PIGDAssess: Wearable Dual-Task Sensing for Self-Administered PIGD Assessment in Parkinson's Disease

YIZHEN ZHANG, The Hong Kong University of Science and Technology, Hong Kong SAR
JINJIAN WANG, The Hong Kong University of Science and Technology, Hong Kong SAR
WENTAO XIE, The Hong Kong University of Science and Technology, Hong Kong SAR
QINGYONG HU, The Hong Kong University of Science and Technology, Hong Kong SAR
HAIYAN HU, The Hong Kong University of Science and Technology, Hong Kong SAR
GUIHUA LI, Jinan University Affiliated Guangdong Second Provincial General Hospital, China
QIAN ZHANG*, The Hong Kong University of Science and Technology, Hong Kong SAR

Postural instability and gait difficulty (PIGD) are leading causes of falls in Parkinson's disease (PD), yet current clinical assessment is infrequent, subjective, and requires expert supervision. Such assessments cannot be safely performed at home. We introduce PIGDAssess, the first wearable system that enables fully self-administered at-home estimation of all four UPDRS–PIGD subitems, including postural stability, without clinician involvement. Patients wear three commodity IMUs (lumbar and both feet) and perform brief sit-to-stand, standing, and walking tasks under both single-task and dual-task (serial–3 subtraction) conditions. To robustly infer clinically meaningful scores from dual-task movement segments, we introduce three key modules: (1) a paired single/dual-task encoder that captures balance-sensitive signatures; (2) a multi-action, multi-task fusion module that jointly predicts all four PIGD subscores from one short protocol; and (3) a prior-guided adaptation stage that reduces subject-to-subject variability while preserving the ordinal 0–3 clinical severity structure. We evaluate PIGDAssess on data from 35 individuals with PD, which were collected in collaboration with a partner hospital in both clinical settings and patients' homes, and were annotated with clinician-scored UPDRS–PIGD labels. Averaged across all four PIGD subitems, the system achieves 0.83 accuracy, 0.87 macro-F1, and 0.25 MAE. For Postural Stability, which normally requires a hands-on pull test, PIGDAssess reaches 0.89 accuracy using only three IMUs in a self-administered protocol. These results suggest a path toward frequent, low-burden, clinician-free PD assessment at home. We further conducted structured interviews with four neurologists, supporting the clinical usefulness and deployment feasibility of PIGDAssess. We have released the first IMU dataset with item-level UPDRS–PIGD annotations to enable reproducible benchmarking and accelerate clinical translation.

CCS Concepts: • **Human-centered computing → Ubiquitous and mobile computing systems and tools**; • **Computing methodologies → Machine learning**; • **Applied computing → Health informatics**.

Additional Key Words and Phrases: Postural Instability and Gait Disorders, Wearable device, Neural network

*Corresponding author.

Authors' Contact Information: Yizhen Zhang, The Hong Kong University of Science and Technology, Hong Kong SAR, yzhangtf@connect.ust.hk; Jinjian Wang, The Hong Kong University of Science and Technology, Hong Kong SAR, jwangjx@connect.ust.hk; Wentao Xie, The Hong Kong University of Science and Technology, Hong Kong SAR, wentaox@ust.hk; Qingyong Hu, The Hong Kong University of Science and Technology, Hong Kong SAR, qhuag@connect.ust.hk; Haiyan Hu, The Hong Kong University of Science and Technology, Hong Kong SAR, hhuap@connect.ust.hk; Guihua Li, Jinan University Affiliated Guangdong Second Provincial General Hospital, Second Clinical College, Southern Medical University, China, guihuali19790302@sina.com; Qian Zhang (corresponding author), The Hong Kong University of Science and Technology, Hong Kong SAR, qianzh@cse.ust.hk.

ACM 2474-9567/2026/6-ART75
https://doi.org/10.1145/3810193

ACM Reference Format:
Yizhen Zhang, Jinjian Wang, Wentao Xie, Qingyong Hu, Haiyan Hu, Guihua Li, and Qian Zhang. 2026. PIGDAssess: Wearable Dual-Task Sensing for Self-Administered PIGD Assessment in Parkinson's Disease. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 10, 2, Article 75 (June 2026), 31 pages. https://doi.org/10.1145/3810193

1 INTRODUCTION

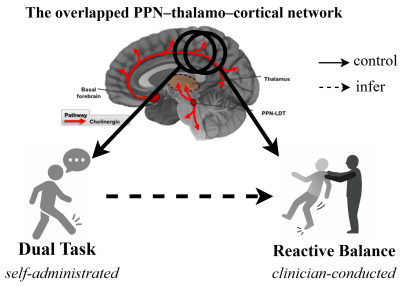
Parkinson's disease (PD) is one of the most prevalent neurodegenerative diseases that affects roughly 10 million people worldwide [8]. With the onset of PD, patients will experience various types of motor disorders, among which, up to 75% of patients develop postural instability and gait difficulty (PIGD) [42], which will largely compromise the patient's ability to stand, walk, and perform other daily activities. Furthermore, PIGD can be fatal because of the sharply elevated fall risk, and according to medical research, about 60.5% of PIGD patients fall annually [9]. To control the symptoms of PIGD and to reduce the risk of falling, clinicians prescribe medication and recommend regular PIGD assessment to adjust the treatment strategy promptly, based on the disease progression.

However, the need for regular PIGD assessment is unmet for many patients due to the substantial time and labor required, especially in developing countries. The conventional way of assessing PIGD is through the Unified Parkinson's Disease Rating Scale (UPDRS) [2], where a clinician needs to guide the patients to conduct a series of predefined maneuvers and rate their symptom severity based on subjective observations of the quality of the maneuvers (see Section 2). This assessment is only available in clinics and hospitals and is administered by a certified clinician, making it hard to access for regular assessment. Also, the complicated action maneuvers involved introduce additional overhead to the patients, restricting the regular conduction of the assessment.

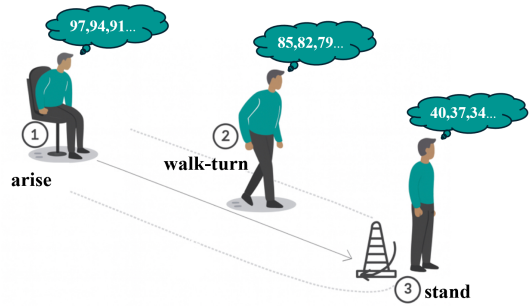
In this work, our goal is to change this picture by designing a home-based system to automatically assess PIGD severity. This allows PD patients to monitor their PIGD status more frequently on their own, then make immediate self-adjustments to their daily activity and fall-prevention strategies, and seek timely medication adjustments from their clinician. In turn, clinicians gain a clearer view of symptom progression between visits. Specifically, we want to assess PIGD against the four tasks listed in the UPDRS that define PIGD severity [33, 36]: (i) ARIS: the arising-from-chair test (item 3.9 in UPDRS, same for the following), (ii) GAIT: the gait test (3.10), (iii) STAB: the postural stability test (3.12), and (iv) POST: the posture test (3.13). However, achieving home assessment of these tests requires special design considerations, as the current PIGD workflow is explicitly clinician-led, where the standardized tasks must be taught and supervised by a clinician. In particular, the postural stability test even requires the clinician to pull the patients backward and examine the compensatory response. This cannot be self-administered at home and, if attempted, would pose safety risks (backward pulling may trigger a fall). Therefore, a new PIGD assessment workflow needs to be established to fit for home-based self-assessment.

Prior work [10, 11, 27, 29, 33] has made important progress toward at-home PIGD assessment, showing that, even without direct supervision, patients can perform self-initiated daily activities that resemble standardized PIGD tasks. These studies provide valuable evidence that simple activities such as standing up and walking may reflect aspects of neurological function [4], and they have helped establish the feasibility of home-based monitoring. At the same time, some components of the UPDRS, particularly those requiring clinician-elicited perturbation (such as the pull test for postural stability), remain difficult to approximate in an unsupervised home setting. Existing studies [27, 29, 33] suggest that free-living gait or standardized self-initiated balance tasks capture postural stability only to a limited extent, with reported classification accuracy around 60%. This likely reflects the inherent difficulty of assessing reactive balance (STAB) using features derived solely from self-generated movements. Therefore, while prior studies provide an important foundation, existing settings and datasets still leave room for improvement in home-based PIGD evaluation.

Nevertheless, medical research reveals that performing a motor task under cognitive load (*i.e.*, *dual-tasking*, such as walking while doing serial–3 subtraction) perturbs balance control and gait regulation using neural resources that overlap with those recruited during a pull test [1, 37]. Moreover, prior work in motor control and



(a) Neuroscience rationale[1]: the overlapping circuitry supports both dual-tasking and reactive balance, so dual-tasking can act as a surrogate for probing reactive balance.



(b) Assessment protocol: a self-administered protocol with three actions: arise, walk-turn, and stand. Each action is performed under cognitive load and recorded by wearable sensors to estimate PIGD severity.

Fig. 1. Based on the rationale illustrated on the left, we designed the self-administered action protocol shown on the right.

neurophysiology has shown that dual-tasking reveals deficits in postural control and stepping responses that resemble what clinicians probe with the backward pull test [24, 32, 38]. Intuitively, this suggests that dual-task behavior during standing and walking may serve as a proxy for reactive balance—and therefore for the postural stability item—without requiring anyone to physically pull the patient (Fig. 1(a)). These observations give us an opportunity to design a home-based action sequence that could reveal PIGD features in a self-service manner (Fig. 1(b)), from which motion sensors can be used to characterize such abnormality and predict PIGD scores.

Building on this insight, we develop PIGDAssess, a wearable sensing and learning system for at-home PIGD self-assessment. To use PIGDAssess, patients only need to wear three off-the-shelf IMUs (on the lumbar and both feet) and perform three short, clinically grounded actions: `arise` (arising from a chair), `walk-turn` (walking with turns), and `stand` (standing). Note that each of the above three actions should be conducted twice, once as a standalone single-task and once with dual-tasking while computing serial–3 subtraction. PIGDAssess ingests the IMU signals during the above actions and estimates *all four* UPDRS–PIGD subitems, including the postural stability test. The goal is continuous, repeatable fall-risk monitoring at home, without requiring a clinician-administered pull test.

Turning that goal into a practical system is non-trivial. We identify three technical challenges that emerge when moving from clinic-based scoring to self-administration with commodity sensors. The **first challenge** is how to extract clinically meaningful balance and gait indicators from raw IMU signals under single/dual task. As discussed in Section 3.3, the performance drop from single-task to dual-task motions (*i.e.*, the dual-task cost) shows higher correlation with PIGD. However, in practice, these two trials are not well-controlled replications of the same action with different cognitive loads. Rather, they are affected by various random factors, such as misalignment in time, especially when the patient is conducting the test at home on their own (see Fig. 3). Therefore, our design needs to reliably capture the decremental motion characteristics brought about by increased cognitive load while combating the randomness between action trials. Even if we can learn clinically meaningful representations from these IMU windows, the **second challenge** is how to map them to the four required UPDRS PIGD subitems. Although our action design discussed above tries to match these UPDRS items as much as possible, each action may reflect motor abnormalities in more than one of these items. For example, the action `walk-turn` carries information about one's gait condition but also implies posture and posture stability. Therefore,

[1]This figure is an illustrative schematic. For a detailed explanation, please refer to the cited reference.

when predicting each of the four UPDRS items, our design needs to strategically fuse the motion readings of the three actions under both single and dual-task conditions. Finally, the **third challenge** is substantial variability between patients. People with PD move very differently, even at the same nominal PIGD severity, with factors and comorbidities such as rigidity, bradykinesia, and compensatory posture also affecting the patient's moving behaviors. Furthermore, dual-tasking amplifies those differences due to the increased cognitive load. Therefore, a deployable design must address this heterogeneity across subjects and focus only on extracting PIGD-focused abnormalities.

We address these challenges with three components in PIGDAssess:

a) **Paired single/dual-task encoding for feature discovery.** To move beyond ad-hoc "dual-task cost" metrics, we introduce a pairing-based encoding module. Each recorded action is segmented into short windows (e.g., 3 s for `arise`, 10 s for `walk-turn` and `stand`); windows from the single-task and dual-task runs are then aligned and paired using three pairing strategies that capture different forms of task-induced perturbation. The resulting paired segments are fed as joint inputs to a shared temporal encoder. This forces the network to learn the patient's own cognitive-load–induced perturbations to balance and gait.
b) **Multi-action, multi-task fusion for comprehensive scoring.** PIGDAssess encodes each action (`arise`, `walk-turn`, `stand`) into an "action token" that contains both single-task and dual-task embeddings. An Action-weighted Task Head then learns, for each PIGD subitem, how much to attend to each action token, and fuses them into a subitem-specific representation. All four UPDRS–PIGD items are predicted jointly. This design directly serves the requirement of full PIGD self-assessment, rather than optimizing for a single item in isolation.
c) **Prior-guided adaptation for cross-subject robustness.** Finally, PIGDAssess applies a prior-guided adaptation stage to bridge patient-to-patient variability. We use a prior-informed conditional alignment step to reduce nuisance variation between subjects while respecting biomechanically plausible similarities, and an ordinal continuity regularizer to preserve the ordered (0–3) severity structure of each subitem. This makes predicted scores more comparable across patients without blurring clinically meaningful progression.

We collaborated with a hospital's neurology department and evaluated PIGDAssess on data collected from 35 PD patients with three commercial wireless IMUs. Each participant performed all three actions under both single-task and dual-task conditions, and each session was annotated with clinician-scored UPDRS–PIGD scores. This dataset captures real movement strategies, compensations, and balance deficits observed in actual patients. We further conducted structured interviews with four neurologists to assess the practical usability of the system at its current accuracy level, providing additional support for its reliability. We will release this dataset—to our knowledge, the first IMU dataset covering all four UPDRS–PIGD items—to enable reproducibility and future work on digital fall-risk biomarkers. We summarize the contributions of this work as follows.

(1) We present the first home-based and self-service PD assessment system that estimates *all four* UPDRS–PIGD items, including the challenging postural stability, using only body-worn IMUs and easy-to-perform home actions—no clinician administration required. The system's practical usability is further examined through structured expert interviews with neurologists.
(2) We propose three technical components that directly address the core requirements of at-home PIGD self-assessment: (i) a paired single-/dual-task encoder for clinically grounded feature discovery (Challenge 1), (ii) a multi-action, multi-task fusion module that delivers all four items simultaneously (Challenge 2), and (iii) a prior-guided adaptation stage that aligns patients while preserving ordinal PIGD severity (Challenge 3).
(3) We collect and analyze a clinically validated dataset from 35 individuals with PD performing standing, gait, and chair-rise behaviors under cognitive load. We release this dataset to support reproducible benchmarking and accelerate research on wearable, clinician-light PIGD monitoring[49].

Table 1. Comparison of gait and postural instability (PIGD-related) assessment across representative approaches.

Criterion	Ours	Safarpour et al. [33]	PDMonitor® [29]	Ma et al. [27]	Vision-based [10, 11]
Sensors	3 IMUs	3 IMUs	5 IMUs	10 IMUs	RGB-Depth camera
Environment	Home / clinic	Home + clinic	Home	Clinic	Home
Test protocol	Easy tasks (~10 min)	Free-living + standardized sway tasks	Free-living	Standardized tasks	Standardized tasks
PIGD item-level	✓	×	×	✓	✓
Balance proxy	Dual task	Postural sway	Gait-based	Gait-based	Proactive stability maneuvers
Balance output	Item-level score estimates	Continuous kinematic metrics	Probability-based indices	Item-level score estimates	Coarse categorical labels

2 BACKGROUND AND RELATED WORK

Before we dive into the details of our design, we first outline the background of this research. We begin with a brief clinical introduction to PIGD assessment, followed by a discussion of related work.

2.1 PIGD Clinical Assessment

Current clinical assessment of PIGD severity relies on the Unified Parkinson's Disease Rating Scale [13, 20], which contains tests to evaluate the patient's ability to stand, walk, and maintain balance and stability, including item 3.9: arising from chair (ARIS), item 3.10: gait test (GAIT), item 3.12: postural stability test (STAB), and item 3.13: posture test (POST). For each test, a clinician will guide the patient to perform a standard action and rate on a scale of 0-4 to represent the severity. All these tests require intense clinician involvement, and in particular, STAB requires the clinician to pull the patient backward and assess the patient's ability to restore balance (*i.e.*, reactive balance), which requires careful administration and assessment skills. Therefore, PIGD assessment is hard to conduct in a home-based and self-administered manner.

2.2 Sensors-based PIGD Assessment

Many recent studies have laid an important foundation for assessing PIGD using wearable IMUs or vision-based sensing systems, offering valuable insights into motor fluctuations and the feasibility of technology-assisted monitoring.

First, existing methods have demonstrated the promise of long-term monitoring for capturing aspects of disease progression, although they have mainly emphasized aggregate scores. For example, Safarpour et al. [33] pioneered the use of week-long continuous monitoring to predict the aggregate PIGD subscore, achieving a correlation of 0.61. Building on this line of work, there is growing clinical interest in directly predicting individual PIGD items (e.g., specific UPDRS PIGD components), as summarized in Table 1. Similarly, the development of commercial wearable systems such as PDMonitor® [29] represents a significant step forward in recording free-living conditions and generating clinician-facing summaries, including an overall UPDRS Part III estimate and probability-based indices of gait and axial impairment. While PDMonitor® provides a postural instability index derived from home monitoring, this probability-based index is inferred from a combination of gait and

motor features rather than directly assessing individual UPDRS PIGD items. This suggests a natural opportunity to extend such probability-based indices toward the characterization of item-level reactive balance impairments. Vision-based approaches using RGB-D cameras can also record standardized motor tasks in the home and apply machine learning to classify task performance [10, 11]. Although these systems can reliably distinguish people with Parkinson's disease from healthy controls, estimating disease severity within the PD population remains more challenging.

Second, assessing reactive balance remains an especially important challenge. The STAB item uniquely probes reactive balance—the component most strongly linked to falls—and therefore requires a clinician-elicited perturbation [6, 31]. Recent work has begun to explore item-level prediction of gait and posture components under controlled clinical conditions. For example, Ma et al. [27] used multi-IMU measurements collected during a standardized walking task to predict MDS-UPDRS III PIGD items, including STAB. While Ma et al. successfully predicted several PIGD items, their findings regarding STAB provide useful evidence that reactive balance is distinct from self-initiated walking. The complex nature of this item—reflected in its moderate association with gait features—highlights the potential value of looking beyond gait-derived proxies toward more specialized assessment strategies. Recent work [11] also relates performance on self-initiated balance tasks (i.e., proactive stability maneuvers) to overall PIGD severity; this represents a meaningful step toward simplified balance assessment. Their framework achieves severity classification with an accuracy of 64.3%–71.4%, providing a valuable basis for future extension from coarse severity categories to the item-level subscores that clinicians rely on for precise treatment guidance.

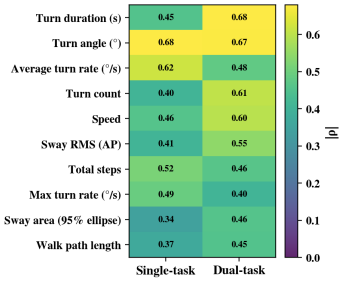
Taken together, these prior efforts provide a strong foundation while also highlighting an important opportunity for a PIGD assessment system that can be deployed in the home, provides fine-grained, item-level information to support clinical decision-making, and, critically, enables the assessment of reactive balance, which is most closely associated with fall risk [6, 31].

2.3 Datasets and Assessment in Parkinson's Disease

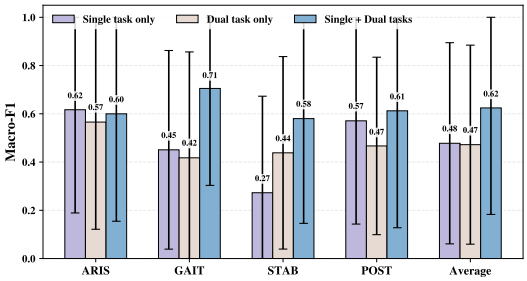
The growing availability of real-world Parkinson's disease datasets has provided a valuable foundation for digital health research. At the same time, aligning these resources with clinical PIGD assessment remains a meaningful and rewarding challenge. As noted by Kirk et al. [23], real-world digital mobility outcomes have shown promising but only weak to moderate associations with clinical motor scores, in part because existing datasets have primarily focused on continuous gait characterization rather than temporally aligned, item-level UPDRS annotations. In addition, relationships with the PIGD subscore have not yet been consistently established, and the reactive balance component assessed by the STAB item remains an important frontier in free-living data, as its assessment typically requires externally induced perturbations that are not easily captured from self-initiated walking alone.

In parallel, a wide range of approaches have been proposed to enable objective and continuous assessment of Parkinson's disease motor symptoms. These include wearable IMUs for gait and balance analysis during standardized tasks [48], vision-based pose estimation from home videos [16, 34], and force plate or radar-based systems for postural sway and gait analysis in controlled settings [15, 47]. At the system level, recent digital health platforms have integrated wearable, smartphone, speech, and video data to support ecologically valid monitoring of Parkinson's disease, primarily for PD detection or estimation of overall motor severity [14, 19, 26, 40, 45, 50]. Together, these advances have substantially improved ecologically valid monitoring for PD detection and overall severity, while also creating valuable opportunities to refine such systems so that they more closely mirror clinical PIGD assessment.

Consequently, these findings motivate the development of new data collection frameworks specifically designed to support the investigation of item-level PIGD assessment in real-world settings.



(a). Top 10 kinematic features correlated with STAB. Color shows $|\rho|$ under single-task and dual-task; many features are more correlated under dual-task.



(b). Macro-F1 under LOSO of Elastic Net + SVM using single-task, dual-task, or both. Combining both gives the best performance for most items.

Fig. 2. PIGD analysis with kinematic features: (a) Dual-tasking strengthens the correlation with STAB. (b) Using both single- and dual-task features improves classification.

2.4 Intended Usage Scenario

Because of limited medical resources, routine care usually includes an initial 2–4 week *medication adjustment period*, followed by clinic reviews every 6–12 months [39]. This schedule is not frequent enough to guide medication changes and often misses worsening between visits, leading to preventable falls, injuries, and higher costs[7, 36].

To address this gap, we enable brief, patient-led self-assessments during daily life. During medication adjustment periods, assessments can be performed daily to capture item-level PIGD fluctuations for clinician review, while weekly assessments during clinically stable periods support the construction of long-term trajectories over months to a year. Trend summaries and threshold-based alerts enable timely remote intervention when deterioration is detected, with the goal of reducing fall-related hospitalizations between scheduled clinic visits.

3 RATIONALES AND CHALLENGES

In this section, we first provide a neuroscience-motivated rationale for why dual-task paradigms can be used to infer reactive balance, a key component of PIGD assessment. We then present a preliminary study demonstrating the feasibility of our approach—namely, replacing clinician-supervised UPDRS tasks with simpler, self-administered tasks. Finally, we discuss the practical challenges involved in realizing such a design in real-world settings.

3.1 Neuroscience Rationale

The above tasks involve the activation of different neural pathways in the brain, and therefore, by asking the patients to do these tasks, a clinician can evaluate different neurological functions. For example, when performing STAB, a pull test, the brain's pedunculopontine nucleus (PPN) activates the thalamo-cortical circuits to manage the subject's responses. If the patient performs poorly in this test, it implies damage to the above neural pathway. However, performing the tasks defined in the UPDRS is not the only way to activate these neural pathways. For example, research has shown that if a subject is asked to walk while conducting cognitive tasks (*i.e.*, dual-tasking), such as doing arithmetic calculations, the neural pathways involved largely overlapped with those used in the response of a pull test [12, 17, 21, 30, 37]. In particular, dual-tasking (simultaneous movement and cognition) elicits EEG patterns resembling reactive postural responses [1]. This pattern is consistent with shared attentional resources [3, 28] and shared cortical dynamics [35]. Intuitively, when a person walks while doing mental math, the brain activates control networks that are also involved in keeping balance when the body is suddenly perturbed. This gives us an opportunity to design alternative physical tasks that, on one hand, are easier to self-administer and, on the other hand, largely preserve the original neural activation.

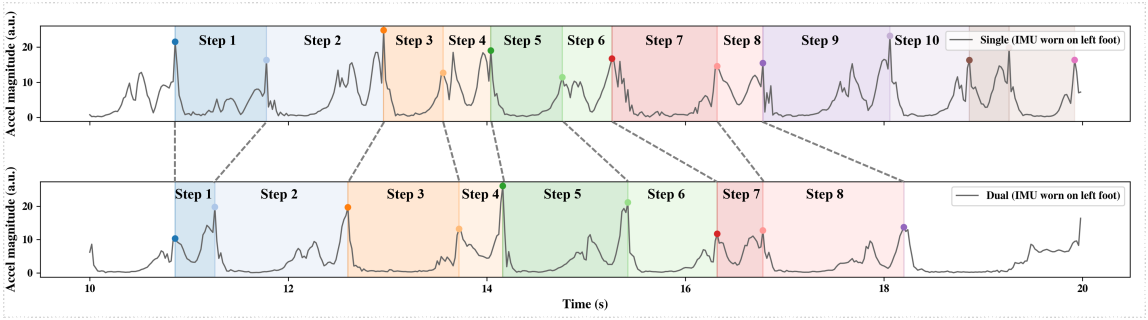


Fig. 3. IMU acceleration magnitude during the walk–turn action for one participant under single-task (top) and dual-task (bottom). Shaded regions mark individual steps. The two trials are temporally misaligned: within the same time window, the participant takes a different number of steps and is in different gait phases.

3.2 Motion Task Design

As discussed above, we want to design an easy-to-perform task set to support self-assessment of PIGD at home, with the help of the dual-task principle. In practice, *serial subtraction* is the most common and sufficiently strong cognitive load while walking or standing [41, 43]. Taking the standard UPDRS-PIGD items and self-administered constraints into account, our **assessment protocol** uses three simple actions—**arise** (rising from a chair), **walk-turn** (walking two laps along a 3 m path with turns), and **stand** (static standing for 30 seconds)—each performed under both single-task and dual-task conditions, without clinician supervision.

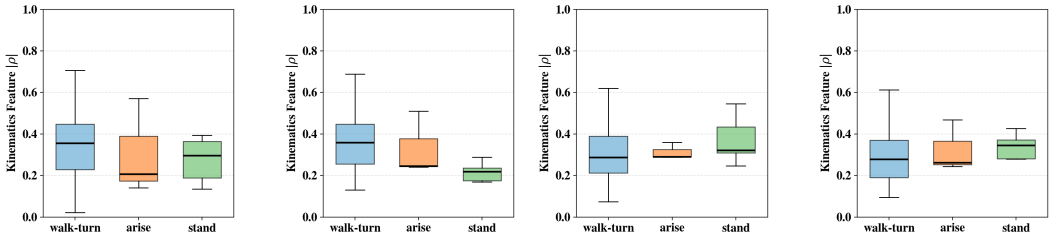
3.3 From Single and Dual Task Motions to PIGD Ratings

We first investigate how single-task and dual-task motions correlate with PIGD ratings. Specifically, we extract kinematic parameters from the IMU readings of 35 patients conducting the above tasks and analyze their correlation with the PIGD score. For each task, we compute IMU-derived kinematic and statistical features under single-task and dual-task conditions and analyze their associations with PIGD via Spearman correlation. A heatmap of the top 10 parameters most correlated with the STAB task (Fig. 1(a)) indicates that dual-tasking increases correlation in the majority of kinematic features such as turn duration and walking speed, suggesting that dual-task motion carries information specific to postural stability. We adopt a conventional pipeline based on hand-crafted kinematic features and classical machine learning models for this preliminary study. We use Elastic Net (EN) for feature selection, and train an SVM classifier under a leave-one-subject-out (LOSO) cross-validation protocol. We compare three input settings with features from single-task trials, dual-task trials, and both. As shown in Fig. 1(b), the combined setting achieves the highest macro-F1 across most UPDRS items, which motivates us to keep both single-task and dual-task inputs in our final design (Design Rationale R1).

3.4 Challenges

Although the previous preliminary study shows good correlation between the IMU readings of the above tasks and PIGD scores under controlled settings, there are a few technical gaps in designing a practical system that can be reliably deployed to patients.

Making use of the single- and dual-tasks. As discussed above, if we intend to use both single-task and dual-task features, a natural idea is to pair them using the same long analysis windows. In practice, however, the two trials are temporally misaligned: within the same fixed window, the single-task trial may contain more steps, and the same timestamps in the two trials may correspond to different gait phases, as shown in Fig. 3. As a result, the paired single-task and dual-task windows are not truly comparable, and this mismatch introduces substantial



(a) ARIS (b) GAIT (c) STAB (d) POST

Fig. 4. Action-specific absolute Spearman correlations ($|\rho|$) between kinematic features and PIGD items. Actions are similarly informative within an item, but the most informative action differs across items, motivating item-specific action weighting.

noise. Such pairing noise is especially harmful for small clinical datasets, and it represents an inherent challenge when leveraging dual-task paradigms for PIGD assessment.

Multi-item assessment across multiple actions. The second challenge is that we need to use three different actions — *arise*, *walk-turn*, and *stand* — to predict the four UPDRS–PIGD items. For a given item, all three actions look similarly informative. For example, for item STAB, the boxplots in Fig. 4(a) show the correlation between each kinematic feature and the clinical score. The correlation level is similar across the three actions, which means no single action is clearly dominant for that item. This motivates us to include all three actions rather than only one. However, when we compare across items (Fig. 4), different items rely on different actions. For example, features from *stand* have relatively low correlation for item GAIT, but much higher correlation for item STAB. This means that each item has its own most informative source, and we cannot use a single shared predictor head that treats all four PIGD items the same. Instead, the model must learn item-specific action weighting, so that each PIGD item can focus on the actions that are most informative for that item (Design Rationale R2).

Heterogeneity. Individuals with Parkinson's disease exhibit distinct gait patterns with varying degrees of tremor, rigidity, and bradykinesia. Clinically, three movement phenotypes are described [18]: (i) PIGD-only (PIGD-O), (ii) tremor-dominant with PIGD (PIGD+TD), and (iii) PIGD with additional motor abnormalities (PIGD+OMA). As illustrated in Fig. 5, adding cognitive load (dual-task walking) generally slows gait, and in general, slower gait is associated with higher STAB. However, phenotype-specific deviations are evident: for instance, a subset of PIGD+TD cases displays slow gait without correspondingly high STAB scores (see the lower-left region, where triangle markers intermingle with circular markers), and conversely, some PIGD+OMA cases show elevated STAB scores at intermediate speeds. Consequently, even patients with the same PIGD severity can move quite differently. As shown in Fig. 6, gait speed tends to decrease as the STAB score increases, for both single-task (blue) and dual-task (red) walking. However, within each score level, the range of gait speeds is still broad, indicating substantial inter-subject variability. Dual-task walking generally shifts gait speed lower and can widen these differences between individuals.

4 SYSTEM DESIGN

4.1 Overview

PIGDAssess is an IMU-based system that assesses postural instability and gait difficulty under single-task and dual-task conditions without clinician involvement. As illustrated in Fig. 8, patients perform at home both a single-task and a dual-task protocol consisting of three actions. IMU signals from three sensors are first preprocessed by aligning timelines and coordinate frames, and then normalizing. The processed data are fed into the PIGDAssess pipeline, which applies pairing-based data augmentation to generate multiple paired movement samples. Each sample is independently processed by a multi-action, multi-task framework, followed by a prior-guided adaptation

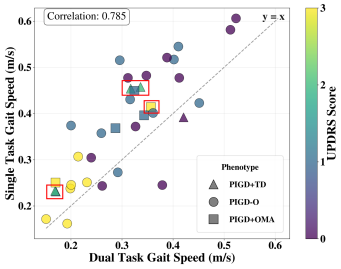


Fig. 5. Gait speed in single-task vs. dual-task walking, colored by STAB score and labeled by phenotype. Cognitive load generally slows gait, but different phenotypes deviate from the overall trend.

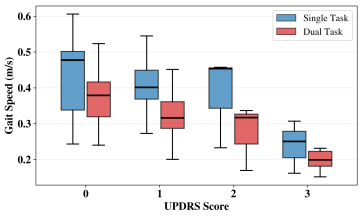


Fig. 6. Gait speed across STAB score levels for single-task (blue) and dual-task (red). Higher scores are linked to slower gait on average, but speed varies widely within each score level.

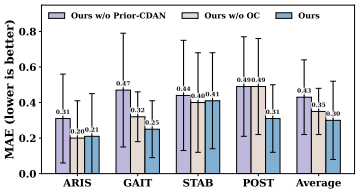


Fig. 7. Ablation of Prior-CDAN and ordinal continuity regularizer (OC). Both modules reduce MAE on all PIGD items; using both gives the lowest average MAE.

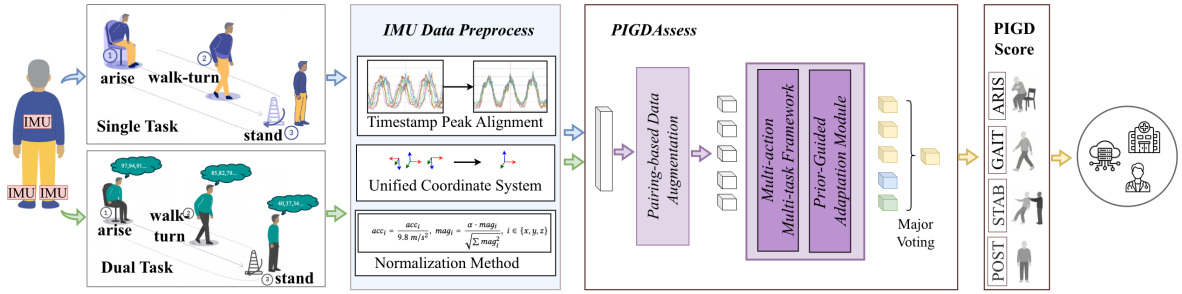


Fig. 8. Overview of PIGDAssess. Patients perform three actions (arise, walk–turn, stand) under single- and dual-task conditions with wearable IMUs. The signals are preprocessed (alignment, coordinate unification, normalization) and fed to our pipeline, which applies pairing-based augmentation, multi-action multi-task learning, and prior-guided adaptation to predict all four UPDRS–PIGD subscores.

module. Sample-level predictions are then aggregated via majority voting to produce a robust trial-level score for each PIGD subitem. These outputs provide quantitative references of the patient's PIGD status that can be synchronized to a clinician dashboard for timely monitoring of disease trends. In the following subsections, we introduce the detailed designs of our framework along with the structures of the three components.

4.2 Data Preprocessing

We collect data using three Bluetooth-connected IMUs mounted on the lumbar and both feet. After acquisition, we perform the following preprocessing. Because the devices transmit data asynchronously, we first synchronize their timestamps by aligning salient motion-induced peaks across sensors to a common reference, resulting in inter-device timestamp discrepancies within 10 ms. Each IMU provides tri-axial linear acceleration and tri-axial angular velocity; since the three IMUs are attached at different locations with arbitrary orientations, we transform these measurements to a unified body coordinate frame (x: forward, y: left, z: upward) via rotation matrices. Finally, following the effective normalization procedure reported in LIMU-BERT [44], we scale linear acceleration by gravitational acceleration and convert angular velocity to radians per second. These steps constitute the preprocessing pipeline for the IMU inputs.

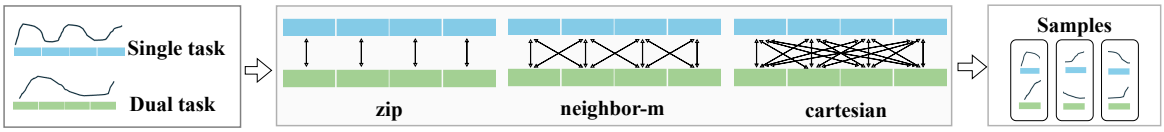


Fig. 9. Pairing-based data augmentation. Single-task and dual-task IMU streams are segmented into short windows, then paired using three schemes: (i) zip, (ii) neighbor-m, and (iii) cartesian, so that multiple paired samples can be generated from one trial to expose diverse dual-task costs.

4.3 Data Augmentation via Pairing

Building on the design rationale R1 in Section 3.3, PIGDAssess should process IMU data from both single-task and dual-task trials. Dual-tasking introduces an implicit *dual-task cost* (i.e., interference), which injects noise and phase misalignment across motion-phase pairs. To mitigate this inherited noise while enlarging the training set, we propose a **pairing-based data augmentation** strategy that can generate up to $m \times n$ paired samples from a single trial. Concretely, we segment the **walk-turn** and **stand** stages using 10 s windows and the **arise** stage using 3 s windows, as this duration is sufficient to capture the complete motion in our setting, thereby producing a large number of samples per trial. Each sample is independently processed by the model, and the final trial-level score is obtained via **majority voting** over all sample-level predictions. This yields diverse realizations of the dual-task cost while providing sufficient supervision for robust learning and noise-tolerant trial-level inference.

We instantiate three pairing schemes (Fig. 9) to better align single- and dual-task segments for actions with different temporal characteristics:

- **zip**: one-to-one, time-aligned pairing in chronological order;
- **neighbor-m**: pairing constrained to segments within $\pm m$ neighboring phases;
- **cartesian**: all-to-all pairing irrespective of temporal order.

We enable all three schemes for every action to leverage richer temporal features. And different motion types prefer different schemes: for example, **stand** is relatively stationary and benefits from *cartesian* pairing, whereas **arise** is transient and is better handled by *zip* pairing.

4.4 Multi-action Multi-task Framework

Based on our analysis in Section 3, we find that all three actions provide useful features for the assessment of multiple UPDRS-PIGD items. Together with the fact that these tasks are also correlated, this motivates a shared feature extractor and a multi-task design rather than four independent models. To this end, we propose a *multi-action, multi-task* framework that learns shared representations from actions while designing dynamic action-weighted heads to produce item-specific predictions, thereby improving sample efficiency, regularization, and consistency across sub-item scores.

As shown in Fig. 10, the multi-action, multi-task module includes two components: (i) three action-specific ResNet1D feature extractors (shared across single task and dual task, one per action) that serve as time-domain encoders, and (ii) four task-specific Action-weighted Task Heads that adaptively weight the action embeddings for each downstream task.

4.4.1 Time-domain ResNet1D Encoders. We use ResNet1D to extract temporal features from raw multi-channel IMU data. Residual connections and normalization layers enable stable training, while adaptive pooling produces fixed-length representations from variable-length signals. For each action, we send both the single-task and dual-task signals into the same encoder, and then concatenate the two embeddings instead of averaging them.

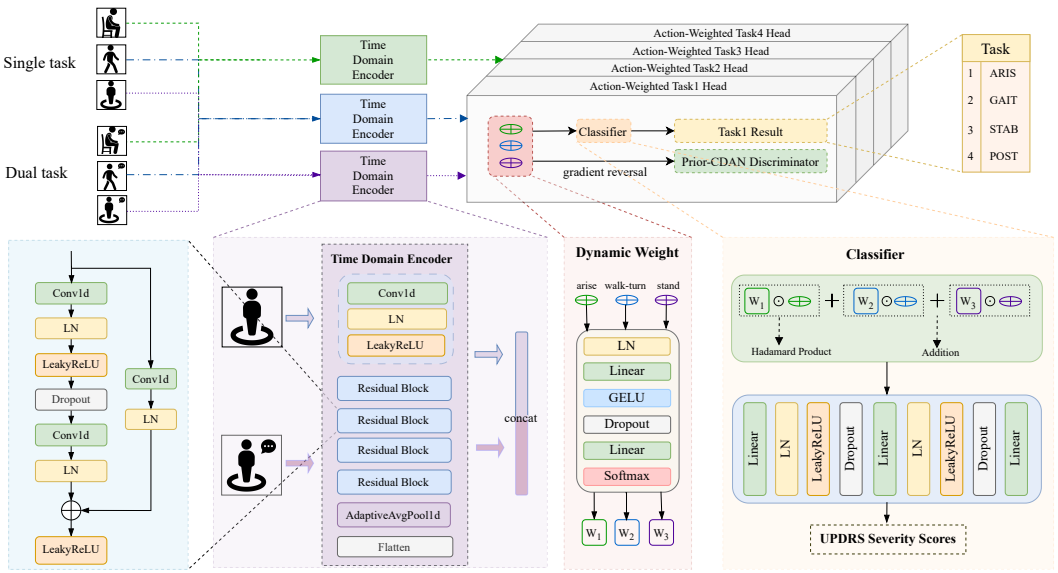


Fig. 10. Multi-action, multi-task architecture. Single-task and dual-task IMU streams from three actions are encoded by shared ResNet1D time-domain encoders, then four Action-weighted Task Heads assign action weights per item and output the four UPDRS–PIGD subscores, with a prior-guided CDAN branch for subject-level adaptation.

This way, we keep the extra motion changes caused by the cognitive task and let the later module decide how much to use from each. Because doing a dual task changes the motion feature compared to the single task, concatenating the two helps later attention/classification modules see those differences.

4.4.2 Action-weighted Task Head (AWTH). Within a multi-task framework, we propose the Action-weighted Task Head, a lightweight module that adaptively weights action embeddings, performs soft fusion, and feeds a task-specific classifier. Drawing on hierarchical attention and attention-based multiple-instance pooling [46], we adapt the scoring and fusion scheme to wearable IMU segments and to the multi-task PIGD assessment setting. Guided by the design rationale R2, AWTH reflects clinical reality: different PIGD items tend to draw their most informative cues from different actions; AWTH learns item-specific action weights to emphasize the most relevant motor cues without diluting them with a single shared head.

As illustrated in Algorithm 1, we first form three action branches. Each branch encodes both *single-task* and *dual-task* segments with a shared ResNet1D time-domain encoder; the two embeddings are then concatenated per action to retain complementary dynamics, yielding $\mathbf{e}^a \in \mathbb{R}^k$ for $a \in \{\texttt{arise}, \texttt{walk-turn}, \texttt{stand}\}$. We stack these into $\mathbf{E} \in \mathbb{R}^{B\times 3\times k}$.

A small two-layer scorer ϕ_t, consisting of LayerNorm followed by a two-layer MLP and shared across the three actions, maps each action embedding to a scalar score. A softmax along the action axis converts these scores into per-action mixture weights $\mathbf{m}$. The fused representation $\mathbf{f}_t$ is the weighted sum of the action embeddings and is then fed to a task-specific classifier to produce logits $\mathbf{z}_t \in \mathbb{R}^{B\times n_{\text{cls}}}$. Because different PIGD sub-items rely on

Algorithm 1: Action-weighted Task Head (AWTH)

Input: Minibatch $\{\mathbf{x}^a_{\text{single}}, \mathbf{x}^a_{\text{dual}}\}_{a\in\mathcal{A}}$, where $\mathcal{A} = \{\texttt{arise}, \texttt{walk-turn}, \texttt{stand}\}$; encoders $\{f^a_\theta\}$, task set $\mathcal{T}$, heads $\text{AWTH}_t = (\phi_t, \psi_t)$.

Output: For each task t: logits $\mathbf{z}_t$, mixture weights $\mathbf{m}_t$, fused feature $\mathbf{f}_t$.

/* Action encoders with single/dual concatenation */

for $a \in \mathcal{A}$ **do**
  $\mathbf{e}^a \leftarrow [\, f^a_\theta(\mathbf{x}^a_{\text{single}});\ f^a_\theta(\mathbf{x}^a_{\text{dual}}) \,]$
$\mathbf{E} \leftarrow [\, \mathbf{e}^{\texttt{arise}}, \mathbf{e}^{\texttt{walk-turn}}, \mathbf{e}^{\texttt{stand}} \,]$

/* Per-task action-weighted fusion and classification */

for $t \in \mathcal{T}$ **do**
  $\mathbf{S}_t \leftarrow \phi_t(\mathbf{E})$ // two-layer scorer, shared across actions
  $\mathbf{m}_t \leftarrow \text{softmax}(\mathbf{S}_t, \text{dim} = a)$
  $\mathbf{f}_t \leftarrow \sum_{a\in\mathcal{A}} \mathbf{m}_{t,a} \cdot \mathbf{e}^a$
  $\mathbf{z}_t \leftarrow \psi_t(\mathbf{f}_t)$
return $\{(\mathbf{z}_t, \mathbf{m}_t, \mathbf{f}_t)\}_{t\in\mathcal{T}}$

different motor cues, $\mathbf{m}$ adapts per item to emphasize the most informative action. In the multi-task setting, we instantiate one AWTH per PIGD item.

4.4.3 Multi-task Objective. As shown in Table 2, the UPDRS severity classes are highly imbalanced (e.g., many more samples in UPDRS 0 than UPDRS 2 for ARIS). To address this, we adopt a class-weighted focal loss within each task. For task t with C_t classes and batch $\mathcal{B}_t$, for each sample $i \in \mathcal{B}_t$, let $\mathbf{z}_{t,i} \in \mathbb{R}^{C_t}$ denote logits, $p_{t,i,c} = \text{softmax}(\mathbf{z}_{t,i})_c$ the predicted probability, and $y_{t,i} \in \{1, \ldots, C_t\}$ the ground-truth label. The per-sample focal loss is

$$\ell_{t,i} = -\,\alpha_{t,y_{t,i}} \left(1 - p_{t,i,y_{t,i}}\right)^{\gamma_t} \log p_{t,i,y_{t,i}}, \qquad L_t \;=\; \frac{1}{|\mathcal{B}_t|} \sum_{i\in\mathcal{B}_t} \ell_{t,i}, \tag{1}$$

where $\gamma_t > 0$ is the focusing parameter and $\alpha_{t,c}$ are class weights (e.g., set inversely proportional to class frequency and normalized).

In the multi-task setting, different sub-items may learn at disparate rates (some improve quickly while others lag). To adaptively balance tasks, we use Dynamic Weight Averaging (DWA) based on recent optimization dynamics. Let $L_t^{(e-1)}$ and $L_t^{(e-2)}$ be the average losses of task t from the previous two epochs.

We define a clipped ratio and compute temperature-scaled softmax weights as

$$r_t = \text{clip}\left(\frac{L_t^{(e-1)} + \varepsilon}{L_t^{(e-2)} + \varepsilon}, [\, r_{\min},\ r_{\max} \,]\right), \quad w_t = \frac{\exp(r_t/T)}{\sum_{t'} \exp(r_{t'}/T)},\ \sum_t w_t = 1. \tag{2}$$

where $T > 0$ is the temperature and $\varepsilon > 0$ prevents division by zero. When $e < 2$ (insufficient history), we use uniform weights $w_t = 1/|\mathcal{T}|$.

The final multi-task loss is the DWA-weighted sum of per-task focal losses:

$$L_{\text{multi-task}} \;=\; \sum_{t\in\mathcal{T}} w_t\, L_t. \tag{3}$$

4.5 Prior-guided Adaptation

We observe that co-occurring motor symptoms can distort the motion patterns used to score PIGD, producing outliers. Moreover, even within the same UPDRS grade, subjects exhibit large intra-class variability, leading

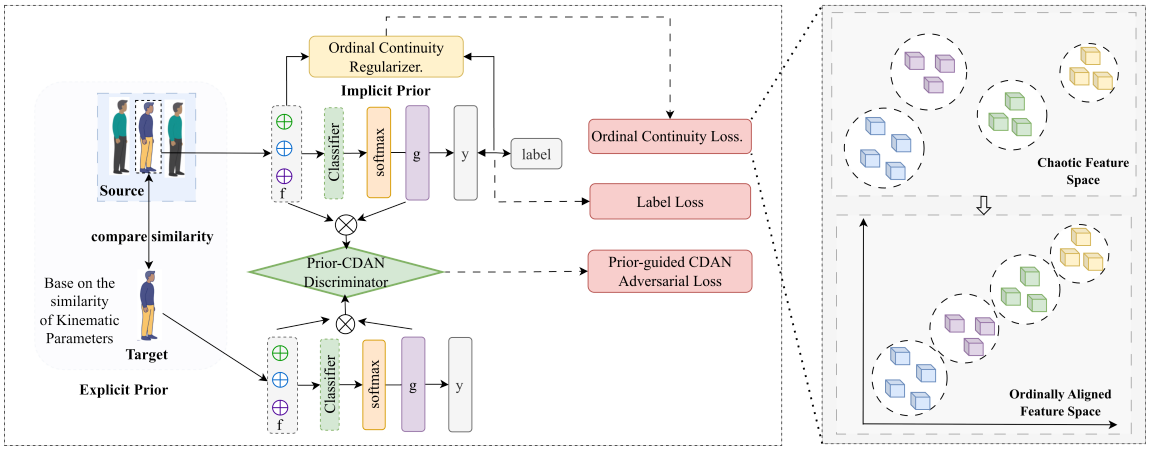


Fig. 11. Prior-guided adaptation. Source samples with similar kinematics are paired with each target sample and aligned by the Prior-CDAN. In parallel, an ordinal continuity regularizer keeps features of neighboring UPDRS levels close. The final loss combines label loss, CDAN loss, and continuity loss.

to blurred boundaries between classes in the motion-parameter space. To address these two challenges, we introduce two prior-driven components. First, we employ a kinematics-informed, prior-guided CDAN to condition domain alignment on clinically meaningful parameters, suppressing subject and symptom specific artifacts while preserving PIGD relevant pathology. Second, we enforce an ordinally aligned feature space that regularizes representations to vary smoothly across adjacent UPDRS levels, sharpening effective decision boundaries while respecting the ordinal structure.

4.5.1 Explicit Prior: Prior-guided Conditional Domain Alignment (Prior-CDAN). To suppress subject-specific and symptom-specific artifacts while preserving PIGD-relevant pathology, we introduce a *prior-guided conditional domain alignment* module. The objective is twofold: (i) to align movement representations across subjects so that the model does not overfit idiosyncratic gait styles, compensatory strategies, or co-occurring motor symptoms that are *not* scored by PIGD, and (ii) to do so without erasing clinically meaningful severity information. We achieve this by combining two components: an explicit kinematics prior and a class-conditional adversarial alignment.

We first reinterpret the two domains of a Conditional Domain Adversarial Network (CDAN)[25] as **seen vs. unseen subjects**. At each training step, minibatches from the training fold and the held-out validation fold are treated as source domain $\mathcal{D}_{\text{src}}$ and target domain $\mathcal{D}_{\text{tgt}}$, respectively. Samples are passed through the shared multi-action encoder and the task-specific AWTH, yielding fused task representations $\mathbf{f}_{t,i} \in \mathbb{R}^k$ and task logits $\mathbf{z}_{t,i} \in \mathbb{R}^{C_t}$ for each PIGD sub-item t. Class posteriors are computed as $\mathbf{g}_{t,i} = \text{softmax}(\mathbf{z}_{t,i})$.

Following CDAN, we construct class-conditioned features via an outer product,

$$\boldsymbol{\phi}_{t,i} = \text{vec}\big(\mathbf{g}^{\sharp}_{t,i} \otimes \mathbf{f}_{t,i}\big) \in \mathbb{R}^{C_t k}, \tag{4}$$

where $\mathbf{g}^{\sharp}_{t,i} = \text{stopgrad}(\mathbf{g}_{t,i})$ optionally prevents gradients from flowing back into the classifier. Stacking over the minibatch yields

$$\boldsymbol{\Phi}_t = [\,\boldsymbol{\phi}_{t,1}; \ldots; \boldsymbol{\phi}_{t,B}\,] \in \mathbb{R}^{(C_t k)\times B}. \tag{5}$$

$\boldsymbol{\Phi}_t$ is passed through a Gradient Reversal Layer (GRL) and then into a task-specific domain discriminator $D_t(\cdot)$, which is trained to distinguish seen from unseen subjects. Because D_t operates on features conditioned on

predicted severity, adversarial alignment is enforced *within* each severity level rather than collapsing all levels together, which helps preserve PIGD-related pathology.

To prevent indiscriminate alignment between fundamentally different movement patterns, we introduce an explicit **kinematic prior**. For each target sample x^{tgt}, we compute a high-dimensional biomechanical descriptor $\mathbf{v} = V(x^{\text{tgt}}) \in \mathbb{R}^D$ (with $D \approx 100$ IMU-derived parameters, such as gait regularity, turn stability, and sit-to-stand effort). We retrieve the top-k most biomechanically similar source samples x^{src} based on cosine similarity,

$$\text{sim}(\mathbf{v}_j, \mathbf{v}_i) = \frac{\langle \mathbf{v}_j, \mathbf{v}_i \rangle}{\|\mathbf{v}_j\|_2 \, \|\mathbf{v}_i\|_2}. \tag{6}$$

Only these matched source–target pairs are used to form adversarial minibatches. In effect, each validation subject is aligned only with training subjects exhibiting comparable movement mechanics, preventing the forced alignment of clearly distinct motor phenotypes (e.g., severely bradykinetic shuffling gait vs. near-normal gait). This prior-guided pairing suppresses nuisance variation such as compensatory posture or unrelated tremor, while retaining movement signatures that truly reflect PIGD severity.

For each sub-item t, the discriminator $D_t(\cdot)$ yields a task-specific adversarial loss $\mathcal{L}_{\text{adv}}^{(t)}$. Averaging across tasks gives

$$\mathcal{L}_{\text{adv}} = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \mathcal{L}_{\text{adv}}^{(t)}. \tag{7}$$

This loss is combined with the supervised objective. Overall, Prior-CDAN promotes invariance to subject identity and off-target symptom patterns, while preserving clinically meaningful PIGD-related impairment.

4.5.2 Implicit Priors: Ordinal Continuity Regularizer. UPDRS scores are discrete but *ordered* ($\{0, 1, 2, 3\}$), and clinically adjacent levels tend to change gradually rather than abruptly. We encode this clinical prior by enforcing smoothness of the learned representations across neighboring severities via a label-distance–weighted continuity regularizer applied *per task* (Fig. 11).

For each task t, we maintain a FIFO memory $\mathcal{Q}_t = \{(\mathbf{f}_t^{(j)}, y_j)\}_{j=1}^{K}$ that stores past fused features and their ordinal labels. Given the current batch features $\mathbf{F}_t \in \mathbb{R}^{B \times k}$ and labels $\mathbf{y} \in \{0, 1, 2, 3\}^B$, all features are L_2-normalized so that dot products correspond to cosine similarities. We first construct an ordinal-distance–weighted soft target distribution,

$$\boldsymbol{\alpha}_{i,\cdot} = \text{softmax}\Big(- |y_i - \mathbf{y}_{\mathcal{Q}_t}| / \tau \Big), \qquad \tau > 0, \tag{8}$$

where $\mathbf{y}_{\mathcal{Q}_t}$ denotes the labels stored in $\mathcal{Q}_t$ and τ controls the emphasis on nearby severities. In parallel, we compute predicted similarity scores

$$\mathbf{S}_t = \mathbf{F}_t \, \mathbf{Q}_t^{\top}, \tag{9}$$

where $\mathbf{Q}_t \in \mathbb{R}^{K \times k}$ stacks the normalized queue features.

The task-level continuity loss and the corresponding global regularizer are defined as

$$\mathcal{L}_{\text{con}}^{(t)} = -\frac{1}{B} \sum_{i=1}^{B} \sum_{j=1}^{K} \alpha_{i,j} \, [\log \text{softmax}(\mathbf{S}_t)]_{i,j}, \quad \mathcal{L}_{\text{con}} = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \mathcal{L}_{\text{con}}^{(t)}. \tag{10}$$

This regularizer penalizes configurations where samples with similar PIGD ordinal scores are far apart in the latent space, yielding representations that vary smoothly with clinical severity. As shown in Fig. 7, this constraint leads to a reduction in mean absolute error (MAE), reflecting smaller prediction deviations.

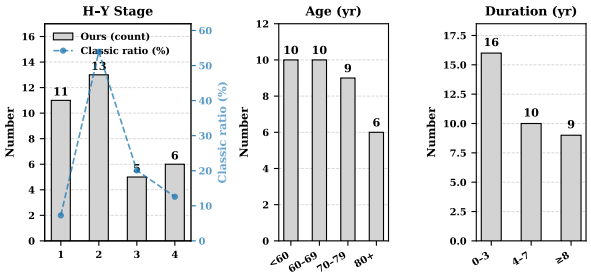


Fig. 12. Demographic and clinical distribution of the cohort (n=35), showing H–Y stages, age groups, and disease duration.

Table 2. UPDRS–PIGD severity distribution for the four assessed subitems (n=35).

PIGD	UPDRS Severity Scores			
	0	**1**	**2**	**3&4**
ARIS	20	9	2	4
GAIT	6	12	12	5
STAB	11	13	3	8
POST	9	4	12	10

4.6 Final Objective

Combining the above components, our overall training objective is

$$\mathcal{L}_{\text{total}} = \underbrace{\sum_{t \in \mathcal{T}} w_t \, L_t}_{\text{multi-task focal loss (class imbalance)}} + \lambda_{\text{adv}} \underbrace{\mathcal{L}_{\text{adv}}}_{\text{prior-guided CDAN alignment}} + \lambda_{\text{con}} \underbrace{\mathcal{L}_{\text{con}}}_{\text{ordinal continuity}} \tag{11}$$

where L_t is the focal loss for task t, w_t are task weights from DWA, λ_{adv} and λ_{con} control the strength of domain alignment and ordinal continuity, and λ_{grl} (inside the GRL) controls the adversarial pressure on the encoder.

5 IMPLEMENTATION

In this section, we describe the implementation of PIGDAssess. First, we present the data and our collection procedure; then we explain the system implementation from both the hardware and software perspectives.

5.1 Data Collection

5.1.1 Patient Recruitment. This study was conducted in collaboration with a tertiary medical center under our institute's IRB approval. Patients were recruited from outpatient clinics and inpatient wards under the supervision of experienced movement-disorder clinicians. Written informed consent was obtained from all participants. We enrolled 35 individuals with Parkinson's disease (20 women, 15 men; mean age 68.2 ± 10.5 years; disease duration 5.29 ± 4.2 years). As shown in Fig. 12, the cohort spans a wide range of Hoehn–Yahr stages (mean 2.17 ± 1.07), largely aligning with the classic clinical distribution [13] while intentionally including a higher proportion of early-stage (Stage 1) and severe (Stage 4) patients to ensure model robustness across the full spectrum of PIGD impairment. Recruiting patients for structured motor assessments posed practical challenges due to mobility limitations, fatigue, and the cognitive burden of repeated task performance. Despite these constraints, all participants completed the full protocol across clinical and home settings, yielding high-quality, longitudinally consistent data.

5.1.2 Ground Truth. Two neurologists independently assessed each patient using the UPDRS–PIGD subscale, and resolved discrepancies through consensus to establish the ground truth (Appendix Table 5). The severity distributions of the four PIGD subitems are summarized in Table 2.

Patients at UPDRS level 4 are typically unable to walk at all—even with assistive devices—due to severe postural instability. As a result, ambulation-based motion data cannot be meaningfully collected for this group. For patients at level 3, walking is possible but requires substantial support (e.g., walkers or caregiver assistance); however, the form and extent of this support vary widely and are difficult to standardize in unsupervised assessments.

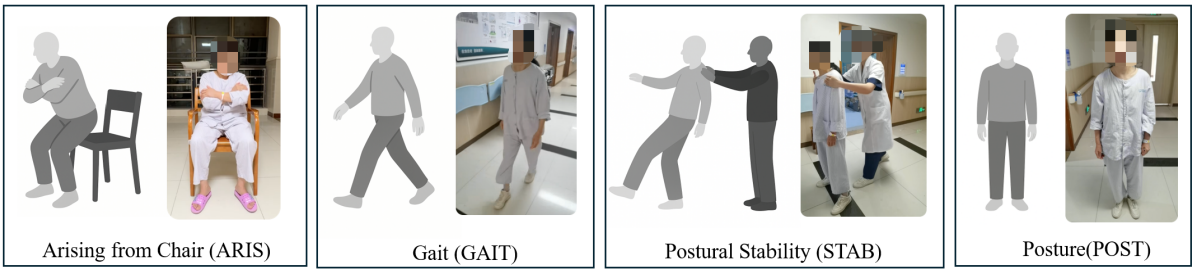


Fig. 13. Clinician-administered UPDRS–PIGD evaluations for the four subitems.

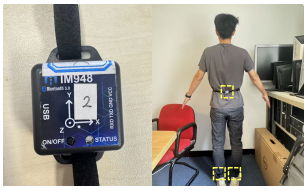


Fig. 14. Wireless IMU(left) and mounting location(right).

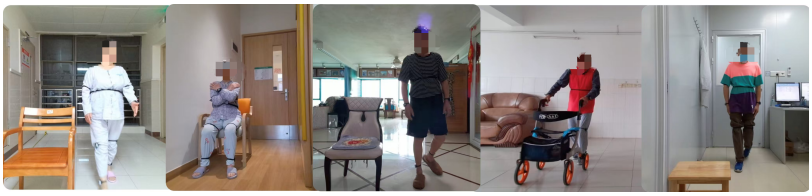


Fig. 15. Example data-collection environments for the self-administered PIGD protocol: outpatient clinic, inpatient unit, and home.

Given that reliable motion data cannot be acquired for level 4, and that level 3 behavior is highly variable and context-dependent, we merged levels 3 and 4 into a single category (3&4). This ensures that our assessment framework remains grounded in observable, quantifiable motor behavior, while avoiding ill-defined or unmeasurable movement patterns.

5.1.3 Experiment Design and Implementation. To enable self-administered or caregiver-supervised PIGD assessments, we designed an assessment protocol that asks patients to rise from a chair, walk two laps along a 3 m path, and stand still for 30 seconds, each under both single-task and dual-task conditions without clinician supervision.

Participants wore three commercial wireless inertial measurement units (IMUs): one on the lower lumbar region secured with an elastic belt and one on each foot (Fig. 14). Donning and doffing the sensors typically required less than 3 minutes and could be performed independently by most participants (with minimal caregiver assistance for the 4 participants at H-Y Stage 4). Prior to data collection, participants received standardized verbal instructions and a brief demonstration of the protocol, no real-time feedback or intervention was provided during task execution. Motion data were recorded continuously during each trial. Each IMU operated on an internal rechargeable battery with a battery life approximately 25 hours during Bluetooth-enabled data acquisition and transmission, which was sufficient for multiple assessment sessions without recharging. Experiments were conducted in outpatient clinics, inpatient units, and home environments (Fig. 15), demonstrating realistic deployment conditions. In total, each participant completed more than three trials under both single-task and dual-task conditions, resulting in over 10 minutes of data per participant and approximately 6 hours of data overall.

5.2 System Implementation

5.2.1 Hardware. We used three off-the-shelf, 9-axis wireless IMUs [5] (each weighing 20 g and equipped with a 250 mAh lithium battery) that are cheap, easy to don, and widely available. Each unit integrates a triaxial accelerometer, a triaxial gyroscope, and a magnetometer, sampled at 50 Hz.

Table 3. Comparison with baseline methods under LOSO. We report mean±std for Acc/F1/MAE. Statistical significance is assessed using BH-adjusted Wilcoxon signed-rank tests on fold-level metrics (one fold per subject), comparing each baseline to **Ours** within each task.

Method	ARIS			GAIT			STAB			POST			AVG		
	Acc↑	F1↑	MAE↓	Acc↑	F1↑	MAE↓	Acc↑	F1↑	MAE↓	Acc↑	F1↑	MAE↓	Acc↑	F1↑	MAE↓
BaselineA	0.45±0.24*	0.58±0.26*	0.90±0.60*	0.29±0.08*	0.44±0.10*	1.03±0.30*	0.28±0.10*	0.43±0.14*	1.23±0.44*	0.26±0.08*	0.41±0.10*	1.31±0.27*	0.32±0.09*	0.46±0.11*	1.11±0.33*
BaselineB	0.56±0.44*	0.60±0.44*	0.54±0.59*	0.67±0.41n.s.	0.71±0.40*	0.36±0.48n.s.	0.53±0.42*	0.58±0.43*	0.62±0.60*	0.61±0.48n.s.	0.61±0.48n.s.	0.45±0.60n.s.	0.59±0.24*	0.62±0.23*	0.49±0.29*
BaselineC	0.61±0.40*	0.66±0.39*	0.44±0.48*	0.62±0.45*	0.64±0.45*	0.38±0.45n.s.	0.60±0.43*	0.63±0.42*	0.52±0.60*	0.66±0.42n.s.	0.69±0.41n.s.	**0.39**±0.50n.s.	0.62±0.22*	0.66±0.22*	0.43±0.28*
BaselineD	0.67±0.41*	0.71±0.40*	0.45±0.65*	0.55±0.41*	0.60±0.40*	0.62±0.64*	0.74±0.37*	0.78±0.35*	0.38±0.58*	0.65±0.40*	0.70±0.38*	0.53±0.62*	0.65±0.23*	0.70±0.21*	0.49±0.44*
BaselineE	0.47±0.40*	0.53±0.39*	0.65±0.64*	0.40±0.40*	0.46±0.42*	0.70±0.52*	0.54±0.45*	0.57±0.45*	0.69±0.79*	0.52±0.39*	0.59±0.39*	0.80±0.81*	0.48±0.26*	0.54±0.25*	0.71±0.47*
Ours	**0.88±0.24**	**0.91±0.21**	**0.16±0.39**	**0.79±0.29**	**0.84±0.26**	**0.29±0.48**	**0.89±0.20**	**0.92±0.17**	**0.16±0.35**	**0.75±0.37**	**0.79±0.35**	0.40±**0.59**	**0.83±0.17**	**0.87±0.14**	**0.25±0.31**

'*' indicates $p < 0.05$, and n.s. denotes not significant.

5.2.2 Model Implementation. We implemented PIGDAssess in PyTorch 2.8.0 and trained it on a single NVIDIA GeForce RTX 4090 D GPU. The model consists of time-domain ResNet1D encoders, an Action-weighted Task Head (AWTH), and two adaptation components.

For each action, a ResNet1D encoder with four stages (channels 64, 128, 256, 512) maps 50 Hz IMU windows to 128-dimensional features. Single-task and dual-task windows share the same encoder and are concatenated to form a 256-dimensional action embedding; the three action embeddings are stacked as $\mathbf{E} \in \mathbb{R}^{B\times 3\times 256}$.

AWTH computes task-specific action weights and produces a fused embedding $\tilde{\mathbf{f}}_t \in \mathbb{R}^{B\times 256}$, which is mapped to logits $\hat{\mathbf{z}}_t \in \mathbb{R}^{B\times C_t}$ with $C_t = 4$ ordinal levels.

To improve robustness, we apply Prior-CDAN with a gradient reversal layer and an ordinal continuity regularizer. The model is trained end-to-end for 50 epochs using AdamW (batch size 128, learning rate 1×10^{-4}).

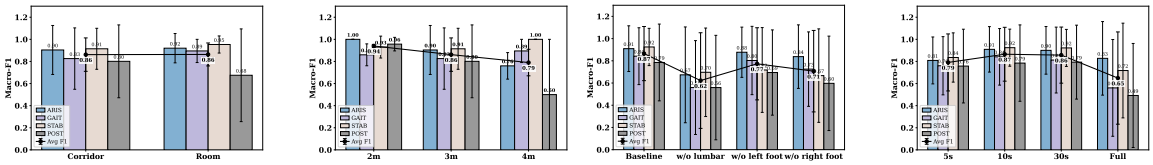
6 EVALUATION

We evaluated PIGDAssess on a cohort of 35 participants using leave-one-subject-out cross-validation. In each fold, one subject was held out as the test set, while data from the remaining 34 subjects were used for training. All trials from the held-out subject formed the test set, with one prediction generated per trial. Evaluation metrics were computed over all test trials in each fold, and final performance was reported as the mean across folds.

Given the four ordinal PIGD score categories, we report performance using Accuracy (Acc), Macro-F1 (F1), and Mean Absolute Error (MAE). Accuracy evaluates the overall correctness of the model, Macro-F1 balances the performance across all categories, making it suitable for imbalanced data, while MAE measures the average deviation between predicted and actual values:

$$\text{Acc} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{1}(\hat{y}_i = y_i), \quad \text{Macro-F1} = \frac{1}{C}\sum_{c=1}^{C}\frac{2\,\text{Prec}_c\,\text{Rec}_c}{\text{Prec}_c+\text{Rec}_c}, \quad \text{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left|\hat{y}_i - y_i\right|.$$

We organize the evaluation into four parts: baseline comparison, robustness analysis, demographic analysis, and ablation study. Baseline, robustness, and demographic analyses are conducted using LOSO cross-validation. For demographic analysis, LOSO predictions are further stratified by participant demographics and evaluated at the subgroup level. The ablation study instead targets sample-level effects. We therefore use 10-fold user-independent cross-validation and report sample-level performance without majority voting to better isolate the contribution of individual components.



(a) Environment-wise (b) Distance-wise (c) Missing IMU sensors (d) Segment-wise

Fig. 16. Robustness of PIGDAssess under different conditions. (a) Environment-wise: corridor vs. room, showing almost no change. (b) Distance-wise: varying walking distances (2–4 m) keeps macro-F1 above 0.79. (c) Missing IMU sensors: performance degrades most when removing the waist IMU, confirming its importance. (d) Segment-wise: a 10 s segment yields the most reliable performance across durations.

6.1 Baselines

We construct five baselines following common practices in prior work. Because current postural stability studies do not yet cover the full set of PIGD items, we instantiate our baselines by adapting the state-of-the-art methods from the closest problem settings:

- **Baseline-A.** Random predictions sampled according to the ground-truth class distribution.
- **Baseline-B.** Extract IMU-derived kinematic (e.g., stride, symmetry) and statistical features from both single-task and dual-task trials. Then perform feature selection with Elastic Net (EN) and train an SVM classifier with an RBF kernel. This baseline reflects a traditional pipeline without deep learning[33].
- **Baseline-C.** Identical to Baseline-B, except that the SVM classifier is replaced with a proportional-odds ordinal regression model to explicitly account for the ordinal nature of UPDRS scores.
- **Baseline-D.** Use a ResNet1D backbone with a multi-action encoder, and directly concatenate the three action embeddings as the input to the classifier, without any action-specific weighting [22].
- **Baseline-E.** Same as Baseline-D, except that the ResNet1D encoder is replaced with a pretrained LIMU-BERT encoder[44], which uses the Transformer architecture and is considered a good backbone.
- **Ours.** Starting from Baseline B, add our proposed modules AWTH, Prior-CDAN, and ordinal continuity regularizer to better fuse multi-action signals, improve domain alignment, and preserve ordinal severity.

As shown in Table 3, our model achieves the best accuracy and macro-F1 on all four UPDRS–PIGD items, indicating that the improvements are consistent across tasks rather than confined to a single item or label space. The gains are also substantial: compared with the strongest baseline on the averaged metrics, our method improves macro-F1 by 0.17, accuracy by 0.18, and reduces MAE by 0.18, well beyond typical fold-to-fold variability. Notably, the improvements are particularly pronounced on STAB, the most clinically challenging subscore, suggesting that action-weighted fusion and prior-guided alignment effectively recover balance-related signals from raw data.

We chose the ResNet1D as our backbone because Baseline-D empirically outperforms the LIMU-BERT baseline (Baseline-E). Although we fine-tuned the pretrained LIMU-BERT model using our data, the backbone, which utilizes motion data from healthy individuals, proved unsuitable for our specific patient scenarios. Furthermore, the performance gain of Ours compared to Baseline-B clearly demonstrates the effectiveness of our designed task-specific modules (AWTH, Prior-CDAN, and the ordinal continuity regularizer).

6.2 Robustness Study

6.2.1 Impact of Environment. We evaluate the system in two physical interaction contexts: (i) a corridor environment—a relatively open hallway suitable for gait assessment, featuring long straight walks, unobstructed turning, and minimal interference (e.g., an inpatient ward); and (ii) a room environment—a confined space resembling a living room or clinical examination room, where furniture, medical equipment, and personnel introduce spatial and visual clutter (e.g., an outpatient clinic or home). As shown in Fig. 16(a), overall performance

remains comparable across environments, with stable mean macro-F1 scores between corridor and room settings. However, item-level variability is observed: POST shows moderate declines in the room environment, likely due to spatial constraints and increased visual clutter affecting upright control. Despite these shifts, performance across all items remains within clinically acceptable bounds, indicating that our model can generate reliable PIGD subscores in real-world home-like environments.

6.2.2 Impact of Distance. We next evaluate robustness with respect to walking distance. In our protocol, participants should perform a walk-turn task. However, depending on different testing setups, *e.g.*, different room sizes, the maximum allowed walking distance will be different in practice. In fact, in our data collection setting, patients were recruited in three different venues (two rooms and one corridor), which allows for three different walking distances: 2 m, 3 m, and 4 m. In this evaluation, we verify whether this maximum walking distance will affect PIGDAssess's performance. The slight macro-F1 dip observed with increasing distance likely stems from increased gait variability (Fig. 16(b)); however, performance remains stable (mean macro-F1 > 0.79) and within clinically acceptable ranges. This enables robust, constraint-free deployment in practical settings, with further potential for enhancement through more diverse training data.

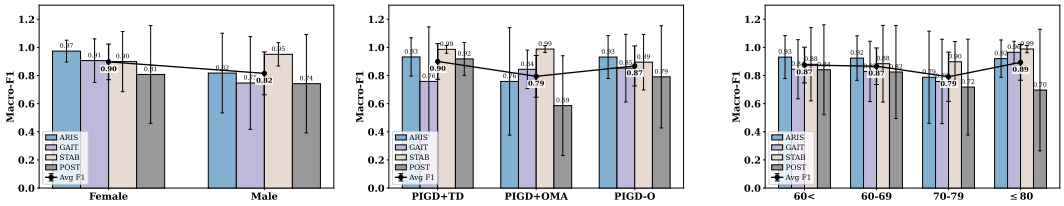
6.2.3 Impact of Missing Sensors. We conduct an IMU ablation study to quantify the contribution of individual wearable sensors by masking one IMU at a time during inference (Fig. 16(c)). With all three IMUs (lumbar and both feet), the system achieves a mean macro-F1 of 0.87. Removing the lumbar IMU leads to the largest degradation (macro-F1 drops to 0.62), highlighting the importance of trunk dynamics for axial and postural assessment. Removing either foot IMU also degrades performance, with macro-F1 decreasing to 0.77 and 0.71, respectively, indicating that foot-level kinematics contribute to gait- and stability-related items but are individually less dominant than the lumbar sensor. Overall, these results justify the three-IMU configuration and demonstrate that each sensor provides complementary clinical information.

6.2.4 Effect of Segment Duration. Beyond robustness, we further analyze the effect of segment duration on model performance. In practical home-based scenarios, walking assessments are typically limited to short distances (2–3 m). Accordingly, we adopt a 10-s walk–turn segment for scoring, which is sufficient to capture turning events for most patients with Parkinson's disease in this setting. We compare this setting with other durations (5 s, 30 s, and the full recording). As shown in Fig. 16(d), performance degrades for 5 s segments, likely because they often fail to capture complete turning events. Performance remains stable at 30 s, whereas using the full recording leads to lower scores due to the lack of segment-based pairing augmentation and increased variability. Overall, these results indicate that a 10-s window strikes a good balance between temporal completeness and modeling stability, supporting reliable real-world deployment.

6.3 Demographic Study

6.3.1 Impact of Gender. We also evaluate performance by gender, grouping participants into female and male, see Fig.17(a). The mean macro-F1 is similar across the two groups, indicating that the model assigns stable scores to both. The female group is slightly higher, which is likely due to sample size (20 females vs. 15 males) rather than systematic bias. Overall, this suggests that sex does not substantially affect model validity and that increasing data in both groups should further stabilize performance.

6.3.2 Impact of PD Subtype. We conduct a subtype-wise analysis by grouping participants into PIGD+TD, PIGD+OMA, and PIGD-O (Fig. 17(b)). Due to smaller sample sizes, PIGD+TD and PIGD+OMA exhibit greater item-level variability. Nonetheless, the average macro-F1 remains comparable across groups, indicating that the model generalizes beyond pure PIGD. Notably, POST shows a pronounced drop in PIGD+OMA, which may reflect increased axial involvement or asymmetry in this subgroup, leading to posture patterns that deviate from



(a) Gender-wise (b) Subtypes-wise (c) Age-wise

Fig. 17. Demographic study of PIGDAssess. (a) Gender-wise: both female and male patients achieve high macro-F1, with females slightly higher. (b) Subtypes-wise: performance is stable across PIGD+TD, PIGD+OMA, and pure PIGD groups. (c) Age-wise: accuracy remains high across age groups.

Table 4. Ablation on five model variants. We add modules step by step. Bold indicates the best result (higher is better for F1, lower is better for MAE), and underlined values indicate the second best.

Method	Pair	AWTH	CDAN	P-CDAN	OC	ARIS		GAIT		STAB		POST		Average	
						F1↑	MAE↓	F1↑	MAE↓	F1↑	MAE↓	F1↑	MAE↓	F1↑	MAE↓
Variant-A	–	–	–	–	–	0.58±0.23	0.43±0.26	0.52±0.17	0.46±0.10	0.57±0.23	0.67±0.42	0.53±0.22	0.70±0.31	0.55±0.21	0.57±0.27
Variant-B	✓	–	–	–	–	0.72±0.22	0.36±0.24	0.62±0.20	0.48±0.22	0.53±0.14	0.63±0.22	0.58±0.19	0.74±0.31	0.61±0.19	0.55±0.25
Variant-C	✓	✓	–	–	–	0.77±0.23	0.31±0.25	0.57±0.19	0.47±0.32	0.70±0.20	0.44±0.31	0.62±0.24	0.49±0.28	0.66±0.21	0.43±0.21
Variant-D	✓	✓	✓	–	–	0.79±0.24	0.21±0.24	0.69±0.19	0.33±0.19	0.57±0.15	0.60±0.24	0.59±0.23	0.51±0.27	0.66±0.20	0.41±0.24
Variant-E	✓	✓	–	✓	–	0.81±0.22	**0.20**±0.21	0.60±0.19	0.32±0.14	0.69±0.17	**0.40**±0.28	0.62±0.22	0.49±0.27	0.68±0.20	0.35±0.13
Ours	✓	✓	–	✓	✓	**0.82**±0.23	0.21±0.24	**0.72**±0.24	**0.25**±0.16	**0.75**±0.17	0.41±0.27	**0.69**±0.22	**0.31**±0.19	**0.75**±0.21	**0.30**±0.22

Pair = pairing-based augmentation, AWTH = adaptive window temporal harmonization, P-CDAN = prior-guided CDAN, OC = ordinal continuity regularizer.

typical Parkinsonian posture. Overall, despite subtype-specific performance differences, the model maintains stable average performance across PIGD phenotypes.

6.3.3 Impact of Age. We also analyze performance by age (Fig. 17(c)), dividing participants into four groups: <60, 60–69, 70–79, and 80+. The mean macro-F1 is broadly similar across age groups, indicating that the model can score PIGD across a wide age range without large degradation. We do observe a dip in the 70–79 group, and more variability in the 80+ group. This is likely because older participants often present more severe or complex motor impairment, which is harder to model consistently. Overall, age does not appear to introduce a major systematic bias, though very advanced age may increase task difficulty.

6.4 Ablation Study

We construct five model variants to systematically evaluate the contribution of each introduced module.

- **Variant-A.** Adopt a ResNet-based multi-action encoder and simply concatenate the three action embeddings, without any action weighting.
- **Variant-B.** Add pairing-based augmentation, which introduces multiple pairing strategies to enrich the training samples.
- **Variant-C.** Add Action-weighted Task Head.
- **Variant-D.** Add standard CDAN for domain alignment.
- **Variant-E.** Modify the standard CDAN with our prior-guided, kinematics-informed alignment method.
- **Ours.** Add the ordinal continuity regularizer, which enforces a smooth progression of latent features across the ordered PIGD severity levels.

6.4.1 Effectiveness of Pairing-based Data Augmentation. Simply concatenating single- and dual-task features introduces systematic bias, as movements under cognitive load differ from baseline movements. With the proposed

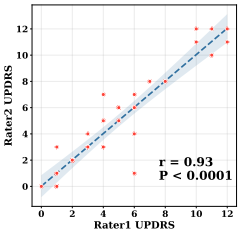


Fig. 18. UPDRS–PIGD score correlation between two clinicians.

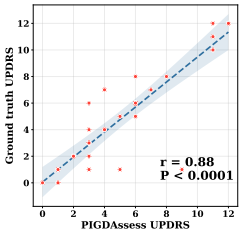


Fig. 19. Correlation between ground truth and system predictions.

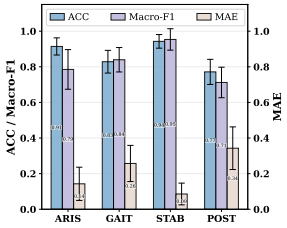


Fig. 20. Subject-level Accuracy, Macro-F1, and MAE across four items.

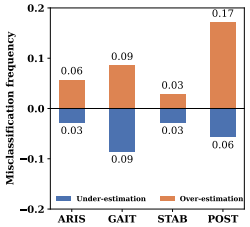


Fig. 21. Subject-level under/overestimation after majority voting.

pairing-based augmentation, Variant B raises the average macro-F1 from 0.55 to 0.61, suggesting that structured pairing both enriches sample diversity and more faithfully models the contrast between baseline and cognitively loaded conditions.

6.4.2 Effectiveness of Action-weighted Task Head. The AWTH enables per-sample adaptive fusion of action embeddings. Compared to Variant-B, Variant-C boosts macro-F1 for STAB from 0.53 to 0.70. Overall macro-F1 rises from 0.61 to 0.66. This confirms that dynamically weighting actions enhances model sensitivity to task-specific motor cues.

6.4.3 Effectiveness of Prior-guided Adaptation. Standard CDAN (Variant-D) improves some items but degrades item STAB and does not yield any increase on POST, likely due to uncontrolled alignment washing out clinical severity structure. Replacing it with biomechanics-informed Prior-CDAN (Variant-E) raises the average macro-F1 to 0.68, with dramatic gains on STAB (macro-F1: 0.57 → 0.69). Adding ordinal continuity further improves average macro-F1 to 0.75. This demonstrates that constraining alignment by kinematic similarity preserves phenotype-level differences, and enforcing ordinal smoothness sharpens decision boundaries toward clinically meaningful neighborhoods.

6.5 Computation Cost

Although training was performed on an NVIDIA RTX 4090D, our system is designed to run inference with low computational overhead. We profiled the deployed model by running 1,000 forward passes on the 4090D and reported the averaged results. The mean inference latency is 19.79 ms per sample, with a peak GPU memory reservation of only 394 MB (388.71 MB peak allocated). These measurements indicate that the model can operate in (near) real time on a single consumer-grade GPU, with a modest memory footprint.

7 CLINICAL USABILITY STUDY

To complement the quantitative results, we conducted a clinician-in-the-loop study to evaluate the clinical validity, reliability, and practical usability of the proposed system. The study consists of (i) a clinician-oriented performance analysis and (ii) semi-structured expert interviews.

7.1 Clinician-oriented Clinical Analysis

Each participant completed multiple dual-task trials. Trial-level predictions were aggregated using majority voting to yield a single subject-level prediction for each item, reflecting realistic clinical use. This resulted in a total of 35×4 subject-level predictions across all participants and UPDRS–PIGD items (see Appendix Table 5).

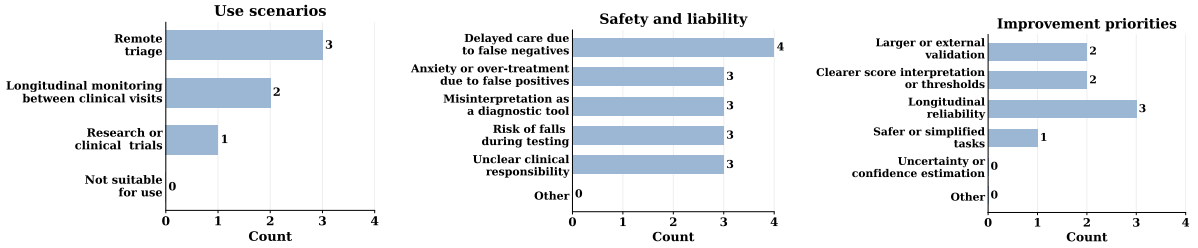


Fig. 22. Expert feedback on application scenarios (Q2), safety and liability concerns (Q6), and improvement priorities (Q8) for *PIGDAssess*. Bars show the number of experts selecting each option (n=4).

Two movement disorder specialists independently scored all participants using UPDRS–PIGD. We evaluated inter-rater reliability between the two clinicians (Fig. 18) and computed the Pearson correlation between system predictions and clinician scores (Fig. 19). The correlation between the system and clinician scores was comparable in magnitude to inter-clinician agreement, indicating that the system's level of agreement falls within the range of clinician variability and supporting its clinical reliability.

Figure 20 summarizes subject-level performance across the four UPDRS–PIGD assessment items in terms of Accuracy, Macro-F1, and MAE. The system achieves an average Accuracy of 0.86, a Macro-F1 of 0.82, and an MAE of 0.21. We further analyzed false positive and false negative rates (Fig. 21). The overall underestimation rate was 8.75% on average, while the false negative rate averaged 5.25%.

7.2 Expert Experience Study

To evaluate the real-world usefulness, safety, and deployment readiness of PIGDAssess, we conducted a semi-structured expert experience study with Parkinson's disease (PD) clinicians.

7.2.1 Study Design. We designed a structured questionnaire to collect expert feedback on the system's clinical value, reliability, error tolerance, safety risks, and future improvement directions. The questionnaire covered eight themes: (Q1) overall usefulness, (Q2) use scenarios, (Q3) result reliability, (Q4) metric adequacy, (Q5) error tolerance, (Q6) safety and liability, (Q7) fall risks, and (Q8) improvement priorities. The full questionnaire is provided in the appendix.

Four experts participated, including two movement-disorder specialists (S1, S2) and two general neurologists (P1, P2). Before completing the questionnaire, experts reviewed per-patient results from 35 PD patients (individual ratings, consensus ground truth, and system predictions), quantitative performance metrics (correlation, Accuracy, Macro-F1, and MAE), and false-positive/false-negative error distributions.

7.2.2 Perceived Usefulness and Application Scenarios (Q1-2). As for Q1, all experts agreed that PIGDAssess is useful for real-world applications, particularly for objective home-based assessment of posture stability and gait. One movement-disorder specialist noted that "*PIGDAssess enables patients to complete objective posture stability and gait assessments at home*" (S1), while another expert described it as "*lightweight and easy to use*" (S2). As shown in the Fig. 22, experts identified appropriate use scenarios, with remote triage most frequently cited (Q2). They emphasized that PIGDAssess outputs should be reviewed by healthcare professionals, as PIGDAssess is designed as an assessment support tool. One neurologist suggested predefined thresholds or longitudinal changes to prompt further evaluation, noting that "*when any sub-item score reaches 3 or higher, or when the score changes by two points between visits, further clinical evaluation should be triggered*" (P1).

7.2.3 Trust in Single-patient Results (Q3). Regarding result reliability, experts reported moderate-to-high confidence in individual patient-level predictions (mean: 4.75/7), viewing system outputs as supportive references that still require clinical interpretation. Confidence was noted to decrease in the presence of comorbidities, motor state fluctuations, or cognitive limitations, as "*the assessment may not reflect typical function*" during off periods or other conditions (S1), or when patients "*cannot properly perform the task due to cognitive impairment*" (P1).

7.2.4 Adequacy of Performance Metrics and Error Sensitivity (Q4-5). All experts agreed that the reported metrics are sufficient to support pilot clinical use as a decision-support tool (Q4). MAE was viewed as a key indicator of clinical agreement, with values below about 0.8 considered acceptable and around 0.5 ideal, reflecting the "*average deviation from the clinical gold standard*" (S1). Errors within the 0.5–1 point range were noted to "*not affect the identification of moderate or more severe impairment*" (P1). For PIGDAssess, the average MAE is 0.21, indicating strong agreement with clinical ratings. For Q5, experts unanimously emphasized that false negatives are less acceptable than false positives, noting that "*missing high-risk patients is more concerning*" (P1). Accordingly, experts emphasized the priority of sensitivity in high-risk detection, considering up to 15–20 false positives but only about 5 false negatives per 100 patients to be acceptable. For PIGDAssess, the average false-negative rate is 5.25%, which aligns closely with this clinical threshold. This strong alignment between our quantitative performance and expert expectations supports the system's viability as an objective screening and longitudinal monitoring tool, facilitating home-based triage rather than serving as a diagnostic replacement.

7.2.5 Safety, Liability, and Future Directions (Q6-8). As shown in the Fig. 22, experts expressed primary concerns about safety and responsibility in home-based self-assessment, with delayed care due to false negatives cited most frequently, followed by misinterpretation of results and fall risk (Q6). These concerns highlighted the need for explicit safeguards and clear non-diagnostic framing. As one expert noted, "*the system should clearly state that it is not a diagnostic tool, with final interpretation remaining with clinicians*" (P1), while another emphasized informed consent, stating that patients should "*acknowledge their understanding of the system's purpose and limitations before use*" (S1). All experts agreed that although the self-testing protocol entails some fall risk—mainly during sit-to-stand, walking, and turning—this risk is primarily attributable to patients' underlying motor impairments, and remains manageable under appropriate conditions, such as caregiver supervision, non-slip flooring, and access to stable support (Q7). Regarding future directions, experts prioritized longitudinal reliability, as reflected in the Fig. 22 (Q8). One expert further suggested cognitive screening prior to home use, proposing that "*a brief cognitive screening could be used before allowing home self-assessment*" (P1), to ensure safe and reliable deployment.

8 DISCUSSION

We discuss key limitations of this work, outline corresponding directions for improvement, and highlight potential extended applications.

Limited and Clinic-centric Data. Our study is based on a single-center cohort of 35 participants, with movement labels derived from clinician ratings obtained under controlled assessment protocols. In the context of IMU-based PIGD assessment with fine-grained, item-level annotations, a cohort of this size constitutes a relatively rare, high-quality aligned dataset. Although all four PIGD subitems are represented, the limited sample size and geographical diversity may constrain the coverage of diverse movement patterns and environmental contexts. In addition, clinic-based assessments may not fully capture symptom progression in daily life. Collecting high-quality at-home motor data from Parkinson's patients remains challenging due to patient burden and the high cost of comprehensive clinical annotation. Despite these challenges, our dataset represents one of the first to pair item-level PIGD labels with wearable IMU signals in real-world settings. As with most within-cohort cross-validation studies, performance estimates may be moderately optimistic compared to real-world deployment

on previously unseen populations. Furthermore, for participants with more severe symptoms (H-Y Stage 3 or 4), caregivers provided physical assistance during walking tasks to ensure safety, which may influence the purity of the recorded gait signals. To address these limitations and improve generalizability, we plan to conduct multi-site studies with larger, more heterogeneous populations and release an IMU dataset with full PIGD item-level labels to support community benchmarking.

Practical Considerations for Real-world Deployment. Our protocol employs dual-task paradigms, such as performing arithmetic during walking or standing, to elicit subtle balance impairments. While supported by motor control literature, performance variability may reflect not only motor deficits but also differences in cognitive load and coping strategies, making fixed-difficulty tasks (e.g., serial subtraction) suboptimal across participants. Future versions could incorporate adaptive cognitive tasks to maintain engagement and improve sensitivity to changes in postural control and gait stability. In addition, the current system requires three IMUs placed on the lumbar region and both feet. Although low-cost and minimally invasive, this configuration poses challenges for long-term home deployment, as sensor misplacement or non-compliance can degrade performance. To improve robustness, we will explore signal imputation and self-supervised pretraining on large-scale unlabeled in-home data, and investigate contactless modalities (e.g., RGB-D sensing) to further reduce user burden and extend applicability.

Clinician Perspective. Clinicians viewed PIGDAssess as most valuable for extending functional assessment beyond the clinic rather than replacing in-clinic UPDRS evaluation. Its subject-level performance, agreement with clinician consensus within inter-rater variability, and false negative rates—while not yet at experts' ideal tolerance levels—were considered sufficient to support screening and longitudinal monitoring under clinical oversight. Repeated home-based assessment over multiple days is an important direction for future work to evaluate test– longitudinal reliability and distinguish true clinical changes from day-to-day variability. Future studies will examine whether the assessment shows high repeatability under stable conditions or systematic variability aligned with symptom fluctuation, which is critical for establishing confidence in objective, home-based longitudinal monitoring.

Extended Applications. This study demonstrates that full UPDRS–PIGD scoring, including STAB, can be performed safely at home using a short and repeatable dual-task protocol. Replacing high-risk maneuvers with low-burden self-tests and reporting item-level scores enables finer longitudinal tracking of axial and gait symptoms between visits. The same framework can extend to other balance- and fall-related disorders and integrate with tele-neurology dashboards, with the potential for deployment on consumer wearables through edge-oriented optimization. More broadly, our results support the idea that cognitive interference during movement can reveal neurodegeneration-related motor deficits, suggesting that common daily dual-task activities may serve as practical probes for passive, unobtrusive home-based monitoring.

9 CONCLUSION

This paper presents **PIGDAssess**, the first wearable system that enables *fully self-administered, at-home* estimation of *all four* UPDRS–PIGD subitems—including postural stability—*without* a clinician-performed pull test. Using only three off-the-shelf IMUs and a brief single/dual-task protocol, our method pairs cognitive-load perturbations with a multi-action multi-task architecture and prior-guided adaptation to deliver reliable item-level scores. In a 35-participant evaluation, PIGDAssess achieves strong accuracy across items and clinician-level performance on reactive balance, pointing toward frequent, low-burden fall-risk monitoring in everyday settings.

Acknowledgments

This research is supported in part by RGC under Contract CERG 16206122, 16204523, 16205824, AoE/E-601/22-R, HB016, RG702, and Contract R8015. Dr. Qian Zhang and Dr. Guihua Li are the co-corresponding authors.

References

[1] Jessica E Bath and Doris D Wang. 2024. Unraveling the threads of stability: A review of the neurophysiology of postural control in Parkinson's disease. *Neurotherapeutics* 21, 3 (2024), e00354.

[2] Bastiaan R Bloem, Johan Marinus, Quincy Almeida, Lee Dibble, Alice Nieuwboer, Bart Post, Evzen Ruzicka, Christopher Goetz, Glenn Stebbins, Pablo Martinez-Martin, et al. 2016. M easurement instruments to assess posture, gait, and balance in P arkinson's disease: Critique and recommendations. *Movement Disorders* 31, 9 (2016), 1342–1355.

[3] Mark D Bogost, Pablo I Burgos, C Elaine Little, Marjorie H Woollacott, and Brian H Dalton. 2016. Electrocortical sources related to whole-body surface translations during a single-and dual-task paradigm. *Frontiers in human neuroscience* 10 (2016), 524.

[4] Janet H Carr and Roberta B Shepherd. 2010. *Neurological rehabilitation: optimizing motor performance.* Elsevier Health Sciences.

[5] Ltd. Chenyi Electronic Technology Co. [n. d.]. IMU948 Introduction. [Online]. https://www.yuque.com/cxqwork/lkw3sg/yqa3e0?.

[6] Augustine Joshua Devasahayam, Kyle Farwell, Bohyung Lim, Abigail Morton, Natalie Fleming, David Jagroop, Raabeae Aryan, Tyler Mitchell Saumur, and Avril Mansfield. 2023. The effect of reactive balance training on falls in daily life: an updated systematic review and meta-analysis. *Physical therapy* 103, 1 (2023), pzac154.

[7] Hany Mohamed Eldeeb and Heba Samir Abdelraheem. 2021. Functional gait assessment in early and advanced Parkinson's disease. *The Egyptian Journal of Neurology, Psychiatry and Neurosurgery* 57 (2021), 1–9.

[8] Parkinson's Europe. [n. d.]. Parkinson's Statistics. [Online]. https://parkinsonseurope.org/facts-and-figures/statistics/.

[9] Alfonso Fasano, Colleen G Canning, Jeffrey M Hausdorff, Sue Lord, and Lynn Rochester. 2017. Falls in Parkinson's disease: a complex and evolving picture. *Movement disorders* 32, 11 (2017), 1524–1536.

[10] Claudia Ferraris, Roberto Nerino, Antonio Chimienti, Giuseppe Pettiti, Nicola Cau, Veronica Cimolin, Corrado Azzaro, Lorenzo Priano, and Alessandro Mauro. 2019. Feasibility of home-based automated assessment of postural instability and lower limb impairments in Parkinson's disease. *Sensors* 19, 5 (2019), 1129.

[11] Claudia Ferraris, Valerio Votta, Roberto Nerino, Antonio Chimienti, Lorenzo Priano, and Alessandro Mauro. 2024. At-home assessment of postural stability in parkinson's disease: A vision-based approach. *Journal of Ambient Intelligence and Humanized Computing* 15, 5 (2024), 2765–2778.

[12] Isobel T French and Kalai A Muthusamy. 2018. A review of the pedunculopontine nucleus in Parkinson's disease. *Frontiers in aging neuroscience* 10 (2018), 99.

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

[14] Tomasz Gutowski, Olga Stodulska, Aleksandra Ćwiklińska, Katarzyna Gutowska, Kamila Kopeć, Marta Betka, Ryszard Antkiewicz, Dariusz Koziorowski, and Stanisław Szlufik. 2025. Machine Learning-Based Assessment of Parkinson's Disease Symptoms Using Wearable and Smartphone Sensors. *Sensors* 25, 16 (2025), 4924.

[15] Cathy C Harro, Alicia Marquis, Natasha Piper, and Chris Burdis. 2016. Reliability and validity of force platform measures of balance impairment in individuals with Parkinson disease. *Physical therapy* 96, 12 (2016), 1955–1964.

[16] Rong He, Zijing You, Yongqiang Zhou, Guilan Chen, Yanan Diao, Xiantai Jiang, Yunkun Ning, Guoru Zhao, and Ying Liu. 2024. A novel multi-level 3D pose estimation framework for gait detection of Parkinson's disease using monocular video. *Frontiers in Bioengineering and Biotechnology* 12 (2024), 1520831.

[17] Shenghong He, Alceste Deli, Petra Fischer, Christoph Wiest, Yongzhi Huang, Sean Martin, Saed Khawaldeh, Tipu Z Aziz, Alexander L Green, Peter Brown, et al. 2021. Gait-phase modulates alpha and beta oscillations in the pedunculopontine nucleus. *Journal of Neuroscience* 41, 40 (2021), 8390–8402.

[18] Talia Herman, Aner Weiss, Marina Brozgol, Nir Giladi, and Jeffrey M Hausdorff. 2014. Gait and balance in Parkinson's disease subtypes: objective measures and classification considerations. *Journal of neurology* 261, 12 (2014), 2401–2410.

[19] Qingyong Hu, Yuxuan Zhou, Jinjian Wang, Zirui Huang, Guihua Li, Qianhui Xu, and Qian Zhang. 2025. mmTremor: Practical Tremor Monitoring for Parkinson's Disease and Essential Tremor in Daily Life. In *Proceedings of the 31st Annual International Conference on Mobile Computing and Networking*. 498–512.

[20] Joseph Jankovic, M McDermott, J Carter, S Gauthier, C Goetz, L Golbe, S Huber, W Koller, C Olanow, I Shoulson, et al. 1990. Variable expression of Parkinson's disease: A base-line analysis of the DAT ATOP cohort. *Neurology* 40, 10 (1990), 1529–1529.

[21] Ashwani Jha, Vladimir Litvak, Samu Taulu, Wesley Thevathasan, Jonathan A Hyam, Tom Foltynie, Patricia Limousin, Marko Bogdanovic, Ludvic Zrinzo, Alexander L Green, et al. 2017. Functional connectivity of the pedunculopontine nucleus and surrounding region in Parkinson's disease. *Cerebral Cortex* 27, 1 (2017), 54–67.

[22] Y. Kim, K. Joa, Haneul Jeong, and Sangmin Lee. 2021. Wearable IMU-Based Human Activity Recognition Algorithm for Clinical Balance Assessment Using 1D-CNN and GRU Ensemble Model. *Sensors (Basel, Switzerland)* 21 (2021). doi:10.3390/s21227628

[23] Cameron Kirk, Emma Packer, Ashley Polhemus, Mhairi K MacLean, Harry Bailey, Felix Kluge, Heiko Gaßner, Lynn Rochester, Silvia Del Din, and Alison J Yarnall. 2025. A systematic review of real-world gait-related digital mobility outcomes in Parkinson's disease. *npj Digital Medicine* 8, 1 (2025), 585.

[24] C Elaine Little and Marjorie Woollacott. 2015. EEG measures reveal dual-task interference in postural performance in young adults. *Experimental brain research* 233, 1 (2015), 27–37.

[25] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. 2018. Conditional adversarial domain adaptation. *Advances in neural information processing systems* 31 (2018).

[26] Mandy Lu, Qingyu Zhao, Kathleen L Poston, Edith V Sullivan, Adolf Pfefferbaum, Marian Shahid, Maya Katz, Leila Montaser-Kouhsari, Kevin Schulman, Arnold Milstein, et al. 2021. Quantifying Parkinson's disease motor severity under uncertainty using MDS-UPDRS videos. *Medical Image Analysis* 73 (2021), 102179.

[27] Lingyan Ma, Shinuan Lin, Jianing Jin, Zhan Wang, Xuemei Wang, Zhonglue Chen, Yun Ling, Fei Zhang, Kang Ren, and Tao Feng. 2025. Objective assessment of gait and posture symptoms in Parkinson's disease using wearable sensors and machine learning. *Frontiers in Aging Neuroscience* 17 (2025), 1618764.

[28] George Mochizuki, Shaun G Boe, Amanda Marlin, and William E McIlroy. 2017. Performance of a concurrent cognitive task modifies pre-and post-perturbation-evoked cortical activity. *Neuroscience* 348 (2017), 143–152.

[29] Caroline Moreau, Tiphaine Rouaud, David Grabli, Isabelle Benatru, Philippe Remy, Ana-Raquel Marques, Sophie Drapier, Louise-Laure Mariani, Emmanuel Roze, David Devos, et al. 2023. Overview on wearable sensors for the management of Parkinson's disease. *npj Parkinson's Disease* 9, 1 (2023), 153.

[30] Martijn LTM Müller, Roger L Albin, Vikas Kotagal, Robert A Koeppe, Peter JH Scott, Kirk A Frey, and Nicolaas I Bohnen. 2013. Thalamic cholinergic innervation and postural sensory integration function in Parkinson's disease. *Brain* 136, 11 (2013), 3282–3289.

[31] Rudri Purohit, Shuaijie Wang, Shamali Dusane, and Tanvi Bhatt. 2023. Age-related differences in reactive balance control and fall-risk in people with chronic stroke. *Gait & posture* 102 (2023), 186–192.

[32] Andrea L Rosso, Massimo Cenciarini, Patrick J Sparto, Patrick J Loughlin, Joseph M Furman, and Theodore J Huppert. 2017. Neuroimaging of an attention demanding dual-task during dynamic postural control. *Gait & posture* 57 (2017), 193–198.

[33] Delaram Safarpour, Marian L Dale, Vrutangkumar V Shah, Lauren Talman, Patricia Carlson-Kuhta, Fay B Horak, and Martina Mancini. 2022. Surrogates for rigidity and PIGD MDS-UPDRS subscores using wearable sensors. *Gait & posture* 91 (2022), 186–191.

[34] Kenichiro Sato, Yu Nagashima, Tatsuo Mano, Atsushi Iwata, and Tatsushi Toda. 2019. Quantifying normal and parkinsonian gait features from home movies: Practical application of a deep learning–based 2D pose estimator. *PloS one* 14, 11 (2019), e0223549.

[35] Teodoro Solis-Escalante, Joris van der Cruijsen, Digna de Kam, Joost van Kordelaar, Vivian Weerdesteyn, and Alfred C Schouten. 2019. Cortical dynamics during preparation and execution of reactive balance responses with distinct postural demands. *NeuroImage* 188 (2019), 557–571.

[36] Glenn T Stebbins, Christopher G Goetz, David J Burn, Joseph Jankovic, Tien K Khoo, and Barbara C Tilley. 2013. How to identify tremor dominant and postural instability/gait difficulty groups with the movement disorder society unified Parkinson's disease rating scale: comparison with the unified Parkinson's disease rating scale. *Movement Disorders* 28, 5 (2013), 668–670.

[37] Samuel Stuart, Rodrigo Vitorio, Rosie Morris, Douglas N Martini, Peter C Fino, and Martina Mancini. 2018. Cortical activity during walking and balance tasks in older adults and in people with Parkinson's disease: A structured review. *Maturitas* 113 (2018), 53–72.

[38] William Wai-Nam Tsang, Vito Wai-Lok Chan, Henry Hei Wong, Tony Wai-Cheong Yip, and Xi Lu. 2016. The effect of performing a dual-task on postural control and selective attention of older adults when stepping backward. *Journal of physical therapy science* 28, 10 (2016), 2806–2811.

[39] Parkinson's UK. [n. d.]. Information And Support. [Online]. https://www.parkinsons.org.uk/information-and-support/your-magazine/experts/adjusting-parkinsons-medication/;https://www.parkinsons.org.uk/information-and-support/nice-guidelines-parkinsons/.

[40] Gaëtan Vignoud, Clément Desjardins, Quentin Salardaine, Marie Mongin, Béatrice Garcin, Laurent Venance, and Bertrand Degos. 2022. Video-based automated assessment of movement parameters consistent with MDS-UPDRS III in Parkinson's disease. *Journal of Parkinson's Disease* 12, 7 (2022), 2211–2222.

[41] Claudia Voelcker-Rehage. 2017. Neural correlates of motor-cognitive dual-tasking in young and old adults. (2017).

[42] R Von Coelln, AL Gruber-Baldini, SG Reich, MJ Armstrong, JM Savitt, and LM Shulman. 2021. The inconsistency and instability of Parkinson's disease motor subtypes. *Parkinsonism & related disorders* 88 (2021), 13–18.

[43] Bettina Wollesen, M Wanstrath, KS Van Schooten, and K Delbaere. 2019. A taxonomy of cognitive tasks to evaluate cognitive-motor interference on spatiotemoporal gait parameters in older people: a systematic review and meta-analysis. *European review of aging and physical activity* 16, 1 (2019), 12.

[44] Huatao Xu, Pengfei Zhou, Rui Tan, Mo Li, and Guobin Shen. 2021. Limu-bert: Unleashing the potential of unlabeled data for imu sensing applications. In *Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems*. 220–233.

[45] Baichen Yang, Qingyong Hu, Wentao Xie, Xinchen Wang, Wei Luo, and Qian Zhang. 2023. PDAssess: A Privacy-preserving Free-speech based Parkinson's Disease Daily Assessment System. In *Proceedings of the 21st ACM Conference on Embedded Networked Sensor Systems*. 251–264.

[46] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*. 1480–1489.

[47] Wenhao Zhang, Haipeng Dai, Dongyu Xia, Yang Pan, Zeshui Li, Wei Wang, Zhen Li, Lei Wang, and Guihai Chen. 2024. mP-Gait: Fine-grained Parkinson's Disease Gait Impairment Assessment with Robust Feature Analysis. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8, 3 (2024), 1–31.

[48] Weishan Zhang, Yun Ling, Zhonglue Chen, Kang Ren, Shengdi Chen, Pei Huang, and Yuyan Tan. 2024. Wearable sensor-based quantitative gait analysis in Parkinson's disease patients with different motor subtypes. *NPJ Digital Medicine* 7, 1 (2024), 169.

[49] Yizhen ZHANG. 2026. PIGDAssess: Wearable Dual-Task Sensing for Self-Administered PIGD Assessment in Parkinson's Disease. https://zenodo.org/records/19450832 [Online].

[50] Yanci Zhang, Zhiwei Zeng, Maryam S Mirian, Kevin Yen, Kye Won Park, Michelle Doo, Jun Ji, Zhiqi Shen, and Martin J McKeown. 2024. Investigating the efficacy and importance of mobile-based assessments for Parkinson's disease: uncovering the potential of novel digital tests. *Scientific Reports* 14, 1 (2024), 5307.

APPENDIX

A PER-SUBJECT PIGD SUBITEM SCORES

This appendix reports subject-level scores for each PIGD subitem. For each participant and each subitem, we list the scores assigned by two clinicians (R1 and R2), the consensus rating reached after discussion and used as ground truth (GT), and the prediction produced by PIGDAssess (PA). All scores follow the MDS-UPDRS ordinal scale (0–4).

Table 5. Per-subject scores for all PIGD-related UPDRS subitems. For each participant and each subitem, we report two independent clinician ratings (R1, R2), the consensus ground truth label (GT) obtained after adjudication, and the prediction from PIGDAssess (PA) produced under the LOSO protocol with majority voting aggregation.

Subject ID	ARIS (3.9)				GAIT (3.10)				STAB (3.12)				POST (3.13)				Total			
	R1	R2	GT	PA	R1	R2	GT	PA	R1	R2	GT	PA	R1	R2	GT	PA	R1	R2	GT	PA
S01	0	0	0	0	1	2	2	1	3	2	3	1	2	2	2	2	6	6	7	4
S02	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	1	0	0	1
S03	0	0	0	0	1	1	1	1	0	0	0	0	0	0	0	0	1	1	1	1
S04	0	0	0	0	1	1	1	1	0	1	0	0	2	2	2	2	3	4	3	3
S05	1	0	1	1	2	2	2	2	1	1	1	1	2	1	2	2	6	4	6	6
S06	0	0	0	0	1	1	1	2	0	1	0	1	0	1	0	2	1	3	1	5
S07	1	1	1	1	2	2	2	2	1	1	1	1	2	2	2	2	6	6	6	6
S08	0	0	0	0	1	2	2	2	1	1	1	1	2	2	2	2	4	5	5	5
S09	0	0	0	0	0	0	0	0	0	0	0	0	2	2	2	2	2	2	2	2
S10	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	1	0	0	1
S11	0	0	0	0	2	2	2	2	4	3	3	3	1	2	2	2	7	7	7	7
S12	1	1	1	1	2	2	2	2	1	1	1	1	1	2	2	2	5	6	6	6
S13	0	0	0	0	2	2	2	2	1	1	1	1	3	3	3	3	6	6	6	6
S14	1	1	1	1	1	2	2	2	1	1	1	1	2	2	2	2	5	6	6	6
S15	0	1	0	0	1	1	1	1	0	0	0	0	2	1	1	2	3	3	2	3
S16	2	1	2	2	3	3	3	3	3	3	3	3	3	3	3	3	11	10	11	11
S17	0	0	0	0	0	0	0	0	2	2	2	2	1	1	1	1	3	3	3	3
S18	0	0	0	0	1	0	1	1	1	1	1	1	2	2	2	2	4	3	4	4
S19	0	0	0	0	1	1	1	1	1	1	1	1	0	0	0	0	2	2	2	2
S20	4	3	3	3	4	4	3	3	4	4	3	3	3	2	3	3	15	13	12	12
S21	1	1	1	1	1	1	1	1	0	0	0	0	1	1	1	1	3	3	3	3
S22	1	1	1	1	2	2	2	0	2	2	2	2	3	3	3	3	8	8	8	6
S23	1	1	1	1	2	2	2	2	1	2	2	2	3	3	3	3	7	8	8	8
S24	0	0	0	0	1	1	1	1	1	1	1	1	2	1	2	2	4	3	4	4
S25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
S26	2	3	3	2	2	3	3	3	3	3	3	3	3	3	3	3	10	12	12	11
S27	3	4	3	3	4	3	3	3	3	3	3	3	2	3	3	3	12	13	12	12
S28	2	3	2	3	2	2	2	2	4	3	3	3	3	3	3	3	11	11	10	11
S29	0	1	0	0	1	2	2	1	1	1	1	1	2	3	3	1	4	7	6	3
S30	0	0	0	0	1	1	1	1	0	0	0	0	0	0	0	0	1	1	1	1
S31	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	3	5	5	5	6
S32	1	0	0	0	0	1	0	2	1	1	1	1	0	0	0	0	2	2	1	3
S33	3	3	3	3	3	3	3	3	3	4	3	3	3	3	3	3	12	13	12	12
S34	3	0	0	3	1	1	1	3	1	0	0	0	1	0	0	3	6	1	1	9
S35	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	0	4	3	4	3

B EXPERT INTERVIEW QUESTIONNAIRE

This questionnaire was designed to collect expert feedback on the usability, safety, and deployment readiness of *PIGDAssess* at its current level of performance.

B.1 Pre-interview Materials

Prior to completing the questionnaire, experts reviewed: (i) per-patient results from 35 Parkinson's disease (PD) patients (R1, R2, GT, PA); (ii) summary performance metrics (correlation, Accuracy, Macro-F1, MAE); and (iii) distributions of false-positive and false-negative errors. Experts were explicitly informed that *PIGDAssess* is an *assessment support tool* rather than a diagnostic system.

B.2 Expert Background

Q0. Professional role (multiple selections): Movement-disorder neurologist; General neurologist; Physical therapist / Occupational therapist (PT/OT); Clinical researcher; Other.

Q0.1 Primary PD assessment context (multiple selections): In-person outpatient visits; Remote follow-up or home assessment; Rehabilitation; Clinical research; Other.

B.3 Questionnaire

Q1. Overall usefulness. Do you consider the current version of *PIGDAssess* useful in real-world clinical or research settings?

Useful in real-world scenarios; Shows potential but is not yet ready for use; Not currently usable

Follow-up: Please briefly explain the primary reason for your selection.

Q2. Use scenarios. If the system were to be used at the current stage, which application scenarios would you consider most suitable for PIGDAssess? (multiple selections):

Remote triage; Longitudinal monitoring between clinical visits; Research or clinical trials; Not suitable for use.

Follow-up: (a) Who should be the primary reviewer of the results? Physician; PT/OT; Nurse; Patient; Other.

(b) Should further clinical evaluation be triggered by threshold exceedance or significant score changes? Yes (please specify criteria) / No.

Q3. Result reliability. After reviewing, for each patient, the predicted scores for ARIS / GAIT / STAB / POST alongside the corresponding ground truth ratings, how would you rate your overall level of trust in the system's results for an individual patient?

Please rate on a 1–7 scale:

- 1–2: Not reliable enough to support any clinical decision
- 3–5: Acceptable as a reference, but requires manual review
- 6–7: Reliable enough to directly support triage or follow-up decisions

Follow-up: Under what conditions would you distrust the system's results?

Q4. Metric adequacy. Given $r \approx 0.88$, along with Accuracy, Macro-F1, and MAE, how would you assess the current performance?

Sufficient for clinical pilot use/ Suitable for research use only/ Insufficient for practical use

Follow-up: (a) What range of MAE would you consider clinically acceptable? (b) Which type of error is more concerning: false negatives or false positives?

Error definitions: False positive (FP): $PA > GT$; False negative (FN): $PA < GT$; Risky FN: underestimation by ≥ 1 point in a high-risk range.

Q5. Error tolerance. Which type of error do you consider more serious? False negatives / False positives / Context-dependent.

Acceptable number of errors per 100 patients: False positives: _____ False negatives: _____

Q6. Safety and liability. If patients complete the assessment independently at home and view the results themselves, which risks would you be most concerned about? (multiple selections):

Delayed care due to false negatives; Anxiety or over-treatment due to false positives; Misinterpretation as a diagnostic tool; Risk of falls during testing; Unclear clinical responsibility; Other.

Follow-up: What statements, warnings, or safeguards should the system include?

Q7. Fall risks. Do you consider the current home-based self-assessment protocol to pose a risk of falls?

No obvious risk; Some risk, manageable with appropriate constraints; High risk, not suitable for home use

Follow-up: (a) Which assessment actions pose the highest fall risk? (b) What safety conditions are required (e.g., supervision, environmental constraints)?

Q8. Improvement priorities. What would be the highest-priority improvements needed to make you more willing to use *PIGDAssess* in your routine practice? (multiple selections):

Larger or external validation; Clearer score interpretation or thresholds; Longitudinal reliability; Safer or simplified tasks; Uncertainty or confidence estimation; Other.

Optional comments:

C IMPLEMENTATION DETAILS

C.1 Network Architecture

We implemented PIGDAssess in PyTorch 2.8.0 and trained it on a single NVIDIA GeForce RTX 4090 D GPU.

Time-domain ResNet1D encoders. For each action, we use a temporal ResNet1D with four stages (channels 64, 128, 256, 512) to encode 50 Hz IMU windows into a 128-dimensional feature. Single-task and dual-task windows share the same encoder, and their two 128-dimensional features are concatenated into a 256-dimensional action embedding $\mathbf{e}^a \in \mathbb{R}^{256}$. The three action embeddings are stacked into $\mathbf{E} \in \mathbb{R}^{B\times 3\times 256}$.

Action-weighted Task Head (AWTH). For each task head, a lightweight attention scorer produces per-action weights $\mathbf{m}_t \in \mathbb{R}^{B\times 3}$. A weighted sum of the three action embeddings yields a fused representation $\tilde{\mathbf{f}}_t \in \mathbb{R}^{B\times 256}$, which is passed to a task-specific classifier to produce logits $\hat{\mathbf{z}}_t \in \mathbb{R}^{B\times C_t}$. All PIGD subitems use $C_t = 4$ ordinal severity levels $\{0, 1, 2, 3\}$.

Prior-guided Conditional Domain Alignment (Prior-CDAN). For each task, the fused embedding $\tilde{\mathbf{f}}_t$ is combined with its class posterior $\mathrm{softmax}(\hat{\mathbf{z}}_t)$ to form a joint feature of dimension approximately $C_t \times 256$. This feature is passed through a Gradient Reversal Layer (GRL) with schedule

$$\lambda(p) = \frac{2}{1 + \exp(-10p)} - 1,$$

and then into a domain discriminator trained to distinguish training and held-out validation data ($n_{\text{domains}} = 2$). The discriminator is optimized using cross-entropy loss; via the GRL, the encoder is encouraged to learn domain-invariant task features.

Ordinal Continuity Regularizer. For each task t, we maintain a FIFO memory of size $K = 4096$ storing past fused features and their ordinal labels $y \in \{0, 1, 2, 3\}$. For each batch sample, a label-aware soft target is constructed with temperature $\tau = 0.25$.

C.2 Training Details

The model is trained end-to-end for 50 epochs with batch size 128 using AdamW, with learning rate 1×10^{-4} and weight decay 1×10^{-4}. In the final training objective (Eq. 11), we set $\gamma = 2.0$, $\lambda_{\text{adv}} = 0.3$, and $\lambda_{\text{con}} = 0.2$.