PathoFM: Toward a Foundation Model for Pathological Gait

Sharmita Dey [*]
SCAI Lab and Sensory Motor Sytems Lab,
D-HEST,
ETH Zurich,
Swiss Paraplegic Research, Nottwil

Diego Paez-Granados
SCAI Lab, D-HEST,
ETH Zurich,
Swiss Paraplegic Research, Nottwil

Abstract

Pathological gait exhibits diverse compensatory strategies that vary across individuals, disease stages, and time. Robust downstream clinical performance can benefit from *foundation models* (FMs) that learn generic, transferable motion representations. *However, an interesting question is what inductive biases prove to be good training objectives for a general recipe to train such FMs.*. We address this with *PathoFM*, an encoder-only pretraining recipe trained on heterogeneous gait cycles from *230* patients, augmented with synthetic generative variants of real trials to broaden coverage of atypical patterns. The recipe blends three complementary objectives: (i) *Local Completion* (recovering continuous segments of the input), (ii) *Temporal Continuity* (predict future segments to enforce dynamic consistency), and (iii) *In-Context Dynamics*, an unsupervised in-context learning objective that encourages relational reasoning from a small support set of exemplars. We evaluate under strict patient (subject) holdout and compare PathoFM against grouping-based pretexts (subject-ID discrimination, InfoNCE contrastive learning, online prototypes) and diffusion variants. Across clinical classification and regression endpoints, PathoFM achieves the best overall balance of performance. These results indicate that dynamics-centric pretraining yields more generalizable clinical timeseries representations than objectives based on grouping or instance discrimination.

1 Introduction

Human gait, especially in pathological conditions and impairments, is a complex time-series signal containing valuable clinical information about health and function [13, 9]. However, developing robust models for gait analysis is challenging due to limited clinical datasets and the diverse ways gait can be impaired [10, 6, 5, 8, 4, 7]. Large-scale foundation models that can be adapted to various tasks offer a promising solution to take advantage of diverse gait data for improved generalization [1]. A key question is: *Which inductive biases should shape the pretraining objectives of a general gait foundation model?*

Based on our analysis, we adopt three key principles: (1) Emphasize information coverage rather than narrow task imitation: learn to complete missing structure and continue dynamics [12] in a way that is agnostic to any single clinical endpoint. (2) Favor relational and contextual reasoning to allow the model to leverage a few relevant gait exemplars provided “in context” to inform its predictions, mirroring a clinician’s ability to compare a patient’s gait with reference patterns. (3) we keep the backbone encoder-centric so most capacity is spent on representation building; small task heads can then adapt with minimal supervision.

[*] corresponding author. SD conceived and led the study, developed methodology, performed experiments, and wrote and revised the manuscript. DPG provided the data, infrastructure and feedback to edit the manuscript.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Workshop on Learning from Time Series for Health.

2 PathoFM: A Multi-Objective Transformer for Multivariate Gait Timeseries

PathoFM aims to (1) capture high-dimensional pathological structure without committing to a single clinical taxonomy; (2) support imputation and continuation as sanity checks for learned structure; and (3) transfer to downstream classification and cross-modal regression with lightweight heads.

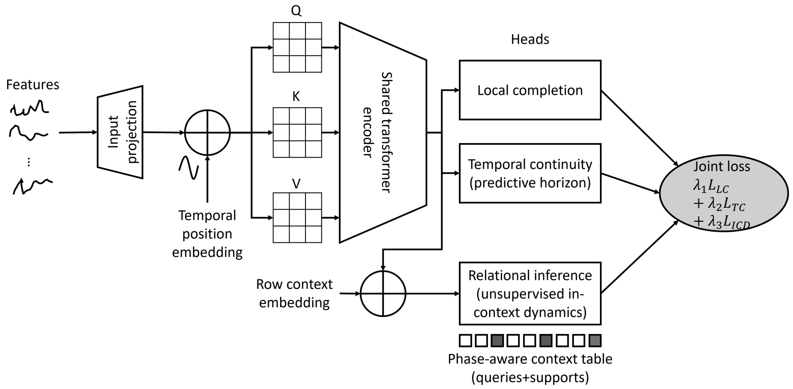
Features
Input projection
Temporal position embedding
Q
K
V
Shared transformer encoder
Heads
Local completion
Temporal continuity (predictive horizon)
Row context embedding
Relational inference (unsupervised in-context dynamics)
Phase-aware context table (queries+supports)
Joint loss
$\lambda_1 L_{LC}$
$+ \lambda_2 L_{TC}$
$+ \lambda_3 L_{ICD}$

Figure 1: Architecture of the proposed pretraining model for multivariate time series. A shared transformer encoder processes projected features enriched with temporal position encodings. Three self-supervised objectives guide representation learning: (i) **Local Completion**, where contiguous masked time steps across the window are reconstructed; (ii) **Temporal Continuity**, where future segments are masked and predicted from past observations; and (iii) **In-Context Dynamics**, where queries are reconstructed from a small support set (phase-aware, subject-balanced). The final pretraining objective is a weighted joint loss over the three terms.

2.1 Pretraining Objectives

We designed PathoFM to combine three complementary objectives (Fig. 1). 1) A local structure completion via masked encoding on contiguous temporal spans, 2) mid-range temporal continuity 3) an unsupervised context-conditioned objective that encourages relational reasoning from a small support set of exemplars.

Local structure completion (LC). This objective, analogous to masked modelling [11], focuses on local structure completion. We mask contiguous spans within the window (covering a certain window of consecutive time steps across all variables), and task the model with reconstructing the missing data from the surrounding context. This enables the encoder to capture local structure, correlations across variables, and phase-aligned morphology.

Temporal continuity (TC). This objective is designed to enforce mid-range dynamic consistency. We mask the latter portion of each gait sequence and ask the model to predict the continuation of the timeseries beyond the observed part.

Unsupervised in-context dynamics learning (uICD). We introduce an unsupervised in-context learning objective designed to favor *relational reasoning* by enabling the model to leverage a few relevant exemplars "in context." Concretely, each training instance is a small table comprising some *query* windows and a *support* set of windows drawn via a *phase-aware sampler* from *subject-balanced* mini-batches (so every subject contributes multiple candidate supports). Our uICD pretext task provides no external labels (no diagnosis, class, event markers, or subject IDs fed into the model). The only supervision signal is the observable timeseries itself. The network processes all rows jointly and attends from the query to its supports, performing non-parametric adaptation: it infers subject- and phase-specific transformations from the supports and applies them to the query. The loss is a conditional masked-reconstruction error computed only on the query, which compels the model to use the relational structure among rows rather than memorize global averages.

Comparison with other objectives. We compare these objective against state-of-the-art benchmarks such as *subject-ID discrimination* (supervised pretext), *contrastive [3]*, *online prototypes* (DINO) [2], and *diffusion-only* [14] and *diffusion-hybrid* variants.

3 Data and Protocol

Gait cycles from *230 Spinal Cord Injury (SCI) patients* were used for pretraining. Trials were augmented with dynamics-preserving variations (e.g., amplitude/phase warps) to improve coverage of atypical pathological regimes while retaining plausibility. Evaluation was performed on strict patient holdout to 10 unseen SCI participants (each $\sim$10 trials, 101 samples), covering *33* kinematic/kinetic variables (such as the left and right ankle, knee, hip joint angles, power, moments along X/Y/Z and foot progression). We report: (i) qualitative sanity via continuation overlays on held-out subjects; and (ii) downstream performance on *classification* (weighted-F1, AUC) of pathology category (tetraplegic vs. paraplegic), gender, and spinal cord independence measurement level (SCIM: high or low) and *regression* to ground reaction forces (GRF)$_{X/Y/Z}$ (Pearson r, RMSE), using lightweight heads for probes on frozen features.

4 Architecture

We use a Transformer-based encoder architecture for multivariate time-series windows $x \in \mathbb{R}^{T \times D}$. Each time step is first mapped to a d_{model}-dimensional embedding by a linear input projection, followed by the addition of a positional encoding. The embedded sequence is then processed by a stack of L Transformer encoder layers. The output of the final encoder layer is mapped back to the original feature space via a linear decoder, across all pretraining objectives. For all experiments we set $d_{\text{model}} = 128$, $L = 8$, number of attention heads $H = 4$, and dropout $p = 0.1$.

5 Results and Conclusion

5.1 Pretext generalization to unseen patient population

We found that the models generalize well for masked encoding and forecasting in patients not seen during the training as given by the high R^2 (local structure completion: 0.90, temporal continuity: 0.85) and r (local structure completion: 0.95, temporal continuity: 0.92) and low RMSE of predictions (local structure completion: 0.034, temporal continuity: 0.041). Qualitatively, predicted trajectories align well in phase and amplitude of the real trajectories (Fig. 2).

5.2 Performance on downstream tasks

We evaluated our proposed model variants through both ablation studies and comparisons against representative benchmarks.

We performed ablation experiments to assess the contribution of local completion (LC), temporal continuity (TC), and unsupervised in-context dynamics (uICD) objectives (Tab. 1). Overall, the combination of all three objectives (LC+TC+uICD) yielded the most balanced improvements across tasks. In classification tasks, the full model achieved best performance in pathology and SCIM level classification. Unsupervised in-context dynamics (uICD) was less informative than generative reconstruction and temporal forecasting for gender classification. On the other hand, the inclusion of uICD was consistently beneficial for prediction of GRF components with the full model or uICD-augmented variants yielding the best performances. These results suggest that uICD contributes significantly to biomechanical signal fidelity, while LC and TC reinforce generalization across classification tasks.

We further compared the best variant (LC+TC+uICD) against state-of-the-art baselines (Table 2). Our method either outperforms the baselines or provided similar performance across tasks. Overall, our approach performed best in at least one of the two metrics reported for each task. For pathology classification, our approach achieved the highest F1 while remaining competitive with the contrastive method in AUC. For GRF prediction, the proposed model consistently delivered either the best or second-best results across components. While diffusion-based methods were competitive in GRF

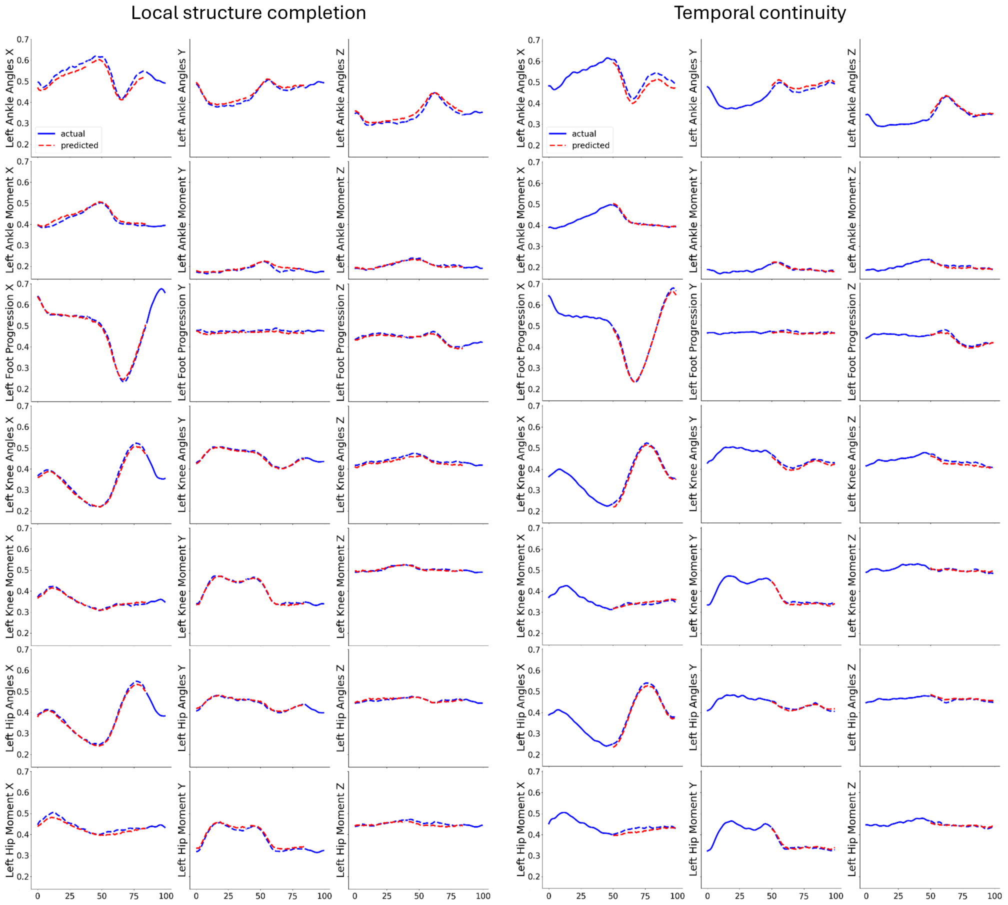

Figure 2: Held-out SCI subject (single trial), per-variable overlay. Ground truth (solid) vs. PathoFM predicted local structure completion and temporal continuation (dashed) across multi-joint angles/moments (X/Y/Z) and progression over a normalized gait cycle (101 samples). The model preserves phase, amplitude, and salient transients across variables, indicating robust local structure completion and temporal continuation on unseen patients.

	Pathology category		Gender		SCIM		GRF_X		GRF_Y		GRF_Z	
Method	F1 ↑	AUC ↑	F1 ↑	AUC ↑	F1 ↑	AUC ↑	ρ ↑	RMSE	ρ	RMSE ↓	ρ ↑	RMSE ↓
TC+uICD	0.68	0.64	0.762	0.54	0.70	0.64	0.65	0.021	0.83	0.033	**0.89**	**0.199**
LC+uICD	0.69	0.64	0.770	0.54	0.70	0.64	0.70	0.020	0.83	0.033	0.89	0.200
LC+TC	0.66	0.64	**0.780**	**0.61**	0.69	0.60	0.64	0.022	0.81	0.035	0.84	0.246
LC+TC+uICD	**0.69**	**0.66**	0.764	0.55	**0.71**	**0.65**	**0.71**	**0.020**	**0.83**	**0.033**	0.88	0.202

Table 1: Performance across downstream classification and prediction tasks when each component of the loss term is ablated. Arrows indicate whether higher (↑) or lower (↓) values are better. The full loss term performs best in four out of six tasks, illustrating the importance of the combined loss in learning generalizable representations.

prediction, they performed weaker on pathology classification task. The grouping-based methods (Dino and contrastive), on the other hand, was competitive in pathology classification, but weaker in GRF prediction, likely because their pretext objective is better aligned with classification rather than continuous signal reconstruction. In contrast to both grouping and diffusion-based objectives, our approach provided a balanced performance across both the pathology classification and GRF prediction tasks. Overall, these results demonstrate that the strength of the combined objective that we propose in offering a general and balanced clinical outcome prediction and biomechanical signal fidelity, outperforming both grouping-based and diffusion-based approaches.

	Pathology category		GRF_X		GRF_Y		GRF_Z	
Method	F1 ↑	AUC ↑	ρ ↑	MSE ↓	ρ ↑	MSE ↓	ρ ↑	MSE ↓
Dino	0.67	0.62	0.67	0.020	0.82	0.035	**0.89**	0.230
Contrastive	0.68	**0.68**	0.63	0.021	0.78	0.038	0.78	0.281
subject identification	0.63	0.65	0.69	0.020	0.82	0.034	0.86	0.224
Diffusion+LC+uICD	0.61	0.64	0.70	**0.019**	**0.83**	**0.033**	0.87	0.212
Diffusion only	0.63	0.65	0.58	0.024	0.73	0.043	0.84	0.230
LC+TC+uICD	**0.69**	0.66	**0.71**	0.020	**0.83**	**0.033**	0.88	**0.202**

Table 2: Benchmark performance across pathology category classification and GRF prediction tasks. Arrows indicate whether higher (↑) or lower (↓) values are better. The combined loss term that we propose performs best in one or both the metrics reported for each task.

References

[1] Rishi Bommasani. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[2] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PmLR, 2020.

[4] Sharmita Dey. Learning-based biomimetic strategies for developing control schemes for lower extremity rehabilitation robotic devices. 2023.

[5] Sharmita Dey and Sarath R Nair. Enhancing joint motion prediction for individuals with limb loss through model reprogramming. *arXiv preprint arXiv:2403.06569*, 2024.

[6] Sharmita Dey and Sarath Ravindran Nair. Remap: Neural model reprogramming with network inversion and retrieval-augmented mapping for adaptive motion forecasting. *Advances in Neural Information Processing Systems*, 37:25195–25227, 2024.

[7] Sharmita Dey, Takashi Yoshida, Robert H Foerster, Michael Ernst, Thomas Schmalz, Rodrigo M Carnier, and Arndt F Schilling. A hybrid approach for dynamically training a torque prediction model for devising a human-machine interface control strategy. *arXiv preprint arXiv:2110.03085*, 2021.

[8] Sharmita Dey, Takashi Yoshida, and Arndt F Schilling. Feasibility of training a random forest model with incomplete user-specific data for devising a control strategy for active biomimetic ankle. *Frontiers in Bioengineering and Biotechnology*, 8:855, 2020.

[9] Alberto Esquenazi. Gait analysis in lower-limb amputation and prosthetic rehabilitation. *Physical Medicine and Rehabilitation Clinics*, 25(1):153–167, 2014.

[10] Elsa J Harris, I-Hung Khoo, and Emel Demircan. A survey of human gait-based artificial intelligence applications. *Frontiers in Robotics and AI*, 8:749274, 2022.

[11] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

[12] Bryan Lim, Sercan Ö Arık, Nicolas Loeff, and Tomas Pfister. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International journal of forecasting*, 37(4):1748–1764, 2021.

[13] Jacquelin Perry and Judith Burnfield. *Gait analysis: normal and pathological function*. CRC Press, 2024.

[14] Yusuke Tashiro, Jiaming Song, Yang Song, and Stefano Ermon. Csdi: Conditional score-based diffusion models for probabilistic time series imputation. *Advances in neural information processing systems*, 34:24804–24816, 2021.