STRIDE: Shift-Tolerant Robust Invariance-gated Dual-pathway Estimation for Cross-Site Parkinsonian Gait Severity

Amirhossein Mohammadi[1,2] and Majid Mohajerani[1,2,3]

[1] McGill University, Montreal, Canada
amirhossein.mohammadi@mail.mcgill.ca
[2] Douglas Hospital Research Centre, Montreal, Canada
[3] Mila – Quebec AI Institute, Montreal, Canada

Abstract. Objective, cross-site scoring of Parkinson's disease (PD) gait severity from 3D pose is undermined by pose encoders that fail *silently* on a clinic they were not trained on. On the CARE-PD leave-one-dataset-out (LODO) benchmark we find one cohort scored far worse than the rest, and trace it to a *covariate shift* that buries the cohort's impaired gait in the healthy (grade-0) region of the frozen pretrained features. We introduce **STRIDE**, a single inductive model that fuses a deep temporal pathway with a site-invariant biomechanical pathway through a per-sample, source-calibrated out-of-distribution (OOD) gate, falling back to the invariant signal only where the deep features are untrustworthy. Trained once on a frozen backbone with no target information, STRIDE reaches macro-F1 44.2/42.2 (three-/four-class): +4.0/+6.4 over its own deep pathway, above the strongest published encoder (36.05/38.75), and nearly doubling its macro-F1 on the previously collapsing cohort ($18 \rightarrow 32$). For cross-site severity, diagnosing where a representation cannot be trusted matters as much as adding capacity.

Keywords: Cross-site gait severity · Covariate shift · OOD gating · Ordinal regression

1 Introduction

Vision-based estimation of the Movement Disorder Society Unified Parkinson's Disease Rating Scale (MDS–UPDRS) gait score promises objective, reproducible motor assessment, but its clinical value hinges on generalizing across sites with unseen capture conditions [1, 2]. The CARE-PD benchmark [1] formalizes this as LODO classification of severity $y \in \{0, 1, 2, 3\}$ over four cohorts (BMCLab, T-SDU-PD, PD-GaM, 3DGait), scored by macro-F1 at 3-class ($F1_{0\text{-}2}$) and 4-class ($F1_{0\text{-}3}$), pairing frozen Human3.6M (H36M) [7] pose encoders [3–6] with a lightweight probe. Reported cross-site gains often lean on *transductive* operations

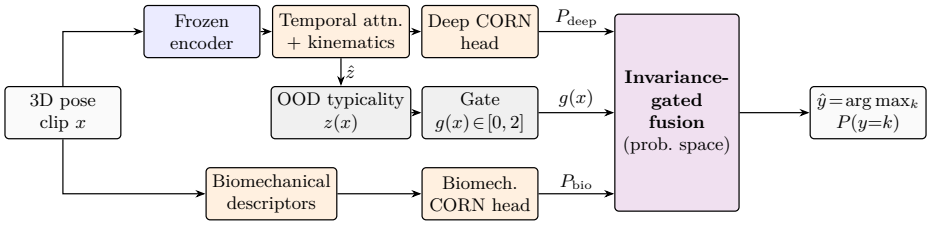

Fig. 1: STRIDE: a frozen-encoder *deep* pathway (blue) and a site-invariant *biomechanical* pathway (orange), each with a CORN head, fused in probability space by a per-sample source-calibrated OOD gate $g(x) \in [0, 2]$. Only the encoder is frozen.

that read target statistics (per-site normalization, test-time adaptation) or on in-domain pretraining. We instead ask, under a strictly *single-model, inductive, leakage-free* protocol with a stock frozen encoder, why one cohort fails so severely and whether it can be recovered without touching the target. We (i) *diagnose* the dominant LODO failure as a covariate shift that collapses an entire cohort into grade-0, and (ii) introduce **STRIDE** (Shift-Tolerant Robust Invariance-gated Dual-pathway Estimation), which recovers it with a per-sample OOD gate that hands off to a site-invariant biomechanical pathway, connecting cross-site robustness to selective prediction [14] and typicality-based OOD detection [11, 13].

2 Diagnosing the Failing Cohort

With the deep pathway trained on the three source cohorts and evaluated on held-out T-SDU-PD (true counts $[96, 118, 167, 0]$), it sends 347 of 381 walks to grade-0, holding grade-1 and grade-2 recall below 8% and collapsing macro-F1 to 18. Restricting predictions to $\{0, 1, 2\}$ does not help (the head predicts almost no grade-3), so the failure is wholesale under-grading, not a grade-3 false-positive or label-prior artifact. A source-calibrated typicality score (the deep features' squared deviation from source statistics, defined in Sec. 3) confirms the mechanism: it is near zero on the two least-shifted held-out cohorts (BMCLab 0.01, PD-GaM 0.04) and large on the two most-shifted (T-SDU-PD 3.02, 3DGait 3.78). Opening a site-invariant biomechanical pathway on this off-manifold cohort pulls predictions back toward the diagonal, raising grade-1 recall 7% $\rightarrow$ 42% and macro-F1 to 32. The transferable severity ordering, in short, resides in the invariant signal rather than in the deep features.

3 STRIDE: The Invariance-Gated Dual-Pathway Network

STRIDE couples two pathways over a single frozen backbone and fuses them with a per-sample gate (Fig. 1); for each LODO fold, *no* statistic of the

Table 1: LODO macro-F1 (%): published frozen H36M encoders (CARE-PD [1]) vs. STRIDE, on the *same* frozen MotionAGFormer. Best **bold**.

Method	3-class F1	4-class F1
MotionBERT [4]	32.25	32.25
MixSTE [5]	35.75	35.75
PoseFormer-V2 [6]	35.75	36.75
MotionAGFormer [3]	36.05	38.75
STRIDE (ours)	**44.2**	**42.2**

held-out cohort (mean, variance, or label prior) enters training or inference, ruling out target BatchNorm, test-time adaptation, and prior correction. The *deep* pathway pools the frozen per-frame features $\phi(x) \in \mathbb{R}^{T\times d}$ (d=512) by temporal self-attention with a kinematic-vigor branch into a rank-consistent CORN [9,10] head P_{deep}; it is strong in-distribution but inherits the fragility of ϕ. The *invariant* pathway maps framerate- and scale-normalized biomechanical descriptors (cadence, arm-swing amplitude/asymmetry, joint range-of-motion, movement vigor, foot clearance), whose handcrafted counterparts rival frozen encoders here [8], through a class-balanced [15] CORN head P_{bio}, decoupled and trained to convergence for a low-variance grader. Each head is decoded from its $K-1$ conditional cut-probabilities $f_k(x) = P(y{>}k \mid y{>}k{-}1)$ into class posteriors by the standard CORN chain rule.

Without using any target statistic, we gate on the deep features' typicality. With the time-pooled, source-standardized feature $\hat{z}$, the typicality $t(x) = \frac{1}{d}\sum_j \hat{z}_j^2$ (≈ 1 in-distribution, inflating off-manifold) is calibrated on the source walks alone into an inductive OOD score and gate,

$$z(x) = \frac{t(x) - m}{s}, \qquad g(x) = g_{\max}\,\sigma\big(a\,(z(x) - \tau)\big), \qquad \sigma(u) = \tfrac{1}{1+e^{-u}}, \tag{1}$$

with m, s the median and standard deviation of t over sources. The decoded posteriors are fused in probability space and renormalized,

$$\begin{aligned} &\tilde{P}(k) = P_{\text{deep}}(k) + g(x)\,P_{\text{bio}}(k) \;\; (k{\in}\{0,1,2\}), \qquad \tilde{P}(3) = \beta\,P_{\text{bio}}(3), \\ &P(y{=}k) = \tilde{P}(k) \Big/ \textstyle\sum_j \tilde{P}(j), \end{aligned} \tag{2}$$

with a fixed $(g_{\max}, a, \tau, \beta){=}(2, 2, 1, 1.5)$ for *all* folds. The large OOD separation makes g a near-hard switch, so the invariant pathway dominates only off-manifold; the ungated β term restores grade-3, which the deep head almost never predicts. Training uses stochastic weight averaging [16].

4 Results and Discussion

Under the CARE-PD LODO protocol (single deterministic run, frozen MotionAGFormer-S, no target statistics), STRIDE reaches 44.2/42.2 macro-F1

Table 2: Per-fold LODO macro-F1 (%) for our pathway variants and pooled-probe baseline, and CARE-PD's Random Forest (RF) where reported. Column-best **bold**; "n/r"=not reported numerically [1].

Metric	Model	BMCLab	T-SDU-PD	PD-GaM	3DGait	Mean
4-class	Repro. baseline	38	28	41	27	33.4
	Biomech. only	33	**33**	**53**	24	35.4
	Deep only	**58**	18	35	**32**	35.8
	+ gate (STRIDE)	**58**	32	50	29	**42.2**
3-class	CARE-PD RF	n/r	**49**	n/r	n/r	n/r
	Repro. baseline	38	28	**54**	36	39.0
	Biomech. only	33	33	53	31	37.4
	Deep only	**58**	18	46	**38**	40.2
	+ gate (STRIDE)	**58**	32	49	**38**	**44.2**

(Table 1), beating the strongest published encoder (+8.2/+3.5) and our matched pooled-probe baseline (+5.2/+8.8); the invariance gate contributes +4.0/+6.4 over the deep pathway alone. The per-fold breakdown (Table 2) shows why per-sample routing is needed: no single pathway wins every cohort. The biomechanical signal leads exactly where the deep pathway collapses (the shifted and grade-3 folds): at 3-class it alone (37.4) already edges the best frozen encoder, and trails it in-distribution, so trusting either everywhere fails. The gate resolves this *selectively*: it opens on the cohort our typicality score flags as most shifted and roughly doubles its macro-F1 ($18 \rightarrow 32$, the recovery of Sec. 2), leaves the in-distribution cohorts untouched, and where the deep head merely misses the rare grade-3 the ungated β term supplies the correction. STRIDE thus attains the best mean not by topping any single cohort but by never collapsing on one.

We report the limits plainly. On the most-shifted 3DGait cohort the biomechanical pathway is itself uninformative, so the gate is mildly net-negative ($32 \rightarrow 29$). A purely handcrafted Random Forest is even stronger on T-SDU-PD (49 vs. our 32, 3-class macro-F1), though CARE-PD reports it "deteriorates in PD-GaM and 3DGait" [1]: the transferable ordering does reside in an invariant signal, which STRIDE captures inside a single frozen-encoder model robust across all four folds rather than a per-cohort feature pipeline. We fix $(g_{\max}, a, \tau)=(2, 2, 1)$ a priori (not the grid-best 45.0), and every setting of the 18-point grid still beats both baselines, so the gain is not target-tuned. Absent the excluded in-domain self-supervised pretraining, the cohort has a genuine low-30s ceiling that STRIDE reaches using no target information. Diagnosing *where* a representation cannot be trusted, rather than adding capacity, thus recovers a cross-site failure a stock encoder cannot, turning a per-sample OOD signal into a principled, leakage-free fallback.

References

1. Adeli, V., Klabučar, I., Rajabi, J., Filtjens, B., Mehraban, S., et al.: CARE-PD: A Multi-Site Anonymized Clinical Dataset for Parkinson's Disease Gait Assessment. arXiv:2510.04312 (2025) 1, 3, 4
2. Sabo, A., Iaboni, A., Taati, B., Fasano, A., Gorodetsky, C.: Evaluating the ability of a predictive vision-based machine learning model to measure changes in gait in response to medication and DBS within individuals with Parkinson's disease. BioMedical Engineering OnLine 22, 120 (2023) 1
3. Mehraban, S., Adeli, V., Taati, B.: MotionAGFormer: Enhancing 3D Human Pose Estimation with a Transformer-GCNFormer Network. In: WACV (2024) 1, 3
4. Zhu, W., Ma, X., Liu, Z., et al.: MotionBERT: A Unified Perspective on Learning Human Motion Representations. In: ICCV (2023) 1, 3
5. Zhang, J., Tu, Z., Yang, J., et al.: MixSTE: Seq2seq Mixed Spatio-Temporal Encoder for 3D Human Pose Estimation in Video. In: CVPR (2022) 1, 3
6. Zhao, Q., Zheng, C., Liu, M., Wang, P., Chen, C.: PoseFormerV2: Exploring Frequency Domain for Efficient and Robust 3D Human Pose Estimation. In: CVPR (2023) 1, 3
7. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments. IEEE TPAMI 36(7), 1325–1339 (2014) 1
8. Adeli, V., Mehraban, S., Ballester, I., et al.: Benchmarking Skeleton-based Motion Encoder Models for Clinical Applications: Estimating Parkinson's Disease Severity in Walking Sequences. In: IEEE FG (2024) 3
9. Shi, X., Cao, W., Raschka, S.: Deep Neural Networks for Rank-Consistent Ordinal Regression Based on Conditional Probabilities. Pattern Analysis and Applications (2023) 3
10. Cao, W., Mirjalili, V., Raschka, S.: Rank consistent ordinal regression for neural networks with application to age estimation. Pattern Recognition Letters 140, 325–331 (2020) 3
11. Hendrycks, D., Gimpel, K.: A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. In: ICLR (2017) 2
12. Lee, K., Lee, K., Lee, H., Shin, J.: A Simple Unified Framework for Detecting Out-of-Distribution Samples and Adversarial Attacks. In: NeurIPS (2018)
13. Nalisnick, E., Matsukawa, A., Teh, Y.W., Lakshminarayanan, B.: Detecting Out-of-Distribution Inputs to Deep Generative Models Using Typicality. arXiv:1906.02994 (2019) 2
14. Geifman, Y., El-Yaniv, R.: Selective Classification for Deep Neural Networks. In: NeurIPS (2017) 2
15. Cui, Y., Jia, M., Lin, T.-Y., et al.: Class-Balanced Loss Based on Effective Number of Samples. In: CVPR (2019) 3
16. Izmailov, P., Podoprikhin, D., Garipov, T., et al.: Averaging Weights Leads to Wider Optima and Better Generalization. In: UAI (2018) 3