MULTI-PERSPECTIVE VIDEOS AND 3D POSES CAPTURING MULTI-TASK PARKINSON'S DISEASE ASSESSMENTS AND DEEP BRAIN STIMULATION TREATMENT EFFECTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Quantitative motor analysis is increasingly central to Parkinson's disease (PD) research and has the potential to transform PD clinical care. However, the field remains bottlenecked by datasets targeting coarse, manually-annotated PD severity scales rather than accurate PD motor feature identification or patient outcome prediction. Here, we introduce TULIP-BENCH, the first PD movement dataset and benchmark to combine raw multi-perspective videos, triangulated markerless 3D pose ground truth, both fine and gross motor tasks with clinical ratings (bilateral hand movements and gait), and within-patient treatment responses to deep brain stimulation (DBS). The TULIP-BENCH benchmark comprises tasks evaluating pose lifting and direct single-camera 3D pose tracking approaches in terms of clinical kinematic feature errors, quantities we found to be weakly associated with standard geometric (millimeter) error measures. We identified a critical perspective-dependent performance gap between single- and multi-camera PD assessments and illustrate how conventional PD clinical scores are unable to resolve DBS-induced gait changes. We also developed a self-supervised cross-cohort model for DBS response scoring based on representations learned from non-DBS PD patients. TULIP-BENCH will catalyze a new generation of methods that emphasize clinical deployability, sensitive treatment-focused measurement, and patient outcome prediction.

1 INTRODUCTION

Parkinson's disease (PD) affects over 10 million people worldwide (Bloem et al., 2021), and the bulk of clinical decision-making depends on the Movement Disorder Society Unified Parkinson's Disease Rating Scale (MDS-UPDRS, shortly UPDRS) (Goetz et al., 2008). The UPDRS is a manually-scored assessment originally developed in the 1980s (Fahn et al., 1987), with a primary focus on assessing assistive device needs. Its limitations are well known: inherently subjective ordinal 0–4 ratings collapse continuous motor variation into five categories that can mask meaningful changes, and inter-rater reliability is inconsistent (Williams et al., 2023). The community has responded with efforts to automate UPDRS scoring using human pose tracking. While making UPDRS scoring automated and reproducible would be an improvement, targeting solely UPDRS prediction pins PD analysis to a fundamentally coarse descriptive measure, thus limiting our ability to model PD motor variation, progression, and treatment responses. Developing methods beyond automated UPDRS scoring requires new publicly available body movement datasets that precisely capture the richness of PD symptoms and pair symptom measurements with objective and actionable clinical outcomes.

Here, we introduce TULIP-BENCH, a publicly available dataset comprising multi-view video recordings of PD motor exam tasks from 66 PD patients, accurate 3D pose ground-truth from multiple cameras, and PD severity ratings from movement disorder experts. A 10-subject subset with a history of deep brain stimulation (DBS) had recordings with their DBS device toggled OFF/ON (Dataset and code are available at: *Dataset link* and *Code link*).

In addition to providing DBS movement data for the first time, TULIP-BENCH also includes several other components that distinguish it from recent public PD pose datasets, such as CARE-PD (Adeli et al., 2025). While most other PD pose datasets treat 3D poses estimated from single-camera video as

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

ground truth for benchmarking, monocular 3D pose estimation is not yet a solved problem. Inferred poses are often noisy, especially when body parts are occluded and when body movement is mostly out of the camera plane. Instead, TULIP-BENCH ground-truth 3D poses are derived via triangulation across 6 hardware-synchronized cameras. Unlike CARE-PD (Adeli et al., 2025), which provides pose data only, TULIP-BENCH includes video recordings in addition to poses. This means that TULIP-BENCH can be used to benchmark end-to-end deployable clinical prediction methods that begin at single- or multi-camera video acquisition. And while most other PD motor datasets focus only on gait, TULIP-BENCH provides data from gait and additional hand motor tasks, which have been deemed essential by movement disorder specialists to assess and monitor fine motor skills.

We designed three benchmark tasks to accompany the TULIP-BENCH dataset. Tasks 1 and 2 quantify the accuracy and clinical utility of approaches for 2D-to-3D lifting (Task 1) and monocular 3D pose estimation directly from video (Task 2). Task 3 quantifies neurostimulation treatment effects detection. Our experiments reveal that methods achieving low metric (millimeter) pose tracking error do not necessarily achieve low error on clinically meaningful movement features. Furthermore, we found that the leading single-camera 3D pose estimation models, including those used to derive "ground truth" 3D poses in previous PD benchmark datasets, are surprisingly error-prone in terms of both geometric precision and clinical feature accuracy. We also show that a representation learned with self-supervision from subjects without DBS implants can be used to establish a continuous PD gait severity scale capable of detecting DBS-induced changes in gait not resolved by UPDRS gait scores. Overall, our tasks establish novel PD movement benchmarks centered on objective measures that are clinically interpretable and treatment-focused.

2 RELATED WORK

General, healthy human motion benchmarks. Elements of TULIP-BENCH were inspired by the large marker-based motion capture datasets, such as Human3.6M (Ionescu et al., 2013), that have driven progress in 3D pose estimation techniques by providing accurate ground truth at scale. Other datasets, such as CMU Panoptic (Joo et al., 2015) and FreeMan (Wang et al., 2024), have extended 3D pose benchmarking to markerless settings using keypoint triangulation across multiple cameras, enabling capture across a wider range of subjects, activities, and environments. However, these datasets focus on healthy subjects performing general activities and lack structured clinical tasks, pathological movements, and expert annotations. Consequently, it remains unclear whether methods developed on these benchmarks faithfully recover clinically relevant kinematics in PD.

Clinical PD motion datasets. PD motor assessment datasets have been released across a range of sensor modalities. Wearable-based datasets (Anderson et al., 2026; Morgan et al., 2023) provide UPDRS annotations in real-world or home-like settings, but lack ground-truth 3D pose. Marker-based motion capture datasets (Shida et al., 2023; Viglialoro et al., 2025) offer precise 3D ground truth, but their instrumentation burden limits scalability. CARE-PD (Adeli et al., 2025) is an impressive resource of 3D gait data from 363 participants assembled mostly from a collection of smaller previously published datasets which were harmonized and converted into a common SMPL mesh representation (Loper et al., 2015; Mahmood et al., 2019). However, CARE-PD does not include any raw video and derives most of the 3D SMPL pose data it uses for benchmarking from monocular RGB, so it does not support vision-based benchmarking or performance assessments with respect to a reliable 3D ground truth. TULIP (Kim et al., 2024) provides multi-camera video recordings of UPDRS motor tasks with expert ratings, but its small sample size (15 subjects) restricts its use for data-driven modeling (Note that the TULIP dataset is drawn from a separate cohort, and is not a subset of TULIP-BENCH). No existing dataset jointly provides raw video, multi-camera triangulated 3D ground truth, and clinical annotations across structured motor exam tasks (see Table 1).

Pose-based treatment monitoring. While prior work has shown that pose estimation can support UPDRS severity classification (Islam et al., 2023; Lu et al., 2021), a more impactful and underexplored application would be to quantify how a given therapy improves patient motor function to inform clinical decision-making and treatment optimization (Di Lazzaro et al., 2021; Lange et al., 2025). For PD, DBS is one of the most important surgical interventions: candidate selection demands extensive preoperative screening, and individual symptom responses to stimulation remain notoriously difficult to predict, especially for gait and midline symptoms (Rajamani et al., 2024; Fasano et al., 2015). TULIP-BENCH fills this gap by providing triangulated 3D ground truth paired with clinical annotations across controlled DBS ON and OFF states, enabling rigorous quantitative analysis of treatment-induced motor changes.

Table 1: **Comparison with existing PD movement datasets.** N denotes the number of subjects; $N_{3D \cap UPDRS}$ denotes the number of subjects with both 3D ground-truth poses and UPDRS scores. Video indicates availability of video data.

Dataset	N	$N_{3D \cap UPDRS}$	Activities	Video	DBS Treatment	Output
WearGait-PD (Anderson et al., 2026)	185	0	Single	×	×	IMU
REMAP (Morgan et al., 2023)	36	0	Single	×	×	IMU + 3D pose
Shida et al. (Shida et al., 2023)	26	26	Single	×	×	3D pose
Viglialoro et al. (Viglialoro et al., 2025)	24	0	Single	×	×	3D pose
CARE-PD (Adeli et al., 2025)	363	23	Single	×	×	3D pose (SMPL)
TULIP-BENCH (Ours)	66	**66**	**Multiple**	✓	✓	**Video + 3D pose**

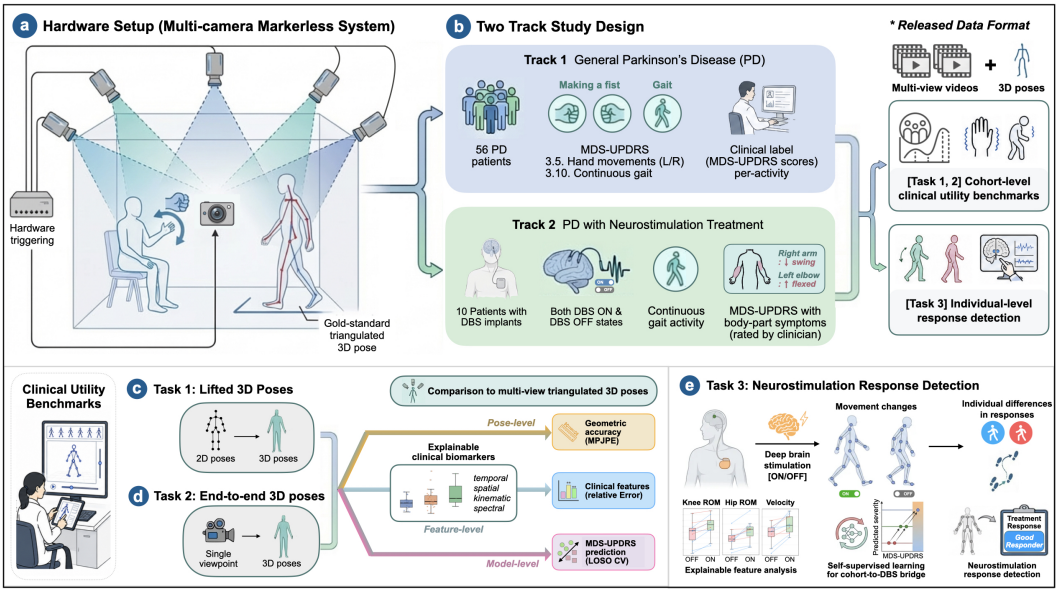

Figure 1: **TULIP-Bench overview.** (a) Six hardware-synchronized cameras capture PD motor tasks. (b) Two-track study design: cohort-level severity analysis ($N = 56$) and individual-level treatment (neurostimulation, DBS) response ($N = 10$). (c, d) Task 1 and 2 measures how 2D-to-3D lifting and end-to-end monocular estimators preserves clinical features relative to triangulated ground truth, across pose and UPDRS prediction levels. (e) Task 3 quantifies whether 3D kinematics detect deep brain stimulation (DBS)-induced motor changes invisible to the ordinal UPDRS-gait score.

3 THE TULIP-BENCH DATASET

TULIP-BENCH is built around two non-overlapping PD patient cohorts. The 56-subject PD cohort, with bilateral hand-closure and gait UPDRS scores for severity analysis and 3D pose validation, and the 10-subject DBS cohort, with paired stimulation OFF/ON gait recordings. In addition to UPDRS ratings, the DBS cohort includes body-part-specific severity annotations made by a movement disorder specialist to support more granular analyses of treatment (DBS) responses. All recordings were IRB-approved. Released data include de-identified multi-view videos, triangulated 3D poses, camera matrices, UPDRS scores, and DBS body-part annotations (see Supplementary Section A.1).

3.1 RECORDING PROTOCOL

All subjects were recorded with the same multi-camera system: six hardware-synchronized Basler cameras (acA1920-155uc, 1920×1200, 80 fps), arranged as one front-facing view approximating a mobile-phone perspective, one top-mounted bird's-eye view, and four elevated corner views. A wired hardware trigger was used for frame-level synchronization, and calibration parameters are included in the dataset. Together these enable recovery of 3D poses via triangulation of 2D pose keypoints without the temporal interpolation required of software-synchronized cameras. Each

subject performed bilateral fist closure (UPDRS 3.5: open and close each hand fully and as quickly as possible) and gait (UPDRS 3.10: walking straight at a self-selected pace) assessment tasks. A movement disorder specialist reviewed each recording to provide UPDRS scores; left and right fist closure received separate scores. By including both fine and gross motor tasks, TULIP-BENCH enables direct cross-scale comparisons that are not possible with existing PD movement datasets.

3.2 TREATMENT (DBS NEUROSTIMULATION) RESPONSE TRIALS

The 10 DBS subjects were enrolled separately and recorded performing gait assessments in both DBS OFF and ON stimulation states. The DBS OFF condition was recorded after a clinically appropriate washout period; the ON condition was reactivated and confirmed by the clinician to have reached stable clinical effect before recording the gait assessment. The same movement disorder specialist assigned UPDRS gait scores for both DBS states and provided structured body-part severity ratings of left and right arm swing, stride length, and elbow flexion. Paired treatment-state recordings are rare because human subjects protocols require that changes to DBS settings must occur under clinical supervision and recording ON and OFF states roughly doubles a patient's time commitment; previous kinematic DBS studies (which have not released their datasets) report cohort sizes that are commensurate with or smaller than ours (Cherif et al., 2025; Mera et al., 2011; Tripathi et al., 2024).

3.3 3D POSE EXTRACTION AND PROCESSING

Ground-truth 3D poses were obtained by triangulating 2D keypoint detections from the six synchronized views, followed by outlier correction and bone-length-constrained optimization; a human annotator inspected and corrected residual tracking failures. Each frame yields 21 hand keypoints (fist) and 33 body keypoints (gait), in millimeter coordinates. We performed a manual consistency check by comparing the reconstructed 3D poses with annotations from an independent human annotator. Additional details on pose processing and validation are provided in Supplementary Section A.1.

4 BENCHMARKS & EXPERIMENTS

4.1 EXPERIMENTAL SETUP

All tasks share a common preprocessing pipeline that converts raw recordings into 3D poses. For Fist activities, consecutive groups of 10 open–close repetitions form sliding-window samples, following the UPDRS instruction. For gait, each stride forms one sample. To isolate pose estimation quality from temporal segmentation accuracy, all pose sources (ground truth, lifting outputs, and end-to-end estimators) share the same segmentation timing derived from the triangulated ground truth.

All evaluations are performed in a root-centered coordinate system (hip for gait; wrist for fist). Pose-level accuracy is reported as MPJPE and PA-MPJPE against the multi-camera triangulated ground truth. Feature-level fidelity is assessed via four clinically established indicators of PD severity (Mancini et al., 2025; Goetz et al., 2008): stride length (StrideLen, spatiotemporal) and arm swing range of motion (ArmROM, kinematic) for gait; middle fingertip speed (TipSpeed, spatiotemporal) and mean knuckle flexion range across fingers and repetitions (FistArc, kinematic) for fist activities. For each feature, we quantify error as the mean absolute percentage error (MAPE): $\frac{1}{S}\sum_{s=1}^{S}\frac{1}{N_s}\sum_{i=1}^{N_s}\frac{|\hat{v}_{si}-v_{si}|}{|v_{si}|}\times 100\%$. Here, S denotes the number of subjects, N_s is the number of sequences (e.g., strides or repetitions) for subject s, and v_{si} and $\hat{v}_{si}$ denote the ground-truth and predicted feature values for the i-th sequence. Sequences where $|v_{si}| < \epsilon$ are excluded to avoid division by zero. Feature extraction details and error calculation methods are provided in Supplementary Section A.2.

4.2 TASK 1: CLINICAL UTILITY OF 2D-TO-3D LIFTING METHODS

Task 1 measures the pose-level geometric accuracy and feature-level clinical fidelity of 2D-to-3D lifting methods given single-camera 2D keypoints as inputs (Supplementary Sections A.1 and A.3). Four representative architectures were benchmarked: SimpleBaseline (Martinez et al., 2017), VideoPose3D (Pavllo et al., 2019), MotionAGFormer (Mehraban et al., 2024), and PoseFormerV2 (Zhao et al., 2023). Models were trained on 46 subjects and evaluated on a held-out test set of 10 subjects.

4.3 Task 2: Clinical Utility of Monocular Pose Estimation

Task 2 measures the clinical utility of monocular pose estimation, evaluated along three levels of increasing clinical relevance: (i) pose-level geometric fidelity (MPJPE), (ii) feature-level clinical fidelity of gait and fist features derived from the estimated poses (Section 4.1), and (iii) downstream UPDRS score prediction accuracy. All pose sources are evaluated on the same single camera view.

We compared three categories of monocular pose sources for gait: (i) **in-domain lifting**—the best lifting model from Task 1, retrained under 5-fold cross-validation; (ii) **off-the-shelf lifting**—the same architecture pretrained on Human3.6M without fine-tuning; (iii) **off-the-shelf end-to-end**—for gait, we evaluated WHAM (Shin et al., 2024), SAM 3D Body (Yang et al., 2026), PARE (Kocabas et al., 2021), 4D-Humans (Goel et al., 2023), MediaPipe Pose 3D (Bazarevsky et al., 2020), Multi-HMR (Baradel et al., 2024), and RTMW3D (Jiang et al., 2024); for the fist task, we evaluated SAM 3D Body, HaMeR (Pavlakos et al., 2024), WiLoR (Potamias et al., 2025), and MediaPipe Hands 3D (Zhang et al., 2020), without task-specific fine-tuning for all methods. For clarity, representative methods are reported in the Table 4 (see Supplementary Table 12 and Supplementary Section A.5 for details).

For the downstream UPDRS evaluation, we first identified the strongest UPDRS classifier based on triangulated 3D poses using leave-one-subject-out (LOSO) cross-validation. Candidate classifiers included five spatio-temporal architectures adapted from published motion encoders (Supplementary Section A.4) and a feature-based baseline using 9 clinically validated fist features and 18 gait features (Supplementary Sections A.2). We then fixed the best-performing classifier used it across all monocular pose sources. We chose this evaluation protocol to isolate the contribution of pose quality from the choice of downstream UPDRS classifier. Two performance bounds serve as useful reference points to contextualize monocular UPDRS classification results: triangulated 3D poses are an upper bound (the ceiling achievable with perfect pose recovery), and raw 2D poses from a single monocular camera are a lower bound (that is, what is achievable without any 3D lifting).

4.4 Task 3: DBS Neurostimulation Response Detection

We designed Task 3 to spur development of novel approaches for detecting and predicting treatment effects from gait 3D poses. This task is especially challenging and important because PD treatments can affect motor function in ways that are not reliably resolved by the coarse UPDRS system.

Task 3 comprises three evaluation components. **(i) DBS OFF vs. ON state detection.** Given the 3D pose sequence for a single gait stride, classify the stride as occurring during DBS OFF vs. ON. We compared four models: logistic regression on 18 hand-crafted features, ridge regression on flattened stride pose, a 1D-temporal CNN, and a spatio-temporal ST-GCN (Yan et al., 2018). **(ii) Representation learning of DBS severity.** The goal is to establish a continuous severity metric that augments coarse UPDRS scores. As a baseline, we employed a self-supervised cross-cohort learning approach that allowed us to leverage our larger 56-subject non-DBS PD dataset. We pre-trained an ST-GCN encoder (the best gait architecture from Task 1) on all unlabeled strides from the 56-subject cohort using masked-coordinate reconstruction, then fit a single linear probe from cohort stride embeddings to cohort UPDRS gait integer scores. We then froze both the ST-GCN encoder and linear probe and applied them to the DBS cohort. Architecture and training details are in Supplementary Section A.7 (iii) **Responder prediction.** As an exploratory benchmark, we assessed whether features only from the DBS OFF state could distinguish clinician-annotated good vs. poor DBS responders.

5 Results & Analysis

5.1 Task 1: Clinical Utility of 2D-to-3D Lifting Methods

Pose accuracy does not predict feature fidelity. Table 2 reveals a substantial gap between pose-level and feature-level performance across the four benchmarked lifting models. On fist activities, MPJPE differences among the top three models were within 2 mm, yet TipSpeed errors diverged sharply: VideoPose3D achieved the lowest MPJPE (16.96 mm) but a 19.11% TipSpeed error, while PoseFormerV2, despite a higher MPJPE (18.69 mm), yielded the lowest TipSpeed error (9.93%). Kinematic errors show another type of disconnect: even when MPJPE values were within a few centimeters of ground truth, kinematic errors remained uniformly elevated (around 20% on fist, 35% on gait), well above the precision typically required for clinical motor assessment (Navarro-

Table 2: **Pose-level and feature-level evaluation of lifting models.** MPJPE and PA-MPJPE are computed against the triangulated 3D poses. Spatiotemporal (ST) and kinematic feature errors are reported as the relative difference (%) between features extracted from estimated and triangulated 3D poses. Spatiotemporal features: TipSpeed (fist) and StrideLen (gait); kinematic features: FistArc (fist) and ArmROM (gait). Best results are shown in **bold**.

Activity	Method	MPJPE (mm)↓	PA-MPJPE (mm)↓	ST Err.↓	Kinematic Err.↓
Fist	SimpleBaseline	21.43	10.74	31.09%	30.94%
	VideoPose3D	**16.96**	**7.53**	19.11%	**19.16%**
	MotionAGFormer	17.22	8.62	17.71%	20.51%
	PoseFormerV2	18.69	8.02	**9.93%**	20.66%
Gait	SimpleBaseline	44.00	32.88	19.29%	42.62%
	VideoPose3D	37.16	26.41	15.21%	38.34%
	MotionAGFormer	**35.24**	**25.80**	**10.01%**	35.01%
	PoseFormerV2	37.22	27.33	13.94%	**34.94%**

Lopez et al., 2022). These results suggest that differences in geometric accuracy do not translate proportionally to clinical utility, and that fine-grained motion features, such as arm swing, remain particularly challenging to recover with existing lifting approaches. For subsequent analyses, we select MotionAGFormer (gait) and VideoPose3D (fist) as lifting architectures based on MPJPE.

5.2 Task 2: Clinical Utility of Monocular Pose Estimation

Table 3: **UPDRS severity prediction** (LOSO, subject-level). Top: pose-based models using raw 3D sequences. Bottom: best feature-based model using 3D clinical features. Sorted by $F1_m$ within each activity. Best results are shown in **bold**. Full feature-based results in Supplementary Section A.2.

Activity	Input	Model	$F1_m$	MAE	κ
Fist	3D Pose	**MotionAGFormer-Lite**	**0.745**	0.494	0.715
	3D Pose	Dual-Stream Transformer	0.724	0.378	0.639
	3D Pose	ST-Transformer	0.689	0.384	0.621
	3D Pose	ST-GCN	0.681	**0.224**	**0.739**
	3D Pose	PoseFormer-v2	0.641	0.450	0.563
	3D Features	Feature-based model	0.483	0.523	0.303
Gait	3D Pose	**ST-GCN**	**0.647**	**0.292**	**0.618**
	3D Pose	PoseFormer-v2	0.608	0.321	0.590
	3D Pose	MotionAGFormer-Lite	0.396	0.577	0.410
	3D Pose	ST-Transformer	0.353	0.512	0.449
	3D Pose	Dual-Stream Transformer	0.276	0.708	0.304
	3D Features	Feature-based model	0.501	0.441	0.276

UPDRS classifier selection. Table 3 compares pose-based models on raw 3D sequences against feature-based models using hand-crafted clinical features, all derived from the triangulated 3D pose, under LOSO cross-validation. Pose-based models substantially outperformed feature-based models on both activities (best $F1_m$: 0.745 vs 0.483 on fist; 0.647 vs 0.501 on gait). For subsequent pose-source comparisons, we selected the best model per activity.

In-domain training is necessary to recover pathological movement patterns. Table 4 compares monocular pose sources across pose, feature, and UPDRS classification metrics. Off-the-shelf methods consistently failed on TULIP-Bench: WHAM and MAGF_H36M both exceeded 189 mm MPJPE on gait with kinematic feature errors above 50%, a failure mode especially consequential for WHAM, on which CARE-PD (Adeli et al., 2025) relies for deriving the 3D poses treated as ground-truth for benchmarking. SAM 3D Body partially closed the pose-level gap by leveraging foundation-model scale (102.81 mm MPJPE), yet its kinematic feature error was the highest among all pretrained methods (105.42%). An architecture-controlled comparison isolated the contribution of in-domain training: training the same MotionAGFormer architecture on TULIP-Bench instead of

Table 4: **Comparison of monocular pose estimation methods.** MAGF_H36M is a MotionAGFormer checkpoint trained on Human3.6M and applied to TULIP-BENCH without fine-tuning; MAGF_TULIP is the same architecture trained on TULIP-BENCH. Best results are shown in **bold**. Dashes indicate metrics not applicable. This table reports a representative subset of the evaluated methods; the complete comparison is provided in the Supplementary Section A.5.

Activity	Pose Input	MPJPE (mm)↓	PA-MPJPE (mm)↓	ST Err.↓	Kinematic Err.↓	$F1_{macro}$↑
Fist	Raw 2D	–	–	–	–	0.723
	SAM3D	36.11	13.93	87.83%	36.36%	0.743
	VideoPose3D	**22.46**	**8.91**	**21.54%**	**22.04%**	**0.775**
	Multi-view 3D	–	–	–	–	0.745
Gait	Raw 2D	–	–	–	–	0.223
	WHAM	197.54	140.88	80.24%	52.81%	0.366
	SAM3D	102.81	63.39	20.89%	105.42%	0.568
	MAGF_H36M	189.94	140.65	50.95%	78.98%	0.457
	MAGF_TULIP	**35.76**	**25.48**	**9.93%**	**33.57%**	0.614
	Multi-view 3D	–	–	–	–	**0.647**

Human3.6M reduced gait MPJPE from 189.94 mm to 35.76 mm and kinematic error from 78.98% to 33.57%, and it lifted F1 from 0.457 to 0.614, approaching the multi-view ceiling (0.647).

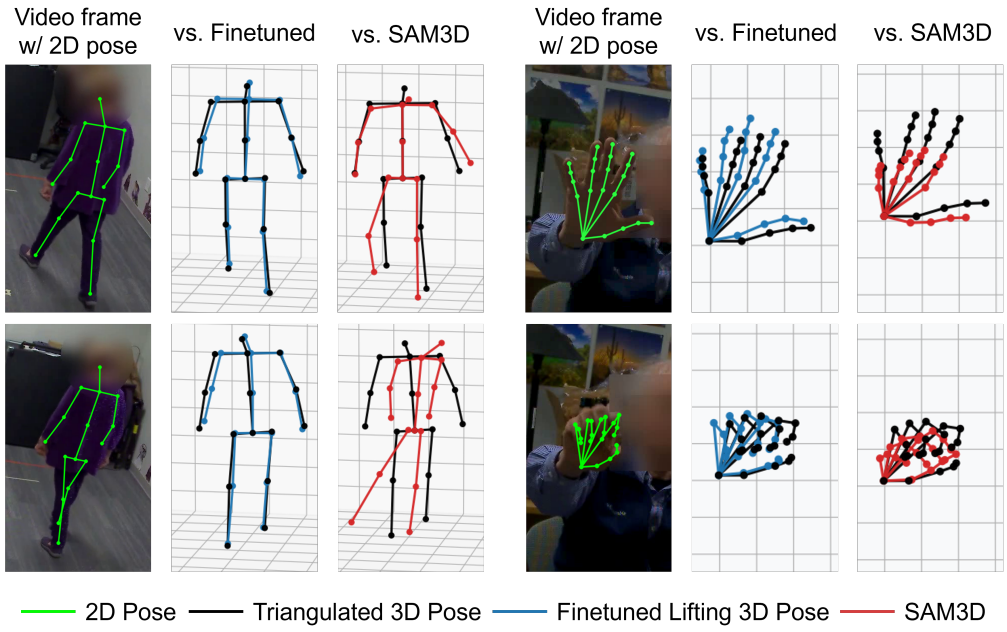

Figure 2: **Frame-level visualization of predicted 3D poses.** Representative frames from gait (left) and fist (right) sequences compare predicted 3D poses against the triangulated 3D poses (black) at two time points. For each frame, we show the input video with overlaid 2D pose (green), the finetuned model prediction (blue) versus triangulated 3D poses, and the SAM 3D Body prediction (red) versus triangulated 3D poses.

Why pose-level accuracy fails to predict feature-level fidelity. Figure 2 illustrates this disconnect using representative frames from gait and fist sequences. Finetuned models closely tracked the triangulated 3D poses with consistent geometry across frames. In contrast, SAM 3D exhibited temporally inconsistent errors: while some frames showed only minor joint offsets, others contain substantial structural distortions. Such errors may appear modest when considered frame-by-frame, and they are further obscured by aggregate metrics such as MPJPE. However, clinical features are computed over entire strides or repetitions and depend on temporally coherent motion patterns. As a result, even infrequent but large deviations can disproportionately affect feature computation, particularly for range-based and phase-dependent measures. This effect is reflected quantitatively:

SAM 3D Body showed 105.42% kinematic error on gait and 36.36% on fist, despite maintaining a competitive MPJPE. These results indicate that pose-level accuracy alone does not capture the temporal consistency required for reliable clinical feature extraction, highlighting the need for benchmarks that explicitly evaluate clinical feature fidelity beyond pose-level geometric accuracy.

2D pose suffices for fist classification but not for feature extraction. On fist, raw 2D pose alone achieved an F1 of 0.723, approaching the multi-view upper bound (0.745), and pretrained SAM 3D Body (0.743) was nearly indistinguishable from the in-domain VideoPose3D model (0.775) on UPDRS classification. This convergence reflects the predominantly in-plane nature of fist closure. Combined with UPDRS' coarse 5-level granularity, even noisy poses contain enough signal to associate with the correct severity level. Yet at the feature level, SAM 3D Body's spatiotemporal error reached 87.83%, compared to 21.54% for the in-domain VideoPose3D, a gap entirely invisible from UPDRS classification metrics alone. For clinical applications that depend on interpretable kinematic measurements or more sensitive detection of kinematic changes, this feature-level gap is consequential even though downstream UPDRS classification appears unaffected.

5.3 Task 3: DBS Neurostimulation Response Detection

a.

Feature	d	95% CI	p	Interpretation
Knee ROM	−1.61	[−3.16, −1.08]	.002**	↑ knee flexion range
Stride length	−1.52	[−2.91, −1.00]	.004**	↑ stride amplitude
Step length	−1.50	[−2.91, −1.00]	.004**	↑ step amplitude
Velocity	−1.49	[−2.86, −0.95]	.002**	↑ walking speed
Hip ROM	−1.34	[−2.23, −0.97]	.002**	↑ hip range
Leg-leg ROM	−1.25	[−2.24, −0.85]	.002**	↑ leg coordination
Step length asym.	+0.70	[+0.29, +1.31]	.049*	↓ asymmetry
Shoulder ROM	−0.62	[−1.18, −0.33]	.037*	↑ arm swing

b.

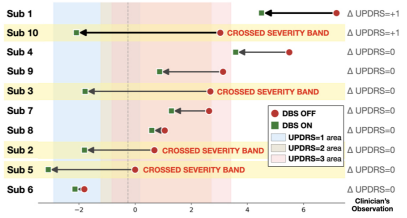

Figure 3: **Continuous 3D pose detects neurostimulation-induced change invisible to UPDRS-gait**. (a) Treatment sensitivity of gait features (DBS OFF vs. ON). Cohen's d: paired effect size (negative = feature increases OFF→ON). Features with $|d| \geq 0.5$ and $p < 0.05$; full table in Supplementary Section A.7. (b) OFF→ON trajectories on a 56 cohort-level feature-space severity axis. X-axis is the severity axis, from PC1 on 6 gait features, 85% var, left→right=more impaired.

Continuous 3D features change with DBS where UPDRS-gait does not. To test for the presence of treatment-induced kinematic changes not captured by UPDRS gait scores, we first extracted 18 explainable gait features – the same set validated as severity-associated in Task 1 – from the 3D poses of subjects within the DBS cohort (Supplementary Section A.2). We then assessed each feature's within-subject treatment sensitivity by computing Cohen's d effect sizes between features in the DBS OFF vs. ON state. Additionally, we computed the correlation between each kinematic feature and the clinician body-part severity annotation (e.g., stride length vs. stride severity), as a per-feature sanity check that automatic features capture what clinicians visually assess (Supplementary Section A.7).

We found that 10 of 18 gait features show medium-to-large treatment effects ($|d| \geq 0.5$), with six demonstrating $|d| > 1.0$ (Figure 3a; full per-feature table in Supplementary Section A.7). The largest effects occurred on knee ROM ($d = -1.61$, $+18\%$ mean DBS OFF→ON change), stride length ($d = -1.52$, $+33\%$), and velocity ($d = -1.49$, $+33\%$). In contrast, temporal features (cadence, stride time) remained stable, consistent with DBS preferentially affecting spatial gait amplitude rather than temporal rhythm (Faist et al., 2001). UPDRS gait scores remained unchanged in 8 of 10 subjects despite these effects (Figure 3b). Comparisons between each kinematic feature and clinician body-part symptom annotations showed high correlation (stride severity vs. stride length $\rho = -0.91$; arm-swing severity vs. shoulder ROM $\rho = -0.83$), further supporting the clinical relevance of our findings (see Supplementary Section A.7 for all per-feature correlation statistics). Together, these results illustrate a dissociation between continuous kinematics and the coarse ordinal UPDRS-gait scale: 3D pose detects DBS-induced kinematic change in 8 of 10 subjects whose UPDRS-gait score does not move, with effect sizes comparable to those of levodopa in PD (Faist et al., 2001; O'Sullivan et al., 1998).

Deep pose models detect DBS OFF vs. DBS ON state from a single stride. Table 5 compares four families of methods DBS state detection from a single shape-normalized stride. 1D-Temporal CNN achieved the best AUC-ROC (0.83), followed by ST-GCN (0.75) and LogReg (0.74). That DBS OFF/ON state is detectable from poses, even in subjects without DBS-induced changes in UPDRS

Method	Median LOSO AUC-ROC	IQR
1D-Temporal CNN	$\mathbf{0.83}$	$[0.66, 0.92]$
ST-GCN	0.75	$[0.58, 0.84]$
Feature LogReg	0.74	$[0.51, 0.91]$
Flat-pose ridge	0.67	$[0.60, 0.84]$

Table 5: **OFF vs. ON state detection from a single stride.** Median LOSO AUC-ROC across 10 DBS subjects ($N = 20$ paired observations), with interquartile range (IQR; 25th–75th percentile across subjects) reflecting heterogeneous treatment-response detectability.

gait score, supports our major claim: continuous 3D pose carries treatment and outcome-associated signals that are not resolved by conventional ordinal clinical scales.

A learned general PD cohort severity scale transfers zero-shot to DBS. Self-supervised learning on the ST-GCN encoder on the 56-subject PD cohort (no labels used) produced an embedding in which a linear probe recovered clinician UPDRS-gait at Spearman $\rho = 0.68$, confirming the embedding encodes severity structure within the cohort. Applying this same probe zero-shot to the patient-disjoint DBS cohort (no DBS data seen at any training stage), we found that predicted continuous severity correlated with clinician UPDRS gait scores at $\rho = 0.52$. Seven of 10 DBS subjects showed predicted severity decreasing OFF→ON (Figure 3b); critically, 6 of 8 subjects whose UPDRS gait scores did not change exhibited decreases on the learned severity scale. Exploratory responder classification reaches 60% accuracy, marginally above chance (Supplementary Section A.7).

5.4 Limitations and Broader Impact

While TULIP-Bench is the largest publicly available dataset of its kind, its relatively small cohort size is a limitation. We mitigated this to an extent through within-subject paired design and cross-cohort transfer. During recruitment, we also excluded patients with severe cognitive impairment or high fall risk for safety considerations. This means our dataset lacks coverage of the most extreme end of the PD severity spectrum. Adding additional clinical sites and subject diversity in the future will promote increased generalizability across recording environments and PD severity levels.

TULIP-Bench is a foundation: extending it with longitudinal follow-up, broader sampling, additional clinical variables and medical history, and data from other sensor types (e.g., wearables) would open the door to solutions for DBS programming optimization (Mera et al., 2011; Pulliam et al., 2015), patient-specific prognosis and treatment decisions, and improved clinical trial endpoints. Realizing a future of advanced prognostic models will require accurate movement quantification benchmarks like those TULIP-Bench provides.

6 Conclusion

TULIP-Bench establishes a multi-view video dataset of PD motor exams with accurate 3D pose ground truth, enabling critical evaluations of pose estimation approaches along new axes of clinical utility. Our experiments reveal three key findings. First, low joint-level error does not imply clinical fidelity; even when MPJPE values were close to the ground truth, errors in clinically meaningful features remained substantially elevated. Second, the importance of 3D quantification is activity-dependent: 3D measurements are essential for gait UPDRS classification, but raw 2D measurements appear sufficient for fist UPDRS classification. However, even for fist, 2D measurements result in poor feature-level fidelity. Third, 3D pose measurements enable more sensitive identification of stimulation-induced changes than UPDRS, highlighting an area where vision-based movement quantification could be particularly valuable as a complement to conventional clinical scoring systems. TULIP-Bench aims to pave the way to precise movement disorder analysis and modeling, enabling deployable systems that will broaden access to objective PD care and improved patient outcomes.

AI Assistance Disclosure

We used generative AI tools to help edit the writing of this paper for grammar and readability. We did not use them for the research itself. This includes the study design, the methods and their implementation, the experiments, the data, the analysis, and the results. The authors reviewed all edits and are responsible for the final content of this paper.

References

Vida Adeli, Ivan Klabucar, Javad Rajabi, et al. CARE-PD: A multi-site anonymized clinical dataset for Parkinson's disease gait assessment. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2025.

Anthony J Anderson, David Eguren, Michael A Gonzalez, et al. Weargait-pd: An open-access wearables dataset for gait in parkinson's disease and age-matched controls. *Scientific Data*, 2026.

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

Sivakumar Balasubramanian, Alejandro Melendez-Calderon, Agnes Roby-Brami, and Etienne Burdet. On the analysis of movement smoothness. *Journal of neuroengineering and rehabilitation*, 12(1): 112, 2015.

Fabien Baradel, Matthieu Armando, Salma Galaaoui, Romain Brégier, Philippe Weinzaepfel, Grégory Rogez, and Thomas Lucas. Multi-hmr: Multi-person whole-body human mesh recovery in a single shot. In *European Conference on Computer Vision*, pp. 202–218. Springer, 2024.

Valentin Bazarevsky, Ivan Grishchenko, Karthik Raveendran, Tyler Zhu, Fan Zhang, and Matthias Grundmann. Blazepose: On-device real-time body pose tracking. *arXiv preprint arXiv:2006.10204*, 2020.

Bastiaan R Bloem, Michael S Okun, and Christine Klein. Parkinson's disease. *The Lancet*, 397 (10291):2284–2303, 2021.

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

Saoussen Cherif, Nicolas Tempier, Mathieu Yeche, et al. Directional subthalamic deep brain stimulation better improves gait and balance disorders in parkinson's disease patients: a randomized controlled study. *Annals of Neurology*, 97(1):149–162, 2025.

Lazzaro Di Biase, Susanna Summa, Jacopo Tosi, et al. Quantitative analysis of bradykinesia and rigidity in parkinson's disease. *Frontiers in neurology*, 9:121, 2018.

Giulia Di Lazzaro, Mariachiara Ricci, Giovanni Saggio, Giovanni Costantini, Tommaso Schirinzi, Mohammad Alwardat, Luca Pietrosanti, Martina Patera, Simona Scalise, Franco Giannini, et al. Technology-based therapy-response and prognostic biomarkers in a prospective study of a de novo parkinson's disease cohort. *npj Parkinson's Disease*, 7(1):82, 2021.

Ruth Djaldetti, Idan Ziv, and Eldad Melamed. The mystery of motor asymmetry in Parkinson's disease. *The Lancet Neurology*, 5(9):796–802, 2006.

Alberto J Espay, Joe P Giuffrida, Robert Chen, et al. Differential response of speed, amplitude, and rhythm to dopaminergic medications in parkinson's disease. *Movement Disorders*, 26(14): 2504–2508, 2011.

Stanley Fahn, Richard L. Elton, and Members of the UPDRS Development Committee. The Unified Parkinson's Disease Rating Scale. In Stanley Fahn, C. David Marsden, Donald B. Calne, and Morris Goldstein (eds.), *Recent Developments in Parkinson's Disease*, volume 2, pp. 153–163, 293–304. Macmillan Healthcare Information, Florham Park, NJ, 1987.

M Faist, J Xie, D Kurz, W Berger, C Maurer, P Pollak, and CH Lücking. Effect of bilateral subthalamic nucleus stimulation on gait in parkinson's disease. *Brain*, 124(8):1590–1600, 2001.

Alfonso Fasano, Camila C Aquino, Joachim K Krauss, et al. Axial disability and deep brain stimulation in patients with parkinson disease. *Nature Reviews Neurology*, 11(2):98–110, 2015.

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

Christopher G Goetz, Barbara C Tilley, Stephanie R Shaftman, et al. Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): Scale presentation and clinimetric testing results. *Movement Disorders*, 23(15):2129–2170, 2008.

Dustin A Heldman, Joseph P Giuffrida, Robert Chen, et al. The modified bradykinesia rating scale for parkinson's disease: reliability and comparison with kinematic measures. *Movement Disorders*, 26(10):1859–1863, 2011.

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

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. Using ai to measure parkinson's disease severity at home. *NPJ digital medicine*, 6(1):156, 2023.

Tao Jiang, Xinchen Xie, and Yining Li. Rtmw: Real-time multi-person 2d and 3d whole-body pose estimation. *arXiv preprint arXiv:2407.08634*, 2024.

Hanbyul Joo, Hao Liu, Lei Tan, et al. Panoptic studio: A massively multiview system for social motion capture. In *Proceedings of the IEEE international conference on computer vision*, pp. 3334–3342, 2015.

Pierre Karashchuk, Katie L Rupp, Evyn S Dickinson, Sarah Walling-Bell, Elischa Sanders, Eiman Azim, Bingni W Brunton, and John C Tuthill. Anipose: A toolkit for robust markerless 3d pose estimation. *Cell reports*, 36(13), 2021.

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

Muhammed Kocabas, Chun-Hao P Huang, Otmar Hilliges, and Michael J Black. Pare: Part attention regressor for 3d human body estimation. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11107–11117. IEEE, 2021.

Florian Lange, Diego L Guarin, Esther Ademola, Dalia Mahdy, Gabriela Acevedo, Thorsten Odorfer, Joshua K Wong, Jens Volkmann, Robert Peach, and Martin Reich. Computer vision uncovers three fundamental dimensions of levodopa-responsive motor improvement in parkinson's disease. *npj Parkinson's Disease*, 11(1):140, 2025.

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. Smpl: A skinned multi-person linear model. *ACM Transactions on Graphics*, 34(6):1–16, 2015.

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. Quantifying parkinson's disease motor severity under uncertainty using mds-updrs videos. *Medical image analysis*, 73:102179, 2021.

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

Martina Mancini, Mitra Afshari, Quincy Almeida, Sommer Amundsen-Huffmaster, Katherine Balfany, Richard Camicioli, Cory Christiansen, Marian L Dale, Leland E Dibble, Gammon M Earhart, et al. Digital gait biomarkers in parkinson's disease: susceptibility/risk, progression, response to exercise, and prognosis. *npj Parkinson's Disease*, 11(1):51, 2025.

Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3d human pose estimation. In *Proceedings of the IEEE international conference on computer vision (ICCV)*, 2017.

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

Thomas Mera, Jerrold L Vitek, Jay L Alberts, and Joseph P Giuffrida. Kinematic optimization of deep brain stimulation across multiple motor symptoms in parkinson's disease. *Journal of neuroscience methods*, 198(2):280–286, 2011.

Anat Mirelman, Paolo Bonato, Richard Camicioli, et al. Gait impairments in Parkinson's disease. *The Lancet Neurology*, 18(7):697–708, 2019.

Catherine Morgan, Emma L Tonkin, Alessandro Masullo, et al. A multimodal dataset of real world mobility activities in parkinson's disease. *Scientific data*, 10(1):918, 2023.

Victor Navarro-Lopez, Diego Fernandez-Vazquez, Francisco Molina-Rueda, Alicia Cuesta-Gomez, Pablo Garcia-Prados, Manuel del Valle-Gratacos, and Maria Carratala-Tejada. Arm-swing kinematics in parkinson's disease: a systematic review and meta-analysis. *Gait & Posture*, 98:85–95, 2022.

John D O'Sullivan, Catherine M Said, Louise C Dillon, Marion Hoffman, and Andrew J Hughes. Gait analysis in patients with parkinson's disease and motor fluctuations: influence of levodopa and comparison with other measures of motor function. *Movement disorders: official journal of the Movement Disorder Society*, 13(6):900–906, 1998.

Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3d with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9826–9836, June 2024.

Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3D human pose estimation in video with temporal convolutions and semi-supervised training. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Rolandos Alexandros Potamias, Jinglei Zhang, Jiankang Deng, and Stefanos Zafeiriou. Wilor: End-to-end 3d hand localization and reconstruction in-the-wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12242–12254, June 2025.

Christopher L Pulliam, Dustin A Heldman, Tseganesh H Orcutt, Thomas O Mera, Joseph P Giuffrida, and Jerrold L Vitek. Motion sensor strategies for automated optimization of deep brain stimulation in parkinson's disease. *Parkinsonism & related disorders*, 21(4):378–382, 2015.

Nanditha Rajamani, Helen Friedrich, Konstantin Butenko, Till Dembek, Florian Lange, Pavel Navrátil, Patricia Zvarova, Barbara Hollunder, Rob MA de Bie, Vincent JJ Odekerken, et al. Deep brain stimulation of symptom-specific networks in parkinson's disease. *Nature communications*, 15(1): 4662, 2024.

Arash Salarian, Heike Russmann, Christian Wider, et al. Quantification of tremor and bradykinesia in parkinson's disease using a novel ambulatory monitoring system. *IEEE Transactions on biomedical engineering*, 54(2):313–322, 2007.

Thiago Kenzo Fujioka Shida, Thaisy Moraes Costa, et al. A public data set of walking full-body kinematics and kinetics in individuals with parkinson's disease. *Frontiers in Neuroscience*, 17: 992585, 2023.

Soyong Shin, Juyong Kim, Eni Halilaj, and Michael J Black. Wham: Reconstructing world-grounded humans with accurate 3d motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2070–2080, 2024.

Richa Tripathi, J Lucas McKay, Stewart A Factor, et al. Impact of deep brain stimulation on gait in parkinson disease: A kinematic study. *Gait & posture*, 108:151–156, 2024.

Rosanna M Viglialoro, Antonia Centrone, Lorenza Mattei, et al. Dataset on gait analysis of parkinsonian subjects: Effect of nordic walking. *Scientific Data*, 2025.

Jiong Wang, Fengyu Yang, Bingliang Li, et al. Freeman: Towards benchmarking 3d human pose estimation under real-world conditions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21978–21988, 2024.

Stefan Williams, David Wong, Jane E Alty, and Samuel D Relton. Parkinsonian hand or clinician's eye? finger tap bradykinesia interrater reliability for 21 movement disorder experts. *Journal of Parkinson's Disease*, 13(4):525–536, 2023.

Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.

Xitong Yang, Devansh Kukreja, Don Pinkus, et al. SAM 3D body: Robust full-body human mesh recovery. *arXiv preprint arXiv:2602.15989*, 2026.

JA Zeni Jr and et al. Two simple methods for determining gait events during treadmill and overground walking using kinematic data. *Gait & posture*, 27(4):710–714, 2008.

Fan Zhang, Valentin Bazarevsky, Andrey Vakunov, Andrei Tkachenka, George Sung, Chuo-Ling Chang, and Matthias Grundmann. Mediapipe hands: On-device real-time hand tracking. *arXiv preprint arXiv:2006.10214*, 2020.

Qitao Zhao, Ce Zheng, Mengyuan Liu, Pichao Wang, and Chen Chen. PoseFormerV2: Exploring frequency domain for efficient and robust 3D human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Wentao Zhu, Xiaoxuan Ma, Zhaoyang Liu, et al. MotionBERT: A unified perspective on learning human motion representations. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2023.

A Appendix

A.1 Dataset Details

A.1.1 Camera Configuration and Keypoint Definitions

The six Basler cameras (acA1920-150uc) are positioned as follows: Cameras 1,2,3, and 6 (four elevated corners at $\sim$2.0 m, angled downward for overlapping full-body coverage), Camera 4 (ceiling-mounted bird's-eye), and Camera 5 (front-facing, $\sim$1.2 m height, $\sim$2.5 m from subject, approximating a mobile phone viewpoint). All cameras share a wired hardware trigger with sub-millisecond synchronization latency. Calibration uses a standard checkerboard pattern with OpenCV's multi-camera pipeline; intrinsic and extrinsic parameters are released.

Hand keypoints (21): MediaPipe Hand convention. 4 joints per finger (MCP, PIP, DIP, tip) plus wrist, for index through pinky and thumb (CMC, MCP, IP, tip).

Body keypoints (33): We release the 33-keypoint skeleton from MediaPipe. For downstream analysis, we map it to the 17-joint Human3.6M (H36M) format for fair comparison. Of MediaPipe's 33 landmarks, 13 joints are mapped directly via one-to-one anatomical correspondence: bilateral shoulders, elbows, wrists, hips, knees, and ankles, along with the nose. Four virtual joints are geometrically derived: the pelvis and thorax as the midpoints of the hip and shoulder pairs, respectively; the spine as the midpoint between pelvis and thorax; and the head as the average of the two ear landmarks.

A.1.2 3D Pose Post-Processing and Data Segmentation

Post-Processing Pipeline Raw triangulated 3D poses undergo four stages: (1) outlier correction (3σ temporal + cross-view consistency), (2) bone-length polishing (15% deviation, iterative projection to anatomical ratios), (3) NaN interpolation (cubic spline, gaps $>$10 frames flagged), (4) re-canonicalization (root-centered, forward-facing).

Hand Closure Segmentation For Fist activities, the first 3 seconds (hand-raising phase) are discarded and the subsequent 27 seconds form the analysis window. A closure score $c(t) \in [0, 1]$ quantifies fist formation at each frame from the mean fingertip-to-wrist distance normalized by palm span (index MCP to pinky MCP):

$$c(t) = 1 - \frac{d_{\text{fingertip-wrist}}(t) - d_{\min}}{d_{\max} - d_{\min}} \tag{1}$$

where $d_{\min}, d_{\max}$ are the 5th/95th percentiles, smoothed with a 100 ms moving average. The first 3 seconds are discarded (hand-raising phase); the subsequent 27 seconds (2,160 frames at 80 fps) form the analysis window.

Individual open-close repetitions are detected as peaks in $c(t)$ using a multi-tier adaptive algorithm with four sensitivity levels (default, sensitive, ultra, maximum). Each tier specifies thresholds for peak prominence ($\geq$0.10 to $\geq$0.02), minimum peak duration (150 to 80 ms), and minimum inter-peak gap (120 to 80 ms). If the default tier detects fewer than 10 repetitions, the algorithm escalates to the next tier; a velocity-based fallback (peaks in dc/dt) is attempted at each tier before escalation. Consecutive groups of 10 repetitions form sliding-window samples (stride=1) following the UPDRS instruction to perform 10 repetitions.

Gait Segmentation Walking direction is estimated per stride via linear regression on ankle trajectories in the x-y plane, and each stride is rotated to align forward motion with the global x-axis. Following Zeni et al. (Zeni Jr & et al, 2008), gait events are detected from projected distances along the walking direction: heel strike at the maximum projected heel position and toe-off at the minimum projected toe position. Features are computed per stride and averaged across strides per subject.

Post processing Raw triangulated 3D poses undergo four sequential stages:
Stage 1: Outlier correction. Per-joint per-frame positions are compared against local temporal neighbors (5-frame window) and triangulation residuals across views. Positions exceeding 3 standard deviations from the local temporal median are replaced by interpolated values.

Stage 2: Bone-length polishing. Per-frame bone lengths are compared against the subject-specific median. Frames with deviations >15% are iteratively adjusted to satisfy reference bone-length ratios, enforcing anatomical plausibility without a rigid skeleton model.

Stage 3: NaN interpolation. Missing keypoints (failed 2D detection in all views) are filled using cubic spline interpolation. Gaps >10 frames (125 ms) are flagged but not interpolated.

Stage 4: Re-canonicalization. Each sequence is re-centered on the root joint (wrist for hand, mid-hip for body) at frame 1 and re-oriented to face the positive x-axis, removing absolute position.

Pose Quality Validation To assess the geometric consistency of the reconstructed 3D poses, we conducted a manual validation on a subset of 20 subjects. For each subject, 10 frames were annotated by an independent human annotator for two activities: gait and right-hand fist making. Annotators labeled keypoints directly on 2D images across all camera views (resolution: 1920×1200 pixels); these 2D annotations were then triangulated into 3D using the same multi-camera calibration pipeline.

Validation proceeded in two stages. First, the 2D annotations were compared against the reprojections of the reconstructed 3D poses onto the image plane, yielding a mean per-joint pixel error of 16.11 px for gait and 7.60 px for fist—corresponding to only 0.84% and 0.40% of the image width, respectively. These values fall below the 1% threshold recommended as indicative of acceptable annotation quality in multi-view pose datasets (Wang et al., 2024), and are also within the range reported for multi-view markerless systems, where joint position errors below 18 px in 90% of frames are considered representative of expected performance (Karashchuk et al., 2021). Second, the reconstructed 3D poses from manual annotations were compared against the 3D poses using per-joint Euclidean distance. The mean 3D discrepancy was 55.07 mm for gait and 21.67 mm for fist, which are within the range typically reported for multi-view markerless pose estimation systems (approximately 35–65 mm) (Joo et al., 2015; Avogaro et al., 2023; Kim et al., 2024; Carrasco-Plaza & Cerda, 2022). Together, these 2D and 3D metrics confirm that the reconstructed poses are geometrically accurate and suitable for downstream kinematic analysis.

A.1.3 Dataset Statistics

The cohort comprises 56 PD patients with three activity recordings each (left fist, right fist, gait), each with an independent UPDRS score. Score distributions are concentrated in UPDRS 2–3 (∼85% of cohort), reflecting the moderate-to-severe profile typical of a movement disorders center (Supplementary Table 6). UPDRS score 4 is not represented.

Table 6: **UPDRS score distribution by activity.** Cohort: 56 PD patients with three independent activity recordings each (Fist L, Fist R, Gait), one UPDRS score per (subject, activity). DBS treatment cohort: 10 patients (separate from cohort) with paired OFF/ON gait recordings and per-state UPDRS scores.

Activity	N	UPDRS 0	UPDRS 1	UPDRS 2	UPDRS 3
Fist L	56	2	5	27	22
Fist R	56	0	7	18	31
Gait	56	0	21	32	3
Gait DBS-OFF	10	0	1	7	2
Gait DBS-ON	10	0	2	7	1

A.1.4 De-Identification and Data Access

Face regions in all video frames are replaced by Gaussian blurring. Raw video is retained under restricted access. **Released data:** De-identified multi-view video (MP4, original resolution/fps), triangulated 3D poses (millimeters), camera matrices, MDS-UPDRS scores (CSV), and activity/DBS metadata. **Access:** Request form for academic research. Data use agreement: non-commercial use, no re-identification, no redistribution, citation required.

A.2 Feature Details

A.2.1 Hand Closure Features (MDS-UPDRS Item 3.5)

The MDS-UPDRS 3.5 scoring anchors explicitly evaluate amplitude, speed, rhythm (including hesitations and arrests), and decrement (Goetz et al., 2008). We design 9 features spanning 5 categories that directly operationalize these clinical dimensions (Supplementary Table 7). TipSpeed is the feature speed_fingertip_mean, and FistArc is the feature amp_mcp_range_rad_mean.

Table 7: **Hand closure features** (9 features, 5 categories).

Feature	Category	Unit	Definition
`speed_fingertip_mean`	Speed	mm/s	Mean 3D speed of middle fingertip across all frames
`amp_mean_closure`	Amplitude	[0,1]	Mean peak-to-trough range of $c(t)$ across repetitions
`amp_max_open_ftw`	Amplitude	ratio	Mean maximum fingertip-wrist distance at "open" phase, normalized by palm span
`amp_mcp_range_rad_mean`	Amplitude	rad	Mean range of MCP joint flexion angle across fingers and repetitions
`smooth_mean_jerk`	Smoothness	mm/s^3	Mean $\|d^3\mathbf{x}/dt^3\|$ of fingertip trajectory
`speed_closure_vel_open`	Speed	ratio	Mean closing velocity / mean opening velocity of $c(t)$
`amp_cv_closure`	Amplitude	–	Coefficient of variation of per-rep closure amplitude
`spec_entropy`	Spectral	nats	Shannon entropy of normalized power spectral density of $c(t)$: $H = -\sum_k P_k \log P_k$
`rhythm_hesitation_count`	Rhythm	count	Number of repetitions with inter-rep interval $> 2\times$ median period

Clinical grounding. Speed features operationalize the "slowing" criterion across all UPDRS 3.5 score levels. Fingertip speed is the primary bradykinesia marker (Espay et al., 2011; Di Biase et al., 2018). Amplitude features capture "amplitude decrement," explicitly named for scores ≥ 1. MCP angle range provides a joint-level complement to the distance-based closure score. Heldman et al. (Heldman et al., 2011) demonstrated that clinicians weight amplitude most heavily when assigning UPDRS scores. Smoothness (mean jerk) quantifies movement fragmentation; smoother movements indicate better motor control (Balasubramanian et al., 2015). Spectral entropy captures the complexity of the movement's frequency structure; healthy repetitive movements show low entropy (concentrated spectral peak), while parkinsonian movements show diffuse, high-entropy spectra (Salarian et al., 2007). Hesitation count directly maps to the MDS-UPDRS anchor language: scores ≥ 2 require "3 to 5 interruptions" and scores ≥ 3 require "more than 5 interruptions or at least one longer arrest."

Table 8: **Full feature-based results: Fist left (LOSO, subject-level).**

Model	ρ	MAE	Acc$_{\pm 1}$	κ	F1$_m$
RF-Cls	0.637	0.450	0.964	0.353	0.387
Ensemble	0.586	0.474	0.929	0.337	0.348
XGB-Cls	0.584	0.475	0.946	0.304	0.338
XGBoost	0.577	0.486	0.946	0.372	0.416
GBR-Cls	0.562	0.498	0.929	0.306	0.338
Lasso	0.517	0.721	0.964	0.008	0.235
ElasticNet	0.496	0.722	0.964	0.037	0.250
GBR	0.491	0.530	0.929	0.220	0.343
OrdinalReg	0.460	0.472	0.911	0.264	0.313
Ridge	0.452	0.744	0.946	0.132	0.286
RF	0.444	0.559	0.911	0.260	0.362
SVR	0.297	0.711	0.875	0.217	0.270
MLP	0.179	0.734	0.857	0.077	0.196

Table 9: **Full feature-based results: Fist right.**

Model	ρ	MAE	$\text{Acc}_{\pm 1}$	κ	F1_m
ElasticNet	0.516	0.556	0.982	0.226	0.457
Lasso	0.488	0.558	0.982	0.250	0.476
Ensemble	0.483	0.541	1.000	0.192	0.422
OrdinalReg	0.460	0.472	0.911	0.264	0.313
Ridge	0.453	0.583	0.982	0.218	0.368
RF-Cls	0.374	0.481	0.929	0.266	0.548
XGBoost	0.275	0.570	0.964	0.186	0.450
GBR	0.273	0.570	0.964	0.210	0.498
RF	0.266	0.560	0.946	0.262	0.559
GBR-Cls	0.200	0.610	0.893	0.142	0.472
SVR	0.179	0.642	0.929	0.108	0.256
MLP	0.168	0.696	0.911	0.055	0.273

Comprehensive Model Comparison for Feature-Based UPDRS Prediction Tables 8–9 report results for left and right fist activities. RF-Cls ($\rho = 0.637$) and XGBoost (F1=0.416) led on Fist Left, while Fist Right was harder across all models (best $\rho = 0.516$, ElasticNet). Tree-based regressors degraded most between sides, while linear models were more stable. Ensembles provided consistent performance on both sides.

Asymmetry Analysis Using $|f_L - f_R|$ feature differences to predict max(UPDRS_L, UPDRS_R) via LOSO, the best model (Ridge) achieves $\rho = 0.293$. Bilateral kinematic asymmetry, while clinically relevant (Djaldetti et al., 2006), is not a strong severity predictor in this cohort. This analysis is uniquely enabled by TULIP-Bench's bilateral recording design.

Paired Wilcoxon Analysis Within-subject Wilcoxon signed-rank tests comparing the more-affected (higher UPDRS) side to the less-affected side for each hand closure feature reveal no features reaching $p < 0.05$ (most lenient: spectral entropy $p = 0.096$). The consistent sign direction (more-affected side: lower amplitude, higher entropy) supports clinical face validity despite non-significance.

A.2.2 Gait Features (MDS-UPDRS Item 3.10)

The MDS-UPDRS 3.10 scoring evaluates stride amplitude, speed, heel strike, arm swing, and turning (Goetz et al., 2008). We design 18 features across 2 categories from established clinical gait analysis conventions (Supplementary Table 10). StrideLen is the feature stride_length, and ArmROM is the feature arm2arm_ROM.

To ensure a fair comparison of features across different lifting and monocular 3D pose estimation methods, all features were computed in a root-centered coordinate system. Each stride sequence was further normalized into a canonical coordinate frame. Specifically, we estimate a vertical (up) axis and a forward walking direction from the 3D pose sequence, and define a lateral axis via their cross product. Each stride is then rotated such that the x-axis aligns with the walking direction and the z-axis corresponds to the vertical direction, ensuring invariance to camera viewpoint and subject orientation.

Stride length is computed along the canonical forward axis (x-axis) in the root-centered coordinate frame. Since stride boundaries are defined by consecutive right heel strikes, the right step length is taken as the forward displacement of the right foot relative to the left foot at the final frame of the stride. The left step length is defined as the maximum forward displacement of the left foot relative to the right foot across all frames within the stride, capturing the moment of greatest left-foot lead. Stride length is then defined as the sum of the two step lengths.

Clinical grounding. Spatial-temporal features (stride length, velocity, cadence) are the most commonly reported gait parameters in PD research (Mirelman et al., 2019). Reduced stride length and walking speed are hallmarks of parkinsonian gait. Joint ROM features capture the amplitude of limb excursion during walking; reduced arm swing (low shoulder/elbow ROM) is one of the

Table 10: **Gait features** (18 features, 2 categories). Asymmetry features (suffix _asy) = $|f_L - f_R|/(f_L + f_R)$.

Feature	Category	Unit	Definition
stride_length	Spatial-temporal	mm	Projected heel-to-heel distance along walking direction
stride_time	Spatial-temporal	s	Duration of one full stride (frames / fps)
step_length	Spatial-temporal	mm	Heel-to-contralateral-heel projected distance
step_time	Spatial-temporal	s	Time between consecutive contralateral heel strikes
velocity	Spatial-temporal	mm/s	stride_length / stride_time
cadence	Spatial-temporal	steps/min	2×60 / stride_time
step_length_asy	Spatial-temporal	–	Left-right asymmetry of step length
step_time_asy	Spatial-temporal	–	Left-right asymmetry of step time
hip_ROM	Joint angle	°	Sagittal-plane hip flexion/extension range per stride
knee_ROM	Joint angle	°	Sagittal-plane knee flexion/extension range
elbow_ROM	Joint angle	°	Elbow flexion/extension range (arm swing proxy)
shoulder_ROM	Joint angle	°	Shoulder flexion/extension range (arm swing amplitude)
arm2arm_ROM	Joint angle	°	ROM of angle between left and right arms
leg2leg_ROM	Joint angle	°	ROM of angle between left and right legs
hip_ROM_asy	Joint angle	–	Left-right asymmetry of hip ROM
knee_ROM_asy	Joint angle	–	Left-right asymmetry of knee ROM
elbow_ROM_asy	Joint angle	–	Left-right asymmetry of elbow ROM
shoulder_ROM_asy	Joint angle	–	Left-right asymmetry of shoulder ROM

earliest gait signs of PD (Mirelman et al., 2019). Asymmetry features capture the motor laterality characteristic of PD, which persists across disease stages (Djaldetti et al., 2006).

Table 11: **Full feature-based results: Gait.**

Model	ρ	MAE	$\text{Acc}_{\pm 1}$	κ	F1_m
Ensemble	0.613	0.385	1.000	0.458	0.497
MLP	0.594	0.404	1.000	0.277	0.317
GBR-Cls	0.551	0.374	0.982	0.431	0.481
SVR	0.552	0.410	1.000	0.385	0.470
ElasticNet	0.546	0.434	1.000	0.277	0.502
XGBoost	0.546	0.406	1.000	0.282	0.431
Lasso	0.542	0.444	1.000	0.203	0.465
GBR	0.540	0.409	0.982	0.374	0.458
Ridge	0.524	0.440	1.000	0.310	0.442
RF	0.510	0.448	0.982	0.242	0.417
OrdinalReg	0.480	0.346	1.000	0.411	0.476
RF-Cls	0.442	0.458	0.982	0.235	0.407

Comprehensive Model Comparison for Feature-Based UPDRS Prediction Table 11 shows more uniform performance across model families. The ensemble achieved the best overall results ($\rho = 0.613$, $\kappa = 0.458$), and all models reached $Acc \pm 1 \geq 0.982$, indicating predictions rarely deviated by more than one UPDRS level. Ordinal regression achieved the lowest MAE (0.346), suggesting gait features are also well-suited to ordinal prediction.

Table 12: Pose-level and feature-level evaluation of off-the-shelf models. MPJPE and PA-MPJPE are computed against the triangulated 3D poses. Spatiotemporal (ST) and kinematic feature errors are reported as the relative difference (%) between features extracted from estimated and triangulated 3D poses. Spatiotemporal features: TipSpeed (fist) and StrideLen (gait); kinematic features: FistArc (fist) and ArmROM (gait). Best results are shown in **bold**.

Activity	Pose Input	MPJPE (mm) ↓	PA-MPJPE (mm) ↓	ST Err. ↓	Kinematic Err. ↓
Fist	SAM 3D Body	36.11	13.93	87.83%	36.36%
	VideoPose3D	**22.46**	**8.91**	**21.54%**	**22.04%**
	HaMeR	70.75	43.29	57.75%	28.93%
	WiLoR	71.60	43.34	56.68%	28.60%
	MediaPipe Hands 3D	97.54	21.21	101.41%	47.46%
Gait	WHAM	197.54	140.88	80.24%	52.81%
	SAM 3D Body	102.81	63.39	20.89%	105.42%
	MAGF_H36M	189.94	140.65	50.95%	78.98%
	MAGF_TULIP	**35.76**	**25.48**	**9.93%**	**33.57%**
	PARE	189.88	139.29	20.18%	79.32%
	4D-Humans (HMR 2.0)	191.75	139.32	45.98%	44.56%
	MediaPipe Pose 3D	171.67	146.49	29.30%	125.56%
	Multi-HMR	178.98	139.74	17.68%	66.80%
	RTMW3D	183.44	139.69	45.50%	201.45%

A.3 Lifting Model Details

All lifting models take 2D detections as input and predict 3D joint positions. For gait, multi-view inputs from four cameras (Camera 1, 2, 3, and 6) are used, while for fist activities, two cameras (Camera 4 and 5) are used. Subjects are split into 46 training and 10 testing subjects using a stratified scheme based on the distribution of UPDRS severity levels. We evaluate a simple baseline and three temporal models. The simple baseline (CNN) is a frame-wise fully connected network (Martinez et al., 2017), trained for up to 80 epochs with early stopping. The temporal models include VideoPose3D (Pavllo et al., 2019), MotionAGFormer (Mehraban et al., 2024), and PoseFormerV2 (Zhao et al., 2023). For all temporal models, the input sequence length is 81 frames, and training is performed for up to 30 epochs with early stopping. All training settings follow those of the original papers and public implementations.

A.4 Pose-Based Prediction: Architecture and Training Details

Four architectures adapted from published models with reduced dimensions ($\sim$56 subjects per activity):

ST-Transformer: Stacked spatial + temporal MHSA blocks (embed=128, depth=4, heads=8), following DSTformer (Zhu et al., 2023). **MotionAGFormer-Lite:** Dual-stream transformer + GCN with adaptive fusion (embed=128, depth=4) (Mehraban et al., 2024). **ST-GCN:** Graph conv + temporal conv, learnable adjacency (channels=[64,128,256]) (Yan et al., 2018). **Dual-Stream Transformer:** Parallel S-T and T-S streams with gate fusion ($\sim$1.2M params).

Input: 80 fps $\rightarrow$ 30 fps, 243-frame sliding windows (50% overlap). Training: focal loss with capped inverse-frequency weights, AdamW (lr=10^{-3}), cosine schedule, early stopping (patience=15), 3 seeds. Test: mean softmax probability aggregation $\rightarrow$ argmax.

A.5 Complete Off-the-Shelf Model Comparison

We further evaluated a broad set of off-the-shelf monocular 3D pose estimation methods to assess their suitability for quantitative analysis of pathological movements in PD. These methods span recent parametric human mesh recovery, direct 3D pose estimation, and hand-specific reconstruction approaches. All off-the-shelf models were applied without task-specific fine-tuning.

As shown in Table 12, the limitations observed with the representative models in the main text persist across this broader set of methods. For gait, all evaluated off-the-shelf methods exhibit substantially larger pose errors than the in-domain MAGF_TULIP model. Although some methods

Table 13: Effect of Human3.6M pretraining under different amounts of labeled TULIP-BENCH training data. *Scratch* denotes MotionAGFormer trained from random initialization, while *Finetuned* denotes initialization from a Human3.6M-pretrained checkpoint followed by training on TULIP-BENCH. Best results within each data setting are shown in **bold**. Lower values are better for all metrics.

Labeled Training Data	Method	MPJPE (mm) ↓	PA-MPJPE (mm) ↓	ST Err. ↓	Kinematic Err. ↓
10%	Scratch	52.97	33.66	16.86%	38.04%
	Finetuned	**45.39**	**29.69**	**13.49%**	**36.20%**
25%	Scratch	42.07	29.11	12.70%	36.41%
	Finetuned	**39.77**	**27.50**	**12.15%**	**35.75%**
50%	Scratch	37.89	26.66	11.53%	34.02%
	Finetuned	**36.87**	**25.84**	**11.42%**	**33.65%**
100%	Scratch	35.76	25.48	**9.93%**	33.57%
	Finetuned	**35.59**	**24.99**	10.21%	**33.47%**

achieve relatively low error for individual clinical measures, their performance is inconsistent across pose-level and clinically derived metrics. A similar pattern is also observed for the fist task

Overall, these results indicate that strong performance on generic human pose and mesh reconstruction benchmarks does not directly translate to reliable measurement of pathological movements in PD. Off-the-shelf models exhibit substantial errors in both reconstructed pose and clinically relevant motion features. These findings support the need for domain-specific adaptation or training when applying monocular 3D pose estimation to PD motor analysis.

A.6 Effect of Human3.6M Pretraining

To examine whether a general human-motion prior remains beneficial after adaptation to pathological PD movements, we conducted an additional pretraining and fine-tuning experiment using the same MotionAGFormer architecture. In the main comparison, MAGF_H36M denotes a Human3.6M-pretrained model applied directly to TULIP-BENCH without adaptation, whereas MAGF_TULIP denotes the same architecture trained from scratch using TULIP-BENCH. These two settings represent the off-the-shelf and fully in-domain endpoints, respectively, but do not isolate whether Human3.6M pretraining provides additional benefit once the model is adapted to the target domain.

We therefore introduced a third setting in which MotionAGFormer was initialized from the Human3.6M-pretrained checkpoint and subsequently fine-tuned on TULIP-BENCH. To further characterize how the benefit of pretraining depends on the amount of in-domain supervision, we compared training from scratch and Human3.6M initialization using 10%, 25%, 50%, and 100% of the available TULIP-BENCH training data. Both settings used identical five-fold splits and training protocols.

As shown in Table 13, Human3.6M pretraining is most beneficial when limited in-domain training data are available. With 10% of the TULIP-BENCH training data, pretraining reduces MPJPE from 52.97 mm to 45.39 mm and improves all clinical feature metrics. The benefit progressively decreases as more in-domain data are used, and the scratch and finetuned models achieve nearly identical performance with the full training set. These results indicate that Human3.6M pretraining primarily improves data efficiency by providing a useful initialization, while sufficient in-domain training largely closes the gap.

A.7 Task3 Details

A.7.1 Body-part severity: convergent validity and prediction

Convergent validity between kinematic features and clinician annotations. Direct correlation between clinician-rated body-part severity and the kinematic feature it should target, pooled across all 20 OFF/ON observations (Spearman ρ) (Supplementary Table 14):

Table 14: Convergent-validity correlations between clinician body-part severity and target kinematic features.

Body-part rating	Target feature	Pooled ρ	p
Stride severity (max)	Stride length	-0.91	$< 10^{-6}$
Stride severity (max)	Velocity	-0.86	$< 10^{-5}$
Stride severity (max)	Hip ROM	-0.85	$< 10^{-5}$
Stride severity (max)	Leg-leg ROM	-0.83	$< 10^{-4}$
Stride severity (max)	Knee ROM	-0.68	$< 10^{-3}$
Arm-swing severity (max)	Shoulder ROM	-0.83	$< 10^{-4}$
Arm-swing severity (max)	Arm-arm ROM	-0.83	$< 10^{-5}$
Arm-swing severity (max)	Elbow ROM	-0.54	0.015
Elbow flexion (max)	Elbow ROM	-0.62	0.004

Negative ρ values reflect the convention that higher clinician severity scores correspond to reduced kinematic ROM. Strong negative correlations across all six body-part ratings confirm that automatic features capture the same motor phenomena clinicians visually assess.

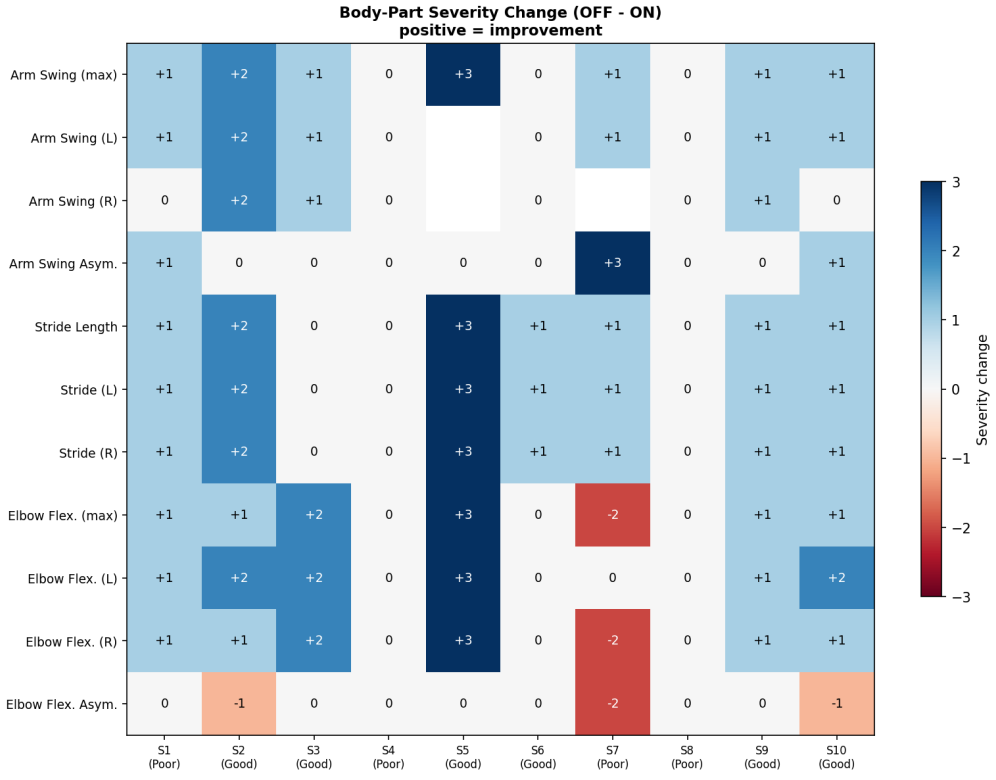

Figure 4: **Body-part severity heatmap across the 10 DBS subjects.** Rows: clinician body-part severity ratings (arm-swing L/R, stride L/R, elbow-flex L/R); columns: 10 DBS subjects, separated into OFF and ON sub-columns. Cell color: ordinal severity score (0–4). Per-subject improvements are heterogeneous: some patients improve primarily in stride severity, others in arm-swing severity, supporting the multi-dimensional nature of DBS-induced kinematic change.

Body-part severity regression (exploratory). We additionally evaluated whether ML models can predict the six clinician body-part severity scores given features or pose input, in a LOSO setting at $N = 20$ paired observations. (see Supplementary Table 15)

The regression analysis is included for completeness and as an open problem for future work with larger cohorts. We do not promote it to main text because (a) the input feature representation is not

Table 15: Per-body-part Spearman ρ across four baselines for body-part severity regression. The 18-feature input is not body-part-specific (e.g., it includes leg features when predicting arm swing); the result therefore reflects across-subject severity correlation as much as body-part-specific signal. Median across body parts in the right column.

Method	Arm-L	Arm-R	Stride-L	Stride-R	Elbow-L	Elbow-R	Median
Feature Ridge	0.45	0.71	0.80	0.72	0.55	0.52	**0.63**
Flat-pose Ridge	−0.31	−0.14	0.72	0.71	0.29	0.70	0.49
Pose CNN	−0.14	0.39	0.17	0.20	0.26	0.49	0.23
ST-GCN	−0.42	−0.42	−0.18	−0.19	−0.41	−0.16	−0.30

body-part-specific, and (b) at $N = 20$ the result reflects cross-body-part severity correlation more than isolated body-part-specific prediction. The convergent-validity table above is the cleaner version of the same underlying claim.

A.7.2 Full treatment sensitivity table (all 18 features)

Table 16: Full per-feature treatment sensitivity (DBS OFF vs. ON, $N = 10$). Cohen's d negative = feature increases OFF→ON. % change = (mean ON − mean OFF) / mean OFF. `wilcoxon_p` reported.

Feature	Mean OFF	Mean ON	% change	d
Knee ROM (°)	33.78	39.80	+17.8%	−1.61**
Stride length (mm)	770.5	1021.9	+32.6%	−1.52**
Step length (mm)	386.4	511.0	+32.3%	−1.50**
Velocity (mm/s)	637.6	847.3	+32.9%	−1.49**
Hip ROM (°)	24.70	32.75	+32.6%	−1.34**
Leg-leg ROM (°)	24.01	32.05	+33.5%	−1.25**
Step length asym.	0.279	0.187	−33.0%	+0.70*
Hip ROM asym.	0.216	0.183	−15.3%	+0.68
Shoulder ROM (°)	14.31	21.33	+49.1%	−0.62*
Step time asym.	0.113	0.085	−24.4%	+0.50
Arm-arm ROM (°)	9.18	18.06	+96.9%	−0.46
Elbow ROM (°)	21.32	27.74	+30.1%	−0.45
Elbow ROM asym.	0.343	0.331	−3.5%	+0.14
Shoulder ROM asym.	0.370	0.357	−3.4%	+0.14
Step time (s)	0.614	0.605	−1.5%	+0.10
Stride time (s)	1.228	1.210	−1.5%	+0.10
Knee ROM asym.	0.184	0.182	−0.9%	+0.07
Cadence	100.77	101.26	+0.5%	−0.03

** $p < 0.01$ (Wilcoxon); * $p < 0.05$.

Clinical-magnitude interpretation. The six features with $|d| > 1.0$ all involve spatial gait amplitude (stride length, step length, velocity, knee ROM, hip ROM, leg-leg ROM) and improve by 18–34% on average between OFF and ON states. Magnitudes are comparable to levodopa effects in moderate PD (Faist et al., 2001). Temporal features (cadence, stride time) change by less than 2%, indicating DBS's effect on this cohort is predominantly on movement amplitude (hypokinesia) rather than rhythm (bradykinesia). Asymmetry features show modest improvement, suggesting DBS partially restores bilateral coordination.

A.7.3 Self-supervised pre-training and cohort-to-DBS evaluations

Architecture. 3-block ST-GCN with spatial partitioning (self / centripetal / centrifugal), bone-connectivity graph, temporal kernel size 9, channel widths $32 \to 64 \to 64$ (stride 1, 2, 2). A lightweight decoder (two ConvTranspose2d upsamples) reconstructs masked coordinates at the input resolution.

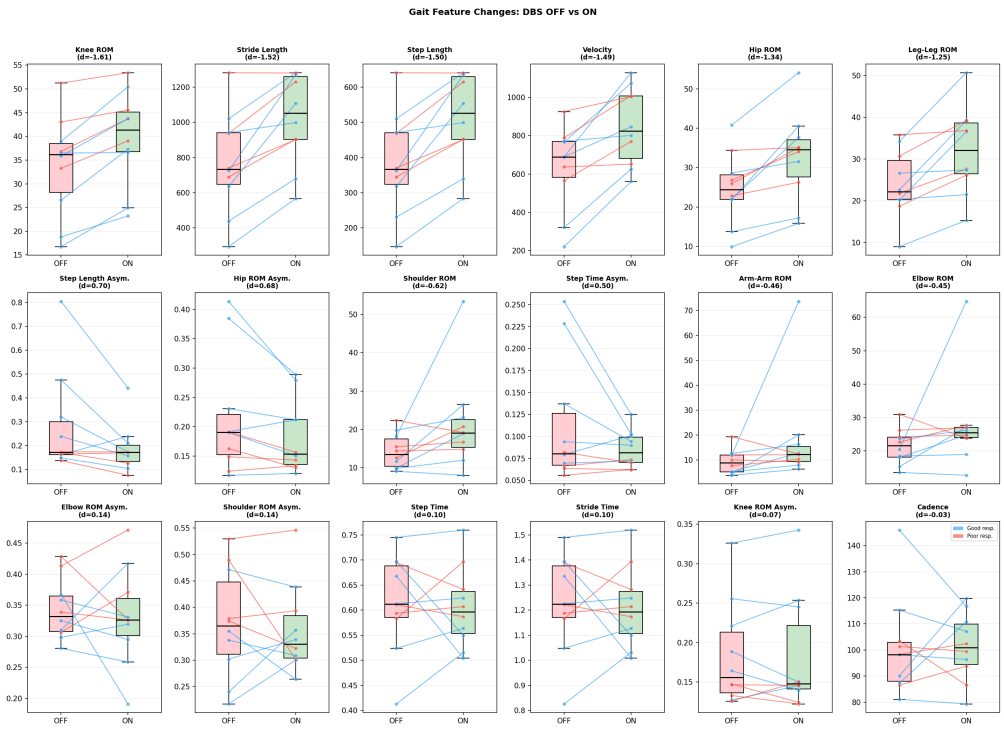

Figure 5: **Paired OFF vs. ON feature distributions for the 10 DBS subjects.** Each panel shows one of the 18 gait features; OFF (red) and ON (green) box plots with per-subject lines connecting paired observations. The six features with $|d| > 1.0$ in Supplementary Table 16 (knee ROM, stride length, step length, velocity, hip ROM, leg-leg ROM) show consistent within-subject increases across most patients; temporal features (cadence, stride time) show negligible OFF/ON differences.

Pretext task. Random 15% (joint, time) masking; input zeroed at masked positions; MSE reconstruction loss averaged over masked positions only.

Training. 100 epochs; batch size 64; Adam (lr $= 10^{-3}$, weight decay 10^{-5}); 10% holdout validation; early stopping by best validation loss. Five random seeds: $\{42, 7, 123, 256, 2024\}$. Best validation reconstruction loss: 0.0002 (consistent across seeds).

Zero-shot severity regressor. On cohort embeddings, we fit a scikit-learn LinearRegression predicting UPDRS-gait integer scores. The fitted regressor is then applied to DBS stride embeddings to produce predicted continuous severity. No DBS labels are used at any stage of pre-training or probe fitting.

Cohort-to-DBS results.

1. *Cohort linear probe.* $\rho = 0.68 \pm 0.03$ (mean $\pm$ SD, 5 seeds). Confirms the SSL representation encodes severity structure in the cohort.
2. *Zero-shot transfer to DBS.* Pooled Spearman $\rho = 0.52$ between predicted continuous severity and clinician UPDRS-gait on DBS subjects (20 observations: 10 subjects $\times$ 2 states).
3. *Per-subject directionality.* 7 of 10 DBS subjects show predicted continuous severity decreasing OFF$\rightarrow$ON. Among the 8 subjects with Δ UPDRS-gait $= 0$, 6 show continuous severity decrease; 0 show clear increase. The remaining 2 (Sub7: $\Delta = +0.007$; Sub4: $\Delta = +0.047$) lie within seed-to-seed noise.
4. *Embedding-space displacement.* Median DBS OFF$\rightarrow$ON displacement is 0.72 cohort-UPDRS-step units (Euclidean), 0.50 (cosine), 0.77 (severity-axis projection). Three of 10 subjects exceed 1 full UPDRS-step; only 2 of 10 have any actual UPDRS-gait change.

Robustness. Three-metric qualitative agreement holds (Euclidean / cosine / severity-axis: 0.72 / 0.50 / 0.77 UPDRS-step units). Jackknife leave-one-out on the cohort reference-step distance gives a $\pm 30\%$ range on the Euclidean reference but does not invalidate the ratio (numerator and denominator rescale together). Excluding Sub1 (imbalanced 6 OFF / 49 ON strides) changes medians by ≤ 0.03.

A.7.4 Additional supervised evaluations on the DBS cohort

Treatment-response magnitude prediction. LOSO across 10 DBS subjects; OFF-state pose as input; OFF-to-ON Δ in six strong kinematic features (stride length, step length, velocity, hip ROM, knee ROM, leg-leg ROM) as target (Supplementary Table 17). We use a residualized target in which the OFF-level baseline is regressed out, so the prediction task cannot be solved by predicting regression-to-mean alone. Reference baselines confirm the residualization matters: a constant-mean prediction yields $\rho = -1.00$ by construction on residualized data, and a $-k$·OFF baseline yields $\rho = -0.91$.

Table 17: Treatment-response magnitude prediction (residualized target). Median Spearman ρ across the six kinematic feature targets.

Method	Median residualized ρ
Predict mean Δ (constant)	-1.00
Predict $-k$·OFF	-0.91
Feature Ridge	-0.16
Flat-pose Ridge	-0.14
1D-Temporal CNN	-0.33
ST-GCN	$\mathbf{+0.27}$

ST-GCN is the only baseline retaining residual signal under residualization. Without residualization, naive raw-Δ 1D-Temporal CNN reaches $\rho = 0.61$, which collapses to $\rho = -0.33$ residualized, confirming that most of the apparent raw-Δ signal was regression-to-mean rather than learned response prediction. The residualized protocol is therefore the headline metric.

Responder classification. OFF-state features predict clinician-rated good ($n = 6$) vs. poor ($n = 4$) responders, LOSO with four classifiers (Logistic Regression, SVM-Linear, SVM-RBF, Random Forest). Best model SVM-Linear, accuracy 60%, $\kappa = 0.167$. Marginally above chance. The benchmark is reported for future treatment cohorts where larger N may permit more reliable response prediction.

A.8 Compute Resources

All experiments were conducted on a single-institution GPU cluster using the following hardware: one NVIDIA RTX A5000 (24 GB), one NVIDIA RTX A6000 (48 GB), two NVIDIA RTX 3090 (24 GB each), and one NVIDIA RTX 4080 (16 GB). No distributed training was used; each experiment ran on a single GPU.

A.9 AI Use Disclosure

Generative AI tools were used solely to edit the manuscript for language, grammar, and readability. They were not used for research design, methodology, data analysis, interpretation of results, or generation of scientific claims or conclusions. All AI-assisted edits were reviewed by the authors, who take full responsibility for the final content of the manuscript.