NeurGait: A Vision-Language Framework for Structured Gait Assessment Support in Parkinson's Disease and Multiple Sclerosis

Guzin Kaya Aytutuldu[1], Ilhan Aytutuldu[2], Nazan Karagoz Sakalli[3], Rabia Gokcen Gozubatik Celik[3], Mesude Tutuncu[3], Zeynep Ozdemir[3], Burcu Ersoz Huseyinsinoglu[4], and Yusuf Sinan Akgul[1]

[1] Department of Computer Engineering, Gebze Technical University, Kocaeli, Turkey
[2] Vocational School of Cybersecurity, Gebze Technical University, Kocaeli, Turkey
[3] Department of Neurology, Bakırköy Prof. Dr. Mazhar Osman Training and Research Hospital for Mental Health and Neurological Disorders, University of Health Sciences, Istanbul, Turkey
[4] Faculty of Health Sciences, Department of Physiotherapy and Rehabilitation, Marmara University, Istanbul, Turkey

Abstract. Clinical gait assessment in neurological rehabilitation depends on the interpretation of specific movement features, yet most video-based approaches provide only diagnostic labels without structured explanations. We propose NeurGait, a vision-language framework for generating rubric-aligned gait annotations from video through knowledge distillation. A large vision-language model generates teacher annotations aligned with a structured gait rubric, which are then used to train a compact student model capable of producing model annotations together with diagnostic predictions. The data comprises gait videos from 77 participants (Parkinson's Disease: $N = 42$, Hoehn & Yahr 1–3; Multiple Sclerosis: $N = 35$, EDSS 0–3), all clinically diagnosed by neurologists. We compared model annotations with independent blinded annotations from physiotherapists with at least ten years of neurorehabilitation experience, using the same structured gait rubric. Inter-rater agreement between physiotherapists for the binary PD/MS diagnosis task was only fair ($\kappa = 0.349$), highlighting the difficulty of distinguishing early-stage PD and MS from gait videos alone. Despite this challenge, agreement between model annotations and physiotherapist annotations for trunk lean, cadence, and walk speed reached moderate levels and approached physiotherapist agreement ranges. NeurGait-LUPI generated model annotations and diagnostic predictions directly from video at inference without requiring manual clinical labeling. The proposed framework achieved moderate agreement with physiotherapists on selected postural gait features, providing preliminary evidence that vision-language distillation can support rubric-aligned gait assessment from video.

Keywords: Parkinson's disease · multiple sclerosis · gait analysis · vision-language models · knowledge distillation · interpretable artificial intelligence

1 Introduction

Neurological disorders such as Parkinson's disease (PD) and Multiple Sclerosis (MS) commonly affect motor control and produce measurable alterations in gait [14, 3]. In clinical practice, gait assessment relies on the observation of specific movement features — such as arm swing, trunk stability, and postural alignment — rather than a binary diagnostic label. Yet most video-based computational approaches reduce this rich clinical reasoning to a classification output, discarding the structured gait features that clinicians rely on for treatment planning and rehabilitation monitoring [12].

Recent work has demonstrated that gait can be analyzed from ordinary video using pose estimation and machine-learning methods, enabling the screening and assessment of neurological disorders such as PD and MS [13, 10, 8]. These findings establish the feasibility of video-based gait analysis, but leave open the question of how to extract clinically interpretable gait features alongside diagnostic predictions.

Recent advances in vision-language models (VLM) have created new opportunities for clinical video analysis by enabling models to capture both visual patterns and clinically meaningful semantic information. In parallel, knowledge distillation offers an effective strategy for transferring knowledge from large foundation models to smaller and more efficient models suitable for real-world deployment [8, 1].

Recent vision-language approaches have begun to move beyond pure classification by providing interpretable explanations of gait patterns from video. In particular, BioGait demonstrated the feasibility of producing clinically meaningful human-readable explanations using large VLMs [4]. However, free-text rationales are difficult to evaluate systematically against clinical standards, and the potential of structured gait feature annotations learned through knowledge distillation remains largely unexplored.

In this study, we propose NeurGait, a vision-language framework for generating rubric-aligned gait annotations from videos of patients with PD and MS. A teacher VLM generates teacher annotations that are distilled into a compact student model capable of producing model annotations and diagnostic predictions at inference. We evaluate model annotations against independent physiotherapist (PT) annotations using the same structured gait rubric and investigate whether privileged clinical observations can improve training through a Learning Using Privileged Information (LUPI) framework. By generating rubric-aligned model annotations alongside diagnostic predictions, NeurGait aims to support more transparent video-based gait assessment.

2 Method

This retrospective study included patients previously diagnosed with MS according to the McDonald criteria by a neurologist and having an Expanded Disability Status Scale (EDSS) score between 0 and 3, and patients aged 40–75 years with

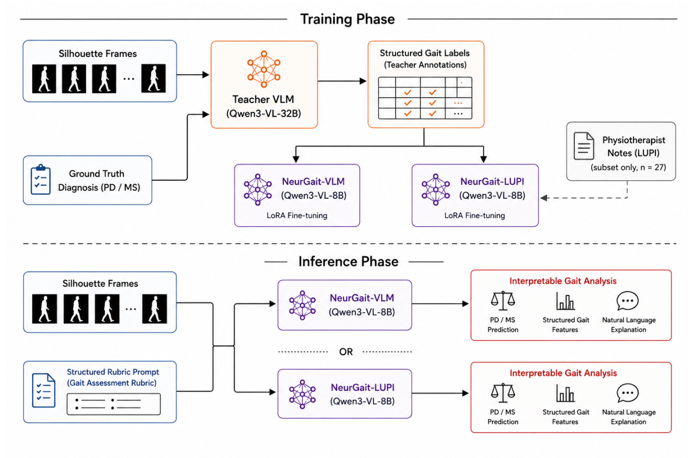

Fig. 1. Overview of the proposed NeurGait framework. Structured gait annotations generated by a Qwen3-VL-32B teacher model are used to train NeurGait-VLM and NeurGait-LUPI. PT observations provide privileged information during training only. At inference, the fine-tuned student produces structured gait annotations together with a diagnostic prediction, enabling direct comparison with PT assessments.

Idiopathic PD who were at Hoehn and Yahr Stages 1-3. All participants had stable medication regimens and a Functional Ambulation Classification score ≥ 3. Demographic and clinical data were extracted from medical records, and previously recorded gait videos were reviewed for analysis.

This study was approved by the institutional ethics committee (Protocol No: anonymized for review). The dataset comprised gait videos from 77 participants (MS: $N = 35$, PD: $N = 42$) recruited from outpatient neurology clinics and community settings. Videos were recorded using handheld cameras at 30 fps during overground walking tasks. A subject-stratified split yielded 61 training subjects. (28 MS, 33 PD; 180 videos) and 16 test subjects (7 MS, 9 PD; 37 videos).

Statistical Analysis. Group comparisons between MS and PD were performed using independent samples t-tests (IBM SPSS 28.0); $p < 0.05$ was considered statistically significant.

Machine Learning (ML) Baselines. We trained Logistic Regression (LR), Support Vector Machine (SVM), and Random Forest (RF) classifiers on two feature sets. First, pose-derived features: two-dimensional body keypoints were extracted using YOLOv8m-pose [7] and gait features were computed and evaluated under Leave-One-Subject-Out (LOSO) ($N = 77$). Second, PT annotations (Physio-GT): the eight rubric features served as input and models were evaluated under LOSO on the annotated subset.

Preprocessing. Background subtraction was applied to each video to isolate the walking subject and reduce scene-dependent visual confounds. The resulting binary silhouette frames were used as input to all VLM-based models.

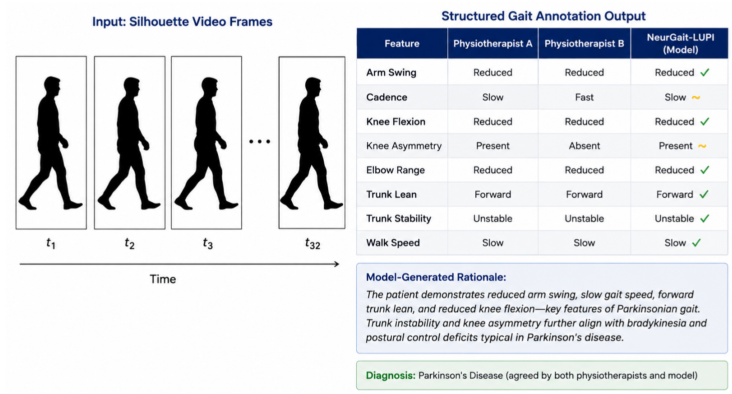

Fig. 2. Representative structured gait annotation generated by NeurGait-LUPI. Left: input silhouette frames. Right: comparison of physiotherapist and model annotations with the generated rationale and diagnosis.

Knowledge Distillation (KD) Pipeline. We adopt a teacher–student paradigm in which a large VLM generates teacher annotations for each training video, and a smaller VLM is subsequently fine-tuned on these teacher annotations via Low-Rank Adaptation (LoRA) [6]. At inference, the fine-tuned student produces model annotations together with a diagnostic prediction, enabling direct comparison with PT annotations (Figure 1) .

Teacher Model. Qwen3-VL 32B was prompted zero-shot to act as a neurorehabilitation specialist. For each training video, frames were uniformly sampled and encoded at 560 px width to control the number of visual tokens. We evaluated 8 and 32 frames per video; 32 frames yielded more consistent rubric annotations and was used for all reported models. The teacher generated annotations together with a free-text clinical rationale and a binary diagnosis. The diagnosis field was subsequently replaced with the ground-truth label before use as a training target.

Structured gait assessment rubric. The rubric comprises eight gait features (*ARM_SWING*, *WALK_SPEED*, *KNEE_ASYMMETRY*, *KNEE_FLEXION*, *CADENCE*, *ELBOW_RANGE*, *TRUNK_LEAN*, and *TRUNK_STABILITY*), each represented by predefined categorical labels. Feature labels capture both qualitative distinctions (e.g., *Stable*/*Unstable* for trunk stability, *Absent*/*Present* for knee asymmetry) and ordinal gradations (e.g., *Slow*/*Normal*/*Fast* for cadence and walk speed, *Reduced*/*Normal*/*Asymmetric* for arm swing). A brief clinical rationale is additionally generated to justify the predicted annotations (Figure 2).

Student Model. Qwen3-VL 8B was fine-tuned via LoRA (r=16, α=32, dropout= 0.05) on the 180 training videos using 3-fold subject-stratified cross-validation. Each fold was trained for 5 epochs with learning rate 2×10^{-4} (cosine schedule) and batch size 1. The full teacher output—teacher annotations, rationale, and diagnosis—served as the training target.

To isolate the contribution of privileged PT observations, we trained two variants using teacher rubrics generated from 32 frames with ground-truth diagnosis labels. *NeurGait-VLM* is trained on silhouette frames and the teacher annotations alone, without access to any clinical notes. *NeurGait-LUPI* additionally incorporates PT observations as privileged information during training: the clinical note is appended to the input at training time but withheld at inference, following the Learning Using Privileged Information (LUPI) paradigm [11, 5]. This design directly tests whether privileged clinical context improves the model's ability to learn discriminative gait features, independent of temporal coverage or label source.

Evaluation. Metrics were computed at the subject level using majority-vote aggregation over a subject's videos. Classification labels were extracted from the model's free-text output; predicted PD probability was estimated from the token-level probability of the first generated token. ML baselines were evaluated under LOSO cross-validation (N=77 for pose-derived features; annotated subset for Physio-GT). Fine-tuned VLM variants were evaluated on the held-out test set (N=16). Primary metrics were AUC, F1, MS sensitivity, and PD sensitivity; 95% bootstrap confidence intervals (2000 iterations, subject-level resampling) were computed for all metrics. Because the primary objective of NeurGait is interpretable gait feature annotation generation, model annotations were additionally evaluated against physiotherapist annotations using the same structured gait rubric.

To establish a human reference, 27 subjects (13 MS, 14 PD) were independently evaluated under blinded conditions by two PTs using the same structured gait rubric. Both physiotherapists routinely assess patients with Parkinson's disease and multiple sclerosis in clinical neurorehabilitation practice. Model predictions were obtained via cross-validation to avoid overlap with the held-out test set. Inter-rater agreement between PTs and agreement between model annotations and PT annotations were assessed using Cohen's κ. A third PT provided unstructured observations that served as privileged training input for NeurGait-LUPI but were excluded from evaluation. This analysis focused on the clinical plausibility of the generated model annotations rather than diagnostic performance alone.

3 Results

A total of 77 participants were analyzed, including recordings from 42 patients with PD and 35 patients with MS. Baseline demographic and clinical characteristics are presented in Table 1. Significant between-group differences were observed for age, BMI, disease duration, and TUG scores ($p < 0.05$). Nevertheless, most participants in both groups were in the early stages of disease and exhibited relatively low levels of disability according to the Hoehn and Yahr and EDSS classifications.

Pose-derived classical ML baselines achieved strong LOSO performance (RF: AUC 0.852, LR: AUC 0.857), confirming that discriminative gait information is

Table 1. Demographic and Clinical Characteristics of Patients with Parkinson's Disease and Multiple Sclerosis

Variable	**PD (n=42)** Mean ± SD	**MS (n=35)** Mean ± SD	**t / p**
Age (years)	59.76 ± 8.31	42.17 ± 9.26	$8.77/ < 0.001$
BMI (kg/m^2)	27.29 ± 3.34	24.79 ± 5.57	2.43/0.01
Disease Duration (years)	5.23 ± 3.94	10.54 ± 7.67	$-3.90/ < 0.001$
TUG	9.46 ± 3.73	7.17 ± 1.69	$3.35/ < 0.001$
Disease Stage			
PD (HY)	**% (n)**	**MS (EDSS)**	**% (n)**
HY 1	11.9 (5)	EDSS 0	45.7 (16)
HY 1.5	7.1 (3)	EDSS 1	45.7 (16)
HY 2	59.5 (25)	EDSS 2	5.7 (2)
HY 2.5	9.5 (4)	EDSS 3	2.9 (1)
HY 3	11.9 (5)	–	–

PD: Parkinson's Disease; MS: Multiple Sclerosis; BMI: Body Mass Index; TUG: Timed Up and Go Test; HY: Hoehn and Yahr stage; EDSS: Expanded Disability Status Scale. Values are presented as mean ± standard deviation (SD) or number (n) and percentage (%).

present in the videos. As expected for an early-stage cohort, gait differences between PD and MS were subtle. Physio-GT baselines — classical ML models trained on PT annotations — performed at chance level (AUC 0.330–0.541; Table 2), suggesting that PT annotations for the limited rubric feature set may be insufficient for reliable PD/MS discrimination in this early-stage cohort, motivating the exploration of learned video representations.

In contrast, VLM-based models operating directly on silhouette video captured information beyond the PT-annotated gait features. Under 3-fold cross-validation on the training set, NeurGait-LUPI achieved an AUC of 0.700 and NeurGait-VLM 0.644; on the held-out test set, NeurGait-LUPI improved to 0.794, suggesting consistent generalization across evaluation protocols. Among the interpretable variants, NeurGait-LUPI outperformed NeurGait-VLM on the held-out test set in AUC (0.794 vs. 0.664) and F1 (0.700 vs. 0.667), suggesting that PT observations provided useful complementary information during training. NeurGait-VLM, however, achieved higher MS sensitivity (0.857 vs. 0.529), reflecting a different classification trade-off between the two conditions (Table 2).

As a diagnosis-only reference, we trained a supervised fine-tuning baseline (Gait-SFT) directly on ground-truth diagnostic labels without structured gait annotation targets. Gait-SFT achieved higher classification performance than the interpretable NeurGait variants (test AUC = 0.921, F1 = 0.783), but does not generate structured gait annotations and was therefore excluded from the primary annotation-agreement analyses.

Unlike classical ML models that yield only a predicted label, both VLM variants produced model annotations directly from raw silhouette video at inference, without any hand-crafted feature engineering.

As NeurGait is designed to support structured gait assessment rather than black-box classification, the primary evaluation focused on the clinical plausibility of model-generated annotations, assessed against independent PT annotations using the same structured rubric. On the blinded evaluation videos, PT A achieved accuracy of 0.556 with balanced sensitivity (MS: 0.538, PD: 0.571),

Table 2. Physio-GT models are trained on physiotherapist-annotated gait features; NeurGait variants generate structured gait annotations directly from video without requiring any annotations, and are evaluated for diagnosis as well. 95% bootstrap confidence intervals are used.

Split	Type	Model	AUC	F1	MS-Sens	PD-Sens	Acc
LOSO	Physio-GT	RF	0.541 [0.302–0.778]	0.538 [0.273–0.750]	0.615 [0.333–0.867]	0.500 [0.231–0.769]	0.556 [0.370–0.741]
		LR	0.505 [0.275–0.744]	0.571 [0.320–0.759]	0.538 [0.250–0.800]	0.571 [0.308–0.824]	0.556 [0.370–0.741]
		SVM	0.330 [0.125–0.574]	0.370 [0.100–0.581]	0.385 [0.133–0.643]	0.357 [0.111–0.625]	0.370 [0.185–0.556]
ZS	VLM	ZS-Qwen-VLM-32B	0.543	0.548	0.571	0.515	0.541
CV	VLM	NeurGait-VLM	0.644 [0.510–0.774]	0.703 [0.576–0.810]	0.464 [0.280–0.652]	0.788 [0.643–0.919]	0.639 [0.525–0.754]
		NeurGait-LUPI	0.700 [0.570–0.818]	0.725 [0.605–0.827]	0.357 [0.185–0.542]	0.879 [0.763–0.971]	0.639 [0.525–0.754]
Test	VLM	NeurGait-VLM	0.664 [0.573–0.875]	0.667 [0.308–0.889]	0.857 [0.545–1.000]	0.556 [0.222–0.875]	0.688 [0.438–0.875]
		NeurGait-LUPI	0.794 [0.164–0.805]	0.700 [0.421–0.900]	0.529 [0.000–0.833]	0.778 [0.444–1.000]	0.625 [0.375–0.875]

LOSO: Leave-One-Subject-Out; CV: 3-fold cross-validation; ZS: zero-shot inference.
Physio-GT: models trained on physiotherapist-provided structured gait annotations.

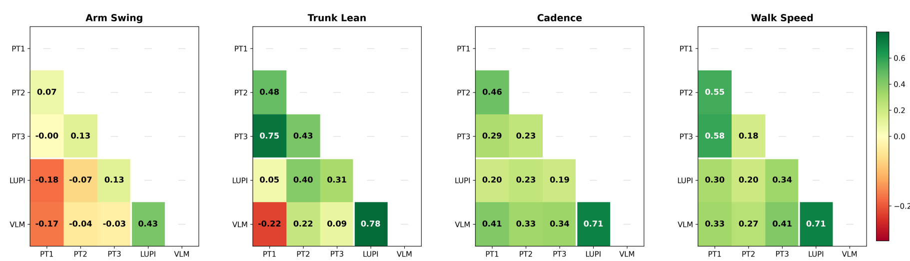

Fig. 3. Feature-level agreement (Cohen's κ) between NeurGait models and physiotherapists. Upper: physiotherapist inter-rater agreement. Lower: model–clinician agreement.

while PT B showed high MS sensitivity (0.923) at the cost of low PD sensitivity (0.286). Inter-rater agreement for the binary PD/MS diagnosis was only fair (κ=0.349) [9], reflecting the inherent ambiguity of distinguishing early-stage PD and MS from gait videos alone. Teacher annotation quality was assessed separately by comparing Qwen3-VL-32B outputs against the same PT evaluations; agreement was modest but within the range observed for more challenging gait features (κ=0.09–0.29), suggesting that teacher annotations captured some aspects of clinician-observed gait features and may provide a useful source of supervision for distillation.

The most clinically relevant findings emerged from the feature-level agreement analysis. Agreement between model annotations and PT annotations was assessed via Cohen's κ across four gait features (Figure 3). For trunk lean, cadence, and walk speed, model annotations from NeurGait-LUPI and NeurGait-

VLM achieved moderate agreement with PT annotations (κ up to $\sim$0.40), approaching inter-rater agreement levels in several rater pairs. NeurGait-VLM showed comparatively lower agreement on these features, particularly for trunk lean where one rater pair yielded negative κ. Arm swing and elbow range of motion remained the most challenging features for both models, with near-zero or negative κ values across all rater comparisons, consistent with the inherent difficulty of visually quantifying subtle upper-limb motion from silhouette video. Overall, inter-rater agreement among PTs was moderate across features (κ ranging from $\sim$0.07 to $\sim$0.75), reflecting the subjectivity of clinical gait assessment and contextualizing the difficulty of the gait feature annotation for automated models. These findings indicate that certain gait features, particularly trunk lean, cadence, and walk speed, can be extracted from video with consistency approaching clinician-level agreement, whereas arm swing remains difficult to assess reliably from silhouette video alone. Elbow range of motion showed a similar pattern, suggesting that upper-limb features are generally difficult to assess reliably from silhouette-based representations.

4 Discussion

The main finding of this study is that structured gait feature extraction from video is feasible and can achieve moderate agreement with PT annotations on selected postural gait features, even when binary diagnosis remains difficult. Inter-rater agreement among PTs for the diagnostic task was only fair ($\kappa = 0.349$), yet NeurGait-LUPI achieved moderate agreement on trunk lean, cadence, and walk speed ($\kappa \sim 0.40$), approaching inter-rater levels on selected features. This level of agreement likely reflects the inherent difficulty of distinguishing early-stage PD and MS based solely on gait videos. The observed variability also highlights the subjective nature of visual gait assessment, even among experienced physiotherapists. This supports our central argument that structured and interpretable gait feature annotations may provide a useful complement to diagnosis-only gait analysis systems.

Recent vision-language approaches such as BioGait have highlighted the potential of multimodal models for interpretable gait analysis [4]. In contrast to generating free-text clinical rationales, NeurGait focuses on gait feature annotation generation using a predefined clinical rubric. This design enables direct feature-level evaluation against PT annotations, which is not possible with diagnosis-only models.

Notably, agreement was not uniform across gait features. Higher agreement was observed for postural characteristics such as trunk lean, whereas features related to limb motion showed greater variability. This pattern is consistent with the challenges of visual gait assessment, where some abnormalities are more readily observable than others [2].

The results also provide preliminary evidence for the usefulness of privileged clinical information during training. Although PT observations were not available at inference time, incorporating these observations through the LUPI framework improved overall performance relative to the corresponding VLM baseline. This finding suggests that expert clinical observations may serve as a useful supervisory signal that may help align model annotations with clinician-observed gait features. Importantly, the strongest diagnosis-only baseline (Gait-SFT) achieved higher classification performance than NeurGait, highlighting a trade-off between predictive performance and structured interpretability. Accordingly, physiotherapist annotations should be interpreted as a clinical reference rather than an absolute ground truth for evaluating model-generated gait annotations. From a clinical perspective, structured gait annotations may provide greater value in rehabilitation settings, where transparent assessment of movement impairments is essential for treatment planning.

This study has several limitations. The cohort is relatively small and focused on early-stage disease, where gait differences are subtle. Privileged PT observations were available only for a subset of participants, and the wide confidence intervals on the held-out test set reflect the small test cohort. Despite these limitations, the results provide proof-of-concept evidence that VLM-based model annotation of gait features can support rubric-aligned gait assessment from video and warrant further evaluation in larger clinical cohorts. Future work will investigate larger cohorts, additional gait features, and the generalizability of the framework across broader neurological populations and clinical settings.

Prospects of Application. NeurGait could be deployed as a clinician training tool for standardizing video-based gait assessment in neurology and physiotherapy education settings. By generating structured, rubric-aligned model annotations from silhouette video, the system supports consistent evaluation without specialized hardware, enabling scalable use in outpatient rehabilitation clinics and tele-rehabilitation programs.

Acknowledgments. This research received no specific grant.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Ali, Z., Moon, J., Gillani, S., Afzal, S., Maqsood, M., Rho, S.: Vision-based approach to knee osteoarthritis and parkinson's disease detection utilizing human gait patterns. PeerJ Computer Science **11**, e2857 (2025)
2. Brunnekreef, J.J., Van Uden, C.J., van Moorsel, S., Kooloos, J.G.: Reliability of videotaped observational gait analysis in patients with orthopedic impairments. BMC musculoskeletal disorders **6**(1), 17 (2005)
3. Chee, J.N., Ye, B., Gregor, S., Berbrayer, D., Mihailidis, A., Patterson, K.K.: Influence of multiple sclerosis on spatiotemporal gait parameters: A systematic review and meta-regression. Archives of Physical Medicine and Rehabilitation **102**(9), 1801–1815 (2021)
4. Chen, E., Ji, Y., Greenberg, J.K., Steel, B., Arkam, F., Lewis, A., Singh, P., Liu, F.: Biogait-vlm: A tri-modal vision-language-biomechanics framework for interpretable clinical gait assessment. arXiv preprint arXiv:2603.08564 (2026)

5. Gao, Z., Cheng, S., Wittrup, E., Gryak, J., Najarian, K.: Learning using privileged information with logistic regression on acute respiratory distress syndrome detection. Artificial Intelligence in Medicine **156**, 102947 (2024)
6. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. Iclr **1**(2), 3 (2022)
7. Jocher, G., Chaurasia, A., Qiu, J.: Ultralytics YOLOv8 (2023), https://github.com/ultralytics/ultralytics
8. Kaur, R., Motl, R.W., Sowers, R., Hernandez, M.E.: A vision-based framework for predicting multiple sclerosis and parkinson's disease gait dysfunctions—a deep learning approach. IEEE journal of biomedical and health informatics **27**(1), 190–201 (2022)
9. Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. biometrics pp. 159–174 (1977)
10. Rupprechter, S., Morinan, G., Peng, Y., Foltynie, T., Sibley, K., Weil, R.S., Leyland, L.A., Baig, F., Morgante, F., Gilron, R., et al.: A clinically interpretable computer-vision based method for quantifying gait in parkinson's disease. Sensors **21**(16), 5437 (2021)
11. Vapnik, V., Izmailov, R.: Learning using privileged information: similarity control and knowledge transfer. The Journal of Machine Learning Research **16**(1), 2023–2049 (2015)
12. Vun, D.S.Y., Bowers, R., McGarry, A.: Vision-based motion capture for the gait analysis of neurodegenerative diseases: A review. Gait & posture **112**, 95–107 (2024)
13. Yang, Y., Liu, P., Sun, Y., Yu, N., Wu, J., Han, J.: A video-based method to classify abnormal gait for remote screening of parkinson's disease. In: 2021 40th Chinese Control Conference (CCC). pp. 3357–3362. IEEE (2021)
14. Zanardi, A.P.J., da Silva, E.S., Costa, R.R., Passos-Monteiro, E., Dos Santos, I.O., Kruel, L.F.M., Peyré-Tartaruga, L.A.: Gait parameters of parkinson's disease compared with healthy controls: a systematic review and meta-analysis. Scientific reports **11**(1), 752 (2021)