TulipTender: A Mobile 3D Parkinson's Disease Assessment System Using a Single Camera

Meng Xue[1,2,†], Yizhen Zhang[2,†], HY Wong[3,‡], Yanni Yang[1], GHF Chan[4], YF Cheung[4], Nancy Y Ip[3,‡,§,¶], and Qian Zhang[2,*]

[1]Shandong University, Jinan, China [2]HKUST, Hong Kong SAR, China
[3]Division of Life Science, HKUST [4]Department of Medicine, Queen Elizabeth Hospital, Hong Kong SAR, China
[1]{xuemeng,yanniyang}@sdu.edu.cn {[2]yzhangtf@connect, [3]annawong, [3]boip, [2]qianzh@cse}.ust.hk [4]{chf862,cyfz02}@ha.org.hk

ABSTRACT

Parkinson's disease (PD) assessment currently relies on subjective, episodic expert observations that often miss critical symptom fluctuations. Existing mobile methods are either impractical for home use or fail to capture the full spectrum of motor deficits. In this paper, we present *TulipTender*, a system that runs entirely on a consumer laptop using only its built-in camera, without any cloud offloading. *TulipTender* introduces a multi-scale 3D human modeling framework that jointly recovers full-body motion and fine hand articulations in metric space from a single video. This enables both millimeter-scale finger-tapping and meter-scale gait analysis. Furthermore, it suppresses visual artifacts, such as specular reflections and floor penetration, in uncontrolled home environments via a physics-aware purification framework. The system then extracts interpretable kinematic variables mirroring Movement Disorder Society Unified Parkinson's Disease Rating Scale (MDS-UPDRS) descriptors. These variables are processed by a lightweight hierarchical cross-modal transformer (HCT) for accurate, explainable score prediction. Evaluated on an RTX 4090 with 70 participants across diverse settings, including 43 inpatients, 20 outpatients, 2 nursing-home residents, and 5 home-based participants, *TulipTender* achieves F1-scores of 82% and 84% for finger tapping and gait, respectively. The full pipeline also runs locally on a MacBook Pro, supporting the feasibility of privacy-preserving assessment on consumer hardware.

CCS CONCEPTS

• **Human-centered computing → Ubiquitous and mobile computing systems and tools**.

KEYWORDS

PD assessment, mobile computing, health monitoring

ACM Reference Format:
Meng Xue, Yizhen Zhang, HY Wong, Yanni Yang, GHF Chan, YF Cheung, Nancy Y Ip, and Qian Zhang. 2026. *TulipTender*: A Mobile 3D Parkinson's Disease Assessment System Using a Single Camera. In *The 32nd Annual International Conference on Mobile Computing and Networking (MobiCom '26), October 26–30, 2026, Austin, TX, USA*. ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3795866.3844160

1 INTRODUCTION

Parkinson's disease (PD) affects over 8.5 million people worldwide [1]. The clinical gold standard, the MDS-UPDRS [9], relies on episodic expert observations in clinics, missing critical symptom fluctuations (e.g., medication ON/OFF states) essential for personalized therapy [3, 5]. Mobile and ubiquitous sensing technologies promise to bridge this gap by enabling frequent, objective assessment of structured motor tasks outside specialist clinics, including at home. However, existing approaches face fundamental limitations that prevent their widespread adoption for comprehensive PD assessment.

Wearable inertial measurement units (IMUs) are a popular mobile sensing modality, yet single-device solutions (e.g., on the wrist) cannot capture full-body kinematics [29], while multi-sensor setups burden patients and restrict natural movement [28, 33, 57]. Moreover, wearables inherently mix voluntary movements with pathological symptoms, complicating interpretation [31]. Wireless sensing techniques

†Meng Xue and Yizhen Zhang contributed equally to this work.
‡Also with Hong Kong Center for Neurodegenerative Diseases.
§Also with HKUST Shenzhen Research Institute.
¶Also with SIAT–HKUST Joint Laboratory for Brain Science.
*Corresponding author.

Please use nonacm option or ACM Engage class to enable CC licenses
This work is licensed under a Creative Commons Attribution 4.0 International License.
MobiCom '26, October 26–30, 2026, Austin, TX, USA
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2505-0/2026/10
https://doi.org/10.1145/3795866.3844160

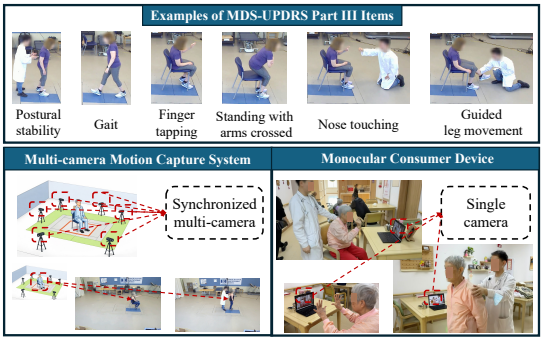


Figure 1: Comparison of motion capture setups: multi-camera system vs. *TulipTender*, targeting MDS-UPDRS.

such as WiFi and radar are contactless and have been explored for PD monitoring [20, 25, 56], but they suffer from multipath interference in cluttered home environments and lack the joint-level resolution needed for detailed motor assessment. Optical methods offer higher resolution: multi-camera systems achieve accurate 3D kinematics but are impractical for home deployment [23]; single-camera solutions are ubiquitous [22, 27, 30, 37, 39] but cannot recover metric multi-scale 3D motion, limiting their ability to quantify both fine hand tremors and whole-body gait from the same video.

To overcome these gaps, we propose a mobile computing solution that leverages the widely accessible commodity device: a laptop with a built-in camera. Laptops are particularly well-suited for this task: modern laptops integrate powerful AI accelerators (e.g., Apple's MPS [2]) that enable real-time on-device inference of deep neural networks, their built-in cameras offer a stable, fixed viewpoint ideal for structured clinical assessments, and their large displays facilitate clear patient instructions—all while preserving privacy by eliminating cloud offloading. We ask: *Can a single laptop camera capture both fine hand tremors and whole-body gait from the same video for comprehensive PD assessment, running entirely on that laptop without cloud offloading*? This is challenging because MDS-UPDRS tasks span vastly different scales, millimeter-level finger tapping vs. meter-level gait, which a fixed camera cannot simultaneously capture. This scale disparity forces prior works into task-specific silos; no existing single-camera approach assesses the full spectrum from one recording. We overcome this by reconstructing the physical scene in 3D on the laptop, obtaining metric-scale camera motion and scene geometry without cloud offloading. In this world space, every body part has a real-world position independent of its 2D projection, enabling quantification of both fine hand and whole-body motions from the same video. Deploying such a system in uncontrolled home environments introduces three challenges.

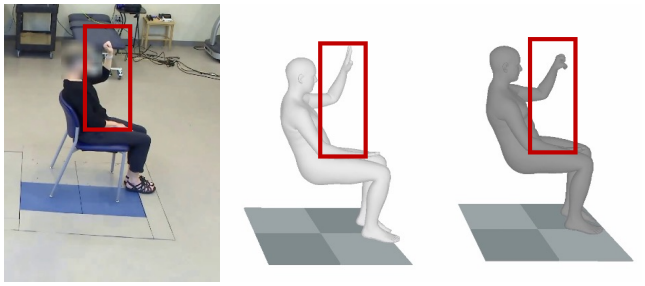
► **Challenge 1: On-device 3D human motion recovery on a laptop.** Recovering metric-scale 3D human motion on a laptop is computationally demanding, especially when needing both whole-body and fine hand details. Our solution integrates SLAM-derived [38] camera trajectories with monocular depth estimation to establish metric scale, then fits a parametric body model (SMPL-X [35]) that jointly represents body and hands. We refine hand articulations using per-frame 2D keypoints and anatomical priors, optimizing for laptop execution (Section 4.2).

► **Challenge 2: Robust perception in uncontrolled home environments.** Home environments introduce interference (illumination, clutter, reflections) that corrupts 3D estimates. We address this with a physics-aware purification framework combining environment-aware preprocessing, multi-person tracking, and differentiable physics constraints (ground plane, foot contact) to rectify implausible reconstructions while preserving true pathological motion. Activity-specific priors further discriminate valid motion from artifacts (Section 4.3).

► **Challenge 3: Bridging kinematics to clinical scores with interpretability.** Direct regression from joint coordinates to UPDRS scores is a black box. We construct a kinematic feature space mirroring UPDRS descriptors, extracting frame-level (e.g., tremor frequency) and event-level (e.g., tapping amplitude decrement) variables. These are processed by a lightweight HCT that fuses macro-level statistics with micro-level dynamics, providing accurate and interpretable score predictions (Section 4.4).

We propose *TulipTender*, a single-camera system that integrates the above components to deliver comprehensive PD assessment in home environments (Fig. 1). After deploying *TulipTender* in various settings, we conducted a user experience survey with 22 Parkinson's disease (PD) patients. The results (detailed in Section 5.7) indicate that the system is perceived as easy to use and well-suited for repeated structured assessment, providing objective feedback to physicians beyond subjective recall. The main contributions are:

- We design and implement *TulipTender*, the first mobile computing system that captures the full spectrum of PD motor deficits, from whole-body gait to fine finger kinematics, from a single RGB camera, via a novel multi-scale 3D human modeling approach that integrates metric-scale body-and-hand recovery with 2D-guided refinement in a unified world coordinate system.
- We develop a lightweight, physics-aware signal purification framework optimized for edge execution, which actively cancels optical noise by combining 2D keypoint constraints, temporal filtering, and differentiable physics priors, enabling clean kinematic extraction from raw 3D estimates and robustness in uncontrolled home environments.
- We introduce an interpretable clinical mapping architecture designed for on-device inference, which predicts MDS-UPDRS classes from aligned event- and frame-level variables rather than raw joint coordinates using an HCT.



(a) Hand movement (b) Tram (c) TulipTender

Figure 2: Hand movement recovered by *TulipTender*.

• We evaluate *TulipTender* with 70 participants across inpatient, outpatient, nursing home, and home environments. The system achieves an average 82% and 84% F1-score for finger-tap and gait analysis. In addition, we release the consented subset of the dataset to support future research [4].

2 RELATED WORK

2.1 PD Assessment in Free-living Scenario

Various sensing modalities have been explored for quantifying PD motor symptoms in free-living settings. IMU-based methods have been extensively investigated for human motion monitoring [18, 24, 32, 34, 47–51, 53]. Single wrist devices enable tremor monitoring [29] but miss gait/posture, while multi-sensor setups capture full-body kinematics [7, 12, 28, 33, 57] yet burden patients and reduce adherence [31]. Moreover, wearable signals mix voluntary and pathological movements, and their separation is hindered by overlapping frequency bands, compromising precision for bradykinesia [31]. Wireless sensing (WiFi/radar/acoustic [13, 19, 40, 42–46, 52, 54, 55]) offers contactless monitoring [20, 25, 56] but suffers from multipath interference and lacks joint-level resolution. Optical sensing provides high spatial resolution: multi-camera systems achieve sub-millimeter accuracy [23] but are impractical for home deployment; single-camera solutions are ubiquitous [22, 27, 30, 37, 39] but struggle to recover metric 3D kinematics, limiting comprehensiveness. This gap motivates the need for monocular 3D sensing tailored to PD.

2.2 Monocular 3D Human Sensing

Recent computer vision advances enable 3D human pose and shape reconstruction from a single RGB image. Parametric models (SMPL [26], MANO [35]) and learning-based methods (HMR2 [16], FrankMocap [36]) infer body/hand models, while depth estimation [8] and visual SLAM [38] assist metric scale recovery. These techniques have been applied to PD assessment, e.g., webcam-based finger tapping [22], 2D-to-3D lifting [39], sit-stand analysis [30], and gait quantification [37]. However, existing works rely on 2D poses or simplified 3D skeletons, lacking full-body high-fidelity reconstruction

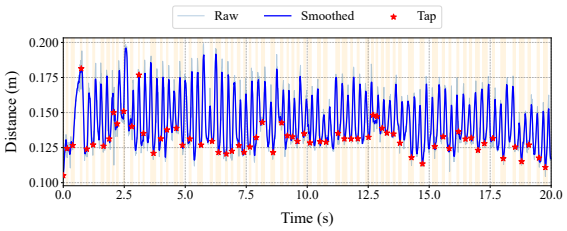


Figure 3: Finger-tap motor captured by *TulipTender*.

with physical scale. They are not optimized for PD-specific symptoms and inadequately address home environmental variability. Thus, robust metric-scale recovery, accurate full-body motion capture, and interpretable mapping to clinical scores remain open challenges.

3 A MEASUREMENT STUDY OF MULTI-SCALE HUMAN POSE

This section introduces the basic components of *TulipTender* and demonstrates the system's ability to recover full-body motion with hand details, a critical enabler for comprehensive PD assessment.

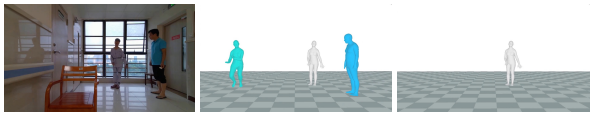
3.1 Basics of Human Mesh Recovery

3D Human Model and Keypoints. We represent the 3D human body using the SMPL model [26]. SMPL is a parametric function $M(\theta, \beta, \gamma, \pi) \in \mathbb{R}^{6890\times3}$ that outputs a mesh, where $\theta \in \mathbb{R}^{23\times3}$ denotes relative rotations of 23 joints, $\beta \in \mathbb{R}^{10}$ controls body shape, $\gamma \in \mathbb{R}^3$ is the root orientation, and $\pi \in \mathbb{R}^3$ is the camera-relative translation. For a temporal sequence, we define the global trajectory at time t as $T_t = \{\gamma_t, \pi_t\}$ and the local pose and shape as $\Theta_t = \{\theta_t, \beta_t\}$ (shape is assumed constant across frames). The 3D joint locations $J \in \mathbb{R}^{J\times3}$ are obtained via a linear regressor from the mesh vertices, following the standard SMPL joint regressor. The mesh is defined in metric units, directly providing physical scale. For 2D pose estimation, we employ AlphaPose [14] to obtain full-body keypoints. It yields 136 keypoints per person (26 body joints, 68 facial landmarks, 21 per hand), providing comprehensive coverage for holistic motor analysis.

Human Scale Estimation. We estimate metric depth from video frames using ZoeDepth [8]. Given keyframe I_i, ZoeDepth predicts depth map $D_i \in \mathbb{R}^{h\times w}$ in meters. Meanwhile, DROID-SLAM [38] provides depth d_i in arbitrary units. By solving for scaling factor α aligning αd_i to D_i, we recover the metric scale of the reconstructed scene and camera trajectory.

3.2 *TulipTender*'s Capability in Recovering Multi-scale Motions in Metric 3D Space

TulipTender unifies the traditionally fragmented pipeline for monocular full-body 3D reconstruction, integrating body and



(a) Visual artifacts (b) Recovered (c) Denoised 3D mesh

Figure 4: Body reconstruction of *TulipTender*.

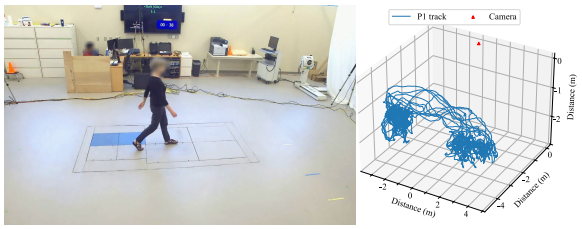


Figure 5: 11 trajectories recovered by *TulipTender*.

hand models (SMPL [26] and MANO [35]) with depth estimation (ZoeDepth [8]) into a coherent framework optimized for laptops. This enables accurate metric-scale full-body recovery directly on a MacBook Pro.

Small-scale Hand Motion Recovery. We evaluate hand motion recovery on the TULIP dataset [23] (Basler camera, 1920×1200, 80 fps). Figure 2 presents 3D hand reconstruction: left shows ground truth, middle is TRAM [41], right is *TulipTender*. TRAM fails to reconstruct detailed hand articulations, whereas *TulipTender* successfully recovers them. Figure 3 visualizes a finger-tapping sequence from the same dataset. The blue line tracks thumb–index distance relative to the wrist axis (peak 10 cm), red stars mark tap events, and yellow boxes indicate the corresponding time windows that denote *TulipTender* can accurately sense fine hand articulations. This comparison shows that the refinement step recovers fine hand articulations that are missed by the baseline reconstruction.

4 *TULIPTENDER* DESIGN

Large-scale Body Motion Recovery. We evaluate body motion recovery using the TULIP dataset [23] and videos collected from PD patients' homes (iPad 9). Figure 4 shows a gait task reconstruction in a hospital setting: left is ground-truth recording, right is reconstructed frame (white: patient, green: artifacts, blue: clinician). Incorporating 2D keypoints and activity priors reduces artifacts; denoising (Sec. 4.3) further removes imperfections. Figure 5 illustrates the gait trajectory from TULIP's P1 (Camera 5), confirming accurate metric 3D motion recovery.

4.1 System Overview

We present *TulipTender*, a framework for comprehensive, precise, and privacy-preserving PD assessment using only a laptop camera. All processing—from video capture to UPDRS scoring—runs locally, ensuring privacy and eliminating network dependency (Fig. 6).

The framework operates in three stages. First, *TulipTender* reconstructs a 3D human mesh from video, capturing fine-grained full-body movement (global trajectory and local motion) in world coordinates. This provides a consistent foundation for UPDRS analysis. Second, to accurately identify UPDRS activity events, it integrates 2D image cues with 3D motion constraints via a multi-dimensional alignment algorithm, mitigating visual artifacts (e.g., projection errors) and enforcing physical plausibility for robust event detection. Finally, to link motion capture to clinical scoring, *TulipTender* extracts kinematic features at event and frame levels (e.g., joint angles, velocities, symmetry). These features feed a Transformer model to predict UPDRS scores, enabling transparent and clinically meaningful assessment. By integrating these stages, *TulipTender* provides metric-scale body-and-hand recovery and clinically aligned scoring within a unified single-camera, on-device pipeline.

4.2 Holistic 3D Human Movement Recovery in World Space

4.2.1 Metric-Scale and Smooth 3D Human Reconstruction. To reconstruct a metric-scale 3D human from monocular video, *TulipTender* combines three components: SLAM for camera motion, metric depth estimation for scale recovery, and a temporally coherent human mesh regressor.

First, DROID-SLAM [38] processes the video to obtain camera poses and a dense 3D scene structure, but with arbitrary scale. It estimates optical flow between frames and optimizes depth and camera parameters via dense bundle adjustment, yielding a geometrically consistent but unscaled reconstruction. Second, to resolve the scale ambiguity, we leverage ZoeDepth to predict metric depth maps (in meters) for each frame. A distance threshold discards unreliable far regions, and temporal median filtering smooths the depth sequence, providing a stable metric reference. Third, for human mesh recovery, we adopt VIMO [41] rather than a single-image method like Human Mesh Recovery (HMR) 2.0 [16]. VIMO incorporates both appearance and motion cues across frames, producing coherent pose and shape estimates, thereby reducing jitter and ensuring physical plausibility. Finally, the camera parameters from SLAM anchor the SMPL human mesh into a consistent world coordinate system with metric scale derived from ZoeDepth. This unified representation decouples the human from the original viewpoint, enabling accurate rendering from arbitrary virtual perspectives and strong generalization.

4.2.2 2D-Guided Fine-Grained Hand Mesh Refinement. To achieve the fine-grained hand details required for PD assessment, we propose a 2D-guided hand refinement module that optimizes SMPL-X hand parameters using per-frame 2D

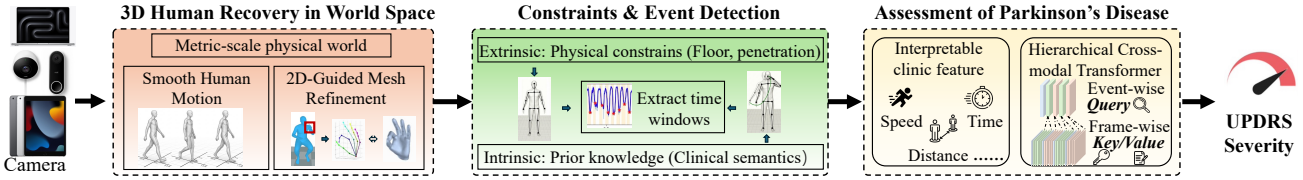


Figure 6: Overview of the *TulipTender* framework, designed to run entirely on a laptop.

keypoints, ensuring that hand articulations are accurately recovered while maintaining coherence with the body model.

2D Full-Body Pose Estimation. Existing HMR methods that recover motion in world coordinates typically estimate only a sparse set of body keypoints, which is insufficient for detecting the full spectrum of motor deficits. To enable comprehensive analysis, we supplement the missing anatomical details—in particular the hands, each represented by 21 keypoints. While OpenPose [10] and Hidalgo [17] offer unified detection of body, face, hand, and foot keypoints for multiple persons, they have notable limitations. OpenPose first detects body keypoints using Part Affinity Fields (PAFs) [11], then applies separate networks for face and hand estimation; this cascaded design is computationally expensive and slows inference. Hidalgo [17] employs a single-step network, but its output resolution is limited, reducing accuracy on fine-grained keypoints such as those of the hands. To address these issues, we adopt AlphaPose [14] for whole-body 2D pose estimation. AlphaPose follows a top-down paradigm: a YOLOv3 [21] human detector first localizes each person with a bounding box, after which a single-person pose estimator processes the cropped region to yield 136 keypoints covering the body, face, and both hands.

3D Fine-Grained Refinement with 2D Guidance. We employ SMPL-X [35], an expressive 3D human body model that extends SMPL with fully articulated hands and a detailed face model, to recover complete human motion.

To obtain a holistic 3D reconstruction in world space, it is necessary to accurately estimate the hand pose parameters within SMPL-X. However, directly fitting the 21 keypoints of each hand is data-intensive and prone to ambiguity. We therefore propose an optimization strategy that leverages 2D keypoints as guidance. Given the 2D full-body pose (including hands) detected by AlphaPose, along with the pre-estimated SMPL body parameters (pose, shape, and global translation) and default SMPL-X hand parameters, we optimize the hand pose parameters to align the projected 3D joints with the detected 2D keypoints.

The optimization utilizes the camera parameters estimated during the 3D reconstruction stage, which include the world-to-camera rotation matrix $R_{\text{world}\to\text{cam}} \in R^{3\times3}$, translation vector $T_{\text{world}\to\text{cam}} \in R^3$, and intrinsic parameters (focal lengths f_x, f_y and principal point c_x, c_y).

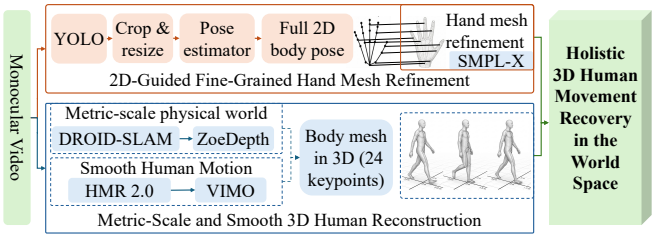


Figure 7: Holistic 3D human movement recovery in the world space.

Prior to optimization, we associate the detected 2D hand keypoints with the corresponding person trajectory. AlphaPose outputs, for each frame, a set of bounding boxes and their associated 136 keypoints, where indices 94–114 correspond to the left hand and 115–135 to the right hand. To assign these detections to the tracked person trajectories, we employ an Intersection-over-Union (IoU) matching strategy. For each bounding box in a trajectory, we compute its IoU with all detected boxes in the same frame: $\text{IoU}(A, B) = \frac{|A\cap B|}{|A\cup B|}$, where A and B are the coordinates of the tracked and detected boxes, respectively. The detection with the highest IoU is assigned to that trajectory. This method robustly handles occlusions and multi-person scenarios, ensuring that the correct 2D hand keypoints are used for optimizing the corresponding SMPL-X hand pose.

Mapping 2D Observations to 3D Space. For each frame $t \in 1, 2, ..., T$ in the video sequence, we formulate an energy-minimization framework to refine the hand pose parameters of the SMPL-X model. This framework optimizes only the hand pose parameters to align with 2D detected keypoints, while keeping other body parameters fixed, namely body pose, global rotation, shape, and global translation.

Optimization variables. At each frame t, the optimization variables are the axis-angle representations of both hands:

$$\Theta_t = \left\{\boldsymbol{\theta}_{\text{left},t} \in \mathbb{R}^{45}, \boldsymbol{\theta}_{\text{right},t} \in \mathbb{R}^{45}\right\}. \tag{1}$$

Each hand contains 15 joints represented by 3D axis-angle vectors, resulting in 45 dimensions per hand. The parameters are initialized using the flat-hand mean pose of SMPL-X.

Projection and Reprojection Error. Given Θ_t, the SMPL-X forward pass produces 21 world-space hand joints. For each joint i, let $\mathbf{j}_{w,i,t}$ denote its world-space coordinate. Its camera-space coordinate is $\mathbf{j}_{c,i,t} = R_{cw,t}\mathbf{j}_{w,i,t} + \mathbf{t}_{cw,t}$, where $R_{cw,t}$ and $\mathbf{t}_{cw,t}$ transform world coordinates to camera coordinates. We

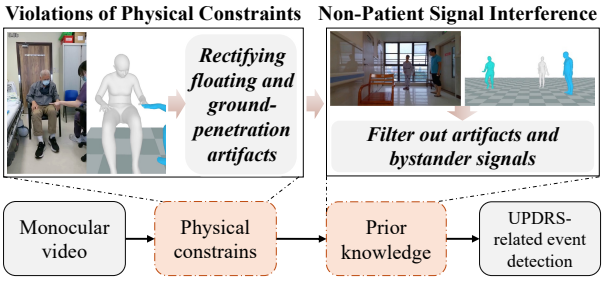


Figure 8: Multi-dimensional activity recognition.

project the joint onto the image plane using $\mathcal{P}$:

$$\mathcal{P}\left(\mathbf{j}_{c,i,t}\right) = \begin{bmatrix} f_x x_{i,t}/z_{i,t} + c_x \\ f_y y_{i,t}/z_{i,t} + c_y \end{bmatrix}. \tag{2}$$

Here, $\mathbf{j}_{c,i,t} = [x_{i,t}, y_{i,t}, z_{i,t}]^\top$, (f_x, f_y) are the focal lengths, and (c_x, c_y) is the principal point.

Frame-wise Objective Function. For each frame t, we refine Θ_t by minimizing

$$\mathcal{E}_t(\Theta_t) = \mathcal{E}_{2D,t}(\Theta_t) + \lambda_{\text{reg}}\mathcal{E}_{\text{reg},t}(\Theta_t). \tag{3}$$

The confidence-weighted reprojection term is

$$\mathcal{E}_{2D,t}(\Theta_t) = \frac{1}{N}\sum_{i=1}^{N} w_{i,t} \left\|\mathcal{P}(\mathbf{j}_{c,i,t}) - \widehat{\mathbf{p}}_{i,t}\right\|_2^2. \tag{4}$$

Here, $N = 21$, $\widehat{\mathbf{p}}_{i,t} \in \mathbb{R}^2$ is the detected 2D keypoint, and $w_{i,t} = \text{clip}(c_{i,t}, 0, 1)$ is its AlphaPose confidence.

A hand-pose regularization term keeps the estimate near the neutral SMPL-X prior:

$$\mathcal{E}_{\text{reg},t}(\Theta_t) = \left\|\boldsymbol{\theta}_{\text{left},t}\right\|_2^2 + \left\|\boldsymbol{\theta}_{\text{right},t}\right\|_2^2, \tag{5}$$

where the L2-norm penalty is weighted by $\lambda_{\text{reg}} = 1 \times 10^{-3}$. The parameters for frame t are obtained by solving: $\Theta_t^* = \arg\min_{\Theta_t} \mathcal{E}_t(\Theta_t)$. We use gradient descent with a learning rate of $\eta = 0.02$ for $T_{\text{iter}} = 80$ iterations. To maintain stability, frames with fewer than 6 confident keypoints are excluded from optimization. The final output for each processed frame includes the refined hand pose parameters $\Theta_{hand,t}$, the corresponding 3D joint positions $J(\Theta_t^*)$, and the palm normal vector.

4.3 Multi-dimensional Activity Recognition

A critical challenge in home-based video analysis is the presence of optical noise—specular reflections, depth inaccuracies, and physically implausible configurations—that corrupt the 3D reconstruction. This section presents our signal purification framework designed to suppress these artifacts and isolate valid patient motion.

To assess motor deficits, the first step is to extract valid activity windows from video data. However, 3D human pose and shape estimates recovered from monocular videos are prone to both visual artifacts (which vary across environments such as homes, sanatoriums, and hospitals) and physical violations (e.g., body parts floating above or penetrating through the floor). These inaccuracies directly hinder the reliable extraction of activity duration. Therefore, prior to segmenting activities, it is essential to enhance the physical plausibility of the estimated 3D human motions. Given that physical violations are often unavoidable in monocular 3D reconstruction, we seek a more robust signal from a lower-dimensional representation: 2D human pose estimation. This approach is motivated by the observation that 2D pose estimation, being a direct image-based prediction, is inherently unaffected by 3D physical constraints and is generally more resilient to certain visual artifacts compared to 3D recovery.

4.3.1 Leveraging 2D Keypoints for Physical Constraints. To mitigate the effects of physical violations (floating/floor penetration), we use the detected 2D keypoints of the target person to guide and constrain the 3D estimates. As described in Section 4.2, we obtain full-body 2D keypoints for each frame. From these keypoints and their corresponding 3D SMPL-X joints, we can derive consistent 2D and 3D bounding boxes for the subject, enabling cross-dimensional consistency checks. We match each 3D track to its corresponding 2D detection using bounding-box overlap (**Pmatch**) and remove low-confidence temporal discontinuities (**Filter**). We then align the supporting foot with the estimated floor (**Align**) and impose contact constraints on the feet during stance or on hand landmarks during touch (**Touch**). Only tracks satisfying these consistency checks are retained for activity-window extraction.

4.3.2 Incorporating Prior Knowledge to Reduce Visual Artifacts. Addressing visual artifacts presents a more fundamental challenge. While using 2D keypoints offers some robustness, artifacts cannot be fully eliminated simply by working in a lower dimension. For the specific goal of recognizing UPDRS activity durations, we can incorporate activity-specific prior knowledge. For instance, for the "finger-to-nose" test (kinetic tremor assessment), we utilize the prior knowledge that the participant's index finger should move between their own nose and the examiner's finger. This prior helps because visual artifacts are often discontinuous and exhibit irregular deformations. By integrating continuous, activity-specific motion patterns, we can more accurately localize the target person's bounding box and filter out inconsistent, artifact-induced noise, thereby isolating the valid activity period. We use analogous task-specific cues for the other activities. Finger tapping, hand movement, and pronation–supination are segmented using extrema in thumb–index distance, hand aperture, and forearm rotation, respectively. Gait and chair-rise windows are identified from pelvis motion, alternating foot contacts, and knee extension. These priors determine

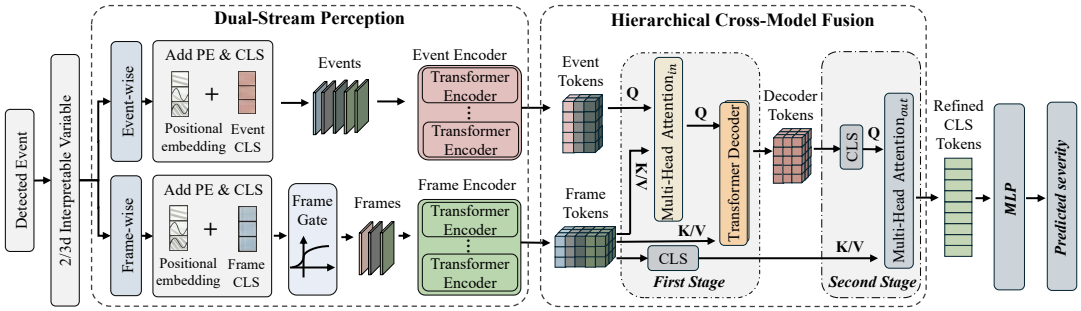


Figure 9: Hierarchical cross-modal transformer.

activity boundaries only; they do not provide symptom severity labels. Task-inconsistent segments are excluded before feature extraction.

4.4 Assessment of Parkinson's Disease

To bridge the gap between 3D kinematics and clinical scores, we design a two-stage interpretable pipeline that operates on MDS-UPDRS-aligned kinematic variables rather than raw joint coordinates: (i) extraction of clinically meaningful variables at both event and frame levels, and (ii) a lightweight Hierarchical Cross-modal Transformer (HCT) that integrates these variables for UPDRS prediction.

4.4.1 Interpretable Variable Extraction. To quantify motor symptoms, we extract interpretable clinical variables at two complementary levels from each detected UPDRS activity window: event-wise summaries and frame-wise kinematic details.

Event-wise variables characterize the overall execution of a complete activity trial. For each trial, we compute summary statistics across temporal (e.g., duration, speed), spatial (e.g., amplitude, path deviation), and regularity (e.g., cycle variability) domains. Frame-wise variables capture detailed kinematics by analyzing the time series of body keypoints. Our general extraction pipeline first establishes a task-specific local coordinate system and normalizes spatial data using anatomical references. Base kinematic features—including joint angles, velocities, and distances—are then computed per frame. From these, we derive higher-order metrics such as tremor frequency power, spectral arc length (smoothness), and inter-limb coordination.

This framework is applied consistently across different motor tasks. For example, in analyzing the finger-to-nose task for kinetic tremor, we compute normalized 2D finger joint angles and extract derived measures of tremor power and smoothness. Similarly, for gait or hand alternating tasks, we adapt the relevant keypoints and compute analogous domain-informed features (e.g., stride length, cadence, or pronation-supination amplitude).

4.4.2 UPDRS Scoring Model. We propose a Hierarchical Cross-modal Transformer (HCT) for fine-grained analysis of hand interaction videos. The model is designed to jointly capture macro-level event statistics and micro-level temporal dynamics. As shown in Fig. 9, HCT comprises two core components: (1) a Dual-Stream Perception that processes both event-level and frame-level feature sequences for each event; (2) a Hierarchical Cross-Modal Fusion that integrates information from both levels.

• **Dual-Stream Perception**. The model input is a two-level hierarchy constructed from a sliding window over the detected event sequence. Each window contains a fixed number of events, denoted as T. For each event t, two feature types are extracted:

▷ Event-level features $\mathbf{e}_t \in \mathbb{R}^{d_e}$: Statistical summaries of the event (e.g., duration, peak touch probability, mean distance). Within a window, these are stacked into $\mathbf{E} \in \mathbb{R}^{T \times d_e}$.

▷ Frame-level features $\mathbf{f}_{t,i}{}_{i=1}^{L_t}$, where $\mathbf{f}_{t,i} \in \mathbb{R}^{d_f}$: Frame-wise kinematic data within event t, such as angle, velocities, and accelerations. Each event has a variable number of frames L_t. The collection for a window is denoted as $\mathcal{F} = \mathbf{F}_t{}_{t=1}^{T}$, where $\mathbf{F}_t \in \mathbb{R}^{L_t \times d_f}$.

To handle variable-length sequences, we set a maximum frame length $L_{\max} = 1024$ and apply padding masks. Both event and frame features are normalized using statistics from the training set.

Feature Projection and Positional Encoding. Event and frame features are projected into a shared hidden dimension $d_{\text{model}} = 128$ via separate linear layers: $\mathbf{H}^e = \mathbf{E}\mathbf{W}^e + \mathbf{b}^e$, $\mathbf{H}_t^f = \mathbf{F}_t\mathbf{W}^f + \mathbf{b}^f$, where $\mathbf{W}^e \in \mathbb{R}^{d_e \times d_{\text{model}}}$, $\mathbf{W}^f \in \mathbb{R}^{d_f \times d_{\text{model}}}$, and $\mathbf{b}^e, \mathbf{b}^f$ are learnable parameters.

Learnable positional encodings are added to preserve temporal order. We define two encoding matrices: $\mathbf{P}^e \in \mathbb{R}^{T_{\max} \times d_{\text{model}}}$ for events ($T_{\max} = 512$) and $\mathbf{P}^f \in \mathbb{R}^{L_{\max} \times d_{\text{model}}}$ for frames. The encoded representations are: $\widetilde{\mathbf{H}}^e = \mathbf{H}^e + \mathbf{P}^e_{:T}$, $\widetilde{\mathbf{H}}_t^f = \mathbf{H}_t^f + \mathbf{P}^f_{:L_t}$. A learnable [CLS] token is prepended to both event and frame sequences to aggregate global information.

Event and Frame Encoder. The event encoder is a standard Transformer encoder with $N_e = 2$ layers. Each layer

contains multi-head self-attention (MSA) followed by a feed-forward network (FFN), with layer normalization (LN) and residual connections applied around each sub-layer: $\mathbf{Z}_l = \text{FFN}(\text{LN}(\mathbf{Z}'_l)) + \text{MSA}(\text{LN}(\mathbf{Z}_{l-1})) + \mathbf{Z}_{l-1}$, where $l = 1, \ldots, N_e$ and $\mathbf{Z}_0 = \widetilde{\mathbf{H}}^e$. The encoder outputs contextualized event representations $\mathbf{Z}^e = \mathbf{Z}N_e \in \mathbb{R}^{T \times d_{\text{model}}}$. We use $h = 4$ attention heads and a dropout rate of 0.1 for regularization. Similarly, the frame encoder is an independent Transformer encoder with $N_f = 2$ layers that processes each frame sequence $\widetilde{\mathbf{H}}t^f$ in parallel for efficiency. The input is reshaped to $(B \cdot T, L_t, d_{\text{model}})$, where B is the batch size. After encoding, we take the [CLS] token representation as the summary for each event's frame sequence, resulting in $\mathbf{Z}^f \in \mathbb{R}^{(B \cdot T) \times d_{\text{model}}}$, which is then reshaped back to $B \times T \times d_{\text{model}}$.

• **Hierarchical Cross-Modal Fusion**. We design a two-layer cross-attention mechanism to fuse event and frame information. First, frame summaries are projected (for generality): $\mathbf{M} = \mathbf{Z}^f \mathbf{W}^m + \mathbf{b}^m$.

The first layer treats event representations $\mathbf{Z}^e$ as queries and frame summaries $\mathbf{M}$ as keys/values: $\mathbf{A}^{(1)} = \text{MultiHeadAttention}(\mathbf{Q} = \mathbf{Z}^e, \mathbf{K} = \mathbf{M}, \mathbf{V} = \mathbf{M})$. A Transformer decoder layer then refines the events: $\mathbf{Z}^{\text{mid}} = \text{TransformerDecoderLayer}(\mathbf{Z}^e + \mathbf{A}^{(1)}, \mathbf{M})$.

The second layer uses only the event [CLS] token $\mathbf{Z}^{\text{mid}}_{[\text{CLS}]}$ as the query to attend to $\mathbf{M}$, producing the final fused representation:

$$\mathbf{A}^{(2)} = \text{MultiHeadAttention}(\mathbf{Q} = \mathbf{Z}^{\text{mid}}_{\text{CLS}}, \mathbf{K} = \mathbf{M}, \mathbf{V} = \mathbf{M}),$$
$$\mathbf{Z}^{\text{fused}} = \mathbf{Z}^{\text{mid}}_{\text{CLS}} + \mathbf{A}^{(2)}. \quad (6)$$

• **MDS-UPDRS Prediction**. The fused vector $\mathbf{Z}^{\text{fused}}$ is passed through a two-layer MLP with GELU activation and dropout to produce logits $\hat{\mathbf{y}} \in \mathbb{R}^C$ for C classes. To address class imbalance, we employ the focal loss with label smoothing: $\mathcal{L}_{\text{focal}} = -\sum_{i=1}^{C} \alpha_i (1 - p_i)^\gamma \log(p_i)$, where p_i is the predicted probability for class i, $\gamma = 1.5$ is the focusing parameter, and α_i is a class-weighting factor set inversely proportional to the class frequency. Label smoothing with a factor of 0.05 is applied to the ground-truth labels to prevent overconfidence.

5 EVALUATION

5.1 Implementation and Experiment Setup

Hardware. We validate the clinical accuracy of *TulipTender* on an RTX 4090 to establish upper-bound performance. To assess deployability on a consumer laptop (target platform), we also profile the system's resource consumption (latency, memory) on a MacBook Pro (Apple M5), demonstrating that the entire pipeline can run locally on a standard laptop without cloud offloading.



(a) UPDRS gait (b) UPDRS finger-tap

Figure 10: Examples of data collection.

Section 5.2 reports RTX 4090 clinical accuracy; MacBook experiments verify full-pipeline execution and profile representative finger-tapping and gait sessions. Both use the same FP32 model. Tested MacBook reconstructions are consistent with server results, and cross-backend numerical differences do not change the discrete MDS-UPDRS class.

MacBook Pro. The MacBook Pro is equipped with an Apple M5 CPU and leverages Apple's Metal Performance Shaders (MPS) framework for GPU acceleration [2]. MPS provides a high-level API and a unified memory architecture, significantly reducing data-copy overhead. Since PyTorch 1.12, official MPS backend support enables tensor and model execution via `torch.device("mps")`, allowing efficient AI model fine-tuning on Apple Silicon Macs. This capability underscores the feasibility of running *TulipTender* on resource-constrained devices, offering a low-barrier alternative to traditional NVIDIA GPU-based setups.

NVIDIA RTX 4090. The server platform utilizes an AMD EPYC 7542 32-Core Processor as the CPU, paired with an NVIDIA RTX 4090 GPU for high-performance computation. This configuration represents a standard high-end setup for comparison with edge-side execution.

Software. We implement the *TulipTender* system in Python for data processing and train the Parkinson's disease assessment model using PyTorch 2.4.0 [6]. The implementation has been open-sourced at [4].

Setup of Data Collection. We conducted a multi-center data collection campaign across diverse environments, including two hospitals, two nursing homes, and patients' private residences. As shown in Figure 10, camera views and setups were tailored to the physical constraints of each site, as our approach does not require a fixed viewing angle. Participants performed instructed, structured MDS-UPDRS Part III tasks in these settings; the dataset does not consist of unconstrained free-living activities. A total of 70 subjects were recruited for this study, with ages ranging from 49 to 91 years, including 30 female participants. The cohort comprised 43 inpatients, 20 outpatients, 2 nursing home residents, and 5 home-based participants.

Within this cohort, 39 subjects contributed videos related to PIGD-relevant assessments, primarily focusing on the gait

	All Subjects	Gait Tasks	Upper-limb Tasks
Subject Number	70	39	44
Age (years)	69.55	68.20	72.04
Sex (Female:Male)	F30:M40	F20:M19	F14:M30
Distance Range (m)	1–5	1–5	1–3

Table 1: Demographics of the study cohort and assessment-specific subsets.

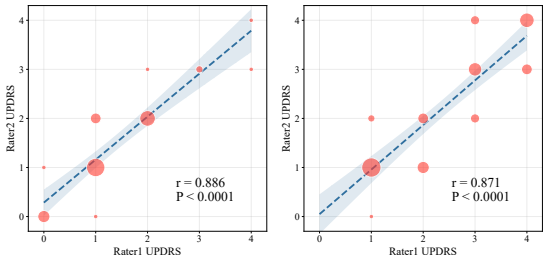
Table 2: Evaluation data by sub-item. S/V denotes subjects/video clips. See [4] for dataset details.

Sub-item	Our (S/V)	TULIP (S/V)
Finger tapping	44/44	12/12
Gait	39/139	12/12
Other evaluated sub-items	–	12/12 each

item of UPDRS Part III. Additionally, 44 subjects participated in upper-limb motor assessments, corresponding to the UPDRS Part III items Finger Tapping, Hand Movements, and Pronation–Supination Movements. Notably, four subjects underwent a complete UPDRS Part III examination.

We collected data using consumer-grade devices (iPhone 15 and iPad 9) at 30 fps. We employed a flexible recording protocol to simulate real-world deployment conditions. Recording distances ranged from 1 m to 5 m, and camera viewpoints were varied randomly to capture complete motion sequences. These choices were adapted to each specific assessment item and to environmental constraints—such as limited space in private homes. Throughout the evaluation, "layout variation" specifically denotes the combined change in camera-to-subject distance (1–5 m), viewing angle, illumination, and background clutter. The resulting variability in lighting, background clutter, and viewing angles provides a rigorous test of the system's generalization ability. While data were collected using various devices (iPhone, iPad) to simulate diverse capture conditions, all subsequent processing for evaluation can be performed on the MacBook Pro, consistent with our target deployment scenario.

Evaluation Data. Finger tapping and gait were evaluated using both our dataset and TULIP. In our dataset, 44 subjects contributed 44 finger-tapping video samples, while 39 subjects contributed 139 gait video samples, as reported in Table 2. We count each separately stored and labeled task recording as one video sample. Multiple recordings from the same participant are counted separately, whereas derived events, overlapping temporal windows, and synchronized TULIP views of the same task performance are not. The remaining evaluated MDS-UPDRS Part III items included hand movements, pronation–supination, toe tapping, leg agility, global spontaneity, arising from chair, postural stability, postural tremor, and kinetic tremor. These items were evaluated only on TULIP, using 12 subjects and 12 video clips per item.



(a) UPDRS gait (b) UPDRS finger-tap

Figure 11: Inter-rater correlation on MDS-UPDRS Subscores. Circle size $\propto$ frequency; blue area: 95% confidence band.

Evaluation Metrics. Following the leave-one-participant-out cross-validation protocol from [23], we evaluate *TulipTender* using two quantitative metrics: i) Accuracy: the proportion of correctly classified test samples. ii) F1-score: the harmonic mean of precision and recall, which accounts for both false positives and missed detections in UPDRS scoring. Each fold holds out all trials and sessions from one participant; normalization is fit only on the remaining participants. Thus, no frame or event from the test participant enters training. Due to the imbalanced class distribution across all UPDRS sub-items, we report the weighted F1-score to compare different models fairly.

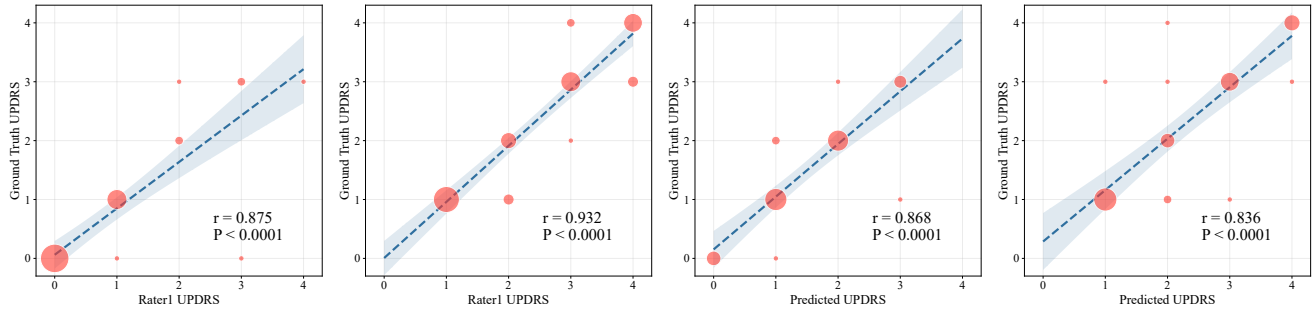
Baselines. We compare *TulipTender* with two baseline methods: TULIP [23], a method trained on a public dataset released by Duke University. Tram [41], which recovers 3D human motion for analysis.

5.2 Overall Performance

5.2.1 Clinical Consistency of Automated UPDRS Assessment. To validate the effectiveness of the proposed method in assessing Parkinson motor symptoms, we compared the automated UPDRS scores predicted by our model with the ground truth ratings provided by clinical experts. We also evaluated inter-rater reliability between two independent neurologists and the agreement between each rater and the consensus ground truth to establish a reference for clinical consistency.

For the gait task, Figure 12(a) shows the correlation between Rater 1 and ground truth (Spearman's $\rho = 0.875$, $p < 0.0001$), while Figure 11(a) illustrates the inter-rater agreement between Rater 1 and Rater 2 ($\rho = 0.886$, $p < 0.0001$). Figure 12(c) presents the correlation between model predictions and ground truth for gait, yielding a coefficient of $\rho = 0.868$ ($p < 0.0001$). The majority of data points cluster tightly around the diagonal dashed line (line of identity), with a narrow 95% confidence band (depicted as the blue shaded region) reflecting high predictive stability.

Similarly, for the finger tapping task, Figure 11(b) demonstrates strong inter-rater consistency between Rater 1 and



(a) Gait: Cons. Vs Clinicians (b) Finger-tap: Cons. Vs Clinicians (c) Gait: Cons. Vs *TulipTender* (d) Finger-tap: Cons. Vs *TulipTender*

Figure 12: Correlation with *TulipTender* and MDS-UPDRS subscores (Cons. denotes Consensus).

Table 3: Comparison of performance on TULIP and our datasets.

Task	Model	TULIP datasets		Our datasets	
		Acc.	F1	Acc.	F1
Fingertap	TULIP	0.70	0.69	-	-
	TulipTender	0.85	0.87	0.78	0.82
Gait	TULIP	0.73	0.72	-	-
	TulipTender	0.79	0.87	0.80	0.84

Rater 2 ($\rho = 0.871$, $p < 0.0001$). The agreement between Rater 1 and ground truth is shown in Figure 12(b), and Figure 12(d) displays the model's predictions versus ground truth, achieving a high correlation of $\rho = 0.836$ ($p < 0.0001$).

Notably, although the final predicted MDS-UPDRS labels are discrete, the model produces continuous-valued class probabilities, which provide additional information about prediction confidence and subtle variations in symptom presentation. Collectively, these results demonstrate that *TulipTender* achieves strong clinical consistency with expert neurologists across both fine (millimeter-scale finger tapping) and gross (meter-scale gait) motor tasks, supporting the feasibility of larger-scale clinical validation.

5.2.2 *Multi-Scale UPDRS Assessment: Capturing Hand and Gait Movements with a Single Camera.*

Performance of Finger-tap. The fine-grained UPDRS scoring performance of *TulipTender* was evaluated on the dataset from Section 5.1. This assessment was conducted across ten diverse real-world settings, encompassing private residences, nursing facilities, and both outpatient and inpatient hospital departments. In addition, we further evaluated the system on an open benchmark dataset, namely TULIP [23].

As shown in Table 3, *TulipTender* demonstrates strong performance on finger-tapping tasks compared to the TULIP baseline. On the TULIP dataset, *TulipTender* achieves an accuracy of 0.85 and an F1-score of 0.87, outperforming the original TULIP baseline, which reports an accuracy of 0.70 and an F1-score of 0.69. These results indicate that *TulipTender* generalizes well to public benchmark data despite using only a single-camera setup. Although fine-grained hand reconstruction from monocular video remains challenging, UPDRS scoring emphasizes clinically relevant temporal patterns, such as amplitude decrement and rhythm variability, rather than geometric accuracy alone. The gain therefore reflects the combination of recovered 3D motion, MDS-UPDRS-aligned variables, and HCT, rather than superior monocular geometry.

On our dataset, collected under more diverse and unconstrained real-world conditions, *TulipTender* attains an accuracy of 0.78 and an F1-score of 0.82 for finger-tapping assessment. Although slightly lower than the results on the TULIP dataset, this performance remains competitive given the greater variability in camera viewpoints, recording distances, and environmental settings.

Overall, the results demonstrate that *TulipTender* enables accurate and reliable UPDRS scoring for fine-grained whole-body motor assessment using a monocular 3D human motion analysis system.

Performance of Gait. We evaluate the UPDRS scoring performance of *TulipTender* on gait-related sub-items that rely on whole-body motion information, using both our dataset (Section 5.1) and the TULIP benchmark [23].

As shown in Table 3, *TulipTender* consistently outperforms the TULIP baseline on gait assessment. On the TULIP dataset, *TulipTender* achieves an accuracy of 0.79 and an F1-score of 0.87, compared to the baseline's reported accuracy of 0.73 and F1-score of 0.72.

On our dataset, *TulipTender* attains an accuracy of 0.80 and an F1-score of 0.84 for gait assessment. The superior performance on our dataset is achieved through our data analysis methodology, namely multi-dimensional activity recognition coupled with a hierarchical cross-modal transformer model. Furthermore, these results underscore *TulipTender*'s capability to stably recover global 3D human motion and effectively capture the gait characteristics defined by the

UPDRS, enabling robust scoring across diverse real-world conditions.

Overall, these results demonstrate that *TulipTender* delivers accurate and reliable UPDRS scoring for gait-related movements using a monocular 3D human motion analysis system.

5.2.3 Evaluation of Other UPDRS Sub-items. Figure 13 summarizes the performance of *TulipTender* on additional UPDRS Part III sub-items evaluated on the TULIP datasets. Overall, *TulipTender* demonstrates stable and competitive performance across multiple motor domains defined by the UPDRS, including bradykinesia, gait and postural control, and tremor.

Performance of Bradykinesia-related tasks. For upper- and lower-limb bradykinesia assessments, *TulipTender* achieves strong results. Specifically, the system attains an accuracy of 0.85 and an F1-score of 0.87 on *finger tapping*, 0.74 accuracy and 0.79 F1-score on *hand movements*, and 0.91 accuracy with 0.93 F1-score on *pronation*. For lower-limb bradykinesia, *TulipTender* achieves an accuracy of 0.75 and an F1-score of 0.86 on *toe tapping*, as well as 0.67 accuracy and F1-score on *leg agility*. The system further achieves an accuracy of 0.62 and an F1-score of 0.72 on *global spontaneity of movement*.

Performance of Gait and postural stability tasks. For whole-body and transitional movement assessments, *TulipTender* achieves an accuracy of 0.72 and an F1-score of 0.76 on *arising from chair*, 0.79 accuracy and 0.87 F1-score on *gait*, and 0.68 accuracy with 0.73 F1-score on *postural stability*, indicating robust performance in capturing clinically relevant gait and balance characteristics.

Performance of Tremor-related tasks. In tremor assessment, *TulipTender* demonstrates particularly strong performance. The system attains an accuracy of 0.81 and an F1-score of 0.85 on *postural tremor*, and achieves near-perfect results on *kinetic tremor*, with an accuracy of 0.98 and an F1-score of 0.99.

The limited number of samples per subject likely explains the lower performance in tasks like leg agility and global spontaneity of movement, which are associated with bradykinesia. In contrast, tremor-related tasks benefit from a larger number of temporal samples, leading to more stable and reliable performance.

Overall, these results indicate that *TulipTender* is capable of accurately handling a diverse set of UPDRS Part III motor sub-items, with performance variations primarily driven by data availability rather than inherent limitations of the proposed model.

5.3 Evaluation of Holistic 3D Human Movement Recovery

To systematically assess the necessity of 3D human motion recovery in UPDRS scoring, we perform a modality analysis comparing 2D pixel-level features with 3D world-space motion representations. Specifically, we use the gait sub-item as a representative task for modality ablation.

As shown in Fig. 14, incorporating 3D motion information yields a clear improvement over 2D-based representations. On the TULIP benchmark, the F1-score increases from 0.41 with 2D features to 0.72 when using 3D features, confirming the importance of view-invariant 3D motion cues. This strong 3D performance is expected, as TULIP relies on a multi-camera motion capture system that provides accurate 3D reconstruction.

In contrast, the 3D representation in our method is recovered from monocular video and achieves an F1-score of 0.66, which is slightly lower than the multi-view TULIP 3D result due to the inherent difficulty of monocular 3D reconstruction. Nevertheless, by jointly integrating complementary 2D and 3D features, our full model further boosts the F1-score to 0.84, substantially outperforming both the 2D and 3D baselines. These results demonstrate that effective fusion of 2D image cues with monocular 3D motion recovery enables robust and highly accurate UPDRS scoring.

5.4 Evaluation of Feature Interpretability

Clinical interpretability is as important as classification accuracy in UPDRS assessment. To analyze the contribution of different motion characteristics, we group the extracted features into three categories: (A) event-level spatiotemporal features (e.g., stride duration and rhythm), (B) event-level kinematic features (e.g., joint range of motion), and (C) frame-level dynamic features (e.g., instantaneous velocity and acceleration).

We conduct an ablation study by selectively enabling each feature group. As shown in Fig. 15, using a single feature group yields limited performance, with F1 scores of 0.49, 0.41, and 0.54 for spatiotemporal, kinematic, and dynamic features, respectively. In contrast, integrating all feature groups significantly boosts performance to an F1 score of 0.84. Notably, kinematic features play a key role in gait-related assessment, which is consistent with clinical observations that reduced movement amplitude (hypokinesia) is a core manifestation of Parkinson's disease. Categories (A) and (B) summarize event-level spatiotemporal and kinematic signatures, whereas Category (C) preserves frame-level dynamics. Their complementary contributions motivate the two-stream HCT design and support the interpretability of predictions based on clinically motivated features rather than raw joint coordinates.

5.5 Evaluation of Architecture Selection

To determine the optimal modeling strategy for mapping extracted gait features to UPDRS scores, we benchmarked our proposed HCT architecture against established machine

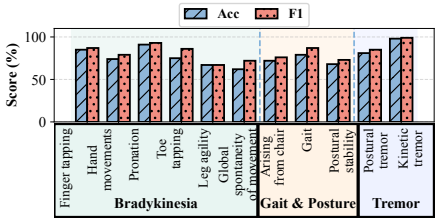


Figure 13: Evaluation of other MDS-UPDRS sub-items.

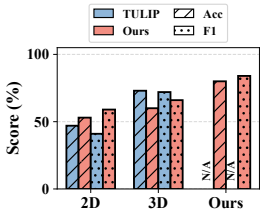


Figure 14: Modality analysis (2D vs 3D).

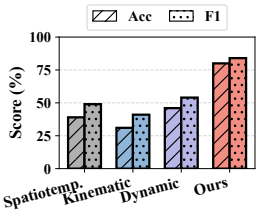


Figure 15: Feature ablation results.

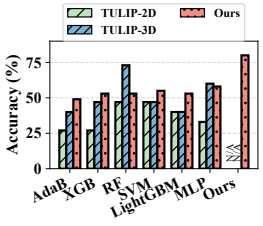


Figure 16: Model architecture comparison.

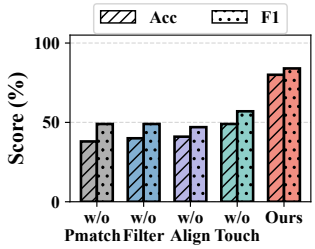


Figure 17: Ablation on physical constraints

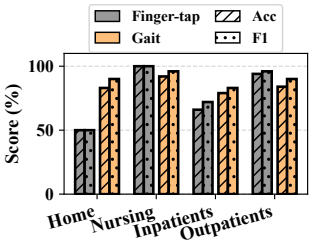


Figure 18: Robustness across scenes

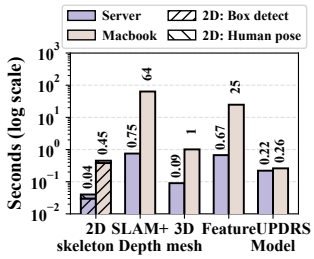


Figure 19: System overload time

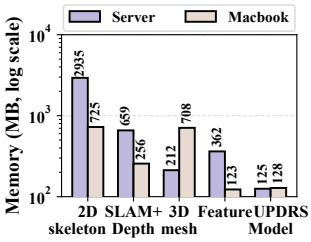


Figure 20: System overload memory

learning models using our dataset. We compared our approach with Adaptive Boosting (AdaBoost), Extreme Gradient Boosting (XGBoost), Random Forest (RF), Support Vector Machine (SVM), Light Gradient Boosting Machine (LightGBM), and Multi-Layer Perceptron (MLP) using leave-one-participant-out cross-validation. All models are evaluated using the same purified activity windows, participant folds, and evaluation metrics. The conventional baselines receive event-level summaries and pooled frame-level statistics, whereas HCT retains the event- and frame-level sequences and fuses them through cross-attention.

As shown in Figure 16, our HCT outperforms all baseline models. Among these baselines, MLP achieves the highest accuracy, attributed to its ability to effectively model both event-wise and frame-wise features, which provides an advantage over tree-based and kernel methods. Additionally, 3D-based feature extraction consistently yields superior performance across all model architectures compared to 2D approaches.

5.6 Impact Factors

In this subsection, we examine the robustness and practical viability of *TulipTender* under real-world conditions, with an emphasis on visual artifacts, environmental variations.

Effectiveness of Visual Artifacts. We systematically evaluate the optical noise cancellation capability of *TulipTender*. Visual artifacts such as occlusions, motion blur, and specular reflections are common in unconstrained home environments and can severely degrade monocular 3D motion recovery, leading to inaccurate UPDRS scoring. We ablate the four purification components described in Section 4.3 (**Pmatch**, **Filtering**, **Align**, and **Touch**) to measure their individual contributions to scoring robustness.

As shown in Fig. 17, removing any single constraint leads to a noticeable performance drop. In particular, excluding point matching or filtering significantly increases sensitivity to visual noise, resulting in F1-scores below 0.50. While the **w/o Touch** variant performs relatively better, its F1-score (0.57) still remains substantially lower than that of the full model. By integrating all physical constraints, *TulipTender* achieves an F1-score of 0.84, demonstrating strong robustness to visual artifacts without requiring manual frame curation.

These results indicate that the proposed physically constrained optimization is essential for stabilizing monocular 3D motion recovery under real-world visual disturbances, enabling reliable and fully automated UPDRS assessment.

Robustness Across Scenes. Figure 18 shows *TulipTender* performance across four typical scenes for Finger-tap and Gait tasks. In home environments, Finger-tap performance drops to 50% (accuracy and F1-score) due to substantial layout and lighting variations, while Gait remains more robust at 83% accuracy. In inpatient wards, background interference reduces Finger-tap accuracy to 66%, while Gait reaches 79%.

Conversely, in structured settings such as nursing facilities and outpatient clinics, *TulipTender* achieves consistently strong performance, with Finger accuracy reaching up to 94% and Gait exceeding 84%. Finger-tapping performance

Table 4: User feedback on *TulipTender* system

Usability dimension: mean score (1–5), N = 22 (PD)	
Ease of use: 4.13	**Long-term daily use**: 4.45
Safety & privacy: 4.32	Feedback clarity: 4.11
Home integration: 4.27	Maintenance ease: 4.34

is more sensitive to distance, oblique views, occlusion, and uneven lighting because fine hand motion occupies fewer image pixels. Gait is more stable because it relies on larger body landmarks and longer temporal context. Closer hand views, basic framing and lighting checks, and rejection of low-confidence windows can mitigate these failures. Performance in unrestricted home settings requires further validation.

5.7 *TulipTender* System Feedback

To evaluate the practical feasibility and patient acceptance of *TulipTender* in home environments, we conducted a user experience (UX) survey with 22 Parkinson's disease (PD) patients using a 5-point Likert scale (1: Strongly disagree, 5: Strongly agree) [15].

As shown in Table 4, *TulipTender* received a high mean score of 4.13 for 'Ease of use', which encompasses instant use, setup ease, and independent use. The high long-term daily use score (4.45) indicates participants' willingness to perform repeated structured assessments at home. The consistently high ratings across all dimensions (all at or above 4.11) support the usability and acceptance of *TulipTender* in this cohort, but do not establish clinical effectiveness for unrestricted long-term deployment.

5.8 System Overhead of *TulipTender*

We evaluate *TulipTender*'s overhead on a MacBook Pro (M5), our target platform, using an RTX 4090 server as reference. Latency is averaged over 1,000 iterations per module; memory is measured end-to-end.

The pipeline has five components: 2D skeleton, SLAM+ depth, 3D mesh, activity recognition, and UPDRS scoring. The first three are per-frame; activity and scoring are per-video/sample. FLOPs are dominated by geometric reconstruction: SLAM+ depth (13.59 G), 3D mesh (6.54 G), and UPDRS scoring (7.44 G); activity recognition is negligible (0.09 M).

Latency measurements (Fig. 19) show that on the server all modules run in <1 s. On the laptop, 3D modules have higher latency, while lightweight ones remain fast, consistent with FLOPs. Memory usage (Fig. 20) is moderate on both; server GPU peaks higher but laptop fits comfortably. For a 10-s gait video (300 frames), total laptop processing is 240 s (4 min)—acceptable for offline home use.

Both platforms use the same trained model, but the clinical-accuracy results in Section 5.2 are measured on the RTX 4090. Representative MacBook runs completed the full pipeline and produced 3D motion consistent with the server results. Small numerical differences between the CUDA and MPS backends did not change the predicted MDS-UPDRS classes in these tests. These results demonstrate the feasibility of running the complete pipeline locally on consumer hardware.

6 DISCUSSION AND LIMITATIONS

TulipTender evaluates guided MDS-UPDRS Part III tasks in a 70-participant feasibility study with uneven sub-item coverage (Table 2); it does not assess unrestricted free-living behavior or replace neurologist evaluation. Fine hand tasks remain more sensitive than gait to distance, viewpoint, illumination, and occlusion (Fig. 18). Larger-scale validation requires clinician-adjudicated labels, broader recruitment across disease states and environments, and multi-site ethics, privacy, and data-use approvals. Prospective longitudinal studies are needed before routine clinical deployment.

7 CONCLUSION

We present *TulipTender*, a mobile single-camera 3D system for structured Parkinson's disease assessment. It integrates metric-scale body-and-hand recovery, 2D-guided refinement, physics-aware purification, and MDS-UPDRS-aligned variables with a lightweight HCT for interpretable scoring. Clinical evaluation on the RTX 4090 achieves 82% and 84% F1-scores for finger tapping and gait, respectively. Representative MacBook tests confirm that the full pipeline runs locally, supporting the feasibility of privacy-preserving laptop assessment under guided capture conditions.

While smartphones offer greater ubiquity, current hardware limits full-body 3D reconstruction on such devices. Our laptop-based system balances accuracy and portability today, and with continued advances in mobile AI accelerators, the proposed methods can be migrated to smartphones in the near future.

ACKNOWLEDGMENTS

We thank all reviewers and our shepherd for their insightful comments. This research is supported in part by RGC under Contract CERG 16205824, 16204523, SZSTI25EG20, SRFS2425-6S05, AoE/E-601/22-R, HB016, and Shandong Science Fund under Contract ZR2026QC1090.

REFERENCES

[1] 2025. Parkinson disease. https://www.who.int/news-room/fact-sheets/detail/parkinson-disease.

[2] 2026. Accelerated PyTorch training on Mac. https://developer.apple.com/metal/pytorch/.

[3] 2026. Adjusting Parkinson's medication. https://www.parkinsons.org.uk/information-and-support/your-magazine/experts/adjusting-parkinsons-medication/.

[4] 2026. Dataset for TulipTender. https://zenodo.org/records/17964593.

[5] 2026. Information and resources. https://www.parkinsons.org.uk/information..

[6] 2026. PyTorch documentation. https://docs.pytorch.org/docs/stable/index.html.

[7] Jamie L Adams, Karthik Dinesh, Christopher W Snyder, Mulin Xiong, Christopher G Tarolli, Saloni Sharma, E Ray Dorsey, and Gaurav Sharma. 2021. A real-world study of wearable sensors in Parkinson's disease. *npj Parkinson's Disease* 7, 1 (2021), 106.

[8] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. 2023. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288* (2023).

[9] 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.

[10] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2021. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE transactions on pattern analysis and machine intelligence* 43, 1 (2021), 172–186.

[11] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2017. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7291–7299.

[12] Carlotta Caramia, Diego Torricelli, Maurizio Schmid, Adriana Muñoz-Gonzalez, Jose Gonzalez-Vargas, Francisco Grandas, and Jose L Pons. 2018. IMU-based classification of Parkinson's disease from gait: A sensitivity analysis on sensor location and feature selection. *IEEE journal of biomedical and health informatics* 22, 6 (2018), 1765–1774.

[13] Laura Dodds, Tara Boroushaki, Kaichen Zhou, and Fadel Adib. 2025. Non-line-of-sight 3D object reconstruction via mmWave surface normal estimation. In *Proceedings of the 23rd Annual International Conference on Mobile Systems, Applications and Services*. 445–458.

[14] Hao-Shu Fang, Jiefeng Li, Hongyang Tang, Chao Xu, Haoyi Zhu, Yuliang Xiu, Yong-Lu Li, and Cewu Lu. 2023. AlphaPose: Whole-Body Regional Multi-Person Pose Estimation and Tracking in Real-Time. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 6 (2023), 7157–7173.

[15] Pere J Ferrando, Fabia Morales-Vives, José M Casas, and José Muñiz. 2025. Likert Scales: A Practical Guide to Design, Construction and Use. *Psicothema* 37, 4 (2025), 1–15.

[16] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. 2023. Humans in 4d: Reconstructing and tracking humans with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 14783–14794.

[17] Gines Hidalgo, Yaadhav Raaj, Haroon Idrees, Donglai Xiang, Hanbyul Joo, Tomas Simon, and Yaser Sheikh. 2019. Single-network whole-body pose estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*. 6982–6991.

[18] Zhiqing Hong, Zelong Li, Shuxin Zhong, Wenjun Lyu, Haotian Wang, Yi Ding, Tian He, and Desheng Zhang. 2024. Crosshar: Generalizing cross-dataset human activity recognition via hierarchical self-supervised pretraining. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8, 2 (2024), 1–26.

[19] Haozheng Hou, Bowen Zheng, Sitong Cheng, Xiaoguang Zhao, Peiheng Wu, Lixing He, Yunqi Guo, Guoliang Xing, and Zhenyu Yan. 2025. AquaScan: A Sonar-based Underwater Sensing System for Human Activity Monitoring. In *Proceedings of the 31st Annual International Conference on Mobile Computing and Networking*. 438–452.

[20] 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.

[21] Muhammad Hussain. 2023. YOLO-v1 to YOLO-v8, the rise of YOLO and its complementary nature toward digital manufacturing and industrial defect detection. *Machines* 11, 7 (2023), 677.

[22] Md Saiful Islam, Wasifur Rahman, Abdelrahman Abdelkader, Sangwu Lee, Phillip T Yang, Jennifer Lynn Purks, Jamie Lynn Adams, Ruth B Schneider, Earl Ray Dorsey, and Ehsan Hoque. 2023. Using AI to measure Parkinson's disease severity at home. *NPJ digital medicine* 6, 1 (2023), 156.

[23] Kyungdo Kim, Sihan Lyu, Sneha Mantri, and Timothy W Dunn. 2024. TULIP: Multi-Camera 3D Precision Assessment of Parkinson's Disease. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 22551–22562.

[24] Yang Liu, Zhenjiang Li, Zhidan Liu, and Kaishun Wu. 2019. Real-time arm skeleton tracking and gesture inference tolerant to missing wearable sensors. In *Proceedings of the 17th Annual International Conference on Mobile Systems, Applications, and Services*. 287–299.

[25] Yingcheng Liu, Guo Zhang, Christopher G Tarolli, Rumen Hristov, Stella Jensen-Roberts, Emma M Waddell, Taylor L Myers, Meghan E Pawlik, Julia M Soto, Renee M Wilson, et al. 2022. Monitoring gait at home with radio waves in Parkinson's disease: A marker of severity, progression, and medication response. *Science Translational Medicine* 14, 663 (2022), eadc9669.

[26] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. 2023. SMPL: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*. 851–866.

[27] 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.

[28] 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.

[29] Nikhil Mahadevan, Charmaine Demanuele, Hao Zhang, Dmitri Volfson, Bryan Ho, Michael Kelley Erb, and Shyamal Patel. 2020. Development of digital biomarkers for resting tremor and bradykinesia using a wrist-worn wearable device. *NPJ digital medicine* 3, 1 (2020), 5.

[30] Deval Mehta, Umar Asif, Tian Hao, Erhan Bilal, Stefan Von Cavallar, Stefan Harrer, and Jeffrey Rogers. 2021. Towards automated and marker-less Parkinson disease assessment: predicting UPDRS scores using sit-stand videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 3841–3849.

[31] 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.

[32] Larsson Omberg, Elias Chaibub Neto, Thanneer M Perumal, Abhishek Pratap, Aryton Tediarjo, Jamie Adams, Bastiaan R Bloem, Brian M Bot, Molly Elson, Samuel M Goldman, et al. 2022. Remote smartphone monitoring of Parkinson's disease and individual response to therapy. *Nature biotechnology* 40, 4 (2022), 480–487.

[33] M Pastorino, MT Arredondo, J Cancela, and S Guillen. 2013. Wearable sensor network for health monitoring: the case of Parkinson disease. In *journal of physics: Conference series*, Vol. 450. IOP Publishing, 012055.

[34] Yanzhi Ren, Yingying Chen, Mooi Choo Chuah, and Jie Yang. 2014. User verification leveraging gait recognition for smartphone enabled mobile healthcare systems. *IEEE Transactions on Mobile Computing* 14, 9 (2014), 1961–1974.

[35] Javier Romero, Dimitrios Tzionas, and Michael J Black. 2022. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics (TOG)* (2022).

[36] Yu Rong, Takaaki Shiratori, and Hanbyul Joo. 2020. Frankmocap: Fast monocular 3d hand and body motion capture by regression and integration. *arXiv preprint arXiv:2008.08324* (2020).

[37] 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.

[38] Zachary Teed and Jia Deng. 2021. Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras. *Advances in neural information processing systems* 34 (2021), 16558–16569.

[39] 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.

[40] Yida Wang, Yu Lu, Yuxuan Zhou, Yifei Shen, Lili Qiu, Zeyuan Lai, Yi-Chao Chen, Hao Pan, Juntao Zhou, Dian Ding, et al. 2025. High-resolution mmwave imaging using metasurface and diffusion. In *Proceedings of the 23rd Annual International Conference on Mobile Systems, Applications and Services*. 319–332.

[41] Yufu Wang, Ziyun Wang, Lingjie Liu, and Kostas Daniilidis. 2024. TRAM: Global Trajectory and Motion of 3D Humans from in-the-wild Videos. In *European Conference on Computer Vision*. Springer, 467–487.

[42] Binbin Xie, Minhao Cui, Deepak Ganesan, Xiangru Chen, and Jie Xiong. 2023. Boosting the long range sensing potential of lora. In *Proceedings of the 21st Annual International Conference on Mobile Systems, Applications and Services*. 177–190.

[43] Binbin Xie, Minhao Cui, Deepak Ganesan, and Jie Xiong. 2024. Wall matters: Rethinking the effect of wall for wireless sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 7, 4 (2024), 1–22.

[44] Binbin Xie, Weizheng Wang, Deepak Ganesan, Lili Qiu, and Jie Xiong. 2025. Cross-Technology Sensing: Leveraging LoRa Signals to Empower WiFi Sensing. In *Proceedings of the 31st Annual International Conference on Mobile Computing and Networking*. 894–908.

[45] Binbin Xie and Jie Xiong. 2020. Combating interference for long range LoRa sensing. In *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*. 69–81.

[46] Binbin Xie, Yuqing Yin, and Jie Xiong. 2021. Pushing the limits of long range wireless sensing with lora. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 3 (2021), 1–21.

[47] Huatao Xu, Yan Zhang, Wei Gao, Guobin Shen, and Mo Li. 2025. Experience Paper: Adopting Activity Recognition in On-demand Food Delivery Business. In *Proceedings of the 31st Annual International Conference on Mobile Computing and Networking*. 1015–1028.

[48] 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.

[49] Xiangyu Xu, Jiadi Yu, Yingying Chen, Qin Hua, Yanmin Zhu, Yi-Chao Chen, and Minglu Li. 2020. TouchPass: Towards behavior-irrelevant on-touch user authentication on smartphones leveraging vibrations. In *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*. 1–13.

[50] Meng Xue, Yinan Zhu, Wentao Xie, Zhixian Wang, Yanjiao Chen, Kui Jiang, and Qian Zhang. 2025. MobHAR: source-free knowledge transfer for human activity recognition on mobile devices. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 9, 1 (2025), 1–24.

[51] Yilin Yang, Yan Wang, Yingying Chen, Zhengkun Ye, Xin Li, Zhiliang Xia, and Yanzhi Ren. 2024. TouchTone: Smartwatch Privacy Protection via Unobtrusive Finger Touch Gestures. In *Proceedings of the 22nd annual international conference on mobile systems, applications and services*. 141–154.

[52] Enze Yi, Dan Wu, Jie Xiong, Fusang Zhang, Kai Niu, Wenwei Li, and Daqing Zhang. 2024. {BFMSense}:{WiFi} sensing using beamforming feedback matrix. In *21st USENIX Symposium on Networked Systems Design and Implementation (NSDI 24)*. 1697–1712.

[53] Xiaojing Yu, Zhijun Zhou, Mingxue Xu, Xuanke You, and Xiang-Yang Li. 2020. Thumbup: Identification and authentication by smartwatch using simple hand gestures. In *2020 IEEE international conference on pervasive computing and communications (PerCom)*. IEEE, 1–10.

[54] Youwei Zeng, Dan Wu, Jie Xiong, Enze Yi, Ruiyang Gao, and Daqing Zhang. 2019. FarSense: Pushing the range limit of WiFi-based respiration sensing with CSI ratio of two antennas. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 3 (2019), 1–26.

[55] Junbo Zhang, Yiwen Song, and Swarun Kumar. 2025. PolarVisor: Clutter-free, Electronics-free Fiducial Markers for mmWave Radars Printed on Paper. In *Proceedings of the 31st Annual International Conference on Mobile Computing and Networking*. 1–16.

[56] 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.

[57] 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.