REINS: Learning Inertia-Induced Geometry for Physics-Consistent Motion Representation in Clinical Gait Phenotyping

Yiran Ding[1]* **Yufei Zhang**[2] **Zijun Cui**[1]
[1]Michigan State University
[2]Independent Researcher
yiran_ding@alumni.brown.edu, yufeizhang96@outlook.com, cuizijun@msu.edu

Abstract

Parkinson's disease and other gait-related neurological disorders manifest in subtle but biomechanically meaningful changes in how patients move, making computational gait phenotyping a valuable tool for clinical assessment. Yet existing latent representations of motion are typically learned from data alone and miss the biomechanical structure that gives clinical motion its diagnostic meaning, leaving them vulnerable when training cohorts are small or acquisition sites differ. We propose **REINS** (**R**iemannian **E**mbedding for **IN**ertia-induced Motion **S**paces), an Euler-Lagrange informed motion representation that uses the generalized inertia matrix of articulated systems to define a biomechanically grounded geometry on configuration space. Because kinetic energy is itself a quadratic form in velocity weighted by the inertia matrix, this metric is a mechanically determined Riemannian structure, and REINS learns a latent embedding of the resulting structured space. By defining the representation space itself through rigid-body dynamics rather than enforcing physics as a post-hoc penalty, REINS aligns the geometry of motion representation with the same physical laws clinicians rely on to interpret gait. On Parkinson's disease (PD) gait analysis, REINS measures deviation from healthy motion as a Riemannian distance under the inertia-induced metric, and improves healthy-control (HC) vs. PD discrimination and Unified Parkinson's Disease Rating Scale (UPDRS) gait severity correlation. Code is available at https://github.com/DominoAI-Lab/REINS-NeurIPS-2026.

1 Introduction

Human gait and movement carry rich diagnostic information for a range of neurological disorders, including Parkinson's disease (PD), where motor symptoms such as bradykinesia, reduced stride length, and increased stride variability are central to clinical assessment [11, 10]. Recent advances in monocular 3D human pose and motion estimation [30, 24, 31] have made it feasible to assess these behaviors automatically from video, opening the door to scalable, non-intrusive clinical evaluation.

Despite this potential, clinical motion datasets remain limited in scale, cohort diversity, and label quality, which makes purely data-driven motion representations vulnerable to overfitting and poor cross-cohort generalization [1]. Human gait, however, is the motion of an articulated rigid-body system, a biomechanical structure that holds across subjects and sites. Rigid-body dynamics therefore provide a universal, dataset-independent prior that constrains motion representations regardless of how much clinical data is available.

*This work is done when the author is a research intern at MSU.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

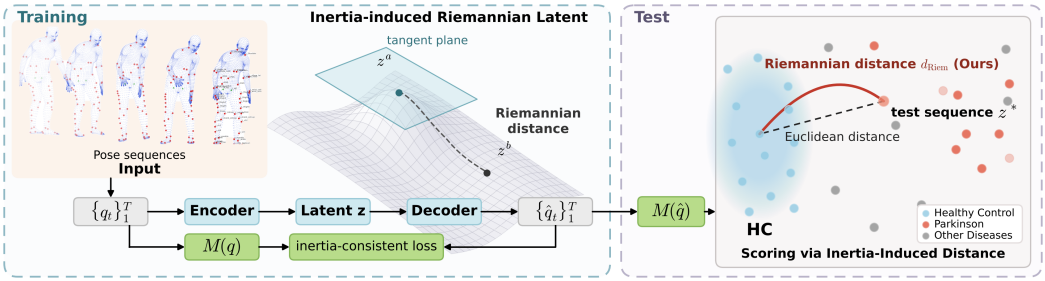

Figure 1: Overview of REINS. Training (left): a pose sequence is encoded into a latent vector and reconstructed by the decoder. The inertia-consistent loss weights the reconstruction velocity error by the body's inertia matrix $M(q)$, so that the latent space inherits the inertia-induced geometry (surface). Distances between embedded sequences z^a and z^b are measured by the Riemannian distance under this geometry. Testing (right): a test sequence is embedded as z^*. The inertia matrix $M(\hat{q})$ of its decoded pose defines the local metric, and the sequence is scored by its Riemannian distance d_{Riem} to the healthy-control (HC) reference embeddings. The score separates subjects with Parkinson's disease from healthy controls and correlates with clinical severity on cohorts not used for training. The Riemannian distance (ours) separates these groups more reliably than Euclidean distance (dashed).

Existing physics-informed motion methods do incorporate physical laws, but typically as a post-hoc trajectory-level regularizer for reconstruction quality [23, 27, 31, 32], without changing the geometry of the representation space itself. A separate line of work in robotics and motion learning has long recognized that the inertia matrix induces a Riemannian metric on configuration space and exploited this geometry for planning, control, and skill learning [12, 14, 5]. Yet using inertia-induced geometry to shape a learned representation of human motion for clinical phenotyping remains largely unexplored. Such a representation is well-suited to clinical phenotyping, where phenotype scores are expected to be grounded in shared mechanical principles, to transfer across acquisition sites and patient cohorts [1], to support subject-level severity tracking when UPDRS ratings are sparse or unavailable [9], and to be obtained from monocular video alone, without requiring specialized motion-capture facilities [24].

In this work, we study physics-informed motion representations through a geometric lens. Because the kinetic energy of a rigid-body system takes the form $\mathcal{T}(q, \dot{q}) = \frac{1}{2}\dot{q}^\top M(q)\dot{q}$, the inertia matrix $M(q)$ is a mechanically determined Riemannian metric on configuration space. We use this inertia-induced geometry as the foundation of representation, rather than enforcing Euler-Lagrange consistency as a post-hoc trajectory constraint. Figure 1 provides an overview of REINS. Our contributions are:

- **A Motion Representation for Parkinson's Gait Phenotyping.** We frame computational gait phenotyping as representation geometry rather than reconstruction quality, and motivate biomechanics as the inductive bias that aligns representation structure with how clinicians read gait.
- **A Geometric Lens of Biomechanics Integration.** We introduce REINS, in which the mechanically determined inertia metric, computed from the body model and pose, structures both the learned motion representation through the inertia-consistent loss and its Riemannian-distance scoring for clinical gait phenotyping.
- **Empirical Evaluation on Clinical PD Gait Phenotyping.** We evaluate REINS on the CARE-PD [1] benchmark across full-data, reduced-data, and cross-cohort transfer regimes, on healthy-control (HC) vs. PD discrimination and UPDRS-gait severity correlation, with a factorial ablation that separates the inertia-consistent loss from the scoring function.

2 Related Works

Computational Gait Analysis for Parkinson's Disease. Automated gait assessment for PD has been studied with various sensing modalities, including inertial measurement units (IMUs) and video cameras [16, 7]. Among these modalities, vision-based methods are non-intrusive, scalable, and more relevant to our setting. They typically extract 2D or 3D pose from clinical recordings and apply supervised classifiers or regressors to predict UPDRS-gait scores [16]. To address label scarcity, some

methods adopt self-supervised pretraining [7], while zero-shot approaches use generative models of healthy motion to flag pathological deviation without severity labels [9]. Beyond single-cohort evaluation, recent work increasingly emphasizes generalization across clinical sites [1]. Nonetheless, existing methods treat motion primarily as a kinematic signal and pursue generalization through architectural and pretraining choices. How mechanical structure, in particular the inertia geometry of rigid-body dynamics, can shape motion representations under clinical data scarcity and cross-site shift remains largely unexamined.

Physics-Informed Human Motion Modeling. Physics provides a strong prior for making estimated, predicted, or generated motion physically plausible. One line of work relies on a physics simulator to incorporate physics, either driving a simulated character to reproduce a target motion with a learned control or imitation policy [27, 28], or refining noisy motion estimates through post hoc trajectory optimization [23, 8]. Another line of work instead integrates physical principles directly into learning models. For example, PhysMoP integrates the rigid-body dynamics equations through neural network design, significantly reducing the error of motion prediction from past observations [32]. PhysPT explicitly models joint torques and contact forces and introduces an Euler-Lagrange residual loss during training to produce physically plausible motion estimates from monocular video [31]. In both cases, the goal is to improve the physical plausibility of motion. They primarily enforce physics as a consistency or regularization constraint at the trajectory level, while only modeling healthy adults. In contrast, we take the latent representation itself as the primary object of study, and ask how the inertia-induced geometry of rigid-body dynamics can structure that representation for clinical motion analysis.

Geometry in Motion Representation Learning. Learning latent representations of motion has been widely studied. For human motion, existing works primarily build on established frameworks whose latent spaces have no mechanical metric [18], such as conditional variational autoencoders that model per-frame transitions [20] or entire sequences [19]. For other dynamical systems, some works constrain the latent states with known dynamics equations for reconstruction and physical parameter estimation [26]. However, these works do not study how distances between latent states should be measured. In robotics, Riemannian metrics derived from the inertia matrix [12, 14] or learned from demonstrations [5] are used for motion learning, control, and planning, but they serve to generate robot motions rather than to compare motions in a learned representation. For learned representations, pullback geometry transfers a data-space metric to the latent space through the decoder [3, 4], but this metric is typically Euclidean, learned, or hand-designed, leaving open which metric suits human motion. To our knowledge, REINS is the first to equip a learned human motion representation with the inertia-induced metric, so that deviation from healthy gait is measured by mechanical effort rather than kinematic extent.

3 Methodology

We consider the task of learning a low-dimensional representation of human motion that is structured by physics rather than by data alone. Given a pose sequence $q_{1:T}$ describing the configuration of an articulated human body across T frames, we seek an encoder ϕ producing a latent $z = \phi(q_{1:T})$ such that distances in latent space reflect physically meaningful deviation between motions. Instead of using physics to refine reconstructed motion as a post-processing step, we use it to shape the geometry of the representation space itself. Section 3.1 reviews the rigid-body dynamics and explains why the inertia matrix defines a Riemannian metric. Section 3.2 defines the inertia-induced motion space. In Section 3.3, we learn a latent embedding of this space with an inertia-consistent reconstruction loss. Section 3.4 defines the score using Riemannian distance in latent space and describes how we use the score for downstream tasks.

3.1 Mechanical Preliminaries

Rigid-Body Dynamics. We treat the human body as an articulated rigid-body system with generalized coordinates $q \in \mathbb{R}^n$ describing root translation and joint angles. Its dynamics are governed by the Euler-Lagrange equation:

$$M(q)\ddot{q} + C(q,\dot{q})\dot{q} + g(q) = \tau + J_C^\top(q)\lambda, \tag{1}$$

where $M(q) \in \mathbb{R}^{n\times n}$ is the generalized inertia matrix, $C(q,\dot{q})\dot{q}$ collects Coriolis and centrifugal forces, $g(q)$ is the gravitational force, τ are the joint torques produced by muscles, and $J_C^\top(q)\lambda$ is the

contact wrench from the environment. The inertia matrix $M(q)$ is the dynamic counterpart of the body's mass distribution at configuration q, since each entry $M_{ij}(q)$ encodes how much of the joint torque τ_i is required to produce a unit acceleration in coordinate j, with M aggregating contributions from all body segments through the kinematic chain.

Inertia as a Riemannian Metric. The same matrix defines the kinetic energy of the system,

$$\mathcal{T}(q, \dot{q}) = \tfrac{1}{2}\dot{q}^\top M(q)\dot{q}. \tag{2}$$

Because $M(q)$ is symmetric and positive definite for any physically realizable body configuration, it defines a Riemannian metric on the configuration space [12]. Geometrically, the unit ball of this metric stretches along directions in which the body has low effective inertia and contracts along directions in which it has high effective inertia, so distances under $M(q)$ are induced by the kinetic-energy metric and measure the mechanical effort of a motion rather than its kinematic extent. In gait, this geometry directly reflects the way a clinician interprets movement: changes in stride length, posture, or limb swing alter the body's effective inertia and therefore the mechanical work required to produce the same kinematic trajectory. At the same angular speed, hip motion has about four times the kinetic energy of shoulder motion and about 260 times that of wrist motion. Euclidean joint-angle distance weights all three equally. Posture also affects the metric: flexing the knee to 90° lowers the effective hip inertia from 2.0 to 1.2 $\mathrm{kg\,m^2}$.

Choice of Inertia Matrix. We use $M(q)$ to define the geometry, rather than torques or muscle quantities, for two reasons. First, this geometry follows directly from rigid-body dynamics. The body is an articulated rigid-body system governed by the Euler-Lagrange equation above. Its kinetic energy is $\frac{1}{2}\dot{q}^\top M(q)\dot{q}$, where $M(q)$ defines a symmetric positive-definite form on velocities and hence a Riemannian metric [2]. Second, $M(q)$ can be computed in closed form from the pose and body model. In contrast, computing torques requires contact wrenches, which can be measured with force plates or estimated with a learned solver [15]. Musculoskeletal models driven by surface EMG [29] require signals that CARE-PD does not provide.

3.2 Defining the Inertia-Induced Motion Space

Limitations of Euclidean Motion Representations. Most motion representations operate in Euclidean configuration space, implicitly assuming an isotropic metric $\|v\|^2 = v^\top v$ in which all directions of motion are weighed equally (Figure 2a). Yet articulated motion does not have this property, since moving the trunk requires substantially more mechanical effort than moving a wrist at the same kinematic speed, because the inertia loaded by trunk motion is far larger than that loaded by wrist motion. As a consequence, two trajectories that look similar in Euclidean coordinate space can correspond to very different mechanical motions (Figure 2b), and trajectories that look different can correspond to the same underlying movement strategy at different body sizes. In clinical settings, the deviations associated with Parkinson's disease, including reduced stride length and altered limb swing, are interpretable as departures from the body's mechanical norms [10], and an isotropic Euclidean distance loses the mass-distribution information that makes such deviations interpretable.

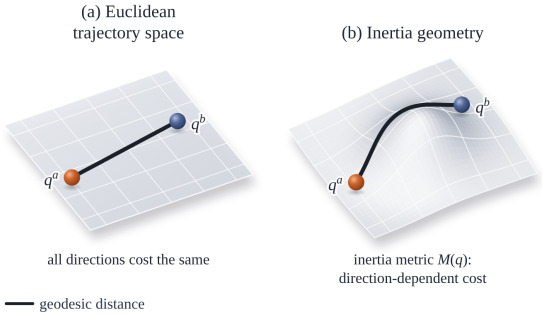

Figure 2: Euclidean versus inertia-induced motion geometry. Two pose sequences, q^a and q^b, are marked in both panels, and the solid curve is the shortest path between them under the metric used in that panel. (a) Euclidean trajectory space treats all directions of motion equally and ignores the articulated mass distribution. (b) The inertia matrix $M(q)$ assigns direction-dependent costs to motion, so both the shortest path between the two sequences and its length change, and the length measures the mechanical effort of transforming one sequence into the other.

We therefore represent motion in the inertia-induced motion space, the physically realizable motions equipped with the inertia metric (Figure 2b), where the distance between two motions reflects the mechanical effort needed to transform one into the other.

3.3 Learning a Latent Embedding of the Inertia-Induced Space

Having defined the inertia-induced motion space (Figure 2b), we seek to learn a latent representation of this space that preserves its inertia geometry.

Latent Embedding. We learn an encoder

$$z = \phi_\theta(q_{1:T}),$$

and a decoder

$$\hat{q}_{1:T} = \psi_\theta(z),$$

where $z \in \mathbb{R}^k$ with $k \ll nT$, the dimension of a pose sequence with n generalized coordinates per frame and T frames. The goal is not merely to reconstruct motion, but to preserve the mechanical structure defined in Section 3.2.

Inertia-Consistent Reconstruction. We measure reconstruction error under inertia geometry:

$$\mathcal{L}_{\text{ine}} = \sum_{t=1}^{T} (\dot{q}_t - \dot{\hat{q}}_t)^\top M(q_t)(\dot{q}_t - \dot{\hat{q}}_t). \tag{3}$$

This is the squared norm of the velocity error under the inertia metric, i.e., twice the kinetic energy of the velocity error evaluated at the ground-truth configuration. Unlike Euclidean reconstruction error, $\mathcal{L}_{\text{ine}}$ weighs each direction of motion by the body's effective inertia in that direction, so reconstruction quality is measured in the same units as the mechanical energy clinicians associate with movement. Unlike learned metrics, the inertia matrix is uniquely determined by rigid-body mechanics.

Overall Objective. The final training objective is

$$\mathcal{L} = \mathcal{L}_{\text{recon}} + \mathcal{L}_{\text{force}} + \mathcal{L}_{\text{contact}} + \lambda_{\text{ine}}\, \mathcal{L}_{\text{ine}}, \tag{4}$$

where $\mathcal{L}_{\text{recon}}$, $\mathcal{L}_{\text{force}}$, and $\mathcal{L}_{\text{contact}}$ are the pose-reconstruction, force, and contact losses inherited from PhysPT [31], and $\mathcal{L}_{\text{ine}}$ is our inertia-consistent reconstruction loss replacing PhysPT's Euler-Lagrange residual term. Under this objective, the learned embedding z serves as a low-dimensional representation that preserves inertia geometry.

3.4 Scoring Function via Inertia-Induced Distance

Scoring Overview. Scoring takes place in the latent space. Let $\mathcal{Z}_{\mathcal{N}}$ be the reference embeddings of the K healthy-control training subjects, and let $z^* = \phi_\theta(q^*_{1:T})$ be the embedding of a test sequence. We define the score of a test sequence as the Riemannian distance from z^* to the nearest healthy reference embedding, measured under the inertia geometry of Section 3.2. A larger score indicates a greater deviation from the healthy reference motions in terms of inertia-weighted motion differences.

The same score serves both tasks. For anomaly detection, we rank subjects by their scores and use AUROC to measure how well the ranking separates subjects with Parkinson's disease or other movement disorders from held-out healthy controls. For severity assessment, we correlate the same scores with clinical UPDRS-gait ratings on cohorts not used for training. In both cases, we compute the scores without clinical labels.

Riemannian Distance as the Score. Formally, the score of a test embedding z^* is its distance to the nearest healthy reference embedding under the inertia geometry,

$$s(z^*) = \min_{z \in \mathcal{Z}_{\mathcal{N}}} d_{\text{Riem}}(z^*, z), \tag{5}$$

where d_{Riem} denotes the Riemannian distance induced by the inertia metric in latent space. It is defined as the length of the shortest curve, the geodesic, connecting the two embeddings,

$$d_{\text{Riem}}(z^*, z) = \min_{\gamma} \int_0^1 \left\| \dot{\gamma}(t) \right\|_{g(\gamma(t))} dt, \tag{6}$$

where $\gamma\colon [0,1] \to \mathbb{R}^k$ connects $\gamma(0) = z^*$ and $\gamma(1) = z$, $\|\dot{\gamma}\|_g = \sqrt{\dot{\gamma}^\top g\, \dot{\gamma}}$ measures each step along the curve under g, and g is the inertia metric in latent coordinates, defined below.

Pulling Back the Inertia Metric. Although we incorporate the inertia loss when learning the latent embeddings (Section 3.3), directly computing the L2 distance between embeddings is still a Euclidean

distance rather than the Riemannian distance defined in Eq. 6. The difference lies in the metric $g(z)$, which determines the length of local displacements along a path in latent space. Unlike the Euclidean metric, which weights all latent directions equally, $g(z)$ accounts for the inertia associated with changes in the decoded poses. Figure 1 (right) illustrates the resulting difference between Euclidean and Riemannian distances. To obtain $g(z)$, we pull back the inertia metric from pose space to latent space through the decoder [3, 4].

Specifically, a latent displacement Δz at an embedding z moves the decoded poses $\hat{q} = \psi_\theta(z)$ by $\Delta q = J(z)\Delta z$, where $J(z)$ is the Jacobian of the decoder at z, computed by forward-mode automatic differentiation. Using the same inertia weighting as in $\mathcal{L}_{\text{ine}}$ (Section 3.3), the squared displacement $\Delta q^\top M(\hat{q})\Delta q$ can be expressed in latent coordinates as $\Delta z^\top g(z)\Delta z$, where

$$g(z) = J(z)^\top M(\hat{q})\, J(z), \tag{7}$$

and $g(z)$ is the inertia metric in latent coordinates at z. It is a $k \times k$ symmetric positive-definite matrix that varies with z because both J and $M(\hat{q})$ depend on z. Here $M(\hat{q})$ is block-diagonal, with one block $M(\hat{q}_t)$ for each of the T decoded frames. The quadratic form $\Delta q^\top M(\hat{q})\Delta q$ therefore sums the per-frame terms, as in $\mathcal{L}_{\text{ine}}$. The matrix $g(z)$ is $k \times k$ and symmetric. It is positive definite if $J(z)$ has full column rank, which we assume for the trained decoder [4]. It varies with z because both J and $M(\hat{q})$ depend on z. This computation is exact and is performed only once for each embedding.

Graph-Based Approximation of Riemannian Distance. Computing the Riemannian distance in Eq. 6 requires optimizing over continuous curves, which is computationally prohibitive in practice. We therefore approximate the Riemannian distance in Eq. 6 using a shortest path on a graph [25, 6].

The graph is a κ-nearest-neighbor graph (κ=20) with the reference embeddings in $\mathcal{Z}_\mathcal{N}$ and the embeddings of one test subject as nodes. The metric g is computed exactly at every node using Eq. 7. Each edge is a straight segment between two nodes z_i and z_j. Its length under g is given by the integral in Eq. 6 along the segment. We approximate this length using the average of the endpoint metrics, yielding $\sqrt{\Delta z^\top \frac{1}{2}\big(g(z_i) + g(z_j)\big)\Delta z}$ where $\Delta z = z_i - z_j$. The sum of edge lengths along the shortest path between two nodes approximates their Riemannian distance. Running Dijkstra's algorithm with all reference embeddings as sources returns the minimum in Eq. 5 for every node in a single pass. We build the graph separately for each test subject so that a subject's score does not depend on other test subjects. All reported REINS results use this procedure. Appendix C describes a constant approximation of g fitted once on the reference embeddings and reports sensitivity to the graph construction.

4 Experiments

Datasets and Tasks. We use CARE-PD [1], a multi-cohort Parkinson's-disease gait benchmark. Its DNE (Digitized Neurological Examination) cohort contains 28 healthy controls (HC), 30 Parkinson's-disease (PD) subjects, and 39 subjects with other movement-related diseases, with no UPDRS-gait severity labels. Four other cohorts, PD-GaM, BMCLab, T-SDU-PD, and 3DGait, have subject-level UPDRS-gait ratings and were collected at different sites under different protocols. We fine-tune the PhysPT encoder-decoder described in Section 3.3 [31] using only DNE healthy controls. We evaluate HC vs. PD discrimination on DNE and, without further fine-tuning, correlate the scores with UPDRS-gait ratings on the four labeled cohorts.

Baselines. Employing the same PhysPT-pretrained encoder-decoder backbone [31], we compare three variations of our approach that differ in the geometry of the resulting latent space and the scoring distance used at evaluation (Figure 1):

- *Data-driven Euclidean* fine-tunes with $\mathcal{L}_{\text{recon}}$ only when obtaining the latent representation and scores each window by its Mahalanobis distance to the mean of the reference windows, with covariance estimated from the same windows using Ledoit-Wolf shrinkage [13].
- *Physics-supervised Euclidean* fine-tunes with the original PhysPT objective ($\mathcal{L}_{\text{recon}} + \mathcal{L}_{\text{force}} + \mathcal{L}_{\text{contact}} + \mathcal{L}_{\text{euler}}$), with the same Euclidean read-out.
- *REINS (ours)* fine-tunes with the REINS objective ($\mathcal{L}_{\text{recon}} + \mathcal{L}_{\text{force}} + \mathcal{L}_{\text{contact}} + \lambda_{\text{ine}}\mathcal{L}_{\text{ine}}$, $\lambda_{\text{ine}} = 0.1$), in which our inertia-consistent loss $\mathcal{L}_{\text{ine}}$ replaces the Euler-Lagrange residual. Our REINS defines an inertia-induced Riemannian latent that is scored by Riemannian distance under the pulled-back inertia metric of Section 3.4.

We additionally consider three published motion encoders, VPoser [18], GaitForeMer [7], and MotionBERT [33], with frozen backbones, scored as described in Appendix D.

Implementation Details. All conditions are fine-tuned from the PhysPT checkpoint pretrained on AMASS [31, 17], using Adam with learning rate 10^{-6}, weight decay 10^{-4}, batch size 16, and a step learning-rate schedule (decay factor 0.8 every 5 epochs). All conditions are fine-tuned for up to 50 epochs, with early stopping on held-out HC mean per-joint position error (MPJPE) (patience 10, minimum 20 epochs). All optimization hyperparameters are held fixed across conditions and dataset-size regimes so that downstream differences are attributable to the loss configuration or training-set size rather than tuning. Experiments use a single NVIDIA RTX 3090 GPU. Every recording is split into 16-frame windows at 30 fps [31].

For input, each window is one input sequence $q_{1:T}$ with $T{=}16$ for both fine-tuning and evaluation. At evaluation, every window of a subject is embedded and scored as in Section 3.4, and the subject's score is the median of its window-level scores. For latent space, before scoring we reduce the latent dimensionality using PCA fitted on the reference embeddings $\mathcal{Z}_{\mathcal{N}}$, retaining 95% of the variance, and remove the leading component, which corresponds to global translation along the walking direction.

Evaluation Protocol. Of the 28 DNE healthy controls, 8 are held out for testing, and K of the remaining 20 are used for fine-tuning and to form the reference embeddings $\mathcal{Z}_{\mathcal{N}}$ defined in Section 3.4. PD and other-disease subjects are used only as anomalies at evaluation. Subject-level AUROC is computed on the 8 held-out HC against the 30 PD subjects (HC/PD) and against the 30 PD plus 39 other-disease subjects (HC/all). The main results use $K{=}20$, and Section 4.1 reduces K to study data scarcity. We use one split of the healthy controls for the main results (the main split). Table 1 reports 95% bootstrap confidence intervals over the test subjects, with the model held fixed (Appendix B). In Section 4.1, we vary the training subjects across random splits. Appendix B describes the bootstrap and random-split procedures and reports confidence intervals, HC/all results, and the factorial ablation for all K, together with a paired bootstrap test of the differences between methods. On the four UPDRS-labeled cohorts we report Spearman ρ and Kendall τ-b between subject-level scores and UPDRS-gait ratings to assess whether the score increases with clinical severity.

Experiments Overview. In Section 4.1, we test whether the Riemannian score separates PD from HC better than the Euclidean readouts and three published motion encoders (Table 1), disentangle the contributions of loss and scoring (Table 2), and evaluate performance under limited healthy training data (Table 3). In Section 4.2, we correlate the resulting scores with UPDRS-gait ratings for patients from four cohorts in a zero-shot setting (Table 4).

4.1 Pathological Motion under Riemannian vs Euclidean Latent Geometry

TL;DR: REINS changes two things relative to a standard latent space: the encoder is fine-tuned with the inertia-consistent loss, and subjects are scored by Riemannian rather than Euclidean distance. We test whether either change alone separates subjects with Parkinson's disease from healthy controls on DNE. The REINS encoder with Euclidean scoring reaches 0.62 AUROC and the standard PhysPT encoder with Riemannian scoring reaches 0.59, while combining the two changes gives 0.91 (Tables 1 and 2). For Riemannian scoring, using the decoder Jacobian alone gives 0.73 AUROC, which rises to 0.91 when the inertia matrix is included. The separation holds even with a single healthy training subject (Table 3). Detailed experimental results and analyses are presented below.

Comparison with Baselines. Table 1 (left) shows that on the main split REINS separates PD from HC, with AUROC 0.9083 on HC/PD and 0.8768 on HC/all, while the two Euclidean readouts of the same backbone remain near chance. The physics-supervised readout is near chance despite being trained with the full PhysPT physics objective, suggesting that trajectory-level physics losses alone do not make pathological motion separable in the latent space. On the main split, REINS improves AUROC over MotionBERT by 0.14 on HC/PD and 0.12 on HC/all.

Sensitivity to the Subject Split. The main split uses 20 healthy controls for training and holds out 8 for testing. To assess sensitivity to this split, we repeat the $K{=}20$ experiment on ten random splits and fine-tune the model on each split. The right half of Table 1 reports the mean and standard deviation of AUROC across the ten splits. Across the tested splits, REINS achieves mean AUROCs of 0.8804 on HC/PD and 0.8663 on HC/all, with standard deviations of 0.027 and 0.020, respectively, compared with mean AUROCs of 0.7621 and 0.7451 for MotionBERT.

Table 1: Anomaly detection AUROC with $K=20$ HC training subjects. Left: the main split, with 95% bootstrap confidence intervals over the evaluation subjects. Right: mean $\pm$ standard deviation over ten random splits of the healthy controls into training and held-out subjects. HC/PD = HC vs PD, HC/all = HC vs PD+other. Riemannian scoring in our M-induced Riemannian latent space outperforms Euclidean baselines and three published motion encoders.

	Main split [95% CI]		**Ten random splits** (mean $\pm$ sd)	
Method	HC/PD	HC/all	HC/PD	HC/all
Euclidean latent space distance: Euclidean				
Data-driven	0.5750 [0.3792, 0.7625]	0.6087 [0.4312, 0.7736]	0.5633 $\pm$ 0.050	0.5988 $\pm$ 0.034
Physics-supervised	0.5167 [0.3167, 0.7125]	0.5380 [0.3587, 0.7156]	0.5242 $\pm$ 0.049	0.5442 $\pm$ 0.033
Published motion encoders distance: Riemannian				
VPoser [18]	0.5708 [0.3750, 0.7708]	0.6051 [0.4312, 0.7790]	0.5888 $\pm$ 0.046	0.5998 $\pm$ 0.031
GaitForeMer [7]	0.7125 [0.5250, 0.8792]	0.7337 [0.5815, 0.8768]	0.7071 $\pm$ 0.033	0.7205 $\pm$ 0.036
MotionBERT [33]	0.7667 [0.5458, 0.9208]	0.7572 [0.5870, 0.8986]	0.7621 $\pm$ 0.025	0.7451 $\pm$ 0.029
Riemannian latent space (M-induced, ours) distance: Riemannian				
REINS	**0.9083** [0.8042, 0.9833]	**0.8768** [0.7736, 0.9529]	**0.8804** $\pm$ 0.027	**0.8663** $\pm$ 0.020

Table 2: Factorial ablation of the fine-tuning objective and the scoring distance. Subject-level HC/PD AUROC at $K=20$ with 95% bootstrap intervals. The rows correspond to the physics-supervised encoder (without $\mathcal{L}_{\text{ine}}$) and the REINS encoder (with $\mathcal{L}_{\text{ine}}$) in Table 1. The $J^\top J$ Riemannian distance uses the same procedure without the inertia matrix.

Fine-tuning	**Euclidean**	$J^\top J$ **Riemannian**	$J^\top M J$ **Riemannian**
Physics-supervised (w/o $\mathcal{L}_{\text{ine}}$)	0.5167 [0.3167, 0.7125]	0.5542 [0.3583, 0.7750]	0.5917 [0.3833, 0.7708]
REINS (w/ $\mathcal{L}_{\text{ine}}$)	0.6208 [0.4250, 0.8000]	0.7333 [0.5375, 0.8667]	**0.9083** [0.8042, 0.9833]

Separating the Loss from the Scoring. Table 1 changes the fine-tuning objective and the scoring distance together, so it is unclear how much each contributes to the gain. In Table 2, we evaluate each encoder with three scoring distances. The two encoders are the physics-supervised encoder, trained with the original PhysPT objective, and the REINS encoder, trained with $\mathcal{L}_{\text{ine}}$. The highest AUROC requires both the inertia loss for training and the inertia metric for scoring. Scoring the physics-supervised encoder with the Riemannian distance raises AUROC from 0.5167 to 0.5917. The REINS encoder with the Euclidean readout gives 0.6208, while combining the REINS encoder and Riemannian distance gives 0.9083. For the REINS encoder, the $J^\top J$ and $J^\top M J$ Riemannian distances yield AUROCs of 0.7333 and 0.9083. Adding the inertia matrix therefore improves AUROC by 0.18, compared with 0.11 from replacing the Euclidean readout with the $J^\top J$ Riemannian distance. This is consistent with the clinical interpretation of bradykinesia and reduced stride length as departures from the body's mechanical norms [10], which the inertia geometry encodes through direction-dependent motion costs. Appendix A compares pose and kinetic-energy errors to assess how the fine-tuning objective affects the decoded motion. The additional conditions include training from scratch, using the pretrained checkpoint without fine-tuning, and fine-tuning with the Euler–Lagrange residual removed from the PhysPT objective. Replacing the residual with $\mathcal{L}_{\text{ine}}$ reduces kinetic-energy error from 193.4% to 22.7% and lowers MPJPE by 0.19 mm.

Physical Prior under Clinical Data Scarcity. A defining property of clinical motion data is that it is fundamentally limited in scale: enrolling additional subjects requires clinical recruitment, and per-subject recording is constrained by patient comfort and protocol cost. The relevant question for clinical deployment is therefore not how a method performs at the size of a curated mocap corpus, but how it degrades as the available training data shrinks. We evaluate this by re-training each method on progressively reduced fractions of the DNE training cohort and re-measuring downstream HC vs. PD discrimination. The inertia matrix $M(q)$ is mechanically determined and is not estimated from the training set. We hypothesize that it remains a useful inductive bias as the number of healthy training subjects decreases, when purely data-driven fine-tuning has less data from which to learn structure.

Table 3: Data efficiency: HC/PD AUROC across HC training-subject count K. Even at $K{=}1$, our REINS (M-induced Riemannian) substantially outperforms both Euclidean baselines. $K{=}20$ is the main reported setting (Table 1). Table 6 reports the confidence intervals and HC/all results.

Method	$K{=}20$	$K{=}15$	$K{=}10$	$K{=}5$	$K{=}3$	$K{=}1$
Euclidean latent space	distance: Euclidean					
Data-driven	0.5750	0.5708	0.5333	0.5250	0.5375	0.4292
Physics-supervised	0.5167	0.5250	0.5125	0.4583	0.4792	0.4417
Riemannian latent space (M-induced, ours)	distance: Riemannian					
REINS	**0.9083**	**0.8458**	**0.8125**	**0.7792**	**0.7583**	**0.7375**

Following the experimental protocol described above, we draw K training subjects from the 20 non-held-out HC, and their windows form the reference embeddings $\mathcal{Z}_{\mathcal{N}}$. For each K, the model is fine-tuned from the same PhysPT pretrained checkpoint, with all training and optimization hyper-parameters held fixed across K so that any difference in downstream performance is attributable to training-set size rather than tuning. We use $K \in \{1, 3, 5, 10, 15, 20\}$, where the training subjects are nested ($K{=}1 \subset K{=}3 \subset K{=}5 \subset K{=}10 \subset K{=}15 \subset K{=}20$) so that smaller-$K$ runs see a strict subset of the data seen by larger-K runs and cross-K comparisons are not confounded by sampling variance. Table 3 reports HC/PD AUROC for each K, with $K{=}20$ the main setting of Table 1.

Across all training-set sizes, REINS yields higher AUROC than both Euclidean baselines (Table 3). The gains range from 0.22 to 0.33 over the data-driven baseline and from 0.28 to 0.39 over the physics-supervised baseline, with no monotonic trend in K. Four Euclidean results fall below 0.5 (0.43–0.48). For 8 HC and 30 PD test subjects, the approximate 95% null range for AUROC is $[0.28, 0.72]$. These results are therefore consistent with chance, and their confidence intervals include 0.5 (Table 6). REINS, by contrast, retains AUROC 0.7375 at $K{=}1$. Since the inertia matrix $M(q)$ is determined by rigid-body mechanics rather than estimated from data, a representation grounded in this geometry continues to organize motion meaningfully even when the optimizer sees almost no healthy examples. At $K{=}1$, the reference embeddings $\mathcal{Z}_{\mathcal{N}}$ contain 97 windows from a single subject selected using the split seed, so the reported result depends on the choice of reference subject. In a clinical-deployment regime, where each additional healthy control subject costs recruitment and protocol time, this property is what makes the representation usable.

4.2 Zero-Shot Severity Correlation across Heterogeneous Cohorts

A clinically deployed gait phenotyping system is evaluated on cohorts and acquisition sites unseen at training time. We test whether REINS, fine-tuned only on DNE healthy controls, produces severity scores that correlate with UPDRS-gait ratings on four external UPDRS-gait-labeled cohorts, PD-GaM, BMCLab, T-SDU-PD, and 3DGait [1], without any fine-tuning, label exposure, or per-cohort calibration on these target cohorts. The hypothesis is that a representation grounded in body-level inertia, which is shared across cohorts and sites, transfers more reliably than representations whose structure is learned from cohort-specific kinematic patterns. REINS scores correlate positively with UPDRS-gait ratings on all four cohorts, with significant Spearman correlations on three. Neither Euclidean readout shows a significant correlation on any cohort (Table 4).

Zero-Shot Protocol. For each test subject in the four target cohorts, we extract REINS embeddings using the same DNE-fine-tuned encoder as in Section 4.1, and take the median of the window-level Riemannian distances to the DNE HC reference embeddings $\mathcal{Z}_{\mathcal{N}}$ as the subject-level score. Each target subject's graph contains only the reference windows and that subject's windows, as in Section 3.4. We then measure the rank correlation between subject-level scores and ground-truth UPDRS-gait ratings using Spearman ρ and Kendall τ-b. We report the latter because UPDRS-gait is ordinal and contains ties at each integer level. No supervision from the four cohorts enters the encoder or the score. Asterisks in Table 4 indicate rank correlations significantly different from zero ($p{<}0.05$, Appendix B).

Results. Table 4 shows that the Riemannian score correlates with UPDRS-gait severity on cohorts not used for training. REINS produces positive rank correlations on all four cohorts, while both Euclidean baselines, despite sharing the same encoder backbone, hover in a substantially lower range. For the Riemannian score, Spearman correlation is significant ($p{<}0.05$) on PD-GaM, BMCLab

Table 4: Zero-shot UPDRS-gait severity correlation on four heterogeneous UPDRS-gait-labeled cohorts. The model is fine-tuned only on DNE healthy controls (K=20) with no labels or samples from the target cohorts entering the encoder. Subject-level Spearman ρ and Kendall τ-b between Riemannian distance to the DNE HC reference and ground-truth UPDRS-gait scores. Asterisks indicate rank correlations significantly different from zero ($p<0.05$, Appendix B). The subject is the unit of analysis, and n denotes the number of subjects in each cohort.

Method	**PD-GaM** (n=30)		**BMCLab** (n=23)		**T-SDU-PD** (n=14)		**3DGait** (n=43)	
	ρ	τ	ρ	τ	ρ	τ	ρ	τ
Euclidean latent space	distance: Euclidean							
Data-driven	0.2802	0.1648	0.2200	0.1972	0.2569	0.2405	0.2138	0.1891
Physics-supervised	0.2510	0.1405	0.2895	0.2648	0.2337	0.1243	0.1700	0.1567
Riemannian latent space (M-induced, ours)	distance: Riemannian							
REINS	**0.5649***	**0.4582***	**0.5742***	**0.4400***	**0.4308**	**0.3078**	**0.3154***	**0.2132**

and 3DGait, but not on T-SDU-PD (n=14). Kendall τ-b is significant on PD-GaM and BMCLab, whereas no Euclidean baseline shows a significant correlation on any cohort. The fact that this gap holds on every target cohort indicates that the underlying severity signal carried by the Riemannian distance is not specific to the cohort the encoder was trained on. The correlations also do not collapse despite substantial heterogeneity across PD-GaM, BMCLab, T-SDU-PD, and 3DGait in subject count, recording protocol, and severity distribution. This is the behaviour one would expect if the score reflects departures from the body's mechanical norms [10] rather than cohort-specific kinematic patterns, since rigid-body dynamics depend on the articulated body itself rather than on the camera or laboratory in which it is recorded.

5 Conclusion

We present REINS, a motion representation that preserves the biomechanical structure of motion in the learned latent space by using the Riemannian geometry induced by articulated rigid-body inertia. On the CARE-PD [1] benchmark, REINS substantially outperforms data-driven and physics-supervised Euclidean baselines in distinguishing healthy controls from subjects with Parkinson's disease. To test whether the advantage holds when healthy training data are scarce, we reduce the number of healthy training subjects from 20 to one. REINS still outperforms the Euclidean baselines with a single training subject. Its scores also correlate with UPDRS-gait ratings in zero-shot evaluations on four heterogeneous external cohorts without per-cohort calibration. These results therefore support using the inertia metric as a common basis for comparing gait across training-set sizes and acquisition sites. We compute the inertia matrix from the body model and pose using rigid-body mechanics, independently of the training data. REINS uses this geometry to characterize how a person's gait departs from healthy movement, a comparison that is central to clinical gait assessment.

Limitations and Future Directions. We compute the inertia matrix using the mean SMPL shape, as CARE-PD does not provide subject-specific shape parameters [1]. Finer-grained personalization of mass distribution from clinically measurable anthropometrics (limb-length, body composition) may further improve the calibration of motion costs at the subject level. Our scoring is computed against a healthy-control reference cohort drawn from one dataset (CARE-PD/DNE), and broader reference cohorts spanning age, body size, and recording protocol would further strengthen the population-level calibration of the resulting severity score. Further work could quantify how uncertainty in monocular pose estimates [24] affects $M(q)$ and the resulting scores. Our evaluation uses one benchmark and one fine-tuned backbone [31], and the protocol can be extended to other motion priors. Beyond the inertia matrix studied here, muscle moment arms $R(q) = \partial l_{MT}/\partial q$, computed from pose in a scaled musculoskeletal model [22], could be used to construct an additional mechanically meaningful quadratic form. Finally, the representation is currently evaluated through subject-level summary scores aggregated as the median of window-level Riemannian distances. Per-window or temporally resolved variants of the same Riemannian distance may surface stride-level patterns that carry clinical information but are averaged out at the subject level.

Acknowledgments and Disclosure of Funding

This work was supported by Michigan State University.

References

[1] Vida Adeli, Ivan Klabucar, Javad Rajabi, Benjamin Filtjens, Soroush Mehraban, Diwei Wang, Hyewon Seo, Trung-Hieu Hoang, Minh N Do, Candice Muller, et al. Care-pd: A multi-site anonymized clinical dataset for parkinson's disease gait assessment. *arXiv preprint arXiv:2510.04312*, 2025.

[2] Vladimir Igorevich Arnold, Karen Vogtmann, and Alan Weinstein. *Mathematical methods of classical mechanics*, volume 60. Springer, 1989.

[3] Georgios Arvanitidis, Lars Kai Hansen, and Søren Hauberg. Latent space oddity: on the curvature of deep generative models. *arXiv preprint arXiv:1710.11379*, 2017.

[4] Georgios Arvanitidis, Søren Hauberg, and Bernhard Schölkopf. Geometrically enriched latent spaces. *arXiv preprint arXiv:2008.00565*, 2020.

[5] Hadi Beik-Mohammadi, Søren Hauberg, Georgios Arvanitidis, Gerhard Neumann, and Leonel Rozo. Learning riemannian manifolds for geodesic motion skills. *arXiv preprint arXiv:2106.04315*, 2021.

[6] Nutan Chen, Francesco Ferroni, Alexej Klushyn, Alexandros Paraschos, Justin Bayer, and Patrick van der Smagt. Fast approximate geodesics for deep generative models. In *International Conference on Artificial Neural Networks*, pages 554–566. Springer, 2019.

[7] Mark Endo, Kathleen L Poston, Edith V Sullivan, Li Fei-Fei, Kilian M Pohl, and Ehsan Adeli. Gaitforemer: Self-supervised pre-training of transformers via human motion forecasting for few-shot gait impairment severity estimation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 130–139. Springer, 2022.

[8] Erik Gärtner, Mykhaylo Andriluka, Hongyi Xu, and Cristian Sminchisescu. Trajectory optimization for physics-based reconstruction of 3d human pose from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13106–13115, 2022.

[9] Xiaodong Guan, Robert James Gray, Ashwani Jha, and Parashkev Nachev. Zero-shot gait classification with diffusion models. In *Medical Imaging with Deep Learning-Short Papers*, 2025.

[10] Jeffrey M Hausdorff. Gait dynamics in parkinson's disease: common and distinct behavior among stride length, gait variability, and fractal-like scaling. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 19(2), 2009.

[11] Jeffrey M Hausdorff, Merit E Cudkowicz, Renée Firtion, Jeanne Y Wei, and Ary L Goldberger. Gait variability and basal ganglia disorders: stride-to-stride variations of gait cycle timing in parkinson's disease and huntington's disease. *Movement disorders*, 13(3):428–437, 1998.

[12] Noémie Jaquier and Tamim Asfour. Riemannian geometry as a unifying theory for robot motion learning and control. In *The International Symposium of Robotics Research*, pages 395–403. Springer, 2022.

[13] Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of multivariate analysis*, 88(2):365–411, 2004.

[14] Yiming Li, Jiacheng Qiu, and Sylvain Calinon. A riemannian take on distance fields and geodesic flows in robotics. *The International Journal of Robotics Research*, page 02783649261420233, 2024.

[15] Xinpeng Liu, Junxuan Liang, Zili Lin, Haowen Hou, Yong-Lu Li, and Cewu Lu. Imdy: Human inverse dynamics from imitated observations. In *International Conference on Learning Representations*, volume 2025, pages 1209–1229, 2025.

[16] Mandy Lu, Kathleen Poston, Adolf Pfefferbaum, Edith V Sullivan, Li Fei-Fei, Kilian M Pohl, Juan Carlos Niebles, and Ehsan Adeli. Vision-based estimation of mds-updrs gait scores for assessing parkinson's disease motor severity. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 637–647. Springer, 2020.

[17] 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*, pages 5442–5451, 2019.

[18] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10975–10985, 2019.

[19] Mathis Petrovich, Michael J Black, and Gül Varol. Action-conditioned 3d human motion synthesis with transformer vae. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10965–10975. IEEE, 2021.

[20] Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J Guibas. Humor: 3d human motion model for robust pose estimation. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11468–11479. IEEE, 2021.

[21] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1010–1019, 2016.

[22] Michael A Sherman, Ajay Seth, and Scott L Delp. What is a moment arm? calculating muscle effectiveness in biomechanical models using generalized coordinates. In *Proceedings of the... ASME Design Engineering Technical Conferences. ASME Design Engineering Technical Conferences*, volume 2013, page V07BT10A052, 2013.

[23] Soshi Shimada, Vladislav Golyanik, Weipeng Xu, and Christian Theobalt. Physcap: Physically plausible monocular 3d motion capture in real time. *ACM Transactions on Graphics (ToG)*, 39 (6):1–16, 2020.

[24] 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*, pages 2070–2080, 2024.

[25] Joshua B Tenenbaum, Vin de Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323, 2000.

[26] Tsung-Yen Yang, Justinian Rosca, Karthik Narasimhan, and Peter J Ramadge. Learning physics constrained dynamics using autoencoders. *Advances in Neural Information Processing Systems*, 35:17157–17172, 2022.

[27] Ye Yuan, Shih-En Wei, Tomas Simon, Kris Kitani, and Jason Saragih. Simpoe: Simulated character control for 3d human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7159–7169, 2021.

[28] Ye Yuan, Jiaming Song, Umar Iqbal, Arash Vahdat, and Jan Kautz. PhysDiff: Physics-guided human motion diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16010–16021, 2023.

[29] Jie Zhang, Yihui Zhao, Fergus Shone, Zhenhong Li, Alejandro F Frangi, Sheng Quan Xie, and Zhi-Qiang Zhang. Physics-informed deep learning for musculoskeletal modeling: Predicting muscle forces and joint kinematics from surface emg. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:484–493, 2022.

[30] Yufei Zhang, Hanjing Wang, Jeffrey O Kephart, and Qiang Ji. Body knowledge and uncertainty modeling for monocular 3d human body reconstruction. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8986–8998. IEEE, 2023.

[31] Yufei Zhang, Jeffrey O Kephart, Zijun Cui, and Qiang Ji. Physpt: Physics-aware pretrained transformer for estimating human dynamics from monocular videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2305–2317, 2024.

[32] Yufei Zhang, Jeffrey O Kephart, and Qiang Ji. Incorporating physics principles for precise human motion prediction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6164–6174, 2024.

[33] Wentao Zhu, Xiaoxuan Ma, Zhaoyang Liu, Libin Liu, Wayne Wu, and Yizhou Wang. Motionbert: A unified perspective on learning human motion representations. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15039–15053. IEEE, 2023.

A Ablation Study

We ablate the contribution of the inertia-consistent reconstruction loss $\mathcal{L}_{\text{ine}}$ by comparing six conditions that vary which losses are active during fine-tuning. All conditions share the same encoder–decoder architecture; conditions C, D, and F correspond to the three setups defined in Section 4, while A, B, and E provide additional reference points.

- *A: From scratch.* No pretraining; trained on DNE with $\mathcal{L}_{\text{recon}} + \mathcal{L}_{\text{force}} + \mathcal{L}_{\text{contact}} + \mathcal{L}_{\text{euler}}$.
- *B: Pretrained, no fine-tuning.* Frozen PhysPT pretrained checkpoint, evaluated directly on DNE.
- *C: Pretrained + $\mathcal{L}_{recon}$ only.* Equivalent to *Data-driven Euclidean*.
- *D: Pretrained + PhysPT objective.* $\mathcal{L}_{\text{recon}} + \mathcal{L}_{\text{force}} + \mathcal{L}_{\text{contact}} + \mathcal{L}_{\text{euler}}$. Equivalent to *Physics-supervised Euclidean*.
- *E: Pretrained + PhysPT objective without $\mathcal{L}_{euler}$.* $\mathcal{L}_{\text{recon}} + \mathcal{L}_{\text{force}} + \mathcal{L}_{\text{contact}}$, isolating the contribution of the Euler–Lagrange residual loss.
- *F: Pretrained + REINS objective (ours).* $\mathcal{L}_{\text{recon}} + \mathcal{L}_{\text{force}} + \mathcal{L}_{\text{contact}} + \lambda_{\text{ine}}\mathcal{L}_{\text{ine}}$. Equivalent to *REINS (ours)*.

For each condition we report physical reconstruction quality (MPJPE, ACCEL, Euler-Lagrange residual, kinetic-energy (KE) error) on the DNE held-out test set. The relative KE error is $|\bar{E}_{\text{pred}} - \bar{E}_{\text{GT}}|/\bar{E}_{\text{GT}}$, where $\bar{E}_{\text{pred}}$ and $\bar{E}_{\text{GT}}$ are the mean predicted and ground-truth kinetic energies over the test set, respectively.

In Table 5, there are three patterns that need to be discussed. First, comparing condition C (recon-only, 82.1% KE relative error) with condition F (REINS, 22.7% KE relative error) shows that adding $\mathcal{L}_{\text{ine}}$ reduces the predicted-vs.-GT kinetic-energy gap by roughly a factor of four, with a 0.74 mm increase in MPJPE (9.05 vs. 9.79 mm). Second, the original PhysPT objective (condition D) is the worst on KE (193.4%) despite achieving the lowest Euler-Lagrange residual (32.41), indicating that minimizing the Euler-Lagrange residual at the trajectory level does not by itself produce decoded motion whose kinetic energy matches the body's. Third, removing $\mathcal{L}_{\text{euler}}$ (condition E, 101.3% KE relative error) recovers part of the gap, and replacing it with our inertia-consistent loss (condition F, 22.7% KE relative error) closes most of the remainder. This phenomenon suggests that enforcing physics as a trajectory-level penalty leaves the latent geometry unconstrained, while supervising reconstruction under the inertia-weighted norm aligns the decoder with the actual mass distribution of an articulated body, recovering kinetic-energy fidelity at the comparable pose-reconstruction quality. The effect on the latent geometry is measured directly by Table 2.

B Statistical Protocol and Additional Results

This appendix gives the bootstrap and split procedures behind the intervals and the mean $\pm$ sd in the main tables, and reports results for all K. The main split is the default split in the released code. To compute the percentile bootstrap confidence intervals, we keep the model fixed, draw subjects with replacement separately from the held-out HC and the positive group, and recompute AUROC. We repeat this procedure 2000 times. The $K{=}20$ setting is repeated over ten random splits of the healthy controls, with the model fine-tuned anew on each split and reported as mean $\pm$ standard deviation. Rank correlations in Table 4 are tested against the null of zero correlation using the p-values returned

Table 5: Ablation on training conditions A-F. Conditions defined in Section A.

Exp	Val MPJPE (mm)	Test MPJPE (mm)	PA-MPJPE (mm)	ACCEL (mm/s^2)	EL Residual	Kinetic Energy (J)		
						Pred	GT	rel error
A	30.8	32.44	28.67	3.19	36.51	16.76	9.94	68.6%
B	–	31.60	21.66	3.21	39.52	44.98	9.94	352.5%
C	8.4	9.05	7.07	2.50	38.42	18.10	9.94	82.1%
D	9.5	9.98	7.91	2.63	32.41	29.17	9.94	193.4%
E	9.4	9.81	7.67	2.54	38.31	20.02	9.94	101.3%
F	9.4	9.79	7.73	2.53	37.10	**7.68**	9.94	**22.7%**

by SciPy's `spearmanr` and `kendalltau` (τ-b), with the subject as the unit. Table 6 reports the confidence intervals and HC/all results corresponding to Table 3. Table 7 reports the factorial ablation for each K. Across all K, AUROC is highest with the $J^\top M J$ Riemannian distance, followed by the $J^\top J$ Riemannian distance and the Euclidean readout. The encoder trained without $\mathcal{L}_{\text{ine}}$ performs near chance with all three distances. Table 8 reports paired bootstrap differences between REINS and each Euclidean baseline, using the same drawn subjects for both methods in each comparison. All intervals exclude zero.

Table 6: Data efficiency with 95% bootstrap intervals: HC/PD and HC/all AUROC for different numbers of HC training subjects K.

Method	K=20	K=15	K=10	K=5	K=3	K=1
HC/PD						
Data-driven	0.5750 [0.3792, 0.7625]	0.5708 [0.3500, 0.7750]	0.5333 [0.3208, 0.7458]	0.5250 [0.3083, 0.7292]	0.5375 [0.3417, 0.7417]	0.4292 [0.2292, 0.6417]
Physics-supervised	0.5167 [0.3167, 0.7125]	0.5250 [0.3250, 0.7250]	0.5125 [0.3167, 0.7250]	0.4583 [0.2750, 0.6500]	0.4792 [0.2708, 0.6500]	0.4417 [0.2542, 0.6458]
REINS	**0.9083** [0.8042, 0.9833]	**0.8458** [0.6917, 0.9458]	**0.8125** [0.6208, 0.9250]	**0.7792** [0.6208, 0.9792]	**0.7583** [0.6125, 0.9000]	**0.7375** [0.5125, 0.9292]
HC/all						
Data-driven	0.6087 [0.4312, 0.7736]	0.5707 [0.3986, 0.7736]	0.5344 [0.3261, 0.7283]	0.5163 [0.3351, 0.7246]	0.5217 [0.3496, 0.7101]	0.4149 [0.2283, 0.6812]
Physics-supervised	0.5380 [0.3587, 0.7156]	0.5145 [0.3352, 0.7029]	0.5036 [0.3315, 0.6938]	0.4620 [0.3062, 0.6341]	0.4511 [0.2953, 0.6087]	0.4149 [0.2428, 0.6214]
REINS	**0.8768** [0.7736, 0.9529]	**0.8351** [0.6920, 0.9330]	**0.7917** [0.6522, 0.8768]	**0.7428** [0.6250, 0.8406]	**0.7264** [0.6087, 0.8279]	**0.7011** [0.5833, 0.8225]

Table 7: Factorial ablation across the number of HC training subjects K. Subject-level HC/PD AUROC with 95% bootstrap intervals.

Fine-tuning	**Distance**	K=20	K=15	K=10	K=5	K=3	K=1
Without $\mathcal{L}_{\text{ine}}$ (D)	Euclidean	0.5167 [0.3167, 0.7125]	0.5250 [0.3250, 0.7250]	0.5125 [0.3167, 0.7250]	0.4583 [0.2750, 0.6500]	0.4792 [0.2708, 0.6500]	0.4417 [0.2542, 0.6458]
	$J^\top J$ Riemannian	0.5542 [0.3583, 0.7750]	0.5583 [0.3625, 0.7667]	0.5375 [0.3375, 0.7542]	0.5292 [0.3042, 0.7208]	0.4917 [0.3167, 0.6958]	0.4625 [0.2417, 0.6875]
	$J^\top M J$ Riemannian	**0.5917** [0.3833, 0.7708]	**0.5875** [0.3667, 0.7500]	**0.5750** [0.3542, 0.7625]	**0.5375** [0.3292, 0.7208]	**0.5458** [0.3208, 0.6917]	**0.5000** [0.2958, 0.6500]
With $\mathcal{L}_{\text{ine}}$ (F)	Euclidean	0.6208 [0.4250, 0.8000]	0.6125 [0.4042, 0.7750]	0.5917 [0.3583, 0.7583]	0.5792 [0.3250, 0.7833]	0.5625 [0.3292, 0.6958]	0.5417 [0.3458, 0.7000]
	$J^\top J$ Riemannian	0.7333 [0.5375, 0.8667]	0.7083 [0.4708, 0.8250]	0.6875 [0.4250, 0.9083]	0.6667 [0.4125, 0.8667]	0.6542 [0.5125, 0.8042]	0.6250 [0.4333, 0.8083]
	$J^\top M J$ Riemannian	**0.9083** [0.8042, 0.9833]	**0.8458** [0.6917, 0.9458]	**0.8125** [0.6208, 0.9250]	**0.7792** [0.6208, 0.9792]	**0.7583** [0.6125, 0.9000]	**0.7375** [0.5125, 0.9292]

C Graph Construction, Metric Form, and Shortest Path

Table 9 varies the graph construction and the metric form in the main setting (REINS, K=20). The first row builds one graph over the 2,193 reference windows and the 6,886 windows of all held-out HC, PD, and other-disease subjects, so a test subject's shortest path can pass through another test subject. The second row follows the protocol in Section 3.4, in which the graph contains only the

Table 8: Paired bootstrap differences in subject-level AUROC, REINS minus baseline, with 95% confidence intervals. Both methods are evaluated on the same resampled subjects, and intervals are computed from their AUROC differences. An interval excluding zero indicates a statistically significant difference. HC/PD is reported for each number of HC training subjects K, HC/all for the main setting $K=20$.

Task	K	**REINS − Data-driven**	**REINS − Physics-supervised**
HC/PD	20	+0.3333 [0.1201, 0.5945]	+0.3917 [0.1765, 0.5108]
	15	+0.2750 [0.0837, 0.4932]	+0.3208 [0.1286, 0.4873]
	10	+0.2792 [0.0674, 0.5086]	+0.3000 [0.1043, 0.4791]
	5	+0.2542 [0.0418, 0.4867]	+0.3208 [0.1127, 0.5184]
	3	+0.2208 [0.0236, 0.4593]	+0.2792 [0.0719, 0.4968]
	1	+0.3083 [0.0927, 0.5461]	+0.2958 [0.0846, 0.5037]
HC/all	20	+0.2681 [0.0792, 0.4551]	+0.3388 [0.1433, 0.5290]

reference windows and the windows of the current subject. Of the 6,886 test windows, 269 have no reference window among their 20 neighbors and receive one edge to their nearest reference window. The third row removes the shortest-path step and scores each test window by its direct distance to the nearest reference window. The constant form $G = \bar{J}^\top \bar{M} \bar{J}$ replaces $J(z_i)$ in Eq. 7 with a single linear map $\bar{J}$ from latent space to pose coordinates, fitted by ridge regression on the reference windows. It also replaces $M(\hat{q}_i)$ with its mean $\bar{M}$ over the same windows, giving the same metric at every node. The κ nearest neighbors of each window are selected under G. Across both tasks, removing the other test subjects changes AUROC by at most 0.0127, and removing the shortest path changes it by at most 0.0042. The pointwise metric gives an AUROC up to 0.0217 higher than the constant form. On an NVIDIA RTX 3090, computing the metric at all 9,079 nodes takes 1,991 s for the pointwise form and 0.49 s for the constant form.

Table 9: Subject-level AUROC for REINS in the main setting ($K=20$) with 95% bootstrap intervals, under different graph constructions and metric forms. All other REINS results use the graph construction in the second row with the pointwise metric.

Task	**Graph**	constant G	pointwise $g(z)$
HC/PD	Reference and all evaluation windows	0.8958 [0.7833, 0.9792]	0.9125 [0.8083, 0.9833]
	Reference and current subject	0.8917 [0.7667, 0.9875]	0.9083 [0.8042, 0.9833]
	No shortest path	0.8875 [0.7625, 0.9875]	0.9042 [0.8000, 0.9833]
HC/all	Reference and all evaluation windows	0.8678 [0.7609, 0.9475]	0.8804 [0.7772, 0.9565]
	Reference and current subject	0.8551 [0.7482, 0.9565]	0.8768 [0.7736, 0.9529]
	No shortest path	0.8533 [0.7464, 0.9547]	0.8750 [0.7717, 0.9511]

D External Baseline Protocol

All three external encoders are evaluated with frozen backbones, following the CARE-PD convention for pretrained encoders [1]. The 20 HC training subjects define the reference set and are not used to adapt any of the models. Each recording is divided into 16-frame windows at 30 fps. For each window we extract a latent with the frozen encoder, score it with the procedure in Section 3.4 using the constant metric in Appendix C, and take the median over a subject's windows as the subject-level score. For VPoser we use the V02_05 checkpoint pretrained on AMASS [17]. It is a per-frame pose prior, so each frame is encoded independently. We encode the 21 SMPL body-pose joints, take the 32-dimensional posterior mean, and average over the 16 frames of a window. For GaitForeMer we use the released checkpoint pretrained on NTU RGB+D [21], feed it the 3D Human3.6M-format joints provided by CARE-PD after reindexing them to the checkpoint's joint ordering, and take the temporally pooled encoder output as the window latent. GaitForeMer was pretrained with 40-frame inputs, and we feed 16-frame windows so that all methods use the same window protocol. For MotionBERT, we use the released checkpoint [33], which is identical to the one distributed with CARE-PD. We follow CARE-PD's preprocessing to project the Human3.6M-format joints to 2D

and normalize the resulting coordinates. We set the confidence channel to one. We average the per-joint representations over the 16 frames and concatenate them across the 17 joints to obtain an 8704-dimensional embedding for each window. MotionBERT was pretrained on 243-frame inputs, and we keep the first 16 positions of its temporal position embedding.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract introduces REINS as an inertia-induced Riemannian motion representation and reports results on HC vs. PD discrimination and UPDRS-gait severity correlation. The introduction lists three corresponding contributions, and Sections 4.1 and 4.2 present the experiments on CARE-PD, with the factorial ablation in Table 2.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paragraph "Limitations and Future Directions" following the Conclusion discusses the mean-shape body model, the healthy reference set from a single dataset, subject-level evaluation, the effect of pose-estimation error on $M(q)$, the use of a single benchmark and backbone, and the use of the mass matrix alone.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper does not include theoretical results requiring formal proofs.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Section 4 reports the optimizer, learning rate, batch size, schedule, window length, epoch budget, early-stopping criterion, hardware, and loss settings for each condition. Section 3.4 and Appendix C describe the scoring procedure. Sections 4.1 and A specify the data splits, values of K, and ablation conditions. The external-baseline and statistical protocols are described in Appendix D and Appendix B, respectively.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Code is available at `https://github.com/DominoAI-Lab/REINS-NeurIPS-2026`. The repository contains the code for fine-tuning with the REINS objective (Section 3.3) and for scoring with the pointwise metric (Section 3.4). It also includes the subject list for the main split and the environment specification. The README explains how to obtain CARE-PD [1] and the PhysPT checkpoint [31] and lists the commands to reproduce the REINS results for the main split in Table 1. Code for the Euclidean baselines, the external encoders (Appendix D), the ablation conditions (Appendix A), the random splits and bootstrap intervals (Appendix B), and the zero-shot evaluation (Section 4.2) will be added after publication.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Training and evaluation hyperparameters, data splits, and protocol details are reported in Sections 4, 4.1, and A.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Every AUROC in Tables 1, 2, 6, and 7 is reported with a 95% confidence interval computed from 2000 bootstrap resamples of the evaluation subjects, sampling healthy controls and the positive class separately with replacement. The main setting is also repeated over ten random train/test subject splits, with mean $\pm$ standard deviation (not standard error) reported in Table 1. Rank correlations in Table 4 are tested against the null hypothesis of zero correlation using p-values returned by SciPy's `spearmanr` and `kendalltau`. The protocol is described in Appendix B.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: Section 4 states that all experiments are run on a single NVIDIA RTX 3090 GPU. Each fine-tuning run on the full DNE training cohort completes within several hours; the K-sweep and ablation conditions add a comparable per-run cost.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: The work uses the publicly available CARE-PD benchmark [1] and does not involve new data collection from human subjects. The research follows the NeurIPS Code of Ethics.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: REINS could support gait phenotyping from monocular video and reduce the cost of objective motor assessment in Parkinson's-disease care. The inertia matrix is computed from a generic SMPL body model, so potential differences in score calibration across body types and demographic groups require further evaluation. The score is intended to support, not replace, clinical assessment.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper does not release datasets or pretrained models with high misuse risk.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.

- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite the datasets and models used in this work: CARE-PD [1] for the benchmark, PhysPT [31] for the pretrained backbone, AMASS [17] for pretraining data, WHAM [24] for pose estimation in CARE-PD, SMPL for the articulated body model, and the released VPoser [18] and GaitForeMer [7] checkpoints as external baselines. Each is used under its publicly stated terms.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release new datasets or pretrained model checkpoints.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or new human-subject data collection. All experiments use the publicly available CARE-PD benchmark [1].

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve new human-subject data collection. IRB approval was not required for this work.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs were not used as a component of the proposed method or experimental pipeline.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.