Alzheimer, the precise etiology triggering the loss of dopamine neurons remains unknown, making the disease idiopathic. As a consequence, no definitive biological marker currently exist, leaving diagnosis to rest exclusively on the subjective clinical evaluation of motor signs. [1].
MDS-UPDRS (Movement Disorder Society-Unified Parkinson's Disease Rating Scale) is the current scale used by neurologists to assess the severity of PD. It consists of 4 parts, each evaluating different aspects of the disorder:[2]

- *Part I: Non-Motor Experiences of Daily Living* assesses the non-motor sympto influencing the patient's ability to carry out everyday activities. Specifically, it assesses problems relating to cognitive function, sleep, mood, psychosis, fatigue, constipation, and urinary problems;
- *Part II: Motor Experiences of Daily Living* evaluates the patient's subjective perception of their motor capabilities in functional tasks, such as speaking, eating, dressing, and hygiene;
- *Part III: Motor Examination*is the objective, clinician-rated component of the scale. The patient undergoes physical tasks in order to assess specific motor signs such as tremor, bradykinesia, muscle rigidity, and coordination;

- *Part IV: Motor Complications* covers the complications arising from long-term therapy, specifically the duration of "OFF" periods -that is, when the medication has worn off or failed to kick in- and the presence of dyskinesias (involuntary movements induced by medication).

Among all parts, MDS-UPDRS Part III serves as the main assessment for clinicians to tailor the therapy, directly correlating medication adjustments with observed improvements in motor performance. Each task has 0-4 ratings, reflecting severity of symptoms: 0: (normal, no symptoms), 1 (slight, with minor symptoms), 2 (mild, noticeable symptoms) 3 (moderate, significant symptoms), 4 (severe symptoms).

## 1.2 Limitations of MDS-UPDRS assessment

Despite its widespread use and status as gold standard for PD assessment, MDS-UPDRS suffers from significant limitations that compromise its reliability and reproducibility. The assessment is intrinsically subjective, relying heavily on the clinician's experience, perceptual judgment, and interpretation of subtle motor signs. Studies showed that even when the same clinician evaluates the same patient repeatedly, scores reveal considerable variation, with wide repeatability limits. Moreover, systematic differences in scoring patterns exist between nurse practitioners, residents in neurology, a movement disorders specialist (MDS), and a senior MDS, introducing consistent bias across the rating spectrum. [3] [4]. This limitation is further compounded by the fact that while physical tasks rated 3 or 4 represent severe motion impairment and hence are more discernible, on the other hand, scores from 0 to 2 characterize a gradual decline in mobility, where differences about tremor, stride length, coordination, and gait become more subtle, making it challenging to distinguish the classes of severity. [5] These inherent constraints in objectivity, granularity, and temporal sampling underscore the urgent need for complementary objective measurement technologies that can provide continuous, quantitative, and rater-independent assessments of parkinsonian motor symptoms.

## 1.3 Machine Learning for Motion Analysis and Domain Gap

Recent studies in computer vision and deep learning have introduced a new paradigm for human motion analysis through skeleton-based motion encoder networks. These networks operate on skeletal representations of human movement, where the body is modeled as a graph of interconnected joints extracted from video or motion capture data. Unlike traditional video-based methods that process raw pixel data, this approach offer a more compact and interpretable representation that is invariant

to appearance and background. [6]. Trained on large-scale datasets of human movement, these encoders have demonstrated remarkable capabilities in capturing and representing complex motion patterns [7].[8]
However, despite their success in general motion recognition tasks, their application in clinical contexts has not yet reached the same level of performance. Firstly, the available datasets of human movements consists primarily of healthy individuals, leaving little representation to Parkinsonian motion patterns. Additionally, collecting labeled clinical data can be challenging, as disease progression monitoring requires repeated assessments over months, clinical evaluations are time-consuming and labour intensive, and due to strict privacy regulations limiting data sharing between institutions. Lastly, though recent studies have shown promising results for severity classification task, results showed that cross-dataset generalizaton gap, rooted in sensitivity to domain-specific factors such as camera configurations, sensor modalities (RGB cameras vs. marker-based motion capture), and recording environments. [9]

## 1.4 Thesis Objectives

This thesis will aim to bridge this gap, analyzing which methods and strategies can provide clinically significant insights for assessing Parkinsonian, leveraging on the small, limited clinical data. We focus specifically on gait analysis, as walking sequences provide extended temporal data that enable deeper analysis of diagnostically relevant features including freezing of gait (FoG), tremor, stride length, posture, and balance. Since patients with MDS-UPDRS gait scores of 3 and 4 typically exhibit severe walking impairments requiring assistive support—and are consequently underrepresented in publicly available datasets—our experiments focus on discriminating between severity classes 0, 1, and 2. We conduct experiments using two state-of-the-art skeleton-based motion encoder models: PoseFormerV2 [10] and MotionBERT [11]. Our analysis encompasses three datasets with diverse data acquisition modalities: one based on marker-based motion capture and two utilizing markerless pose estimation for skeleton extraction. To identify optimal adaptation strategies for limited clinical data, we evaluate multiple fine-tuning approaches, ranging from end-to-end full fine-tuning to parameter-efficient methods such as Low-Rank Adaptation (LoRA) [12]. Additionally, we investigate the effect of self-supervised pretraining tasks designed to capture gait dynamics more effectively, specifically employing masked motion modeling with Taylor series expansion [13]to learn spatiotemporal patterns in human movement. Finally, to rigorously assess the clinical viability and cross-dataset generalization of our optimized models, we benchmark our best-performing configurations against the recently introduced CARE-PD benchmark suite [9], utilizing its standardized state-of-the-art evaluation

protocols.

# Chapter 2

# Background studies

## 2.1 Pose estimation

Recovering the structure of the human body from visual data represents a complex challenge, due to the high number of degrees of freedom in human kinematics and the inherent ambiguities of inferring 3D motion from its two-dimensional projections. To address this challenge, the field has traditionally relied on Optical Motion Capture (MoCap) systems.

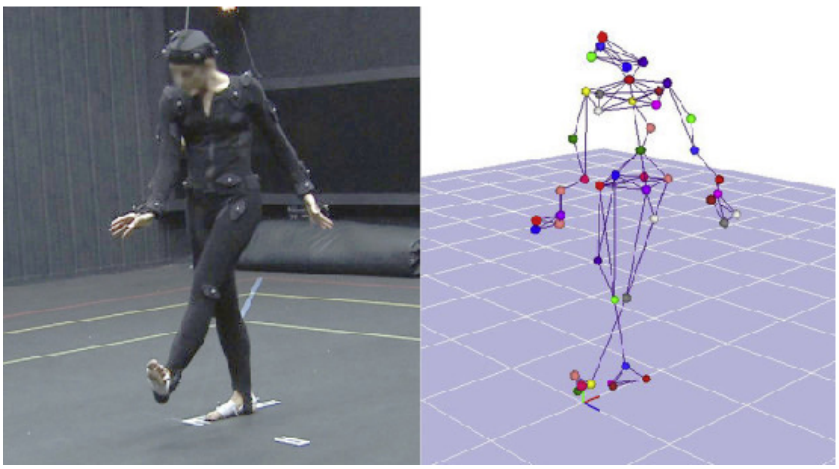


**Figure 2.1:** Example of an Optical Motion Capture (MoCap) setup. The subject (left) wears a suit equipped with markers, which are tracked by cameras to reconstruct a highly accurate skeletal representation (right).

MoCap systems consist of a set of markers attached to predefined locations of the human body. Furthermore, movements are tracked using multiple cameras,

allowing analysis from different viewing angles. A dedicated software program uses the trigonometric relations among the markers and the camera positions to calculate the positions and orientations of the body joints. This system is considered the most accurate and reliable for tracking, as many studies showed exceptional results. [14][15][16] However, the prohibitive costs and limitation of the laboratory environment have led to the urgent demand for an accessible, markerless alternative.

### 2.1.1 YOLO-pose

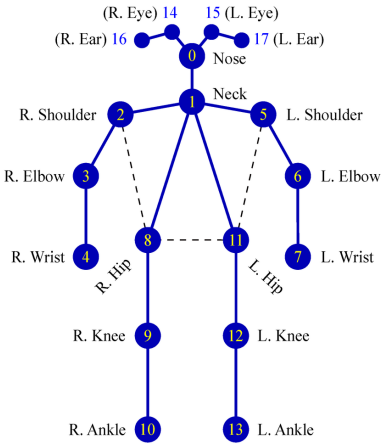
To overcome the limitations of MoCap systems, YOLO-Pose [17] has emerged as a state-of-the-art architecture for the recovery of human pose information directly from standard RGB videos. Built upon the YOLO (You Only Look Once) object detection framework [18], YOLO-Pose employs a single-stage, top-down architecture.

More precisely, YOLO-Pose formulates human pose estimation as a direct regression problem, predicting the 2D spatial coordinates of anatomical keypoints directly from the input RGB image. The architecture processes an RGB input (a photo or a video frame) through a shared backbone (CSP-Darknet53) that generates feature maps at multiple scales. These features are then fused by PANet (Path Aggregation Network), which brings rich semantics from deep features to shallow features, and spatial details from shallow features to deep features. The 17-keypoint skeleton format is determined by the COCO (Common Objects in Context) dataset [19], on which YOLO-Pose is trained. This standardized format defines specific body landmarks including nose, eyes, ears, shoulders, elbows, wrists, hips, knees, and ankles, as illustrated in Figure 2.2a.

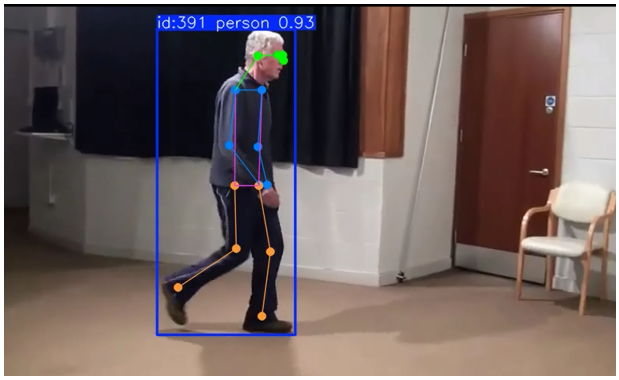
This single-stage, unified design fundamentally differs from traditional top-down approaches, where person detection and pose estimation occur sequentially, therefore it enables faster processing.
In particular, YOLO-Pose demonstrates robustness in the detection of occluded joints. Such result is achieved not by explicit modules, but by providing visible and occluded keypoints the same confidence ground truth, forcing the model to learn anatomical patterns by means of numerous annotated occlusions and of the analysis of thousands of pose configurations. In the context of gait analysis, such capability becomes significant for assessing gait turns - where the torso obscures the trailing arm and leg - providing complete kinematic data for assessing disease-relevant features such as turning duration, step patterns, and trunk-limb coordination.

These strengths reflect the state-of-the-art performance in pose estimation tasks,

motivated by the achievement of 90.2% AP50 on the COCO validation set and 90.2% AP50 on the test-dev set [17], where AP50 measures the proportion of correct predictions, that is, if Object Keypoint Similarity (OKS) between predicted and ground truth keypoints is at least 0.5. OKS quantifies the spatial agreement between predicted and annotated keypoints, accounting for the scale of the person and the natural variability of each keypoint type.




**(a)** Key points definition (R/L: right/left)




**(b)** Estimation example

**Figure 2.2:** Visualizing keypoint estimation: (a) The standard COCO human pose format diagram; (b) Example of inference generated by YOLOpose.

### 2.1.2 SMPL

While YOLO-Pose excels at 2D keypoint detection, gait analysis requires 3D spatial information—step length, stride width, joint angles—that cannot be extracted from image-plane projections. Depth information is particularly critical for analyzing parkinsonian gait features such as reduced stride length, which occurs primarily along the camera's depth axis.

To recover 3D body geometry from monocular video, parametric models such as SMPL (Skinned Multi-Person Linear Model) [20] provide a framework for 3D pose and shape reconstruction. SMPL (Skinned Multi-Person Linear Model) [20] represents the human body as a function $M(\boldsymbol{\beta}, \boldsymbol{\theta}) : \mathbb{R}^{10} \times \mathbb{R}^{72} \rightarrow \mathbb{R}^{6890 \times 3}$ that maps shape parameters $\boldsymbol{\beta}$ and pose parameters $\boldsymbol{\theta}$ to a 3D mesh with 6,890 vertices. Shape parameters control anthropometric variations (body proportions), while pose parameters encode the 3D orientation of 24 body joints.

The model applies pose-dependent blend shapes and linear blend skinning to generate realistic body deformations.

WHAM (World-grounded Humans with Accurate Motion) [21] further improves standard SMPL-based reconstruction by introducing global spatial context. Specifically, WHAM employs a transformer-based motion encoder that processes video sequences and predicts: per-frame SMPL parameters $(\boldsymbol{\beta}, \boldsymbol{\theta})$, global root translation $\boldsymbol{t} \in \mathbb{R}^3$, and global root orientation $\boldsymbol{r} \in SO(3)$. By integrating motion knowledge learned from large-scale motion capture datasets, WHAM produces 3D reconstructions with accurate global trajectories, enabling extraction of spatially grounded gait parameters essential for parkinsonian movement analysis. This is particularly beneficial for clinical studies where videos may be captured from different camera angles. In fact, WHAM produces spatially consistent 3D reconstructions regardless of camera position, as all estimates are anchored to a common world coordinate frame.

As a form of data privacy and data protection, we set shape parameters $\boldsymbol{\beta} = \mathbf{0}$ in all reconstructions, using a neutral body template that removes identifiable anthropometric features while maintaining accurate joint positions and motion trajectories necessary for kinematic analysis.

**Figure 2.3:** Example of WHAM pose estimation. On the left, a neutral body mesh is applied to the subject. On the right, the same body mesh is analyzed with a simulated ceiling mounted camera, with a back right angle view.

## 2.2 Motion Encoders

Having established how pose estimation methods extract skeletal coordinates from visual input, we now turn to motion encoders, which represent the next critical stage in the analysis pipeline. Motion encoders are designed to solve the fundamental problem of 2D-to-3D pose lifting, that is, to reconstruct the true three-dimensional geometry of the human body from two-dimensional pixel coordinates. Motion encoders take as input the skeletal sequences represented as three-dimensional tensors of shape $T \times J \times C$, where $T$ denotes the temporal length (number of frames), $J$ represents the number of body joints, and $C$ specifies the coordinate dimensions

per joint. To accurately lift these flat representations into 3D space and filter out detector noise, modern motion encoders predominantly employ Spatio-Temporal Transformer architectures [22, 7, 23]. Unlike traditional convolutional or recurrent neural networks that process information sequentially or within local windows, transformers use attention mechanisms capable of simultaneously considering relationships between all elements in a sequence. For human motion analysis, this is achieved through two complementary attention streams:

- **Spatial Attention:** It models relationships between different body joints within a single time step. It learns to identify which joints are geometrically and functionally related (e.g., how the shoulder's position influences the wrist).

- **Temporal Attention:** It models how individual joints evolve over time across the entire sequence. It identifies motion patterns such as periodic movement (e.g., the cyclical swing of the leg) or temporal coordination.

In this section, we examine two state-of-the-art motion encoder architectures designed for 2D-to-3D pose lifting: PoseFormerV2 and MotionBERT. While both share the core foundation of spatio-temporal attention, they represent two distinct architectural philosophies. PoseFormerV2 tackles the computational bottlenecks of transformers by utilizing the frequency domain to compress sequences and filter noise. MotionBERT, in the other hand, introduces a universal foundation model approach, deeply intertwining spatial and temporal streams to build a robust understanding of human kinematics.

### 2.2.1 PoseFormerV2

To address the computational intensity of standard spatio-temporal transformers while maintaining high accuracy in 2D-to-3D lifting, PoseFormerV2 introduces a highly optimized frequency-based architecture. Built upon PoseFormerV1, PoseFormerV2 consists of two main modules, that is, the spatial encoder for single-frame joint correlation modeling, and the temporal encoder for cross-frame human motion modeling.

The input is a 2D skeleton sequence $\mathbf{x} \in \mathbb{R}^{F \times J \times 2}$, where F and J respectively denote the sequence length and the number of joints representing the body. The coordinates of the pelvis are identified and substracted by every other joint in the frame. Such process -namely, root-relative normalization - shifts the pelvis to $(0, 0)$, while the other joints now represented by its distance and direction relative to the pelvis, rather than its absolute pixel location on the screen.hence, making the model translation-invariant. Two operations, in parallel, are applied to $\mathbf{x}$.

First, we sample $F'$ frames such that they are around the centre of the sequence and such that $F' \ll F$; the new sequence $\mathbf{x}' \in \mathbb{R}^{F' \times J \times 2}$ will be the input to the spatial encoder. This strategy ensure that the model avoids the massive computational bottleneck of analyzing the full, long sequence frame-by-frame. Moreover, we focus on the sequence center as it most crucial, fine-grained spatial information. Next, a linear projection to a $c$-dimensional vector is applied to all coordinates $(x, y)$ in each frame, hence obtaining $\mathbf{z'}_0 \in \mathbb{R}^{F' \times J \times c}$, to which we add a learnable spatial positional embedding $\mathbf{E}_{SPos} \in \mathbb{R}^{1 \times J \times c}$ in order to capture joint-specific characteristics. For each frame $\mathbf{z'}_0^i \in \mathbb{R}^{F' \times J \times c}$. we apply the self-attention mechanism, allowing the model to learn the spatial relationships between different body parts within that same instant. The obtained per-frame representations are then flattened and concatenated as $\mathbf{z'}^{Time} \in \mathbb{R}^{F' \times (J \cdot c)}$, i.e. frame-level features in the time domain.

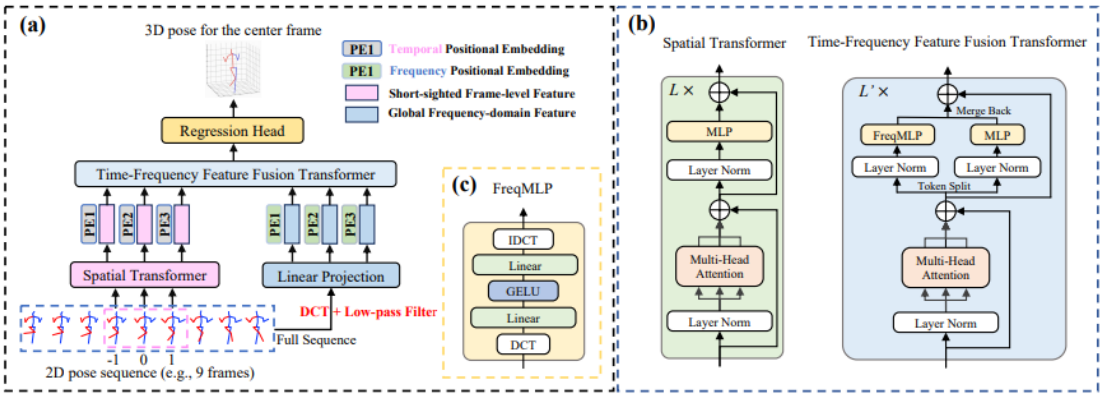



**Figure 2.4:** Overview of the PoseformerV2[10] architecture, illustrating the process of converting 2D joint skeletons to 3D poses and joint trajectories using spatial, temporal, and frequency-domain transformers.

In parallel, we convert the full sequence $\mathbf{x}$ into $\mathbf{C} \in \mathbb{R}^{F \times\ J \times 2}$ by means of the Discrete Cosine Transform (DCT). DCT is a mathematical technique used in signal processing and data compression, which converts a signal(e.g. an image or a time-series sequence) from its original spatial or temporal domain into the frequency domain, therefore expressing it as a sum of cosine waves oscillating at different frequencies. Then, we keep only the first $N \ll F$ coefficients $\mathbf{C}' \in \mathbb{R}^{N \times\ J \times 2}$ of $\mathbf{C}$ by applying a low-pass filter for every joint trajectory. These filtered coefficients preserve the majority of the temporal information of $\mathbf{x}$, while removing high-frequency noise. In fact, most of a human's actual motion data is concentrated in low frequencies. By retaining only a handful of low-frequency DCT coefficients, PoseFormerV2 is able to process much longer videos with a fraction of the computational cost of standard all-frame self-attention. Next,

the coefficients $\mathbf{C}'$ are flattened and linear projected to $\mathbf{z}^{Freq} \in \mathbb{R}^{N\times(J\cdot c)}$, which is summed with a learnable frequency positional embedding $\mathbf{E}_{FPos}$. The concatenation $\mathbf{z} = [\mathbf{z}^{Time}; \mathbf{z}^{Freq}]$ constitutes the input of the Time-Frequence Feature transformer.

In this module, the whole $\mathbf{z}$ goes through the same self-attention layer. However, $\mathbf{z}^{Time}$ and $\mathbf{z}^{Freq}$ use different feed-forward network. While $\mathbf{z}^{Freq}$ is fed into a simple MLP layer (specifically, Linear $\rightarrow$ GeLU $\rightarrow$ Linear layers), $\mathbf{z}^{Time}$ goes through a FreqMLP, where DCT and IDCT (Inverse Discrete Cosine Transform) are applied before and after the Multi-Layer Perceptron. In this phase, a blunt low-pass filter is not needed, as the sequence is already compressed and computational time has already been optimized; instead the objective is to remove noise while still maintaining eventual fast motions, which have high frequency.

Lastly, after concatenating again $\mathbf{z}^{Time}$ and $\mathbf{z}^{Freq}$, a 1D convolutional layer is used to gather temporal information, and subsequently a linear projection is applied, to obtain the final 3D pose representation $\mathbf{y} \in \mathbb{R}^{1\times(J\cdot 3)}$.

The overall results in a computational efficient and noise resistant model. The robustness and efficiency of PoseFormerV2 have enabled its adoption in real-world applications beyond traditional pose estimation benchmarks. Notably, Pepper-Pose [24], a companion robot system for full-body pose estimation, employs Pose-FormerV2 as its core pose estimation module. In this system, a mobile Pepper robot actively tracks users and refines its viewpoint to capture optimal pose estimations while they move and perform diverse actions in open spaces. The robot processes captured frames through PoseFormerV2 to obtain real-time 3D full-body pose estimates, demonstrating the architecture's capability to operate in unconstrained, real-world scenarios for applications in sports, fitness, and healthcare monitoring.

### 2.2.2 MotionBERT

While PoseFormerV2 optimizes the 2D-to-3D lifting task through frequency domain filtering, it processes spatial frames and global temporal frequencies as separate domains. MotionBERT [7], proposed by Zhu et al., approaches the same fundamental problem through the lens of a unified foundation model, treating space and time as deeply intertwined, parallel streams that continuously share information. The input for MotionBERT is the 2D skeleton sequence $\mathbf{x} \in \mathbb{R}^{R\times J\times C_{in}}$ -where $C_{in}$ represents the number of input- which is projected to a high dimensional feature $\mathbf{F}^0 \in \mathbb{R}^{T\times\ J\times C_f}$, to which we add a learnable spatial and termporal positional encoding, denoted respectively with $\mathbf{P}^T_{pos}, \mathbf{P}^S_{pos} \in \mathbb{R}^{1\times\ J\times C_f}$. Next, we feed this feature to the *Dual-stream Spatio-temporal Transformer* (DSTformer), a neural network whose objective is to capture the long-range relationship among skeleton points. The structure is as follows: DSTformer uses two parallel streams where a spatial and temporal block are stacked in different order. In one stream, the spatial block is followed by the temporal block, and is responsible of understanding how

human poses evolve over time. In the other stream, the order is reversed, focusing on analyzing how movement paths correlate spatially. The output features of the two branches are fused using adaptive weights predicted by an attention regressor. The dual-stream-fusion module is then repeated for N times.

Specifically, the Spatial block with Multi-Head Self-Attention (S-MHSA) models the relationship among the joints within the same time step. It temporarily slices $\mathbf{F}^0$ into $T$ per-frame spatial features $\mathbf{F}_S \in \mathbb{R}^{J \times C_e}$, where $C_e$ is the number of embeddings. Given the projection matrices, $\mathbf{W}_S^{(Q,i)}, \mathbf{W}_S^{(K,i)}, \mathbf{W}_S^{(Q,i)} \quad \forall i \in 1, \ldots, h =$ number of heads, we compute via self-attention query, key, and value for each head:

$$\mathbf{Q}_s^i = \mathbf{F}_S \mathbf{W}_S^{(Q,i)}, \mathbf{K}_s^i = \mathbf{F}_S \mathbf{W}_S^{(K,i)}, \mathbf{V}_s^i = \mathbf{F}_S \mathbf{W}_S^{(V,i)}, \qquad \forall i \in 1, \ldots, h$$

We then compute the heads

$$\text{head}_i = \text{softmax}\left(\frac{\mathbf{Q}_s^i (\mathbf{K}_s^i)'}{\sqrt{d_k}}\right)\mathbf{V}_s^i \qquad \forall i \in 1, \ldots, h \qquad d_k = \text{feature dimension of } \mathbf{K}_s$$

and the multi-head self-attention:

$$S - MHSA(\mathbf{Q}_s, \mathbf{K}_s, \mathbf{V}_s) = [\text{head}_1, \ldots, \text{head}_h]\mathbf{W}_s^P$$

We apply residual connection and layer normalization to the spatial block's output, followed by MLP and again residual connection and layer normalization.

The Temporal Multi-Head Self-Attention (T-MHSA) block follows a similar process as S-MHSA, with the difference that the per-joint temporal feature $\mathbf{F}_t \in \mathbb{R}^{T \times C_e}$ and parallelized over the spatial dimension.

The weights are determined by multiplying the output feature of each branch with a matrix that is updated during training.

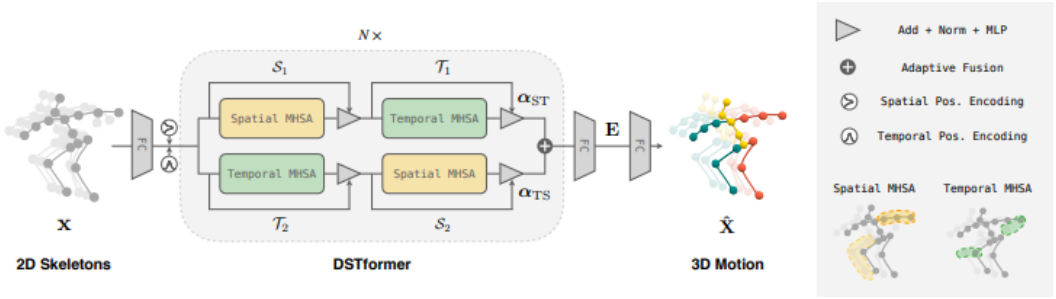



**Figure 2.5:** Overview of the MotionBERT architecture [7]. The framework processes 2D skeletal sequences through a Dual-stream Spatial-Temporal Transformer (DSTformer) to extract comprehensive kinematic representations. These pre-trained representations are then fine-tuned for downstream tasks such as 3D pose estimation and action recognition.

The motion representations acquired by MotionBERT incorporate geometric, and physical knowledge about human motion which can be easily transferred to various downstream tasks. Indeed, Zhu et al., demonstrated that rather than training isolated models from scratch, it is sufficient to add an extra linear layer or an MLP with one hidden layer, and then finetune network end-to-end.

In particular, for 3D human pose estimation, no additional head is added and the network is finetuned, where the input 2D skeletons are estimated from videos without the implementation of additional masks.

For the downstream task of action recognition, the network is trained with cross-entropy classification loss. A global average pooling operation is performed over the temporal and person dimensions, and the output is then fed into an MLP with one hidden layer.

Lastly, for human mesh recovery, SMPL model [20] is used for representing the human mesh, which consists of shape parameters $\beta \in \mathbb{R}^{10}$ and pose parameters $\theta \in \mathbb{R}^{72}$, and calculates the 3D mesh as $M(\theta, \beta) \in \mathbb{R}^{6890\times3}$. To regress the pose parameters for each frame, motion embeddings are fed to an MLP with one hidden layer sequence.

Across all three tasks, MotionBERT achieves state-of-the-art performance, demonstrating that the spatio-temporal representations learnt by DSTformer represent a general-purpose motion representation transferable to a broad range of human motions videos.

## 2.3 State of the Art in Computer Vision for Clinical Gait Assessment

The transition from restrictive, marker-based Motion Capture and wearable sensors to accessible, vision-based gait analysis has been a major focus of recent literature. Early vision-based approaches relied primarily on traditional, rule-based motion features and traditional machine learning pipelines.

Among early works, Sabo et al. [25] utilized the markerless motion capture sensor of Microsoft Kinect to analyze the gait of parkinsonian patients. From human pose tracking, they extracted a combination of 2D and 3D kinematic features. Using multivariate logistic regression, they managed to successfully correlate specific gait features with Parkinsonian severity scores, rather than training a predictive model.

Taking a different approach, Rupprechter et al. [26] leveraged on datasets collected using commercially available phones and tablet. To characterize the

gait, they manually defined five features: speed, step frequency, arm swing, postural control, and roughness of movement. By training an ordinal random forest classification model, evaluated via 10-fold cross-validation, they achieved reliable estimations of the UPDRS.

Despite the success of these early vision-based systems, their scalability and robustness have historically been bottlenecked by the scarcity of large-scale, annotated medical datasets. To address this limitation, Chavez et al. [27] proposed the generation of synthetic data through linear combinations of normal and Parkinsonian gaits, using these data to train and to investigate the performance of k-Nearest Neighbors (KNN), Support-Vector Machine (SVM), and Gradient Boosting (GB) algorithms, achieving exceptionally high accuracy in the severity classification task.

The emergence of deep learning-based motion encoders marked a paradigm shift in gait analysis by enabling automatic feature learning from raw skeletal data.

Dadashzadeh et al. [28] introduced PECoP (Parameter-Efficient Continual Pretraining), addressing the domain gap between general action recognition datasets and specialized action quality assessment tasks. Their approach inserts lightweight 3D-Adapter modules into pretrained 3D CNNs, and learns domain-specific spatiotemporal knowledge by means of self-supervised learning. During domain-specific pretraining, only the adapter parameters are updated while the original pretrained weights are frozen. This strategy demonstrated significant improvements on action quality assessment benchmarks, while reducing computational costs compared to full model fine-tuning. Moreover, the authors of PECoP introduced the PD4T dataset, consisting of videos of Parkinsonian patients performing gait, finger tapping, hand movement, and leg agility tasks annotated with UPDRS scores. Their work established that continual pretraining on domain-specific unlabeled data, when combined with parameter-efficient adaptation, can effectively bridge the gap between general action recognition and clinical assessment tasks.

In 2024, Adeli et al. [29] pioneered the application of general-purpose motion encoders within a clinical setting. Recognizing the success of transformer-based architectures, such as MotionBERT and PoseFormerV2, in general human motion analysis, they evaluated whether these pretrained models could effectively interpret complex gait patterns associated with Parkinson's disease without requiring manual feature engineering. Their study established a comprehensive benchmark by feeding skeletal sequences directly into pretrained motion encoders, followed by lightweight classification heads for severity prediction. This approach demonstrated that motion encoders pretrained on large-scale general human motion datasets (e.g., Human3.6M) could transfer to clinical domains, learning discriminative representations of parkinsonian movement without explicit programming of biomechanical

rules. The work provided meaningful insights into which encoder architectures and training strategies were most effective for clinical gait assessment.

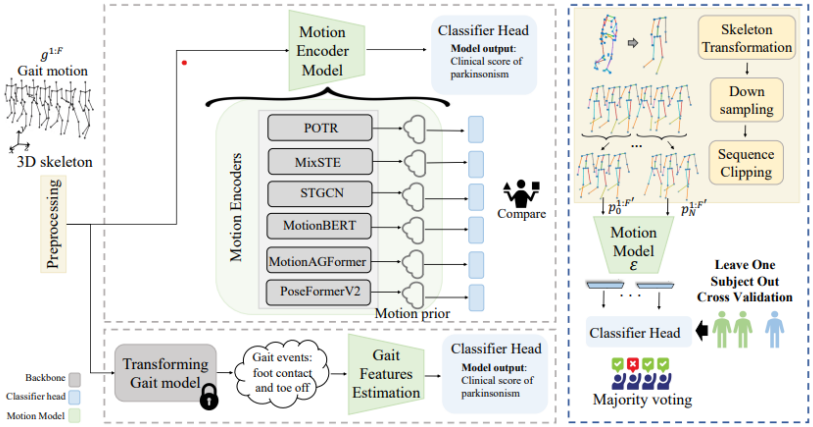



**Figure 2.6:** The benchmarking framework proposed by Adeli et al.(2024) [29], detailing the spatial alignment and multi-view projection pipeline. While their benchmark evaluates a broad spectrum of both video-based and skeleton-based architectures, this thesis specifically isolates and expands upon PoseFormerV2 and MotionBERT.

Building upon this foundation, in late 2025 Adeli et al. [9] introduced the Clinical Assessment Rating Encoder for Parkinson's Disease (CARE-PD), extending the motion encoder framework with improved preprocessing pipelines and systematic evaluation protocols. CARE-PD introduced 9 datasets collected from 8 clinical centres, where video sequences go through a harmonization process consisting of data cleaning, SMPL mesh extraction, and data anonymization. The study benchmarked multiple motion encoder architectures on the introduced datasets under four protocols (within-dataset, cross-dataset, leave-one-dataset-out, and multi-dataset in-domain adaptation), establishing performance baselines and best practices for clinical deployment.

This thesis builds upon the motion encoder framework established by Adeli et al. (2024) and extends it through the investigation of training strategies, multi-dataset integration, and parameter-efficient adaptation techniques. The architecture adaptation methodology and experimental protocols follow the framework introduced in their 2024 work, while our final benchmark comparison adopts the CARE-PD preprocessing pipeline to ensure fair evaluation against the current state-of-the-art.

# Chapter 3

# Methodology

## 3.1 Datasets

We performed our experiments using three main datasets:

- **AAP**: we denote with AAP a dataset of 145 recordings of 37 patients. The dataset was kindly offered by the association *Associazione Amici Parkinsoniani Piemonte*. Patients were captured walking towards a consumer-grade, static camera. However, 45 videos were discarded as no clinical evaluation was provided.

- **PD-GaM**: we denoted with PD-GaM the PD4T dataset provided by the authors of PECoP [28]. It consists of 426 videos of 30 patients, who were recorded from a left lateral perspective using a dynamically tracking camera. Subjects walked back and forth, resulting in four continuous walking segments per recording. To ensure patient safety, a medical staff member remained on standby throughout the trials, intervening strictly in the event of a potential fall.

- **BMCLab**: we denote with BMClab the dataset collected by the *Laboratory of Biomechanics and Motor Control at the Federal University of ABC* in Brazil [30]. The dataset includes 885 walking sequences using a 3D Motion-Capture system, where 26 patients walked barefoot. However, 3 patients were excluded from our analysis because of motion issues.

While the primary experimental evaluations were conducted on AAP, PD-GaM, and BMCLab datasets, our final benchmarking comparison against CARE-PD [9] uses the following datasets:

**T-SDU and T-LTC** [25][31] The datasets were collected during prospective observational studies on gait changes in a specialized dementia unit and a long-term care facility in Toronto, Canada. T-SDU and T-LTC consists respectively of 53 and 14 patients. The data was recorded via ceiling-mounted cameras, and the sequences were curated by the benchmark authors to retain only clean walking segments, explicitly excluding turns and stops.

**T-SDU-PD** A subset of T-SDU consisting of 14 patients is defined, where gait annotations and clinical evaluations are defined.

**DNE** [32][33] DNE was collected across multiple clinical sites in the United States using smartphone video recordings for neurological assessments, with a total of 97 patients. We utilized the stand-up and walk task subset, where participants walked back and forth. Similar to PD-GaM, the benchmark focuses exclusively on clean, extracted walking segments with per-walk labels.

**3DGait** [34] This dataset of 43 patients was collected at the University of Strasbourg, France, using an RGB camera as patients walked across an 8-meter GAITRite walkway. We used the specific subset of trials featuring UPDRS-labeled straight walks from individuals with Parkinson's disease.

**KUL-DT-T** [35][36] The dataset was recorded at the Movement Disorders Clinic of the University Hospital Leuven, Belgium. An 8-camera 3D optical motion capture system tracking 34 markers was used to record 29 patients.

**E-LC** [37][38] The E-LC dataset, curated by the e Emory Movement Disorders Clinic (Atlanta, USA) uses A 14-camera 3D optical motion capture system with 60 markers. It consists of 59 patients.

Among these datasets, only AAP, PD-GaM, BMClab, T-SDU-PD, and 3DGait contain severity labels. In particular, only AAP, PD-GaM, and 3DGait have patients whose Parkinson severity has been labeled 3. Let us note that the presence of few sequences with severity 3 and the complete lack of sequences labeled as severity of 4 is motivated by the fact that such severity implies severe walking impairment, to the point that the subject is unable to walk even short distances without external support.

## 3.2 Data pre-processing

The following operations were applied before feeding the data into the motion encoders.

**Conversion to skeletal data.** BCMLab, already had skeletal data, as it is a MoCap based-system where 26 markers were used. Conversely, because AAP and PD-GaM are RGB video datasets, we extracted skeletal data using YOLOpose for pose estimation.

**Integration of SMPL Features for Benchmarking.** To ensure a fair and comprehensive benchmarking of our proposed architecture against CARE-PD [9], we implemented a secondary preprocessing pipeline to extract and handle SMPL data from the RGB videos. Specifically, to ensure strict patient privacy and data anonymization, the extracted SMPL body shape coefficients, represented by the $\beta$ parameters, were systematically set to zero. This operation strips away specific physical features, such as height and body proportions, that could retrace back to the patient.

**Conversion to clear, patient-focused sequences.** Our YOLOpose implementation demonstrated strong performance, with bounding box confidence scores exceeding 90%. However, some sequences exhibited flickering, due to missing pose estimations between frames. To mitigate this issue, third-order polynomial interpolation was applied to the joint coordinates. Moreover, two aspects were addressed. First, in the AAP dataset, patients walked directly toward the camera, meaning that by the end of the sequences, only their torso remained visible. While YOLOpose is capable of estimating obstructed keypoints, this feature relies on training data primarily derived from healthy subjects. Inferring joint positions becomes significantly more complex when limbs are bilaterally occluded, as a joint's position cannot be deduced from its symmetric counterpart. Because YOLOpose was not trained on sufficient clinical data to accurately represent pathological movements (such as Parkinsonian stride or Freezing of Gait) we chose not to rely on this feature. Consequently, we trimmed each video to the longest continuous sequence in which all joints remained visible. Second, within the PD-GaM dataset, YOLOpose tracked also the medical staff assisting the patients during the tasks. To isolate the patient's movements, we assigned unique subject IDs during the tracking phase. We then filtered out the pose data associated with any ID exhibiting minimal movement throughout the sequence, effectively removing the staff members from the analysis.

**Conversion to H36M format.** Both PoseFormerV2 and MotionBERT require skeletal input data in the Human3.6M (H36M) format. However, our datasets use different conventions: YOLOpose outputs data in the COCO format, while the BMCLab dataset provides 44 distinct biomarkers (detailed in the Appendix). Therefore, a joint mapping operation is necessary to harmonize the data. While some joints share the exact same position across formats (e.g., the knees) or simply require renaming (e.g., mapping COCO ankles to H36M feet), the positions of completely missing joints must be geometrically derived. The complete mapping and specific calculations used to compute these missing joints are detailed in Table 3.1.

**Class imbalance mitigation.** The underrepresentation of class of severity 3 tipically leads to a classifier with poor generalization. Considering that class 2 and class 3 exhibit similar pathological features, we addressed this class imbalance by merging class 3 into class 2 during the data loading phase, allowing to mantain statisical robustness. As a result, the model evaluates a three-class framework, which now can be interpreted as light (0), mild (1), and severe (2).

## 3.3 Architecture adaptation

Let us recall that PoseformerV2 and MotionBERT are designed with 2D-to-3D pose lifting as a core task. Hence, in order to adapt them for the downstream task of PD's evalutation, we adopt the benchmarking framework introduced by Adeli et al. (2024) [29]. That is, given the feature embedding extracted by the motion encoders

$$e_i = E(p_i^{1:F'})$$

where $p_i^{1:F'}$ is a preprocessed sequence clip of length $F'$, a classifier head $C$ is added to map the embedding to UPDRS-gait score prediction.

$$\text{Score}_i = C(e_i)$$

## 3.4 Pretraining: Kinematic-Aware Taylor Maskin

To explicitly enforce the learning of higher-order movement dynamics, we implemented a Taylor Reconstruction Loss, adapting the temporal modeling concepts introduced by Wang L. et al [13].

Let us recall that in mathematical analysis, the Taylor series of a function $f(x)$ is the sum of of terms that are expressed in terms of the function's derivatives at a single point $x_0$. When the series is truncated at the $n$-th degree, the resulting

**Table 3.1:** Mapping from PD markers and COCO keypoints to Human3.6M (H36M) keypoints.

| **H36M Keypoint** | **Derivation from PD Markers** | **Derivation from COCO Keypoints** |
|---|---|---|
| Bottom Torso | $\frac{\text{L. Asis+R. Asis}}{4} + \frac{\text{L. Psis+R. Psis}}{4}$ | $\frac{\text{Left Hip+Right Hip}}{2}$ |
| Left Hip | $\frac{\text{Left Asis+Left Psis}}{2}$ | Left Hip |
| Left Knee | Left Knee | Left Knee |
| Left Foot | Left Ankle | Left Ankle |
| Right Hip | $\frac{\text{Right Asis+Right Psis}}{2}$ | Right Hip |
| Right Knee | Right Knee | Right Knee |
| Right Foot | Right Ankle | Right Ankle |
| Upper Torso | $\frac{\text{C7+Clav}}{2}$ | $\frac{\text{Left Shoulder+Right Shoulder}}{2}$ |
| Central Torso | $\frac{\text{Strn+T10}}{2}$ | $\frac{\text{Bottom Torso+Upper Torso}}{2}$ |
| Right Shoulder | Right Shoulder | Right Shoulder |
| Right Elbow | $\frac{\text{Right El+Right Em}}{2}$ | Right Elbow |
| Right Hand | $\frac{\text{Right Wl+Right Wm}}{2}$ | Right Wrist |
| Left Shoulder | Left Shoulder | Left Shoulder |
| Left Elbow | $\frac{\text{Left El+Left Em}}{2}$ | Left Elbow |
| Left Hand | $\frac{\text{Left Wl+Left Wm}}{2}$ | Left Wrist |
| Neck | Upper Torso + $[0.27, 57.48, 11.44]$ | $\frac{\text{Upper Torso+Nose}}{2}$ |
| Head | Upper Torso + $[-2.07, 165.23, 34.02]$ | Nose |

polynomial is an accurate approximation of the function in the neighbourhood of $x_0$, with an approximation error of $o((x - x_0)^n)$.

$$f(x) = \sum_{k=0}^{n} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k + o((x - x_0)^n)$$

Building upon this concept, the original authors used a Taylor series expansion to approximate an implicit motion-extraction function, aiming to isolate dominant motion patterns from temporal sequences and filter out static elements. In their

work, they applied this to grayscale video temporal blocks, extracting motion by calculating iterative finite differences of pixel intensities (where difference maps, velocity, and acceleration were defined as first, second, and third-order differences, respectively). In our work, since we operate directly on skeletal joint coordinates, we adapt their mathematical framework to align with standard Newtonian kinematics. This allows us to create a domain-specific pretraining objective where the finite differences correspond to physically meaningful kinematic properties of the human body.

This choice of reconstruction targets is tightly coupled with our masking strategy. Rather than masking individual joints, which allows the model to trivially reconstruct missing values by attending to the same joint in adjacent visible frames [39], we mask 50% of the frames. This design is consistent with the spatial-temporal masking paradigm introduced in skeleton-based masked autoencoders [40], and ensures that the encoder infers missing frames from the global motion context, making velocity and acceleration reconstruction meaningful pretraining signals.

Let be predicted sequence tensor $\hat{\mathbf{P}}$ the predicted sequence tensor and let be the target sequence tensor $\mathbf{P}$ the target sequence tensor of shape $(B, T, J, C)$ (Batch, Frames, Joints, Channels). We define the positional loss $L_{pos}$ as the standard Mean Squared Error (MSE) of the zeroth-order spatial coordinates:

$$L_{pos} = \text{MSE}(\hat{\mathbf{P}}, \mathbf{P})$$

We then compute the first-order finite differences to represent the velocity $\mathbf{V}$, where $\mathbf{V}_t = \mathbf{P}_{t+1} - \mathbf{P}_t$. The velocity reconstruction loss is calculated as:

$$L_{vel} = \text{MSE}(\hat{\mathbf{V}}, \mathbf{V})$$

To capture the smoothness and force changes in the movement, we compute the second-order finite differences to represent the inter-frame acceleration $\mathbf{A}$, where $\mathbf{A}_t = \mathbf{V}_{t+1} - \mathbf{V}_t$. The acceleration reconstruction loss is:

$$L_{acc} = \text{MSE}(\hat{\mathbf{A}}, \mathbf{A})$$

We define the pretraining objective as a weighted combination of these kinematic levels:

$$\mathcal{L}_{Taylor} = \lambda_0 L_{pos} + \lambda_1 L_{vel} + \lambda_2 L_{acc}$$

where $\lambda_0, \lambda_1$, and $\lambda_2$ are empirically chosen hyperparameters. By explicitly weighting these derivative terms, the network is forced to learn a representation space that is highly sensitive to abrupt shifts in movement speed and fluidity—features that are paramount for accurate clinical motor assessment in Parkinson's disease.

## 3.5 End-to-end finetuning

To adapt the pre-trained models to the downstream task of PD severity estimation, an end-to-end fine-tuning approach is considered. End-to-end fine-tuning unfreezes the parameters within the pre-trained backbone ($\Theta_{backbone}$) and the newly initialized classification head ($\Theta_{head}$).During the backward pass, the optimization is governed bythe following loss function:

$$\mathcal{L}_{total} = \mathcal{L}_{cls}(Y, \hat{Y}) + \gamma \|\Theta\|_1$$

where $\mathcal{L}_{cls}$ is the weighted Cross-Entropy loss between the ground truth severity labels $Y$ and the model predictions $\hat{Y}$, while $\|\Theta\|_1$ is an optional $L_1$ regularization term, scaled by $\gamma$, applied to all trainable parameters to encourage weight sparsity and prevent overfitting on small clinical datasets. The gradients of the loss, $\nabla\mathcal{L}_{total}$, are backpropagated through the classification head and flow continuously into the transformer blocks. The AdamW optimizer updates the weights. By applying this strategy, the model shifts its focus away from general action recognition and learns to specifically target and evaluate Parkinsonian movement patterns.

## 3.6 LoRA finetuning

While full end-to-end fine-tuning maximizes architectural flexibility, it introduces a risk of overfitting. In light of these observation, our work additionally explores Low-Rank Adaptation (LoRA) [41] as an alternative optimization strategy. LoRA freezes the pre-trained transformer weights and injects trainable low-rank decomposition matrices into the self-attention layers. This approach reduces the computational costs and preserves the network's foundational understanding of human biomechanics trainable parameter count by orders of magnitude. Therefore, LoRA acts as a strong structural regularizer, stabilizing the optimization process and enabling the extraction of subtle Parkinsonian kinematic features without sacrificing the robust, generalized knowledge learnt during pre-training.

## 3.7 Metrics

To asssess the performance and generalization capabilities of the proposed architecture, the following metrics are considered:

**Precision.** The ratio of correctly predicted positive cases to the total predicted positives.

**Recall.** It calculates the ratio of correctly predicted positive cases to the total number of actual positive cases.

**F1-score.** The harmonic mean of precision and recall. Because our clinical datasets suffer from class imbalance, we chose the weighted F1-score as the primary classification metric. This ensures that the performance on the minority classes is accurately represented without allowing the majority class to artificially inflate the overall score.

**Spearman correlation.** Although the usual evaluation metrics consider the severity levels (0, 1, 2) to be completely independent, the UPDRS gait scores are actually ordinal because they represent a progressive deterioration of the motor impairment (Light, Mild, Severe). For example, a Severe (2) case being predicted as Light (0) is a much more severe error than a Severe (2) case being predicted as Mild (1). To capture this progressive relationship, we employ the Spearman Rank Correlation Coefficient $\rho$, which assesses monotonic relationships by comparing the ranks of the predicted and ground truth values. This makes it mathematically ideal for evaluating ordinal clinical scales. A high Spearman score indicates that as the true severity increases, the model's predicted severity reliably increases as well. Alongside the correlation coefficient, the $p$-value is computed to assess statistical significance. A $p$-value below the standard threshold $p < 0.05$ allows us to reject the null hypothesis, confirming that the observed monotonic relationship between the model's predictions and the clinical ground truth did not occur by random chance.

**t-SNE plots.** Besides using quantitative metrics, it is essential to ensure that the network is learning generalized and domain-independent pathological features, as opposed to memorizing dataset-specific backgrounds and camera views. To visually assess this, we examine the embeddings of the data before they have been fed into the classification head. Using t-Distributed Stochastic Neighbor Embedding (t-SNE), the 512-dimensional latent representations are projected into a 2D scatter plot. The benefit of this visual evaluation is to further understand the magnitude of the performance strength model. In order to effectively evaluate the domain shift, the internal weights of the fully trained network are frozen, and forward inference pass is then performed on both the training and testing data. In a successful model, the t-SNE projection will reveal distinct spatial clusters corresponding to the target severity classes, proving the feature extractor has learned to separate patients based on clinical impairment, without making distinction based on the dataset to which they belong, hence, highlighting robust generalization skills.

# Chapter 4

# Experiments and Results

This chapter presents the experimental configurations to assess the models' generalization ability. For the sake of brevity, only the most representative latent space visualizations are presented in this section.

## 4.1 Cross-Dataset Generalization and In-Domain Adaptation

In the initial phase of our experiments we investigate the generalization capabilites of the motion-encoders across different datasets.

Following the benchmarking established by Adeli et al. [29], we selected the three highest-performing -in terms of weighted F1-score- model configurations as our baselines: MotionBERT, PoseFormerV2, and PoseFormerV2 finetuned. Two scenarios can be distinguished:

**Strict Cross-dataset Generalization.** The model is completely blind to target test domain during training. This evaluates zero-shot transferability of the extracted features to an unseen dataset. Hence, we ran the following experiments.

- Training on BMClab and AAP, and testing on the PD-GaM's test split.
- Training on BMClab and PD-GaM, and testing on the AAP's test split.

**In-Domain Adaptation.** In this scenario, a portion of the target domain is introduced into the training phase, In this way, it is possible to analyze how much the models benefit from seeing domain-specific instances. Therefore, we ran the following experiments:

- Training on BMClab, AAP, and PD-GaM's train split, and testing on the PD-GaM's test split.
- Training on BMClab, PD-GaM, and AAP's train split, and test on the AAP's test split.

To prevent data leakage, the train and test sets are partitioned strictly at the patient level. The AAP dataset is divided into 22 training patients and 9 testing patients, which corresponds to a sequence-level split of 70% for training and 30% for testing. Similarly, the PD-GaM dataset is split into 22 training patients and 8 testing patients, yielding a sequence distribution of 73% and 27%, respectively. On the other hand, throughout all configurations, the BMCLab dataset is used exclusively for training. This methodological decision is driven by two primary considerations. First, Spatio-Temporal Transformers need massive amount of data in order to obtain good performances; because BMCLab contains the highest volume of pose sequences among the three datasets, it provides the critical mass of data necessary to stabilize optimization and prevent overfitting. Second, BMCLab is the only dataset in this study acquired via Motion Capture (MoCap). These pristine representations of human kinematics allow the network to establish an ideal, mathematical baseline for gait. Simultaneously, injecting the RGB-based datasets (AAP and PD-GaM) during training exposes the model to the natural jitter and estimation errors inherent to YOLOpose. This hybrid approach teaches the network to effectively distinguish between actual pathological movement deviations and mere recording artifacts

**Table 4.1:** Quantitative performance metrics evaluating cross-dataset generalization and in-domain adaptation. In order, weighted Precision, weighted Recall, weighted F1-score, Spearman correlation and its p-value are shown. The training configurations explicitly distinguish between the training and testing partitions of the target datasets.

| Model | Train Set | Test Set | Prec | Rec | W-F1 | Spearman ($\rho$) | $p$-value |
|---|---|---|---|---|---|---|---|
| **MotionBERT** | BMCLab + PD-GaM | AAP (Test) | 0.07 | 0.27 | 0.12 | 0.2286 | 0.2244 |
| | BMCLab + PD-GaM + AAP (Train) | AAP (Test) | 0.07 | 0.27 | 0.12 | 0.2286 | 0.2244 |
| | BMCLab + AAP | PD-GaM (Test) | 0.11 | 0.21 | 0.10 | -0.2217 | 0.0167 |
| | BMCLab + AAP + PD-GaM (Train) | PD-GaM (Test) | 0.66 | 0.66 | **0.66** | **0.6305** | 3.35E-14 |
| **PoseFormerV2** | BMCLab + PD-GaM | AAP (Test) | 0.62 | 0.37 | 0.29 | 0.3155 | 0.08944 |
| | BMCLab + PD-GaM + AAP (Train) | AAP (Test) | 0.50 | 0.43 | 0.29 | 0.2286 | 0.2244 |
| | BMCLab + AAP | PD-GaM (Test) | 0.21 | 0.32 | 0.22 | 0.3135 | 0.0006 |
| | BMCLab + AAP + PD-GaM (Train) | PD-GaM (Test) | 0.55 | 0.53 | 0.54 | 0.5255 | 1.39E-06 |
| **PoseFormerV2 (FT)** | BMCLab + PD-GaM | AAP (Test) | 0.28 | 0.30 | 0.18 | 0.1480 | 0.435 |
| | BMCLab + PD-GaM + AAP (Train) | AAP (Test) | 0.49 | 0.37 | 0.30 | 0.1790 | 0.344 |
| | BMCLab + AAP | PD-GaM (Test) | 0.41 | 0.41 | 0.34 | 0.2279 | 0.0138 |
| | BMCLab + AAP + PD-GaM (Train) | PD-GaM (Test) | 0.55 | 0.53 | 0.54 | 0.5051 | 7.36E-09 |

Table 4.1 describes the comparative performance of the three baseline architectures. The results reveal a stark contrast in model behavior depending on the

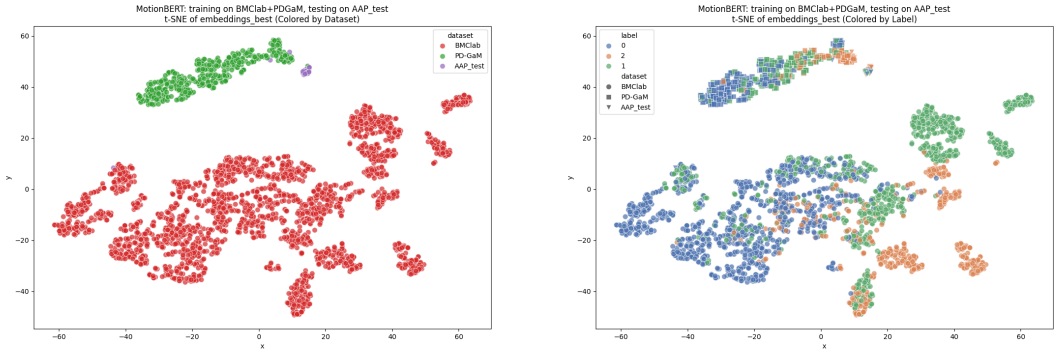



**Figure 4.1:** MotionBERT Cross-Dataset Evaluation on AAP.

target domain, heavily underscoring the complexities of clinical domain shift.

The AAP dataset demonstrated to be a highly challenging target domain, with all architectures struggling to achieve consistent performance in the zero-shot evaluation. Most notably, none of the models achieved a statistically significant Spearman correlation when evaluated on AAP, regardless of whether in-domain training data was provided. Furthermore, the introduction of AAP training data completely failed to shift the classification equilibrium for the base architectures, as both MotionBERT and PoseFormerV2's Weighted F1-scores remained stagnate at 0.12 and 0.29 respectively. When visually analysing the zero-shot evaluation on AAP in MotionBERT (Figure 4.1), we can see that the left t-SNE plot, colored by source dataset, reveals the unresolved domain gap. There is a sharp separation between the MoCap data (BMCLab) and the video-based RGB data (PD-GaM and AAP).

However, when coloring the latent space by ground truth clinical severity (Figure 4.1, right) we can see that although the global division of classes is not sharp, a visible ordinal gradient is still present within the isolated dataset clusters. Indeed, on both MoCap and RGB clusters, it is visible a transition from healthy and mild cases (class 0) on the left, to moderate cases (class 1) in the center, toward severe cases (class 2) on the right. This indicates that while the model successfully learns to represent the progressive severity of PD, it is penalized by the domain shift.

The fine-tuned PoseFormerV2 saw an improvement of the weighted F1-score from 0.18 to 0.30 with in-domain data, though ordinal ranking capability remained statistically insignificant ($\rho = 0.1790$, $p = 0.344$).

Conversely, the architectures showed good adaptability when evaluated on the PD-GaM dataset. Although zero-shot cross-dataset evaluation on PD-GaM yielded poor results across the models, with MotionBERT exhibiting negative correlation ($\rho$ = -0.2217), the introduction of in-domain training data triggered a massive performance improvement. In particular, MotionBERT emerged as the

best architecture in this scenario, achieving a weighted F1-score of 0.66, proving its ability to effectively handle class imbalances, while also achieving a highly significant Spearman correlation of $\rho = 0.6305$ ($p < 0.001$). This indicates that, given a small amount of in-domain calibration, MotionBERT is highly capable of bridging the domain gap between MoCap data and RGB videos in the clinical context of Parkinsonian gait assessment.

The t-SNE plots (Figure 4.2) of this top-performing configuration, illustrates how the model surpasses the severe MoCap-to-RGB domain gap. The dataset projection (left) shows that while some BMCLab data remains isolated, a tightly integrated manifold emerges for the RGB data. In particularly, PD-GaM test and training samples perfectly intertwine, confirming the network leveraged the PD-GaM training data to align unseen sequences in the same feature space. In the t-SNE plot colored by label, the BMCLab embeddings are heavily fractured into distinct groups of severity, likely due to the high precision of MoCap data allowing the network to separate individual subjects. However, within the targeted PD-GaM and AAP video cluster (lower-left), the embeddings form a smooth transition from mild to severe cases.

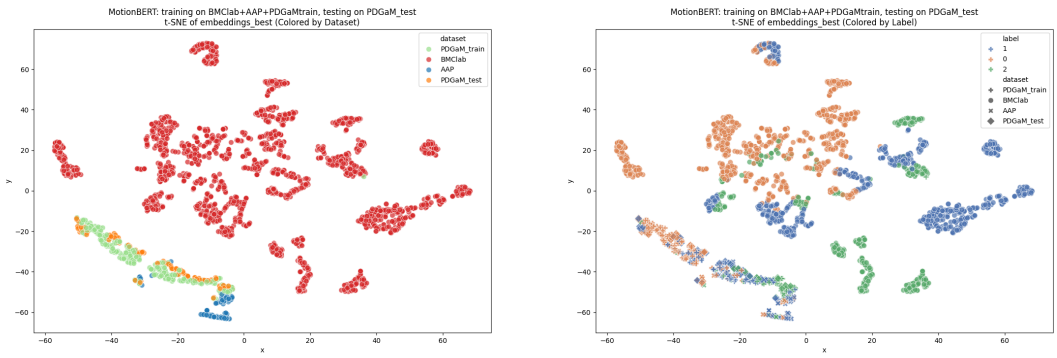



**Figure 4.2:** MotionBERT In-Domain Adaptation on PD-GaM.

The PoseFormerV2 variants also showed robust, statistically significant adaptation on PD-GaM ($\rho > 0.50, p < 0.001$), though they do not to match MotionBERT's quality of performance.

## 4.2 Implementation of Self-Supervised Learning with Taylor Pre-training

The next phase of experiments investigated the impact of Self-Supervised Learning (SSL) on the motion encoders, where all baseline architectures were evaluated using a self-supervised Taylor pre-training strategy.

Let us recall that in Taylor pre-training, the networks are tasked with predicting masked joints, velocity, and acceleration at frame-level.
The hyperparameters $\lambda_0, \lambda_1, \lambda_2$ (corresponding to $L_{pos}, L_{vel}, L_{acc}$) were empirically chosen based on the characteristics of each backbone. Since these losses operate at different scales and temporal granularities, different weights were needed to ensure the velocity and acceleration reconstructions are not overshadowed by the easier positional reconstruction.

We observed that MotionBERT and Poseformerv2 manifest different sensitivities to temporal noise.

Considering MotionBERT's proficiency in capturing motion context, there was a possibility of over-smoothing the local kinematics, therefore, we assigned a high weight to position and velocity ($\lambda_0 = 10.0$, $\lambda_2 = 8.0$). Conversely, a relatively lower weight was assigned to the acceleration ($\Lambda_2 = 3.0$), as the model might otherwise overfit to the jitter found in the noisy RGB-to-Pose skeletal data.

In contrast, we discovered that PoseFormerV2's representation tends to be dominated by the positional term. For this reason, we heavily penalize velocity and acceleration error, leading to $\vec{\lambda} = [1, 20.0, 10.0]$.

Furthermore, given that MotionBERT demonstrated the most robust performance in the previous experiments, in this setting we also investigate the performance of the fine-tuned variant of the MotionBERT architecture. This inclusion allows to determine whether combining SSL pre-training with targeted clinical fine-tuning can push its already superior classification and ranking capabilities even further.

Table 4.2 summarizes the results of the self-supervised Taylor pre-training pipeline.

The fine-tuned MotionBERT architecture exhibited extreme behavior, demonstrating both the highest performance and catastrophic failures. The model collapsed in three out of four configurations, yielding NaN for the Spearman correlation. Such scenario occurs when the predicted variable has zero variance, i.e. the model adopted a fallback strategy consisting in predicting a single class for all patients.

By contrast, the same architecture obtained the best overall performance, specifically in the in-domain adaptation of PD-GaM. Although it achieved the same weighted F1-score of 0.66 as the best performer of the previous setting, it established a strong and highly Spearman correlation, yielding $\rho = 0.6699$ ($p < 0.0001$).

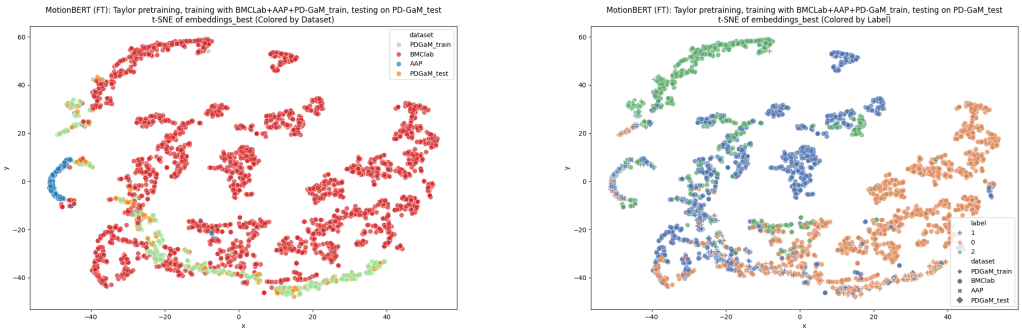
**Table 4.2:** Results of Taylor pretraining pipeline

| Model | Train Set | Test Set | Prec | Rec | F1 | Spearman ($\rho$) | $p$-value |
|---|---|---|---|---|---|---|---|
| **MotionBERT** | BMCLab + PD-GaM | AAP (Test) | 0.41 | 0.30 | 0.18 | 0.2286 | 0.2244 |
| | BMCLab + PD-GaM + AAP (Train) | AAP (Test) | 0.54 | 0.30 | 0.22 | 0.1092 | 0.5656 |
| | BMCLab + AAP | PD-GaM (Test) | 0.39 | 0.39 | 0.23 | 0.1309 | 0.1615 |
| | BMCLab + AAP + PD-GaM (Train) | PD-GaM (Test) | 0.53 | 0.52 | 0.46 | 0.5267 | 1.254E-09 |
| **MotionBERT (FT)** | BMCLab + PD-GaM | AAP (Test) | 0.07 | 0.27 | 0.11 | NaN | NaN |
| | BMCLab + PD-GaM + AAP (Train) | AAP (Test) | 0.16 | 0.4 | 0.23 | NaN | NaN |
| | BMCLab + AAP | PD-GaM (Test) | 0.14 | 0.38 | 0.21 | NaN | NaN |
| | BMCLab + AAP + PD-GaM (Train) | PD-GaM (Test) | 0.67 | 0.66 | **0.66** | 0.6699 | 1.978e-16 |
| **PoseFormerV2** | BMCLab + PD-GaM | AAP (Test) | 0.41 | 0.27 | 0.28 | -0.383 | 0.0367 |
| | BMCLab + PD-GaM + AAP (Train) | AAP (Test) | 0.43 | 0.43 | 0.29 | 0.2514 | 0.1802 |
| | BMCLab + AAP | PD-GaM (Test) | 0.3 | 0.42 | 0.30 | 0.315 | 0.00057 |
| | BMCLab + AAP + PD-GaM (Train) | PD-GaM (Test) | 0.51 | 0.44 | 0.34 | 0.4169 | 3.24E-06 |
| **PoseFormerV2 (FT)** | BMCLab + PD-GaM | AAP (Test) | 0.27 | 0.27 | 0.24 | -0.205 | 0.9142 |
| | BMCLab + PD-GaM + AAP (Train) | AAP (Test) | 0.67 | 0.5 | 0.44 | 0.3919 | 0.03218 |
| | BMCLab + AAP | PD-GaM (Test) | 0.22 | 0.35 | 0.27 | 0.1758 | 0.05901 |
| | BMCLab + AAP + PD-GaM (Train) | PD-GaM (Test) | 0.62 | 0.62 | 0.61 | 0.5972 | 1.48E-12 |

However, analysis of the base MotionBERT architecture reveals a more nuanced impact of the Taylor pretraining. The new approach improved the zero-shot cross-dataset generalization. In particular, when trained on BMCLab and AAP and tested on PD-GaM's test portion, the weighted F1-score more than doubled from 0.10 to 0.23. whereas when also injecting the PD-GaM's training portion, the weighted F1-score decreased from 0.66 to 0.46.
This dualism suggests that the Taylor reconstruction loss forces the network to learn generalized kinematic rules that aid in cross-domain transfer, but this same strictness may constrain the model's capacity to fit the in-domain features when considerable amount training data is available. In effect, Taylor pretraining acts as a strong regularizer that sacrifices peak in-domain performance for better generalization.

Figure 4.3 shows t-SNE projections of the learned representations from the best-performing configuration (MotionBERT FT, F1=0.66, $\rho = 0.6699$). The left panel, colored by dataset, reveals that while the encoder preserves dataset identity for the AAP dataset, the BMCLab and PD-GaM samples are well-intertwined, indicating signs of good generalization.

The right panel, colored by severity labels, explains the strong Spearman correlation observed. The embedding space preserves information regarding the progression of disease severity. The regions with class boundaries overlap reflect the light motor impairments are inherently ambiguous and hence it remains a challenge




**(a)** Colored by dataset

**(b)** Colored by severity label

**Figure 4.3:** t-SNE visualization of motion encoder representations for MotionBERT (FT) with Taylor pretraining in the best-performing configuration (F1=0.66, $\rho = 0.6699$). Left: samples colored by source dataset. Right: samples colored by Parkinson's severity label.

for both machine and clinicians.

By contrast, PoseFormerV2 architecture exhibited a much more stable relationship with the Taylor pretraining, as in no configuration the model collapsed. In terms of weighted F1-score, the performance on AAP test configurations remained stable (F1 $\approx$ 0.29).
When tested on PD-GaM, a similar regularizing effect seen in MotionBERT can be observed: in the zero-shot evaluation, weighted F1-score and Spearman correlation increased from 0.22 to 0.30 and from 0.2286 to 0.2514 respectively, while still maintaining a very significant p-value of 0.00057; in the in-domain adaptation, weighted F1-score dropped from 0.54 to 0.34, and $\rho$ decreased from 0.65255 to 0.4164, though the p-value remained highly significant.

However, the fine-tuned PoseFormerV2 proved to mainly benefit from the Taylor features. When the target evaluation dataset is included during training, the Taylor pretraining consistently boosts performance. For example, when testing on the AAP dataset, the F1-score improved from 0.30 to 0.44, and resulted in a significant Spearman correlation ($\rho = 0.3919$, p-value $= 0.032$). Similarly, when testing on PD-GaM, the F1-score increased from 0.54 to 0.61 and the correlation increased from $\rho = 0.5051$ to $\rho = 0.5972$ ($p < 0.0001$). Nonetheless, this improvement comes at the cost of zero-shot generalization to unseen datasets. Though zero-shot evaluation on AAP improved the F1-score from 0.18 to 0.24, it yielded negative Spearman correlation, whereas zero-shot evaluation on PD-GaM the weighted F1-score declined from 0.34 to 0.27.

In conclusion, Taylor pretraining is not a universally optimal strategy, but rather a highly context-dependent approach that dictates a strict trade-off between generalization and specialization. For base models, the Taylor reconstruction enhances cross-dataset generalization at the expense of in-domain fitting capacity. Conversely, task-specific fine-tuning shifts this dynamic towards hyper-specialization; it enables peak in-domain performance but severely penalizes zero-shot adaptability.

## 4.3 Dataset Composition Analysis: The Role of AAP in Training

In this section, we address a methodological consideration regarding the composition of the AAP dataset. The previous experiments consistently demonstrated poor performance when testing on AAP. However, the more important consideration is whether including AAP in the training set provides beneficial cross-dataset regularization or introduces noise that contaminates the learning process for PD-GaM assessment instead.

In light of these considerations, we compare configurations with and without AAP across three experimental settings, all exclusively testing on PD-GaM. First, we conduct standard training comparisons with baseline architectures. Second, we evaluate whether AAP's contribution changes when models undergo Taylor self-supervised pretraining. Third, we investigate a decoupled pretraining-training approach, where Taylor pretraining is performed on BMCLab and AAP, but trained on. This final configuration tests whether AAP benefits general kinematic learning during pretraining but should be excluded during task-specific adaptation.

**Table 4.3:** Impact of AAP dataset on training: All configurations tested on PD-GaM

| Model | Train Set | Prec | Rec | F1 | Spearman ($\rho$) | $p$-value |
|---|---|---|---|---|---|---|
| **MotionBERT** | BMCLab | 0.14 | 0.38 | 0.21 | NaN | NaN |
| | BMCLab + AAP | 0.06 | 0.24 | 0.09 | NaN | NaN |
| **MotionBERT (FT)** | BMCLab | 0.32 | 0.41 | 0.36 | NaN | NaN |
| | BMCLab + AAP | 0.14 | 0.38 | 0.21 | NaN | NaN |
| **PoseFormerV2** | BMCLab | 0.05 | 0.22 | 0.09 | -0.2286 | 0.01356 |
| | BMCLab + AAP | 0.11 | 0.21 | 0.10 | -0.2284 | 0.01366 |
| **PoseFormerV2 (FT)** | BMCLab | 0.06 | 0.24 | 0.09 | NaN | NaN |
| | BMCLab + AAP | 0.21 | 0.25 | 0.12 | 0.0162 | 0.8626 |

Table 4.3 reveals that introducing AAP in cross-dataset evaluation worsen the performance or provides negligible benefit. For MotionBERT, adding AAP causes severe performance collapse: the base model drops from a weighted F1-score of 0.21 to 0.09, while the fine-tuned variant reduces from a weighted F1-score of 0.36 to 0.21. In particular, in all settings MotionBERT produce NaN Spearman correlations, indicating the model predicts a single class for all samples regardless of actual severity.

PoseFormerV2 demonstrates greater stability, though no evident benefit can be attributed to AAP. In the base architecture, injecting AAP slightly improves the weighted F1-score (from 0.09 to 0.10), though it has not resolved the negative Spearman correlations issue. The fine-tuned variant shows very marginal improvement with AAP (weighted F1 score from 0.09 to 0.12) but produces near-zero correlation ($\rho = 0.016$, $p = 0.86$), indicating predictions bear no meaningful relationship to disease severity.

These results provide strong evidence against including AAP in training for PD-GaM assessment, as AAP introduces distribution shift that degrades generalization. The consistent negative impact across both architectures and training strategies suggests incompatibility between AAP and PD-GaM, likely derived from differences in recording protocols, patient populations, or severity annotation procedures.

**Table 4.4:** Impact of AAP with Taylor pretraining: All configurations tested on PD-GaM

| Model | Train Set | Prec | Rec | F1 | Spearman ($\rho$) | $p$-value |
|---|---|---|---|---|---|---|
| **MotionBERT + Taylor** | BMCLab | 0.14 | 0.38 | 0.21 | NaN | NaN |
| | BMCLab + AAP | 0.39 | 0.39 | 0.23 | 0.1309 | 0.1615 |
| **MotionBERT (FT) + Taylor** | BMCLab | 0.29 | 0.41 | 0.32 | -0.1422 | 0.1278 |
| | BMCLab + AAP | 0.14 | 0.38 | 0.21 | NaN | NaN |
| **PoseFormerV2 + Taylor** | BMCLab | 0.05 | 0.22 | 0.09 | -0.2286 | 0.001356 |
| | BMCLab + AAP | 0.3 | 0.42 | 0.3 | 0.315 | 0.00057224 |
| **PoseFormerV2 (FT) + Taylor** | BMCLab | 0.06 | 0.24 | 0.09 | NaN | NaN |
| | BMCLab + AAP | 0.22 | 0.35 | 0.27 | 0.1758 | 0.05901 |

In Table 4.4 we can observe Taylor's pretraining positive impacts on AAP's contribution to model training.

For base MotionBERT with Taylor pretraining, adding AAP now improves performance (w-F1=0.21 to F1=0.23) and, prevents model from falling to zero prediction variance ($\rho = 0.1309$, p-value $= 0.1615$. By contrast, fine-tuned MotionBERT shows the opposite pattern, though, based on the analysis made in Section 4.2, such decline is likely rooted from the incompatibility between Taylor's pretraining and MotionBERT full finetuned architecture.

PoseFormerV2 exhibits substantial improvement when AAP is included with

Taylor pretraining. The base model jumps from weighted F1=0.09 with negative correlation ($\rho = -0.23$) to weighted F1=0.30 with strong positive correlation ($\rho = 0.315$, $p < 0.001$). Similarly, the fine-tuned variant improves from complete collapse (F1=0.09, NaN) to weighted F1=0.27 with marginally significant correlation ($\rho = 0.18$, $p = 0.059$).

However, these performance levels are nearly identical to those achieved in the full Taylor pretraining experiments where AAP was included during both pretraining and training. Since both factors -AAP inclusion and Taylor pretraining- are changed simultaneously, we cannot isolate their individual contributions from these results alone.

**Table 4.5:** Decoupled pretraining-training: Taylor pretraining on BMCLab+AAP, training on BMCLab only (tested on PD-GaM)

| **Model** | **Prec** | **Rec** | **F1** | **Spearman ($\rho$)** | **$p$-value** |
|---|---|---|---|---|---|
| **MotionBERT** | 0.06 | 0.24 | 0.09 | NaN | NaN |
| **MotionBERT (FT)** | 0.06 | 0.24 | 0.09 | NaN | NaN |
| **PoseFormerV2** | 0.06 | 0.22 | 0.09 | -0.1859 | 0.04576 |
| **PoseFormerV2 (FT)** | 0.29 | 0.41 | 0.32 | -0.1422 | 0.1278 |

To separate these effects, Table 4.5 presents a decoupled experimental design: Taylor pretraining is performed on BMCLab and AAP, but training is performed using only BMCLab.

The decoupled results reveal dramatic performance collapse across all configurations. Similarly to the first scenario in Table 4.3, both MotionBERT architectures collpase, yielding NaN Spearman correlations and only weighted F1-score of 0.09. PoseFormerV2 base achieves F1=0.09 with weak negative correlation ($\rho = -0.19$, $p < 0.05$), while the fine-tuned variant reaches F1=0.32 but with negative correlation ($\rho = -0.14$, $p = 0.13$). These results are substantially worse than the Taylor+AAP configurations (F1=0.30 and F1=0.27 with positive correlations) and, critically, similar to or worse than training without AAP at all.

The decoupled experiments hence show that AAP must be included during task-specific training alongside Taylor pretraining to provide benefit— using AAP only during pretraining offers no advantage.

Based on these results, in all following experiments we test exclusively on PD-GaM (Test). This approach uses AAP's regularization benefits when paired with Taylor pretraining while focusing on the main clinical goal: automated severity assessment for PD-GaM patients. Since testing on AAP consistently produced poor results across all conditions, we concentrate our evaluation efforts on PD-GaM to enable more focused analysis of the target task.

## 4.4 Parameter-Efficient Fine-Tuning with LoRA

The collapse observed in fully fine-tuned models across zero-shot configurations suggests that updating all network parameter is excessively powerful for the limited clinical datasets available. This lack of robustness is incompatible with clinical deployment requirements, where models must maintain reliable performance across varying patient demographics, and different recording protocols, without requiring complete re-training.

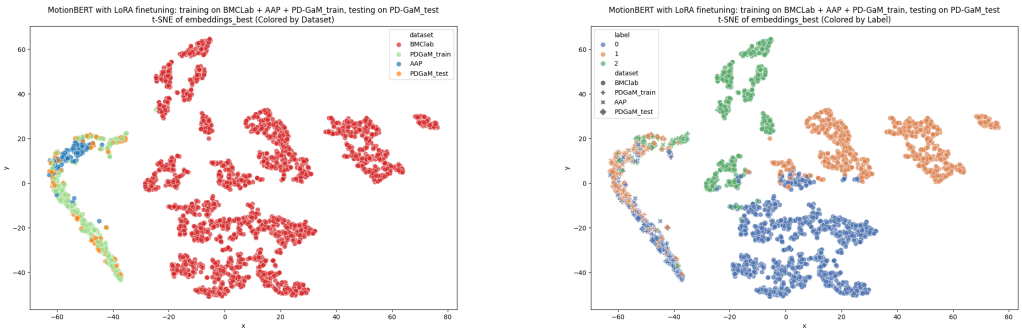
To address this limitation, we investigate Low-Rank Adaptation (LoRA) [41], a parameter-efficient fine-tuning technique that restricts model adaptation to low-rank weight updates rather than modifying all parameters. By constraining the fine-tuning process, LoRA aims to preserve the general motion representations learned during pretraining while allowing controlled adaptation to task-specific patterns. This section evaluates whether LoRA can achieve competitive performance without the brittleness observed in full fine-tuning.

**Table 4.6:** Results of LoRA fine-tuning pipeline

| Model | Train Set | Test Set | Prec | Rec | F1 | Spearman ($\rho$) | $p$-value |
|---|---|---|---|---|---|---|---|
| **MotionBERT** | BMCLab | PD-GaM (Test) | 0.04 | 0.16 | 0.06 | -0.431 | 1.366E-06 |
| | BMCLab + AAP | PD-GaM (Test) | 0.23 | 0.39 | 0.26 | 0.3212 | 0.0004389 |
| | BMCLab + AAP + PD-GaM (Train) | PD-GaM (Test) | 0.73 | 0.72 | 0.71 | 0.7241 | 4.11E-20 |
| **PoseFormerV2** | BMCLab | PD-GaM (Test) | 0.15 | 0.23 | 0.11 | -0.0904 | 0.3345 |
| | BMCLab + AAP | PD-GaM (Test) | 0.15 | 0.23 | 0.11 | -0.0904 | 0.3345 |
| | BMCLab + AAP + PD-GaM (Train) | PD-GaM (Test) | 0.61 | 0.61 | 0.60 | 0.6144 | 2.22E-13 |

Table 4.6 shows the results obtaining by injecting LoRA finetuning into a simple train/test split configuration.LoRA was applied with rank $r = 32$, scaling parameter $\alpha = 64$, and dropout rate of 0.1, constraining the adaptation to low-rank updates while preserving pretrained representations.

MotionBERT with LoRA achieves exceptional results in the full training configuration (BMCLab + AAP + PD-GaM Train), reaching weighted F1 score of 0.71 with strong Spearman correlation ($\rho = 0.72$, $p < 10^{-19}$). This performance surpasses the best fully fine-tuned result (F1=0.66, $\rho = 0.67$) while maintaining stability across all configurations.




**(a)** Colored by dataset

**(b)** Colored by severity label

**Figure 4.4:** t-SNE visualization of motion encoder representations for MotionBERT with LoRA fine-tuning in the best-performing configuration (F1=0.71, $\rho = 0.72$). Left: samples colored by source dataset. Right: samples colored by Parkinson's severity label.

In Figure 4.4 we can observe how the best configuration's performance is reflected in the embedding representations.

The left panel shows dataset distribution. Similar to full fine-tuning, BMCLab's samples dominate the space due to their larger quantity. However, the smaller datasets show improved organization and integration. Though AAP's samples remain substantially compact, they overlap with PD-GaM's samples rather than forming a completely isolated region.The right plot, colored by severity label, reveals excellent severity separation. Though MoCap and RGB samples are well separated, we can observe a pronounced overall gradient from class 0 to class 2, meaning that LoRA enables the model to better preserve the ranking of disease severity. Embeddings representing patients with mild symptoms (label 0) occupy the lower area of the space. Severely impaired gait samples (label 2) concentrate in the upper regions, showing clear spatial separation from healthy patterns. The slight impairment class (label 1) forms an intermediate zone, distributed between the normal and severe regions.

Figure 4.5 provides additional insight by coloring BMCLab samples by patient ID. Clear patient-based agglomerations emerge, with samples from the same patient clustering together. This patient-specific organization explains the sparse distribution of BMCLab samples observed throughout our experiments, rather than forming a single cohesive cluster based on severity.

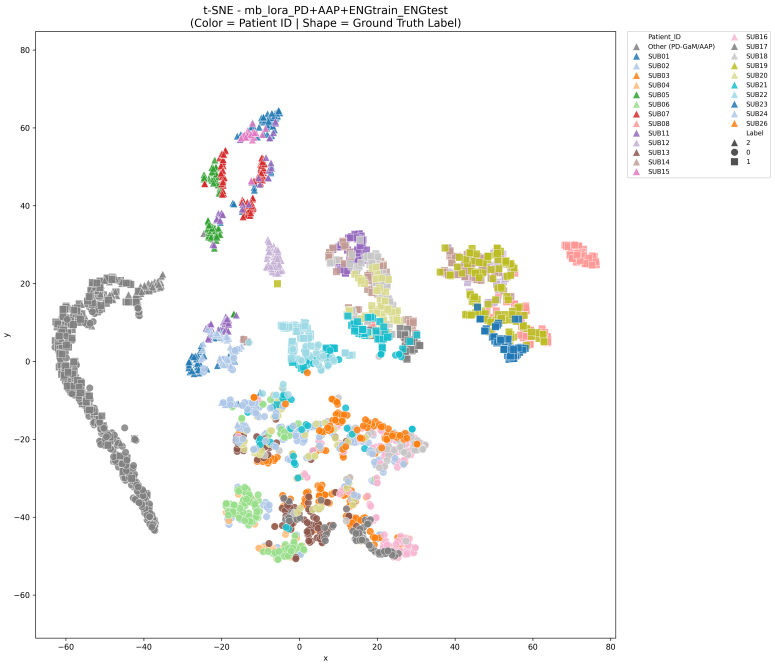



**Figure 4.5:** t-SNE visualization of MotionBERT with LoRA embeddings colored by patient ID and shaped by severity label, revealing patient-specific clustering within the BMCLab region.

When only BMCLab is used for training, the performance is poor, likely due to the gap between MoCap data and RGB data. Indeed, when we add another RGB dataset in the training, i.e. AAP, performance improves substantially from F1=0.06 to F1=0.26 with strongly positive correlation ($\rho = 0.32$, $p < 0.001$), demonstrating that LoRA successfully leverages AAP's parkinsonian patterns for cross-dataset generalization.

For what concerns PoseFormerV2, we can observe that adding AAP into the training configuration provides no improvement with respect to the zero-shot performance, as both achieved F1=0.11 with weak negative correlation ($\rho = -0.09$, $p = 0.33$), indicating PoseFormerV2 with LoRA cannot extract useful features from AAP without in-domain PD-GaM data. However, with full training data (BMCLab + AAP + PD-GaM Train), PoseFormerV2 reaches F1=0.60 with strong correlation ($\rho = 0.61$, $p < 10^{-12}$). This suggests PoseFormerV2 with LoRA requires in-domain data to learn effectively, while MotionBERT benefits from multi-dataset training even without target domain data.

Comparing with full fine-tuning results, LoRA provides critical advantages.

While full fine-tuning achieved F1=0.66 but collapsed in three out of four cross-dataset scenarios, LoRA maintains stability across all conditions while achieving higher peak performance (F1=0.71). The constrained adaptation prevents overfitting to dataset-specific artifacts, enabling the model to preserve pretrained motion representations while learning task-specific severity patterns.

Having established LoRA's effectiveness in standard training configurations, we now investigate whether combining LoRA with Taylor pretraining can further enhance performance by leveraging both parameter-efficient adaptation and kinematic reconstruction objectives. The results can be observed in Table 4.7.

**Table 4.7:** Results of Taylor pretraining + LoRA fine-tuning pipeline

| Model | Train Set | Test Set | Prec | Rec | F1 | Spearman ($\rho$) | $p$-value |
|---|---|---|---|---|---|---|---|
| **MotionBERT** | BMCLab | PD-GaM (Test) | 0.14 | 0.38 | 0.21 | NaN | NaN |
| | BMCLab + AAP | PD-GaM (Test) | 0.14 | 0.38 | 0.21 | NaN | NaN |
| | BMCLab + AAP + PD-GaM (Train) | PD-GaM (Test) | 0.62 | 0.54 | 0.53 | 0.4767 | 6.31E-08 |
| **PoseFormerV2** | BMCLab | PD-GaM (Test) | 0.44 | 0.25 | 0.11 | -0.0238 | 0.7999 |
| | BMCLab + AAP | PD-GaM (Test) | 0.19 | 0.3 | 0.23 | -0.0619 | 0.5091 |
| | BMCLab + AAP + PD-GaM (Train) | PD-GaM (Test) | 0.62 | 0.60 | 0.60 | 0.606 | 5.68E-13 |

For MotionBERT, Taylor pretraining degrades performance. The best configuration (BMCLab + AAP + PD-GaM Train) achieves only F1=0.53 with $\rho = 0.48$ ($p < 10^{-7}$)—markedly worse than LoRA alone (F1=0.71, $\rho = 0.72$). More critically, configurations where BMCLab was used for training and where AAP was injected collapsed, as NaN Spearman correlation was yielded.

PoseFormerV2 shows a different pattern. When training on BMCLab, performance remains poor (F1=0.11) but avoids collapse, producing near-zero correlation ($\rho = -0.02$, $p = 0.80$). Adding AAP yields a weighted F1 score of 0.23 with weak negative correlation ($\rho = -0.06$, $p = 0.51$), which is similar to standalone LoRA, indicating no additional benefit from Taylor pretraining in cross-dataset scenarios. However, the full training configuration (BMCLab + AAP + PD-GaM Train) achieves F1=0.60 with strong correlation ($\rho = 0.61$, $p < 10^{-12}$), matching the standalone LoRA performance exactly. This identical result suggests that for PoseFormerV2, Taylor pretraining provides no additive benefit when combined with LoRA—the performance is determined entirely by LoRA's constrained adaptation.

These results demonstrate that combining Taylor pretraining with LoRA does not yield synergistic benefits. For MotionBERT, the combination is actively harmful, reducing both peak performance and robustness. For PoseFormerV2, Taylor pretraining is redundant. The negative interaction likely derives from from LoRA's low-rank constraints limiting the model's capacity to simultaneously satisfy Taylor's reconstruction objectives and learn task-specific severity discrimination.

With only a small fraction of parameters trainable, the model cannot balance both objectives effectively.

Based on these findings, standalone LoRA (without Taylor pretraining) emerges as the superior approach, achieving the best overall performance (MotionBERT: F1=0.71, $\rho = 0.72$) while maintaining stability and computational efficiency.

## 4.5 Benchmarking Against State-of-the-Art: Comparison with CARE-PD

Our experiments have identified MotionBERT as the superior motion encoder architecture, consistently outperforming PoseFormerV2 across standard training, Taylor pretraining, and LoRA adaptation. In the final phase of this thesis, we evaluate our methodology against the current state-of-the-art. During the progression of this research, a comprehensive new benchmark suite, CARE-PD [9], was published by Adeli et al. (2025), establishing newly standardized baselines and datasets for vision-based Parkinson's disease assessment. To ensure our findings remain relevant and directly comparable to current state-of-the-art baselines, we adapted our final experimental framework to align with this newly established evaluation protocol.

The CARE-PD benchmark assesses multi-view robustness by standardizing camera angles -specifically, side right and back right view - across multiple datasets. To achieve this, sequences are processed through pose estimation and fitted to the SMPL model. From this 3D volumetric mesh, standard Human3.6M 3D skeletal joints are extracted. A sensor harmonization step normalizes the floor height, anchors the initial root position to the origin, and dynamically corrects for curved walking trajectories using Kabsch algorithm [42]. In the experiments that follow, we comply to the CARE-PD's pipeline by training the datasets using their lateral and posterior view independentely, and their softmax outputs are averaged at inference time. Spearman correlation and its p-value are computed on the combined result. Let us note that 3DGait did not have the back view, so the reported results for this dataset refer to said view.

Once the 3D kinematic data is spatially aligned, virtual cameras are positioned to extract standardized 2D views using a perspective projection matrix. To simulate a side view, a virtual camera is placed perpendicular to the walking path at a distance of five meters. To simulate a back view, the camera is positioned two meters behind the subject with a specific 40-degree lateral rotation. The 3D joints are projected back onto a bidimensional plane, obtaining a unified 2D skeletal sequences that isolate specific parkinsonian features such as posture, arm swing, and base of support without noises of the original recording environments.

We structured our final train-test splits to integrate the datasets featured in

their work alongside our primary datasets. Specifically, the training phase involves training partitions of both PD-GaM and BMCLab, along with the AAP dataset. For the final evaluation, we test the model's capabilities across four distinct targets: the test partitions of PD-GaM and BMCLab, T-SDU-PD and 3DGait dataset. When evaluating the test partitions of PD-GaM and BMCLab, our setup corresponds to the Multi-domain In-Domain Adaptation (MIDA) paradigm, as the model is trained on all datasets, in addition to the training portion of the in-domain target dataset. Conversely, when evaluating the entirely unseen T-SDU-PD and 3DGait datasets, our setup aligns with the Leave-One-Dataset-Out (LODO) paradigm, where the model is trained on all datasets except the target dataset. The adoption of these specific data splits allows us to bridge our initial experimental design with the new standardized benchmark.

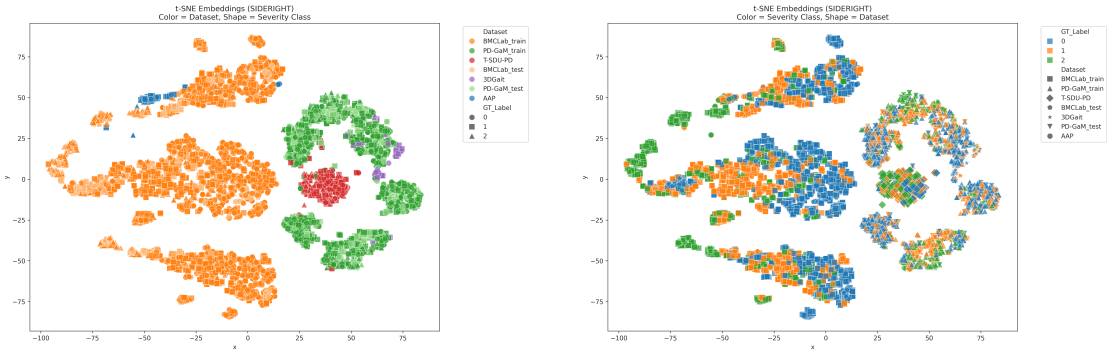
Lastly, to ensure a direct and comparison with the state-of-the-art benchmark, we transition our primary evaluation metric from the weighted F1-score to the macro-averaged F1-score adopted by the authors of CARE-PD. While our preceding experiments used the weighted F1-score to account for the inherent class imbalances within our isolated datasets, consistent with earlier preliminary works in the field, the macro F1-score imposes a significantly more rigorous evaluation standard.

**Table 4.8:** Benchmark results against the CARE-PD suite using standard training

| **Test Dataset** | **Prec (W)** | **Rec (W)** | **F1 (W)** | **Spearman ($\rho$)** | **$p$-value** |
|---|---|---|---|---|---|
| PD-GaM (Test) | 0.61 | 0.62 | 0.60 | 0.61 | 6.85E-48 |
| BMCLab (Test) | 0.66 | 0.65 | 0.66 | 0.68 | 2.30E-84 |
| T-SDU-PD | 0.49 | 0.34 | 0.25 | 0.37 | 4.76E-14 |
| 3DGait | 0.46 | 0.53 | 0.49 | 0.58 | 1.81E-09 |

Table 4.8 provides performance results of the standard training approach across all four test targets. For the in-domain datasets, we achieve strong positive correlations (PD-GaM $\rho = 0.61$, $p < 10^{-47}$; BMCLab $\rho = 0.68$, $p < 10^{-83}$) with weighted F1-scores of 0.60 and 0.66 respectively. These results demonstrate that standard supervised fine-tuning establishes a reliable baseline for severity assessment when the test distribution closely matches the training data.

When evaluation is performed pn 3DGait, standard training achieves a weighted F1-score of 0.49 while mantaining a statistically signifcant Spearman correlation ($\rho = 0.58$, $p < 10^{-8}$), demonstrating reasonable generalization despite never seeing this dataset during training. In contrast, performance on T-SDU-PD degrades, yielding a weighted F1-score of only 0.25, although correlation remains statistically significant ($\rho = 0.37$, $p < 10^{-13}$).




**(a)** Colored by dataset **(b)** Colored by severity

**Figure 4.6:** t-SNE visualizations of learned embeddings in the standard configuration.

The t-SNE plots in Figure 4.6 reveal the representation structure learned by standard training. Examining the plot colored by dataset (left panel), a clear spatial separation emerges between BMCLab and PD-GaM samples, forming two distinct macro-regions in the left and right areas of the embedding space respectively. In particular, T-SDU-PD forms a compact, isolated cluster positioned between the BMCLab and PD-GaM regions without meaningful overlap with either, which reflects the poor performance observed on this dataset. In contrast, 3DGaits samples are integrated within the PD-GaM distribution, particularly in the upper-right area, explaining the reasonable generalization of the dataset.

The severity-based view (right panel) shows the same two-region structure observed in the dataset view, with severity classes distributed independently within each macro-cluster. In the BMCLab-dominated left region, class 2 occupies the leftmost region of the embedding space. Moving from the right, we can observe a shift to class 1 and class 0 embeddings, showing an ordinal though weak, separation. Similarly, in the PD-GaM-dominated right region, the three severity classes occupy overlapping spatial locations without clear progression, as they are intermixed within each sub-cluster. This lack of coherent severity structure across datasets indicates that standard training has learned dataset-specific representations rather than generalizable biomechanical patterns of parkinsonian severity.

**Table 4.9:** In-domain evaluation (Macro F1-score) compared to the CARE-PD MIDA paradigm.

| **Model Configuration** | **PD-GaM (Test)** | **BMCLab (Test)** |
|---|---|---|
| CARE-PD (MIDA Baseline) | 0.59 | 0.74 |
| Ours: Standard Training | 0.43 | 0.66 |
| Ours: LoRA Fine-tuning | **0.63** | 0.50 |
| Ours: Taylor Pretraining + LoRA | 0.59 | 0.61 |

While we adopt the MIDA approach by including in-domain training data, our models are trained on a smaller, modified dataset pool (AAP, and the training portions of BMCLab and PD-GaM) compared to the CARE-PD baseline (BMCLab, PD-GaM, T-SDU-PD, 3DGait).

**Table 4.10:** Zero-shot cross-dataset evaluation (Macro F1-score) compared to the CARE-PD LODO paradigm.

| **Model Configuration** | **T-SDU-PD** | **3DGait** |
|---|---|---|
| CARE-PD (LODO Baseline) | 0.34 | 0.14 |
| Standard Training | 0.29 | **0.36** |
| LoRA Fine-tuning | 0.28 | **0.15** |
| Taylor Pretraining + LoRA | **0.45** | **0.37** |

In the LODO approach, all labeled datasets except the target dataset are used in the training phase. In out experiments, models are trained on a smaller, modified dataset pool (AAP, and the training portions of BMCLab and PD-GaM) compared to the CARE-PD baseline (BMCLab, PD-GaM, T-SDU-PD, 3DGait).

Transitioning to the stricter Macro F1 score, Table 4.9 highlights the limitations of our standard training configuration for in-domain adaptation, as we fail to meet the CARE-PD baselines (0.43 against 0.59 for PD-GaM, and 0.66 against 0.74). This shows that standard training achieves higher performance when using more training datasets. As presented in Table 4.10, standard training evaluated on T-SDU-PD slightly underperforms the CARE-PD benchmark, achieving a Macro F1 score of 0.29 compared to their 0.34. However, on the 3DGait dataset, our standard training approach demonstrates a substantial advantage, reaching a Macro F1 of 0.36 and significantly outperforming the CARE-PD baseline of 0.14. Such outperformance likely stems from the inclusion of the AAP dataset. Although 3DGait patients walk on a GAITrite walkway, both datasets share a similar clinical protocol consisting of continuous, straight-line walking over a short distance.

**Table 4.11:** Benchmark results against the CARE-PD suite using LoRA finetuning

| **Test Dataset** | **Prec (W)** | **Rec (W)** | **F1 (W)** | **Spearman ($\rho$)** | **$p$-value** |
|---|---|---|---|---|---|
| PD-GaM (Test) | 0.64 | 0.64 | 0.64 | 0.64 | 6.85E-48 |
| BMCLab (Test) | 0.55 | 0.49 | 0.50 | 0.59 | 1.05E-58 |
| T-SDU-PD | 0.43 | 0.43 | 0.33 | 0.09 | 0.085 |
| 3DGait | 0.19 | 0.19 | 0.19 | 0.12 | 0.081 |

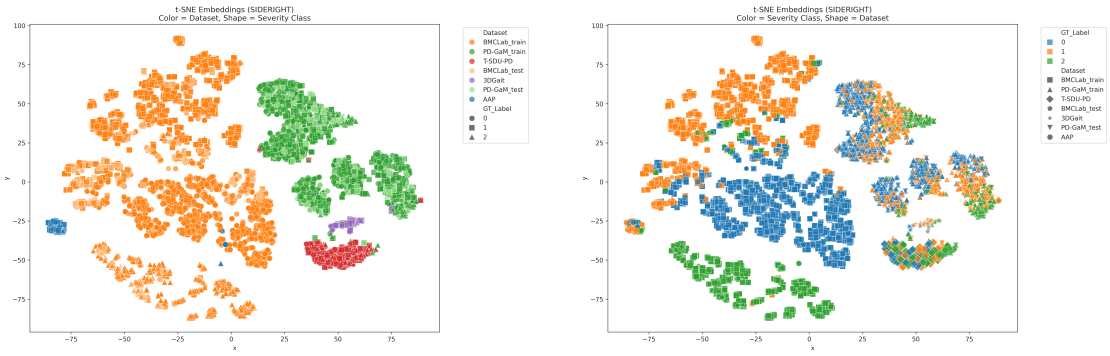
Table 4.11 presents the evaluation metrics for the LoRA fine-tuning configuration across the four target datasets. Following hyperparameter search, we selected rank $r = 16$, scaling parameter $\alpha = 16$, and dropout rate of 0.3 for optimal performance. LoRA provides improvement on the PD-GaM test dataset, increasing the weighted) F1 score from 0.60 to 0.64, and improving the Spearman correlation ($\rho = 0.64$,$p < 0.001$). Performance on the BMCLab test dataset remains solid, with a significant positive Spearman correlation ($\rho = 0.59$, $p < 0.001$), although we can observe a moderate drop in the weighted F1-score to 0.50.
In the other hand, on the zero-shot cross evaluation, we lose statistical relevance of the results as both $p$-values exceed the 0.05 threshold ($p = 0.085$ for T-SDU-PD and $p = 0.081$ 3DGait), and in particular, performance on 3DGait collapses, yielding a weighted F1-score of 0.19, while a slight improvement when testing on T-SDU-PD is observed (weighted F1=0.33).

The t-SNE visualizations in Figure 4.7 reveal how LoRA's parameter-efficient adaptation affects the learned representations. Examining the dataset organization (left panel), the embeddings of training datasets exhibit a more accentuated separation, with AAP and PD-GaM forming in particular compact cluster on the leftmost and rightmost ares in the embedding space respectively, while BMCLab is scattered between the other two datasets. The in-domain test samples are intermixed with their respective in-domiain training samples, reflecting how a moderately good perfromance is mantained in the in-domain evaluations.
In the other hand, the unseen test datasets of T-SDU-PD and 3DGait form compact isolated clusters in the lower right region of the embedding space, which explains the loss of statistical significance in both zero-shot evaluations.

The severity-based view (right panel) reveals that each dataset follows its own ordinal coloration rather than exhibiting a universal severity structure across the embedding space. LoRA maintains severity separation, but this organization is dataset-dependent.

Remarkably, within the BMCLab region, the spatial arrangement shows class 1 in the upper areas, transitioning to class 0 in the central region, and finally class 2

in the lower region, rather than following an ordering based on severity progression. This disrupted ordinal structure likely explains the moderate performance drop on BMCLab compared to standard training performance, as the model has learned a severity mapping that, while statistically significant, is less coherent than the representations learned by standard training. Furthermore, although the PD-GaM region exhibits a more coherent severity organization, four distinct subregions can be identified, each exhibiting more coherent severity organization. These subregions may correspond to patient-specific clusters, suggesting that LoRA organizes representations by individual movement characteristics rather than purely by severity level.




**(a)** Colored by dataset **(b)** Colored by severity

**Figure 4.7:** t-SNE visualizations of learned embeddings in the LoRA finetuning configuration.

Compared to CARE-PD baselines, LoRA's finetuning approach on PD-GaM test set outperforms the current state-of-the-art, achieving a macro F1-score of a.63 against 0.59. This achievement highlights LoRA's ability to boost performance leveraging on a smaller pool of datasets. Conversely, on the MoCap-based BMCLab dataset, LoRA yields a Macro F1 score of 0.50, falling significantly short of the 0.74 baseline, suggesting LoRA's parameter updates lack the capacity to fully capture the complex, multi-domain variance required to match the baseline on BMCLab. On the unseen T-SDU-PD dataset, LoRA's approach achieves a Macro F1 score of 0.28, therefore still underperforming the CARE-PD baseline of 0.34. On the contrary, despite the lack of statistical significance of the evaluation on 3DGait, a marginal improvement is observed with respect to the baseline of 0.14, yielding a macro F1 score of 0.15.

**Table 4.12:** Benchmark results against the CARE-PD suite using the Taylor pretraining combined with LoRA fine-tuning configuration.

| **Test Dataset** | **Prec (W)** | **Rec (W)** | **F1 (W)** | **Spearman ($\rho$)** | ***p*-value** |
|---|---|---|---|---|---|
| PD-GaM (Test) | 0.64 | 0.64 | 0.64 | 0.69 | 1.04E-65 |
| BMCLab (Test) | 0.64 | 0.60 | 0.61 | 0.56 | 4.04E-53 |
| T-SDU-PD | 0.48 | 0.46 | 0.46 | 0.29 | 3.67E-09 |
| 3DGait | 0.51 | 0.48 | 0.45 | 0.66 | 5.98E-13 |

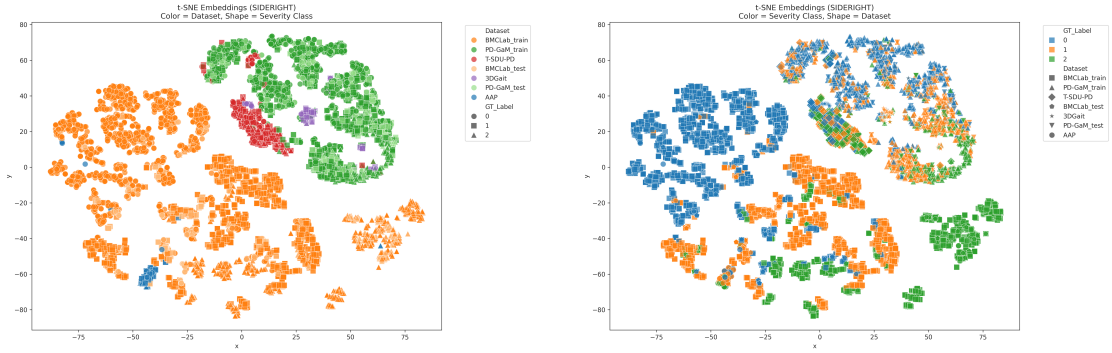
Table 4.12 presents results for the combined Taylor pretraining and LoRA fine-tuning configuration. In this setting, along with the datasets used in the training phase, we injected the unlabeled datasets of the CARE-PD during the pretraining phase: T-SDU, T-LTC, DNE, KUL-DT-T, and E-LC.

On in-domain evaluation, the Taylor+LoRA configuration maintains the weighted F1 score obtained previously in the LoRA-only performance, but strengthens the Spearman correlation ($\rho = 0.69$, $p < 10^{-60}$). When tested on BMCLab performance, adding Taylor pretraining along LoRA boosts the performance, improving the weighted F1 score to 0.50 to 0.61, with $\rho = 0.56$ and $p < 10^{-50}$, though it is not able to match performance in the standard training configuration.

Taylor pretraining enhances LoRA in the zero-shot cross-dataset scenario as well. On T-SDU-PD, it recovers the F1 score from 0.33 (in the standalone LoRA configuration) to F1=0.46 with statistically significant correlation ($\rho = 0.29$, $p < 10^{-8}$). Similarly, on 3DGait, the configuration achieves F1=0.45 with strong correlation ($\rho = 0.66$, $p < 10^{-12}$), recovering from LoRA's complete collapse (F1=0.19, $p = 0.081$).

The t-SNE visualizations in Figure 4.8 reveal the representation structure that enables this improved generalization.

Examining the dataset organization (left panel), BMCLab samples form distinct, cohesive clusters in the left portion of the embedding space, while PD-GaM samples are distributed throughout the central and right region of the embedding space. In both cases, their respective training and testing samples are intermixed, demonstrating good in-domain generalization.




**(a)** Colored by dataset **(b)** Colored by severity

**Figure 4.8:** t-SNE visualizations of learned embeddings for Taylor pretraining + LoRA configuration. Left: samples colored by dataset show BMCLab forming distinct clusters while PD-GaM, T-SDU-PD, and 3DGait are spatially integrated, explaining robust zero-shot generalization. Right: samples colored by severity class demonstrate clear ordinal structure with cross-dataset coherence. .

AAP samples notably overlap with the BMCLab region, suggesting the model recognizes shared motion capture characteristics despite AAP's different clinical protocol. The T-SDU-PD and 3DGait are not isolated as well, as they overlap the PD-GaM distribution. This spatial integration explains the restored statistical significance and strong correlations on these unseen datasets.

The severity-based view (right panel) reveals a more complex organization pattern. Two distinct macro-regions can be distinguished: a left cluster dominated by BMCLab samples and a right cluster dominated by PD-GaM samples. However, within each macro-region, clear ordinal severity structure emerges independently. In both cases, classes maintain spatial progression from Class 0 in the upper areas through Class 1 in the central regions to Class 2 in the lower areas. This severity class organization suggests that while the model has not fully unified the representation space across dataset, it has learned that severity progression follows consistent kinematic patterns regardless of recording protocol. Comparing against CARE-PD baselines (Table 4.9 and Table 4.10), the combination of Taylor and LoRA matches the MIDA in-domain performance on PD-GaM (Macro F1=0.59 for both), demonstrating competitive results despite training on a smaller dataset pool.
For zero-shot evaluation, combining Taylor with LoRA outperforms CARE-PD's LODO baseline on both unseen datasets: Macro F1=0.45 vs. 0.34 on T-SDU-PD (+32% improvement) and 0.37 vs. 0.14 on 3DGait (+164% improvement). This superior cross-dataset generalization, achieved with fewer training datasets,

highlights the effectiveness of combining self-supervised kinematic pretraining with parameter-efficient adaptation.

The synergy between Taylor pretraining and LoRA emerges clearly when comparing all three configurations. While LoRA alone improved in-domain PD-GaM performance but sacrificed zero-shot generalization, and Taylor+LoRA was suboptimal in earlier experiments, the CARE-PD benchmarking context reveals a different pattern. The combination excels specifically when evaluated on completely unseen datasets with different recording protocols. This suggests that the combination's strengths manifest in extreme distribution shift scenarios.

# Chapter 5

# Discussion

This chapter summarized the experimental findings presented in Chapter 4, interpreting the results across multiple training configurations and drawing connections between quantitative performance metrics and learned representation structures. We analyze the trade-offs in different training strategies, identify optimal deployment scenarios for each approach, discuss limitations of the current work, and outline directions for future research.

## 5.1 Summary of Experimental Findings

Our investigation evaluated motion encoder architectures across four distinct training paradigms: standard supervised training, self-supervised Taylor pretraining, parameter-efficient LoRA fine-tuning, and combined Taylor+LoRA approaches. Each strategy revealed different strengths, weaknesses, and suitability for specific clinical scenarios.

**Cross-Dataset Generalization and In-Domain Adaptation** The initial experiments investigated the cross-datasets generalization of MotionBERT, PoseFormerV2, and its finetuned variant. Zero-shot evaluation on the AAP dataset revealed that all architectures struggled to achieve statistically significant performance regardless of training configuration. Even when injecting in-domain AAP training data, MotionBERT and base PoseFormerV2 remained stagnant at F1=0.12 and F1=0.29 respectively, with no correlation achieving statistical significance. This consistent failure across all models suggests that AAP presents unique characteristics, possibly related to recording protocols or patient demographics, that creates a domain gap that the current architectures are not able to bridge. Indeed, t-SNE visualizations of the zero-shot evaluation AAP show a sharp spatial separation between MoCap data (BMCLab) and RGB video data (PD-GaM, AAP), with

AAP samples forming isolated clusters. However, within these isolated regions, clear ordinal severity gradients emerged, demonstrating that the model successfully learned parkinsonian severity patterns.

Conversely, in the in-domain adaptation of PD-GaM, all architectures demonstrated strong performance. MotionBERT emerged as the superior architecture, achieving F1=0.66 with highly significant correlation ($\rho = 0.63$, $p < 10^{-14}$). The t-SNE visualization of this best-performing configuration showed a less harsh separation between MoCap and RGB datasets. In particular PD-GaM training and test samples perfectly intertwined, and a tightly integrated manifold emerged for RGB data.

Two main findings emerged from these initial experiments. First, MotionBERT's superior performance across all configurations identified it as the optimal architecture for parkinsonian gait assessment. Second, in-domain adaptation proved is key for achieving clinically meaningful performance.

**Self-Supervised Taylor Pretraining: The Generalization-Specialization Trade-off** We introduced Taylor pretraining, consisting of the reconstruction of randomly masked joint positions, velocities, and accelerations, forcing networks to learn kinematic rules governing human motion, and potentially extract deeper information about gait features such as tremor, stride length, and bradykinesia. At the same, in light of the observations about the previous experiments, we also investigate the use of finetuned MotionBERT in clinical settings. Taylor pretraining revealed a trade-off: while base MotionBERT's zero-shot performance improved substantially, in-domain adaptation degraded significantly. Fine-tuned MotionBERT with Taylor pretraining achieved peak performance in the best configuration, whereas it collapsed catastrophically in all other scenarios. PoseFormerV2 exhibited similar generalization-specialization trade-offs with modest zero-shot improvements and degraded in-domain performance, though demonstrating greater stability with no collapse in any configuration. These results establish that Taylor pretraining imposes a strict choice: enhanced cross-dataset generalization at the expense of in-domain fitting capacity.

**Dataset Composition: The Context-Dependent Role of AAP** From previous experiments, it emerged that performance remained poor when testing on the AAP dataset. Therefore, we investigated as well whether using AAP in the training phase provides contribution to cross-dataset generalization or contaminates the learning process. Our revealed highly context-dependent effects. In standard training, AAP consistently degraded performance, introducing harmful distribution shift. However, when combined with Taylor pretraining, the introduction of AAP in the training set prevented MotionBERT from collapse, and boosted PoseFormerV2

performance, increasing substantially the F1 score with strong positive correlation.

These findings reveal that dataset composition effects cannot be evaluated in isolation but depend critically on the training strategy employed. From a practical standpoint, subsequent experiments focused on testing on PD-GaM's test portion only.

**Parameter-Efficient LoRA Fine-Tuning.** The instability observed in fine-tuned MotionBERT with Taylor pretraining prompted the investigation of LoRA as alternative, parameter-efficient finetuning approach.

MotionBERT with LoRA achieved the highest overall performance (F1=0.71, $\rho = 0.72$), surpassing the best full fine-tuning result (F1=0.66). For PoseFormerV2, LoRA finetuning did not provide any improvements when in zero-shot evaluation. However, it boosted performance in in-domain adaptation, as the weighted F1 score jumped from 0.11 to 0.60.

Combining with Taylor pretraining with LoRa revealed no synergistic benefits. For MotionBERT, the combination actively degraded performance (F1: 0.71→0.53) and reintroduced collapse in zero-shot scenarios. For PoseFormerV2, Taylor pretraining combined with LoRA exactly matched standalone LoRA in the in-domain adaptation. In zero-shot evalutation, the combination improved F1 score when BMCLab and AAP were used in the training phase, but yielded negative Spearman correlation.

These results establish standalone LoRA as optimal for scenarios prioritizing in-domain performance and stability, while MotionBERT emerges as superior architecture, achieving the best overall performance in all three training configurations.

**Benchmarking Against CARE-PD** The final benchmarking phase evaluated MotionBERT and our train/test dataset configuration against CARE-PD, the current state-of-the-art. For in-domain evaluation, standard training underperformed on both test sets, confirming that without parameter-efficient constraints, larger multi-dataset training is required for competitive performance. LoRA finetuning fine-tuning exceeded CARE-PD's MIDA baseline on PD-GaM (Macro F1=0.63 vs. 0.59) despite training on fewer datasets, demonstrating that parameter-efficient adaptation can compensate for limited data availability. Combining Taylor pretraining and LoRA finetuning matched CARE-PD's baseline on PD-GaM (Macro F1=0.59 for both) but underperformed LoRA alone, demonstrating that the combination sacrifices peak in-domain performance for improved generalization. On BMCLab, Taylor+LoRA achieved moderate performance (Macro F1=0.61) but remained below both the CARE-PD baseline (0.74) and standard training (0.66).

For zero-shot cross-dataset evaluation, on 3DGait, all three configurations outperformed CARE-PD's LODO baseline (0.14), with standard training and Taylor+LoRA achieving the strongest results (Macro F1=0.36 and 0.37 respectively), likely due to protocol similarity with AAP, as both datasets involve continuous straight-line walking over short distances. On T-SDU-PD, only Taylor+LoRA succeeded (Macro F1=0.45 vs. 0.34 baseline), while standard training (Macro F1=0.29) and LoRA alone (Macro F1=0.28) underperformed. These findings demonstrate that coupling self-supervised kinematic pretraining with parameter-efficient adaptation is a highly effective strategy for mitigating extreme distribution shifts.

## 5.2 Limitations

Several limitations affect the generalizability and clinical applicability of our findings:

**Dataset scale and severity distribution.** Clinical datasets are typically characterized by severe data scarcity (e.g. AAP has 31 patients, PD-GaM consists of 30 patients). Moreover, due to the nature of Parkinson's disease - that is, as the disease progresses, the more challenging movement becomes - the datasets lack representation of classes 2 and 3, with the predominance of class 0 and 1. This class imbalance prevents evaluation of fine-grained discrimination in advanced disease stages.

**Persistent domain gap and limited true zero-shot transfer.** In our results emerged a relevant limitation: successful cross-dataset generalization is dependent on either in-domain exposure or recording protocol similarity, rather than genuine learning of universal parkinsonian features. When such similarities are absent, models struggle to maintain performance in zero-shot transfer settings.

This persistent domain gap indicates that current motion encoders, despite pretraining on large-scale general human motion datasets, have not learned fully universal biomechanical representations of parkinsonian severity. The inability to separate parkinsonian factors from protocol-specific elements limits clinical deployment, as there are heterogeneous recording equipment, patient populations, and clinical protocols across different institutions and countries.

**Dataset quantity vs. quality trade-offs.** Our benchmarking against CARE-PD revealed that simply aggregating more datasets does not guarantee improved performance. Despite training on fewer datasets, our approaches achieved competitive or superior results in various evaluation scenarios. This demonstrate that

dataset composition, recording protocol alignment, and training strategy interact in complex ways that cannot be reduced to simple dataset count. Indiscriminately adding datasets may degrade performance by introducing conflicting optimization objectives, as highlighted by the context-dependent effects of multi-dataset training observed throughout our experiments. Effective multi-dataset training requires either careful curation of compatible datasets with similar recording protocols or training strategies that extract generalizable patterns while filtering protocol-specific information.

**Lack of universal training strategy.** In our experiments it emerged that no single training strategy performs optimally across all architectures, dataset combinations, and evaluation scenarios. Taylor pretraining improved base MotionBERT's zero-shot transfer but degraded in-domain performance, while producing catastrophic collapse when combined with full fine-tuning. LoRA achieved peak in-domain performance but underperformed on zero-shot evaluation. Taylor and LoRA succeeded on extreme distribution shift scenarios but underperformed LoRA alone for in-domain assessment. Furthermore, these effects varied by architecture: MotionBERT exhibited both the highest peak performance and the model collapse, whereas PoseFormerV2 demonstrated consistent stability but lower overall performance. This lack of a universal strategy complicates clinical deployment, highlighting the need of more robust training methods that maintain performance across different settings.

## 5.3 Future research directions

Several promising directions emerge from our findings and limitations:

**Task-specific joint and kinematic masking.** Rather than uniformly masking all joints during Taylor pretraining, targeted masking strategies could focus reconstruction on relevant features for gait assessment, Masking strategies should prioritize lower-body joints such as hips, knees, feet, and their associated velocities and accelerations, as they directly capture step length, stride, and walking speed. Additionally, masking elbows and wrists would force the model to learn arm swing patterns and therefore extract useful insights about the smoothness of the gait. This selective masking approach would concentrate the model's learning capacity on clinically relevant kinematic features rather than distributing it uniformly across all body parts.

This biomechanically informed masking could potentially improving cross-dataset generalization by forcing the model to learn clinically meaningful motion patterns rather than arbitrary ones.

**Exploring different SSL pretext tasks.** Future works could explore alternative self-supervised learning tasks that could better refine capture parkinsonian movement patterns. Contrastive learning methods, such as SimCLR [43], could be adapted to learn discriminative features such as walking velocity or stride length. Since both of these factors tend to progressively decrease as Parkinsonian severity advances, formulating positive and negative sequence pairs based on their speed or step amplitude would encourage the model to learn representations that reflect progressive physical impairment.

**Adapter-based finetuning.** While LoRA constrains adaptation through low-rank weight decomposition, adapter-based methods offer an alternative parameter-efficient approach. By inserting small bottleneck modules between frozen pretrained layers, adapters allow task-specific adaptation while preserving the majority of pretrained parameters. Compared to LoRA, applying adapters to motion encoders could potentially provide more targeted control over which representations are modified, with selective insertion at different depths of layers. For instance, early layers might focus on extracting low-level kinematic patterns, whereas deeper layers could be adapted to capture to high-level severity discrimination, hence enabling a hierarchical adaptation strategy.

# Chapter 6

# Conclusion

This thesis investigated the application of motion encoder architectures for automating the assessment of Parkinson's disease severity from gait videos.

## 6.1 Key Contributions

Our work established several significant contributions to vision-based clinical gait assessment.

We identified MotionBERT as the superior motion encoder architecture for parkinsonian gait assessment, consistently outperforming PoseFormerV2 across all training configurations when provided with in-domain data.

We introduced Taylor pretraining, a novel self-supervised strategy, which consists of reconstructing randomly masked joint positions, velocities, and accelerations. While Taylor pretraining enhanced zero-shot transfer capabilities, it degraded in-domain performance when sufficient target data was available. Fine-tuned architectures with Taylor pretraining exhibited brittleness behaviour, achieving peak performance in in-domain configurations but collapsing in all other scenarios.

We identified Low-Rank Adaptation (LoRA) as an effective parameter-efficient fine-tuning method that achieves state-of-the-art in-domain performance and prevents model collapse observed in full finetuning experiments. LoRA's constrained adaptation preserved pretrained motion representations while enabling task-specific severity discrimination, though at the cost of zero-shot generalization capability.

Our findings demonstrated that combining Taylor pretraining with LoRA produces superior cross-dataset generalization under extreme distribution shift, outperforming current state-of-the-art methods on completely unseen datasets while maintaining competitive in-domain performance.

## 6.2 Closing Remarks

This thesis demonstrated that motion encoder architectures, when combined with appropriate training strategies, can achieve clinically meaningful Parkinson's disease severity assessment from video data. The systematic investigation of self-supervised pretraining, parameter-efficient fine-tuning, and multi-dataset integration established clear best practices for different deployment scenarios.

Our results demonstrate that training strategy selection critically determines deployment feasibility: parameter-efficient methods achieve competitive performance with limited data while maintaining stability, and self-supervised pretraining enhances cross-dataset generalization without protocol-specific retraining. These advances address fundamental barriers to clinical adoption by reducing data requirements and enabling deployment across heterogeneous clinical environments.

While significant challenges remain in the persistent domain gap and lack of universal training strategies, this work advances the broader goal of accessible, objective, automated movement assessment for neurological disorders. As vision-based health monitoring is progressing in the medical field, the methods established in this work provide a foundation for robust, clinically validated systems that can operate across the heterogeneous conditions of real-world healthcare delivery.

# Appendix A

# Supplementary Data and Tables

This appendix contains the extended data tables and supplementary information referenced throughout the methodology.

## A.1 BCMLab Anatomical Marker Configuration

The following table, provided by Shida et al. [30], lists the 44 anatomical reflective markers used in the BCMLab dataset to determine the position and orientation of the body segments during walking trials.

**Table A.1:** Details of the 44 anatomical reflective markers used to determine the position and orientation of the body segments during walking trials.

| Label | Name | Description |
|---|---|---|
| R.ASIS | Right Anterior Superior Iliac Spine | Right anterior superior iliac spine |
| L.ASIS | Left Anterior Superior Iliac Spine | Left anterior superior iliac spine |
| R.PSIS | Right Posterior Iliac Spine | Right posterior superior iliac spine |
| L.PSIS | Left Posterior Iliac Spine | Left posterior superior iliac spine |
| R.GTR | Right Greater Trochanter | Most lateral prominence of the right greater trochanter |

Continued on next page

**Table A.1 continued from previous page**

| Label | Name | Description |
|---|---|---|
| R.Knee | Right Knee | Most lateral prominence of the right lateral femoral epicondyle |
| R.Knee.Medial | Right Knee Medial | Most medial prominence of the right lateral femoral epicondyle |
| R.HF | Right Head of Fibula | Proximal tip of the head of the right fibula |
| R.TT | Right Tibial Tuberosity | Most anterior border of the right tibial tuberosity |
| R.Ankle | Right Ankle | Lateral prominence of the right lateral malleolus |
| R.Ankle.Medial | Right Ankle Medial | Most medial prominence of the right medial malleolus |
| R.Heel | Right Heel Bottom | Aspect of the Achilles tendon insertion on the right calcaneous |
| R.MT1 | Right 1st Metatarsal | Dorsal margin of the right 1st metatarsal head |
| R.MT5 | Right 5th Metatarsal | Dorsal margin of the right 5th metatarsal head |
| R.MT2 | Right 2nd Metatarsal | Dorsal margin of the right 2nd metatarsal head |
| L.GTR | Left Greater Trochanter | Most lateral prominence of the left greater trochanter |
| L.Knee | Left Knee | Most lateral prominence of the left lateral femoral epicondyle |
| L.Knee.Medial | Left Knee Medial | Most medial prominence of the left lateral femoral epicondyle |
| L.HF | Left Head of Fibula | Proximal tip of the head of the left fibula |
| L.TT | Left Tibial Tuberosity | Most anterior border of the left tibial tuberosity |
| L.Ankle | Left Ankle | Lateral prominence of the left lateral malleolus |

Continued on next page

**Table A.1 continued from previous page**

| Label | Name | Description |
|---|---|---|
| L.Ankle.Medial | Left Ankle Medial | Most medial prominence of the left medial malleolus |
| L.Heel | Left Heel Bottom | Aspect of the Achilles tendon insertion on the left calcaneous |
| L.MT1 | Left 1st Metatarsal | Dorsal margin of the left 1st metatarsal head |
| L.MT5 | Left 5th Metatarsal | Dorsal margin of the left 5th metatarsal head |
| L.MT2 | Left 2nd Metatarsal | Dorsal margin of the left 2nd metatarsal head |
| CLAV | Incisura jugularis | Deepest point of Incisura Jugularis (suprasternal notch) |
| STRN | Processus xiphoideus | Most caudal point on the sternum |
| C7 | Seventh cervical vertebra | Processus spinosus (spinous process) of the 7th cervical vertebra |
| T10 | Tenth thoracic vertebra | Processus spinosus (spinal process)of the 10th thoracic vertebra |
| RSHO | Right shoulder | Most dorsal point on the acromioclavicular joint (shared with the scapula) |
| RUPA | Right upper arm | Between the elbow and the shoulder markers |
| REL | Right lateral epicondyle | Most caudal point on lateral epicondyle |
| REM | Right medial epicondyle | Most caudal point on medial epicondyle |
| RFRA | Right lower arm | Between the elbow and the wrist markers |
| RWL | Right lateral wrist | Most caudal–lateral point on the radial styloid |
| RWM | Right medial wrist | Most caudal–medial point on the ulnar styloid |

Continued on next page

**Table A.1 continued from previous page**

| Label | Name | Description |
|---|---|---|
| LSHO | Left shoulder | Most dorsal point on the acromioclavicular joint (shared with the scapula) |
| LUPA | Left upper arm | Between the elbow and the shoulder markers |
| LEL | Left lateral epicondyle | Most caudal point on lateral epicondyle |
| LEM | Left medial epicondyle | Most caudal point on medial epicondyle |
| LFRA | Left lower arm | Between the elbow and the wrist markers |
| LWL | Left lateral wrist | Most caudal–lateral point on the radial styloid |
| LWM | Left medial wrist | Most caudal–medial point on the ulnar styloid |

# Bibliography


[1] Lorraine V. Kalia and Anthony E. Lang. «Parkinson's disease». In: *The Lancet* 386.9996 (Aug. 2015), pp. 896–912. DOI: 10.1016/S0140-6736(14)61393-3 (cit. on p. 1).

[2] Christopher G. Goetz et al. «Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): scale presentation and clinimetric testing results». In: *Movement Disorders* 23.15 (Nov. 2008), pp. 2129–2170. DOI: 10.1002/mds.22340 (cit. on p. 1).

[3] Bart Post, Maarten P. Merkus, Rob M. A. de Bie, Rob J. de Haan, and Johannes D. Speelman. «Unified Parkinson's disease rating scale motor examination: are ratings of nurses, residents in neurology, and movement disorders specialists interchangeable?» In: *Movement Disorders* 20.12 (Dec. 2005), pp. 1577–1584. DOI: 10.1002/mds.20640 (cit. on p. 2).

[4] Luc J. W. Evers, Jesse H. Krijthe, Marjan J. Meinders, Bastiaan R. Bloem, and Tom M. Heskes. «Measuring Parkinson's disease over time: The real-world within-subject reliability of the MDS-UPDRS». In: *Movement Disorders* 34.10 (Oct. 2019), pp. 1480–1487. DOI: 10.1002/mds.27790 (cit. on p. 2).

[5] C. Goetz et al. «The MDS-sponsored revision of the Unified Parkinson's disease rating scale». In: 33 (2019) (cit. on p. 2).

[6] Sijie Yan, Yuanjun Xiong, and Dahua Lin. «Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition». In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 32. 1. Apr. 2018 (cit. on p. 3).

[7] 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 (ICCV)*. 2023, pp. 15929–15939 (cit. on pp. 3, 9, 11, 12).

[8] Ce Zheng, Sijie Zhu, Matias Mendieta, Taojiannan Yang, Chen Chen, and Zhengming Ding. *3D Human Pose Estimation with Spatial and Temporal Transformers*. 2021. arXiv: 2103.10455 [cs.CV]. URL: https://arxiv.org/abs/2103.10455 (cit. on p. 3).

[9] Vida Adeli et al. *CARE-PD: A Multi-Site Anonymized Clinical Dataset for Parkinson's Disease Gait Assessment*. 2025. arXiv: 2510.04312 [cs.CV]. URL: https://arxiv.org/abs/2510.04312 (cit. on pp. 3, 15, 16, 18, 38).

[10] Qitao Zhao, Ce Zheng, Mengyuan Liu, 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 (CVPR)*. 2023, pp. 8877–8886 (cit. on pp. 3, 10).

[11] Wentao Zhu, Xiaoxuan Ma, Zhaoyang Liu, Libin Liu, Wayne Wu, and Yizhou Wang. «MotionBERT: A Unified Perspective on Learning Human Motion Representations». In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2023, pp. 15085–15099 (cit. on p. 3).

[12] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. «LoRA: Low-Rank Adaptation of Large Language Models». In: *International Conference on Learning Representations (ICLR)*. 2022 (cit. on p. 3).

[13] Lei Wang, Xiuyuan Yuan, Tom Gedeon, and Liang Zheng. *Taylor Videos for Action Recognition*. 2024. arXiv: 2402.03019 [cs.CV]. URL: https://arxiv.org/abs/2402.03019 (cit. on pp. 3, 19).

[14] Jim G. Richards. «The measurement of human motion: A comparison of commercially available systems». In: *Human Movement Science* 18.5 (1999), pp. 589–602. DOI: 10.1016/S0167-9457(99)00023-8 (cit. on p. 6).

[15] Pierre Merriaux, Yohan Dupuis, Rémi Boutteau, Pascal Vasseur, and Xavier Savatier. «A study of Vicon system positioning performance». In: *Sensors* 17.7 (2017), p. 1591. DOI: 10.3390/s17071591 (cit. on p. 6).

[16] PP Sahoo. «A Comparison of Dual-Kinect and Vicon Tracking of Human Motion for Use in Robotic Motion Programming». In: *IEEE International Conference on Robotics and Automation*. 2017. URL: https://api.semanticscholar.org/CorpusID:28480021 (cit. on p. 6).

[17] Debapriya Maji, Soyeb Nagori, Manu Mathew, and Deepak Poddar. *YOLO-Pose: Enhancing YOLO for Multi Person Pose Estimation Using Object Keypoint Similarity Loss*. 2022. arXiv: 2204.06806 [cs.CV]. URL: https://arxiv.org/abs/2204.06806 (cit. on pp. 6, 7).

[18] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. *You Only Look Once: Unified, Real-Time Object Detection*. 2016. arXiv: 1506.02640 [cs.CV]. URL: https://arxiv.org/abs/1506.02640 (cit. on p. 6).

[19] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. «Microsoft COCO: Common Objects in Context». In: *European Conference on Computer Vision (ECCV)*. Springer, 2014, pp. 740–755 (cit. on p. 6).

[20] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. «SMPL: A Skinned Multi-Person Linear Model». In: *ACM SIGGRAPH Asia 2015 Papers*. SIGGRAPH Asia '15. New York, NY, USA: ACM, Nov. 2015, 248:1–248:16. DOI: 10.1145/2816795.2818013 (cit. on pp. 7, 13).

[21] Soyong Shin, Juyong Kim, Eni Halilaj, and Michael J. Black. *WHAM: Reconstructing World-grounded Humans with Accurate 3D Motion*. 2024. arXiv: 2312.07531 [cs.CV]. URL: https://arxiv.org/abs/2312.07531 (cit. on p. 8).

[22] Ce Zheng, Sijie Zhu, Matias Mendieta, Taojiannan Yang, Chen Chen, and Zhengming Ding. «3D Human Pose Estimation with Spatial and Temporal Transformers». In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2021, pp. 11656–11665 (cit. on p. 9).

[23] Soroush Mehraban, Vida Adeli, and Babak Taati. *MotionAGFormer: Enhancing 3D Human Pose Estimation with a Transformer-GCNFormer Network*. 2023. arXiv: 2310.16288 [cs.CV]. URL: https://arxiv.org/abs/2310.16288 (cit. on p. 9).

[24] Chongyang Wang, Yuntao Wang, Yuan Gao, Tin Lun Lam, and Yuanchun Shi. «PepperPose: Full-Body Pose Estimation with a Companion Robot». In: *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM. 2024, pp. 1–16 (cit. on p. 11).

[25] 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». In: *Journal of NeuroEngineering and Rehabilitation* 17.1 (2020), p. 97. DOI: 10.1186/s12984-020-00728-9. URL: https://doi.org/10.1186/s12984-020-00728-9 (cit. on pp. 13, 17).

[26] Samuel Rupprechter et al. «A Clinically Interpretable Computer-Vision Based Method for Quantifying Gait in Parkinson's Disease». In: *Sensors* 21.16 (2021). ISSN: 1424-8220. DOI: 10.3390/s21165437. URL: https://www.mdpi.com/1424-8220/21/16/5437 (cit. on p. 13).

[27] Jorge Marquez Chavez and Wei Tang. «A Vision-Based System for Stage Classification of Parkinsonian Gait Using Machine Learning and Synthetic Data». In: *Sensors* 22.12 (2022). ISSN: 1424-8220. DOI: 10.3390/s22124463. URL: https://www.mdpi.com/1424-8220/22/12/4463 (cit. on p. 14).

[28] Amirhossein Dadashzadeh, Shuchao Duan, Alan Whone, and Majid Mirmehdi. «PECoP: Parameter Efficient Continual Pretraining for Action Quality Assessment». In: *arXiv preprint arXiv:2311.07603* (2023) (cit. on pp. 14, 16).

[29] 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*. 2024. arXiv: 2405.17817 [cs.CV]. URL: https://arxiv.org/abs/2405.17817 (cit. on pp. 14, 15, 19, 24).

[30] Thiago Kenzo Fujioka Shida et al. «A public data set of walking full-body kinematics and kinetics in individuals with Parkinson's disease». In: *Frontiers in Neuroscience* 17 (2023), p. 992585. DOI: 10.3389/fnins.2023.992585 (cit. on pp. 16, 55).

[31] 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». In: *The Journals of Gerontology: Series A* 75.6 (2020), pp. 1148–1153. DOI: 10.1093/gerona/glz187 (cit. on p. 17).

[32] 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». In: *IEEE Journal of Biomedical and Health Informatics* 26.8 (2022), pp. 4020–4031 (cit. on p. 17).

[33] Trung-Hieu Hoang, Christopher Zallek, and Minh N Do. «Smartphone-based digitized neurological examination toolbox for multi-test neurological abnormality detection and documentation». In: *IEEE Journal of Biomedical and Health Informatics* (2024) (cit. on p. 17).

[34] 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*. Springer. 2023, pp. 72–82 (cit. on p. 17).

[35] 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». In: *Movement Disorders* 25.15 (2010), pp. 2563–2570 (cit. on p. 17).

[36] 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». In: *Journal of NeuroEngineering and Rehabilitation* 19.1 (2022), p. 48 (cit. on p. 17).

[37] 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». In: *npj Parkinson's Disease* 5.1 (2019), p. 25 (cit. on p. 17).

[38] 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». In: *The Journals of Gerontology: Series A* 75.6 (2020), pp. 1148–1153 (cit. on p. 17).

[39] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. *VideoMAE: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training*. 2022. arXiv: 2203.12602 [cs.CV]. URL: https://arxiv.org/abs/2203.12602 (cit. on p. 21).

[40] Wenhan Wu, Yilei Hua, Ce Zheng, Shiqian Wu, Chen Chen, and Aidong Lu. *SkeletonMAE: Spatial-Temporal Masked Autoencoders for Self-supervised Skeleton Action Recognition*. 2023. arXiv: 2209.02399 [cs.CV]. URL: https://arxiv.org/abs/2209.02399 (cit. on p. 21).

[41] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. *LoRA: Low-Rank Adaptation of Large Language Models*. 2021. arXiv: 2106.09685 [cs.CL]. URL: https://arxiv.org/abs/2106.09685 (cit. on pp. 22, 34).

[42] Jim Lawrence, Javier Bernal, and Christoph Witzgall. «A Purely Algebraic Justification of the Kabsch-Umeyama Algorithm». In: *Journal of Research of the National Institute of Standards and Technology* 124 (Oct. 2019). ISSN: 2165-7254. DOI: 10.6028/jres.124.028. URL: http://dx.doi.org/10.6028/jres.124.028 (cit. on p. 38).

[43] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. *A Simple Framework for Contrastive Learning of Visual Representations*. 2020. arXiv: 2002.05709 [cs.LG]. URL: https://arxiv.org/abs/2002.05709 (cit. on p. 52).