Towards Context-Aware Clinical Motion Understanding in Daily Living at Home: Freezing of Gait Detection with Egocentric Vision

Vayalet Stefanova[1]*, Diwas Lamsal[1]*, Margot Genbrugge[2], Maxim Yudayev[1], Christian Schlenstedt[3], Moran Gilat[2], Bart Vanrumste[1]†, and Benjamin Filtjens[4,5]†

[1] Department of Electrical Engineering (ESAT), KU Leuven, Leuven, Belgium
[2] Neurorehabilitation Research Group (eNRGy), Department of Rehabilitation Sciences, KU Leuven, Leuven, Belgium
[3] Institute of Interdisciplinary Exercise Science and Sports Medicine, Medical School Hamburg, Hamburg, Germany
[4] Department of Engineering Systems and Services, Delft University of Technology, Delft, The Netherlands
[5] Institute for Health Systems Science, Delft University of Technology, Delft, The Netherlands
vayalet.stefanova@kuleuven.be, diwas.lamsal@kuleuven.be

Abstract. Understanding motion in daily living requires context beyond kinematics, because similar inertial patterns during activities of daily living (ADLs) can reflect intentional stopping, object interaction, or pathological movement impairment. Egocentric vision provides task-related context that may help disambiguate these cases. We investigate this challenge through freezing of gait (FOG) detection in Parkinson's disease (PD), a symptom strongly influenced by contextual factors during ADLs. Using synchronized egocentric video, wearable IMUs, and expert-annotated FOG labels collected from 13 PD participants in their homes, we evaluate frozen representations from pretrained ego-video and time-series foundation models, alongside an IMU-based TCN trained from scratch, under leave-one-subject-out evaluation. The IMU-based TCN achieved the strongest event-detection performance, reaching 42.3 F1 and 83.0 AUROC, compared with 32.6 F1 and 77.2 AUROC for V-JEPA2 ego-video features. Although ego-video alone did not outperform IMU-based sensing, it showed above-chance discrimination, and qualitative analyses suggest that egocentric vision may capture FOG-relevant information independent of IMUs. Together, these results support the use of pretrained ego-video representations to add contextual information to wearable-sensor-based clinical motion understanding in daily living.

Keywords: Context-aware motion understanding · Freezing of gait · Egocentric vision · Foundation models

*Equal contribution †Equal senior contribution

1 Introduction

Understanding clinical motion in daily living environments is an important challenge for wearable and computer vision systems [11]. Unlike structured laboratory assessments, daily-life activities performed at home are highly variable and shaped by the surrounding environment and ongoing tasks. Commonly used kinematic sensors may be insufficient for determining clinical motion, as similar low-motion inertial patterns can arise from intentional stopping, object interaction, or balance recovery [11,29,47]. This ambiguity is most consequential for clinical applications that rely on accurate motion assessment and real-time closed-loop interventions for gait disorders [9,14,17,25,29]. Examples include continuous estimation of digital mobility outcomes [17], exoskeletons, on-demand cueing, kinematic-based deep brain stimulation [9,14,25], and fall-risk assessment [29].

Egocentric vision (ego-vision) is an emerging modality for capturing contextual motion information during activities of daily living (ADLs) [15]. By recording the wearer's first-person perspective, it directly observes the surrounding environment and ongoing activities. Ego-vision has been successfully applied to tasks such as action recognition and video captioning in home environments and is supported by the availability of large-scale benchmark datasets [8,15]. The modality has also shown value in clinical applications such as fall-risk assessment in older adults and people with PD [28–30]. The findings suggest that contextual information may be particularly relevant for movement disorders that strongly depend on environmental factors.

An example of such disorder is freezing of gait (FOG) in Parkinson's disease (PD), defined as "paroxysmal episodes wherein there is an inability to step effectively, despite attempting to do so" [13]. FOG is a major cause of falls in this population [5]. Although an episodic phenomenon, context during ADLs is particularly important for FOG, as freezing episodes are often triggered by environmental and cognitive factors. Examples that evoke FOG include turning, narrow passages (*e.g.*, going through a doorway), cluttered pathways, and dual-tasking (*e.g.*, carrying a full cup of water while walking) [7]. Many patients also report recurrent FOG at specific locations (*i.e.*, hotspots) within their homes [53]. Despite the importance of context, automatic FOG detection has primarily relied on physiological and kinematic sensing modalities, including electroencephalography (EEG), electromyography (EMG), skin conductance (SC), heart rate signals, and the most commonly used inertial measurement units (IMUs) [4,24]. A key limitation of IMU-based FOG detection is the difficulty of distinguishing FOG episodes from voluntary stopping [47]. Whether ego-vision can address this challenge and support clinical motion understanding for FOG detection remains unexplored.

Another challenge specific to the clinical domain is the scarcity of large labeled home-based datasets, which makes training robust deep learning (DL) models from scratch difficult [23]. Foundation models (FMs) offer a promising alternative by using representations learned from large-scale video and time-series data, potentially reducing the need for task-specific clinical training data [20]. Frozen FM representations also enable lightweight comparisons across modalities

by only training a simple classifier (*e.g.*, linear probe). This allows to evaluate whether general-purpose features encode clinically relevant information while avoiding computationally expensive end-to-end retraining [1,6]. Large-scale egocentric video datasets have further enabled FMs specialized in ego-vision representation learning [31,40]. In the context of FOG, this raises the question of whether pretrained ego-vision representations capture predictive features relevant to freezing.

Our contribution is a problem formulation and empirical study of context-aware clinical motion understanding in daily living at home. First, we introduce egocentric video as a contextual modality for FOG detection during home-based ADL. Second, we evaluate frozen video and IMU FM representations under subject-independent leave-one-subject-out (LOSO) evaluation on synchronized ego-video, IMU, and expert-annotated FOG data. Third, we show that temporal egocentric video representations contain predictive information for FOG, though less strong than IMU representations alone. Finally, through a qualitative analysis of modality-specific errors, we identify cases where visual context could help and where it fails.

2 Related Work

2.1 Automated FOG detection

Expert annotations of video recordings serve as the gold-standard ground truth for identifying FOG episodes [13]. Various machine learning and DL techniques have been proposed to learn representations across modalities, including end-to-end architectures based on Long Short-Term Memory (LSTM) [27], convolutional [46] and transformer networks [18], as well as conventional classifiers such as SVMs and XGBoost [34,35]. Among these, temporal convolutional networks (TCNs) [19] are a popular choice for modeling inertial time-series data. Because of their ease of use most state-of-the-art FOG detection systems rely exclusively on IMU signals commonly acquired from the lower limbs, waist, or trunk [23]. Pose-based approaches, such as those using graph-based architectures [10], instead exploit skeletal structure from marker-based or third-person motion capture to encode joint-level spatial relations beyond what raw IMU signals provide. However, these require an external body-view camera unavailable in egocentric setups such as ours, placing them outside the scope of this paper. Combining multiple sensing modalities has also been studied with the goal to improve detection robustness [26,42,47,50]. While these approaches provide promising results, their generalizability beyond standardized laboratory settings remains unclear. Additionally, to the best of our knowledge, no existing study has explored egocentric vision as a modality for FOG detection.

2.2 Contextual understanding through ego-vision

In recent years, ego-vision has been increasingly adopted to characterize the environmental and behavioural context of daily activities. Previous studies have

used ego-vision to estimate fall risk in older adults by identifying walking surfaces and environmental hazards during daily life [30]. More recently, egocentric videos have been used for assessing fall risk in people with PD to augment IMU-based mobility measures with environmental context [28,29]. The authors fine-tuned a YOLOv8 [39] object detection model to identify the overlap of objects and the direct walking path of the subject to inform fall risk. Within the FOG domain, a recent study explored detection of environmental triggers commonly associated with FOG onset, such as tight turns, narrow passages, and floor pattern changes [32]. The authors finetuned multimodal FMs on egocentric videos and textual descriptions of these trigger scenarios to enable a wearable trigger detection system. This addresses a related but distinct problem from ours: identifying contextual situations that may provoke FOG, rather than detecting the occurrence of a FOG episode itself. Moreover, validation in people with PD was not performed [32].

2.3 Foundation models for video and wearable sensing

FMs have become a common approach for representation learning in both video and time-series analysis. In the visual domain, models such as VideoMAE [38], V-JEPA 2 [3], and VideoPrism [52] are trained on large-scale collections of third-person videos using self-supervised objectives. These models have shown strong transferability across a wide range of downstream tasks, including human activity recognition, captioning, and question answering [49]. However, third-person and egocentric videos differ in viewpoint, motion patterns, and the visual context available to the camera. Consequently, recent works such as EgoVLM [40] and EgoVideo [31] have been pretrained on large-scale egocentric datasets to better capture first-person representations.

Similar advances have been made for processing time-series data. FMs such as UniMTS [51] and Chronos-2 [2] are pretrained on large and diverse collections of temporal signals and have achieved strong performance across tasks including classification, forecasting, and anomaly detection. Importantly, such models have shown the ability to learn transferable representations from wearable sensing modalities such as accelerometer and gyroscope signals, making them a useful tool for IMU-based movement analysis [43,51].

Building on these advances, we evaluate a representative set of FMs covering complementary design choices. For IMU signals, we use UniMTS [51], a motion-specialized model pretrained for human activity recognition, and Chronos-2 [2], a general-purpose multivariate time-series model. For ego-video, we select models differing in temporal coverage, pretraining domain, and capacity: the single-frame image model DINOv3 [37]; the third-person video models VideoMAE-v2 [41] and V-JEPA 2 [3], which encode short and long temporal windows, respectively; and the egocentric video model EgoVideo [31].

Table 1: Participant demographics, clinical characteristics, and dataset statistics ($n = 15$). Values are reported as mean ± standard deviation. The reported time spent frozen (%TF) and number of FOG episodes (#FOG) are calculated across both ON and OFF medication states.

Characteristic	**Mean ± SD**
Age (years)	68.2 ± 8.8
Disease duration (years)	13.5 ± 7.0
MDS-UPDRS III (ON)	32.3 ± 10.9
MDS-UPDRS III (OFF)	39.2 ± 11.0
MoCA	25.8 ± 2.3
NFOGQ	18.7 ± 4.2
Recording duration (min)	13.19 ± 5.63
%TF (ADL)	7.2 ± 6.5 %
#FOG (ADL)	14.4 ± 18.0

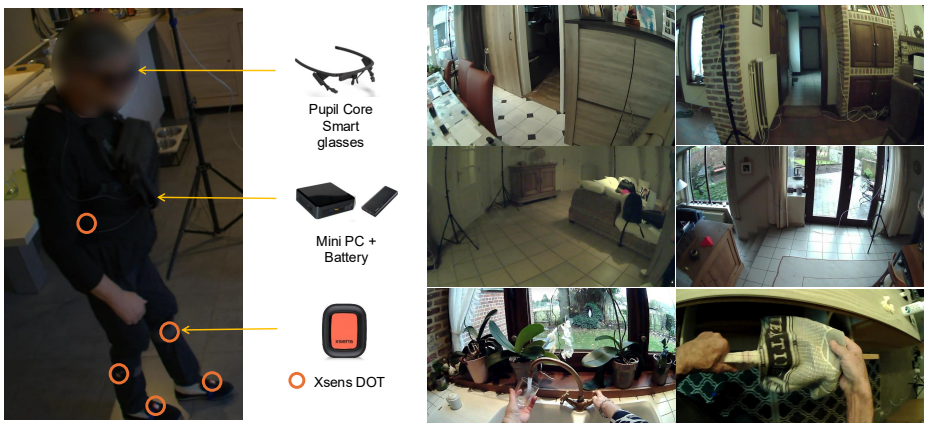
3 Methods

3.1 Dataset

This study recruited 15 individuals with PD (5 female, 10 male) who reported experiencing FOG on a daily basis. Participants provided written informed consent before any study related activities, and the study was approved by the ethical committee of KU Leuven and the university hospital UZ Leuven (EC Research UZ/KU Leuven) with reference number S70220. The cohort had a mean age of 68.2 years and mean disease duration of 13.5 years. Demographic and clinical characteristics are summarized in Table 1. Two participants were excluded from the analysis due to incomplete recordings.

Subjects completed two data collection sessions in their homes, one in the OFF-medication state (following overnight withdrawal) and one ON-medication (1 hour after medication intake). Each session followed the same protocol, comprising five ADL tasks: walking through a doorway, two daily-life tasks (*e.g.*, watering plants, washing dishes) randomly selected from a predefined list of 14 tasks (see Supplementary Material for the full list), and two hotspot tasks corresponding to locations within the home where the participant reported frequently experiencing FOG. These tasks were selected to reflect real-world conditions. Representative egocentric views of the three ADL task types (doorway, hotspot, and daily life) are shown in Fig. 1b.

Data were collected using five Xsens DOT IMUs (60 Hz) positioned on the pelvis, bilateral shins, and the dorsum of both feet (Fig. 1a). Ego-video was captured using Pupil Core smart glasses (30 Hz, 1280×720), while four external cameras (30 Hz, 2560×1440) were used for offline annotation. FOG episodes and tasks were post-hoc annotated by a human expert using the ELAN software [12]. Annotations followed the updated FOG definition, specifically the technical definition for scoring FOG from video recordings [13]. Participants wore a lightweight



(a) Sensor setup. **(b)** Example egocentric views.

Fig. 1: (a) Sensor setup for data collection. Pupil Core smart glasses for egocentric video and five Xsens DOT IMUs for lower-body inertial data, collected via a battery-powered mini PC. (b) Representative egocentric views captured during the ADL tasks: doorway (top), hotspot (middle), daily life (bottom).

chest pack containing an Intel NUC13ANKi7 mini-PC and a portable battery. We used HERMES [48] for data collection, which provided us with synchronized data across all modalities.

The final dataset for processing comprised 13 subjects and 171.6 minutes of recordings. Data were segmented into individual tasks by removing transitions between activities. For each subject, ON and OFF medication sessions were combined to increase the number of windows and better represent real-life conditions, where FOG can occur during ADLs in both states. We observed considerable inter-subject variability in both the frequency and duration of FOG episodes (Fig. 2). While some participants displayed little or no FOG, others experienced frequent and prolonged episodes. Such heterogeneity is in line with the clinical presentation of FOG, but also poses a challenge for training robust subject-independent FOG detection models [45].

3.2 Data preprocessing

The data were segmented into overlapping windows using a fixed stride of 0.5 s (see Supplementary Material for stride-sensitivity ablation). We use a 2 s window as our main configuration, in line with common practices in window-based FOG detection [36], and additionally evaluate 3 s and 10 s windows as ablations on a subset of models; the 10 s length matches the temporal context used by the UniMTS foundation model during pretraining [51].

Each window was assigned a binary FOG label and marked as FOG if it contained at least 0.5 s of annotated FOG, a threshold chosen to preserve short FOG episodes, as the median episode duration in the dataset was 0.9 s. For

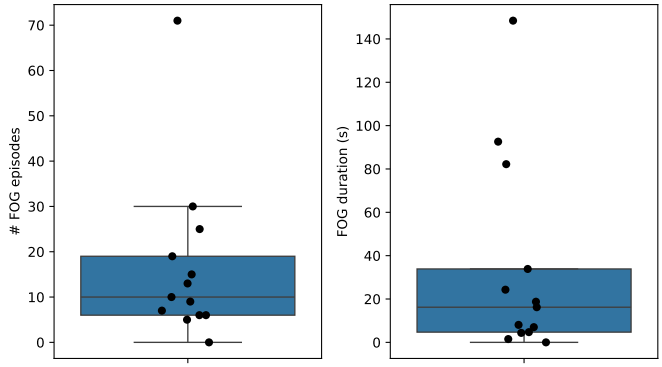


Fig. 2: Inter-subject variability in FOG occurrence during home-based ADL. Each black point represents one participant, aggregated across the three ADL task types (doorway, hotspot, and daily life) and medication states (ON and OFF). Left: total number of FOG episodes per participant. Right: cumulative FOG duration per participant.

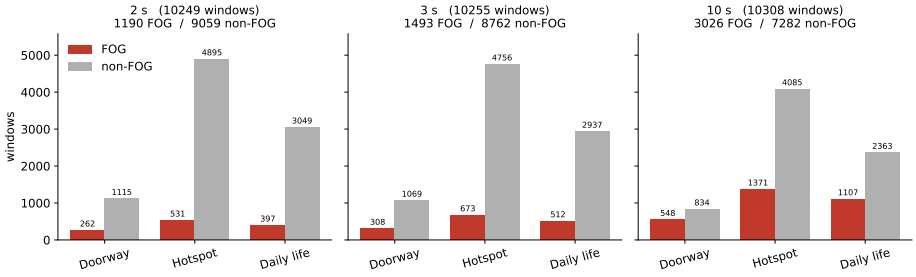


Fig. 3: Window-level FOG vs. non-FOG counts per ADL task (doorway, hotspot, daily life) for the three window lengths, pooled over the 13 cohort subjects.

evaluation, each window was further assigned to whichever ADL task (doorway, hotspot, or daily life) represented more than 50% of its duration. Windows with more than 50% missing annotations, caused by occasional feet occlusion in the external camera views, were discarded.

Figure 3 summarizes the resulting window counts per task and window length across the 13 subjects. Because the 0.5 s FOG criterion is more easily satisfied, the number of FOG windows grows with window length, while the total number of retained windows varies because the task and missing-annotation criteria produce a different number of windows at each length. We therefore treat the three lengths as separate operating points rather than directly comparable conditions.

3.3 Feature extraction

For each window, modality-specific features were obtained from pretrained FMs, using UniMTS [51] and Chronos-2 [2] for the IMU signals and VideoMAE-v2 [41], V-JEPA 2 [3], EgoVideo [31], and DINOv3 [37] for the egocentric video.

IMU features: First, missing IMU samples were linearly interpolated with edge-value extrapolation before resampling the signals to a uniform 60 Hz grid. The resulting signals were then preprocessed according to the requirements of each FM.

UniMTS is designed specifically for motion understanding and human activity recognition, using a spatio-temporal graph encoder that models body kinematics through a skeletal representation. Following the preprocessing recommended by the authors, the five IMU acceleration signals were mapped to the corresponding lower-body joints of the 22-joint SMPL skeletal representation [21] expected by the model and resampled from 60 Hz to 20 Hz using FFT-based resampling. The resulting signals were passed through the pretrained encoder and the output embedding was extracted for each window, yielding an embedding tensor of size (N, D), where N denotes the number of windows and $D = 512$ is the feature embedding dimension.

Chronos-2 was applied to both accelerometer and gyroscope signals. Each window was represented as a multivariate time series of the 30 IMU channels (5 IMUs $\times$ 6 channels) and passed through the pretrained model. This resulted in an embedding tensor $\mathbf{E} \in \mathbb{R}^{N \times C \times P \times D}$, where N denotes the number of windows, $C = 30$ the number of IMU channels, P the number of temporal patches, and $D = 768$ the latent embedding dimension. As Chronos generates a variable number of output tokens depending on the input sequence length, $P \in \{8, 12, 38\}$ for 2 s, 3 s, and 10 s windows, respectively. We mean-pool the embedding over the temporal patch dimension P, yielding a per-window representation of size $C \times D = 30 \times 768 = 23{,}040$.

Ego-vision features: For all video models, the frames of each window were resized along the shorter edge and center-cropped to the native input resolution expected by each model. Frames were sampled at uniformly spaced timestamps across the window, by repeating frames when the window contained fewer than the number required by the model, and encoded jointly, producing a single per-window embedding of size (N, D) aligned with the IMU windows and labels. DINOv3 (ViT-B, $\approx$86 M parameters) is an image model and was therefore applied to the single center frame of each window at 224×224, producing $D = 768$. VideoMAE-v2 (base variant, also ViT-B) encodes 16 frames at 224×224 into $D = 768$, V-JEPA 2 (ViT-L, $\approx$300 M) encodes 64 frames at 256×256 into $D = 1024$, padding by repetition at the 2 s length, and EgoVideo (4-frame variant, InternVideo2 backbone, $\approx$1 B) encodes 4 frames at 224×224 into $D = 512$. Frame count is fixed per checkpoint, so temporal coverage, model size and pretraining data covary. DINOv3 and VideoMAE-v2 are the closest controlled pair with the same ViT-B backbone and $D = 768$, differing mainly in frame count.

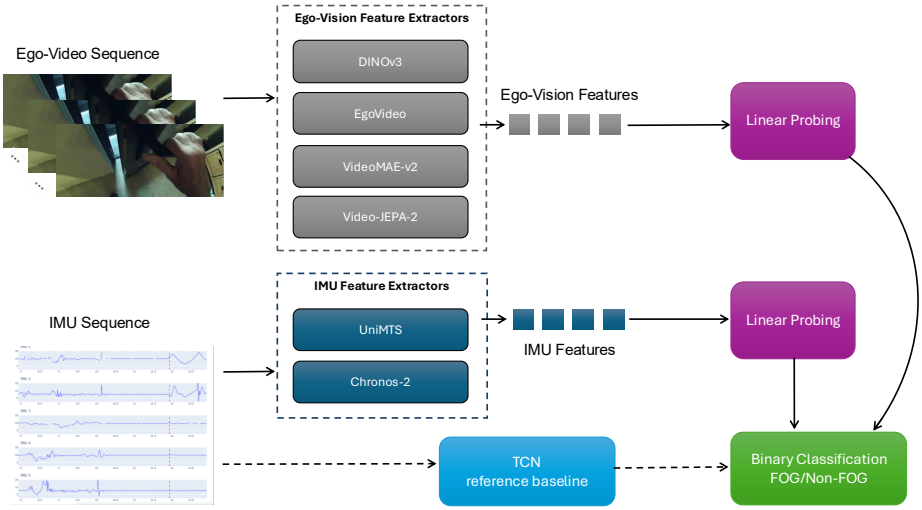


Fig. 4: Experimental Setup. For each window, features are extracted from IMU (UniMTS, Chronos-2) and ego-vision (VideoMAE-v2, V-JEPA 2, EgoVideo, DINOv3) foundation models. All models are evaluated under leave-one-subject-out cross-validation, with a fully trained TCN on the IMU signals as a supervised reference.

3.4 Experimental setup

Figure 4 summarizes our experimental pipeline. All models were evaluated under a leave-one-subject-out cross-validation (LOSO-CV) protocol. In each fold, all recordings of one subject, comprising both the ON- and OFF-medication sessions, were held out for testing while the remaining subjects formed the training set. All metrics were computed separately within each held-out subject and then averaged across subjects. The same per-subject averaging was applied to the per-task breakdown over the three ADL tasks (doorway, hotspot, daily life).

For each FM, we assessed its frozen representations with a linear probe, which provides a uniform comparison across models. The per-window embeddings were standardized to zero mean and unit variance using statistics computed on the training folds. An L_2-regularized logistic-regression classifier (class-balanced) was then fit on the training windows and used to predict FOG on the held-out subject. We report a single regularization strength $C = 10^{-3}$ with all models, chosen from a sweep over $[10^{-6}, 10]$ in which it remained close to the optimum for every model. Ranking the models at their individual optima instead did not change the order on either F1 or AUROC. We also evaluated the attentive probe used in the original V-JEPA 2 evaluation protocol [3], which did not improve on the linear probe, likely because our training set is small relative to its capacity; we therefore report linear-probe results throughout.

Finally, as fully trained baselines, we train two TCNs end-to-end on accelerometer signals after gravity removal via a 0.3 Hz high-pass filter, with one variant additionally including the gyroscope signals. We set the number of di-

lated residual blocks so the receptive field spans the full window (five, six, and eight blocks for the 2, 3, and 10 s windows). The remaining hyperparameters were selected by grid search, giving a residual block width of 32, kernel size 3, dropout 0.3, and weight decay 10^{-3}. The TCN was optimized with a class-weighted cross-entropy loss (AdamW [22], learning rate 5×10^{-4}, cosine schedule) under the same LOSO protocol.

3.5 Evaluation metrics

FOG detection performance was evaluated using multiple metrics, reflecting the complementary aspects of performance captured by each measure [24]. We use F1-score as the primary metric because it balances precision and recall for the minority FOG class, although it can be lower for subjects with few FOG events. Secondary metrics include AUROC, AUPRC, recall, and false positive rate (FPR). AUROC measures class separability and is relatively insensitive to changes in FOG prevalence, but can appear optimistic in imbalanced settings due to the abundance of non-FOG samples. AUPRC focuses on the FOG class and is therefore more informative under class imbalance, but is sensitive to prediction confidence; even when two models produce the same false positives, the model assigning higher scores to those false positives relative to true positives will achieve lower AUPRC. AUROC and AUPRC were computed from predicted probabilities, while F1-score, recall, and FPR were computed after thresholding predictions at 0.5. Metrics that are undefined for a held-out subject are omitted from the corresponding average rather than scored as zero. Two subjects contribute no FOG-positive window under our labelling rule, so F1, recall, AUPRC and AUROC are averaged over the remaining 11 subjects, while FPR is defined for all 13; the same rule applies within the per-task breakdown. Because the same subjects are excluded for every model, this leaves the model rankings and paired comparisons unchanged.

Statistical analyses were performed on the subject-level scores obtained from LOSO-CV. Differences between modalities or models were first assessed using the Friedman test, a non-parametric repeated-measures alternative to ANOVA. Post-hoc pairwise comparisons were conducted using the Wilcoxon signed-rank test. To control for multiple comparisons p-values were adjusted using Holm's correction. Statistical significance was defined as $p < 0.05$.

4 Results

4.1 Ego-vision contains FOG-relevant information

Table 2 reports the main comparison at the 2 s window length across IMU and ego-vision foundation models, alongside the fully trained TCN baselines. Notably, adding gyroscope data to the TCN baseline lowered F1 relative to accelerometer alone (35.5 vs. 42.3), but also reduced the number of false positives, as reflected by the reduced FPR (5.8% vs. 7.2%). Among the IMU FMs,

Table 2: Window-level FOG detection on the ADL tasks. Windows are 2 seconds long. Rows in the foundation-model blocks are linear probes on frozen features; the final block reports fully trained TCN IMU baselines. Models are evaluated leave-one-subject-out and metrics reported are mean (± std) across held-out subjects. All values are %. **Bold**/underline: the best and second-best per column for foundation models.

	Model	F1↑	Recall↑	FPR↓	AUPRC↑	AUROC↑			
						D	H	DL	Overall
Frozen foundation models									
IMU	UniMTS	29.1 (± 15.8)	**56.7 (± 23.8)**	28.8 (± 10.8)*	28.9 (± 16.7)*	75.0 (± 17.2)	72.8 (± 10.6)	76.9 (± 9.2)	73.5 (± 9.4)*
IMU	Chronos-2	**38.7 (± 17.4)**	39.9 (± 18.8)	**6.1 (± 4.0)**	**42.0 (± 20.7)***	**91.1 (± 6.4)**	**81.4 (± 10.9)**	**83.9 (± 6.8)**	**82.9 (± 7.5)**
Vision	DINOv3	17.0 (± 8.9)*	28.1 (± 10.0)	19.0 (± 11.8)*	15.4 (± 10.6)*	56.4 (± 17.6)*	53.1 (± 14.5)*	58.0 (± 10.5)	53.6 (± 7.0)*
Vision	VideoMAE-v2	22.0 (± 14.1)	32.8 (± 22.6)	19.3 (± 13.1)	21.8 (± 14.0)*	79.3 (± 12.2)	65.3 (± 15.8)	68.3 (± 9.5)	68.6 (± 9.1)*
Vision	EgoVideo	27.7 (± 16.4)	52.3 (± 24.6)	26.7 (± 17.2)*	23.9 (± 14.2)*	74.3 (± 11.5)	70.6 (± 11.3)	75.5 (± 9.3)	72.4 (± 7.7)*
Vision	V-JEPA2	32.6 (± 13.6)	47.4 (± 25.7)	17.6 (± 16.7)	33.0 (± 16.9)*	87.1 (± 11.5)	77.6 (± 10.2)	77.1 (± 14.2)	77.2 (± 6.3)*
Fully trained IMU Baseline									
IMU	TCN (acc only)	42.3 (± 19.8)	48.6 (± 24.6)	7.2 (± 6.2)	47.3 (± 20.7)	90.1 (± 10.1)	83.5 (± 10.5)	80.6 (± 11.1)	83.0 (± 8.8)
IMU	TCN (acc+gyro)	35.5 (± 21.6)	36.6 (± 25.8)	5.8 (± 4.1)	40.0 (± 23.4)	87.1 (± 17.7)	78.9 (± 14.2)	80.3 (± 11.3)	83.3 (± 7.1)

Per-task AUROC: D = doorway, H = hotspot, DL = daily life. * indicates a statistically significant difference from TCN (acc only) ($p_{\text{corrected}} < 0.05$, Wilcoxon signed-rank test with Holm correction).

Chronos-2 was the stronger representation (F1 38.7 vs. 29.1) and operated at a much lower FPR (6.1% vs. 28.8%), whereas UniMTS achieved higher recall (56.7% vs. 39.9%). Ego-video representations were predictive of FOG, most clearly for EgoVideo and V-JEPA2. On F1, the single-frame DINOv3 representation was weakest (17.0), significantly below the trained TCN. Comparing DINOv3 with VideoMAE-v2, which share a backbone and output dimensionality, AUROC rose from 53.6 to 68.6 when moving from one frame to sixteen. All three video encoders exceeded the single-frame baseline on AUROC (68.6, 72.4 and 77.2 vs. 53.6; $p < 0.01$ in each case), indicating that egocentric video carries FOG-relevant information beyond a single frame. Capacity does not explain these differences: the largest backbone (EgoVideo, ≈1 B) did not outperform V-JEPA 2 (≈300 M; F1 27.7 vs. 32.6). Across all models, the standard deviations are large relative to the mean scores reflecting high between-subject variability.

Notably, V-JEPA2 was competitive with the IMU FMs. It surpassed the motion-specialized UniMTS on F1 (32.6 vs. 29.1), AUROC (77.2 vs. 73.5), and FPR (17.6% vs. 28.8%). Results suggest frozen egocentric video features alone are predictive of FOG at a level comparable to the IMU FMs, with the fully trained TCN (F1 42.3) remaining the strongest reference.

4.2 Effect of window length

Table 3 examines how the standalone performance of each IMU and ego-video FM changes as the temporal window widens from 2 s to 10 s. Across models, F1-score generally increased with window length (*e.g.*, Chronos-2 38.7 to 42.4, V-JEPA2 32.6 to 36.4), likely due to the higher proportion of FOG windows and the resulting reduction in class imbalance (Figure 3). In contrast, AUROC consistently decreased (*e.g.*, Chronos-2 82.9 to 71.6, V-JEPA2 77.2 to 66.4),

Table 3: Window size ablation for video and IMU foundation models. Values are F1/AUROC (%) averaged over the LOSO subjects. **Bold**: best value per column.

	2 s		3 s		10 s	
Model	F1	AUROC	F1	AUROC	F1	AUROC
IMU foundation models						
UniMTS	29.1	73.5	34.4	73.4	**45.1**	67.7
Chronos-2	**38.7**	**82.9**	**42.0**	**82.2**	42.4	**71.6**
Ego-video foundation models						
DINOv3	17.0	53.6	18.6	57.2	31.4	54.3
VideoMAE-v2	22.0	68.6	29.9	71.4	41.7	65.6
EgoVideo	27.7	72.4	32.4	71.0	35.5	54.4
V-JEPA2	32.6	77.2	34.8	74.5	36.4	66.4

suggesting that longer windows reduce the separability of FOG and non-FOG samples. A likely explanation is that longer windows contain a mixture of FOG and non-FOG behaviors and less precise temporal alignment with FOG events, producing more heterogeneous samples and noisier labels. Consequently, comparisons across window lengths should be interpreted cautiously, as increasing F1 does not necessarily indicate improved discriminative performance. By AUROC, Chronos-2 remained the strongest model at every window length, and V-JEPA2 was the best ego-video representation at the shorter 2 and 3 s windows.

4.3 Qualitative Findings

Representative examples of model predictions are shown in Figure 5. In several tasks, both modalities detected FOG episodes at similar time points (*e.g.*, subjects 007, 010 Doorway) suggesting that ego-vision contains independent predictive information. Individual recordings also show segments where the ego-vision model identifies prolonged stopping better than the IMU models (Fig. 5, middle). This does not hold across the cohort: measuring FPR on windows containing at least 0.5 s of annotated stopping and no FOG, V-JEPA2 produced more false alarms than both Chronos-2 (24.9% vs. 9.7%) and the acc-only TCN (11.0%), and every model false-alarmed more during stopping than on other non-FOG windows (see Supplementary Material). Voluntary stopping is therefore a failure mode shared across modalities, and whether visual context helps appears to vary by recording. The vision model also missed some FOG episodes correctly identified by the IMU models (*e.g.*, subject 008). In a few cases ego-vision performed much worse than the IMU baseline (*e.g.*, subjects 015 and 014). For example, subject 015 had no annotated FOG episodes yet generated many false positives for V-JEPA2, possibly because prolonged stopping occurred during object interaction (*e.g.*, cleaning a cat litter box) while the hands were not visible. Similarly, subject 014 displayed poor performance, potentially due to the low-light recording conditions (Fig. 5, bottom right), which may have degraded the quality of the extracted visual representations.

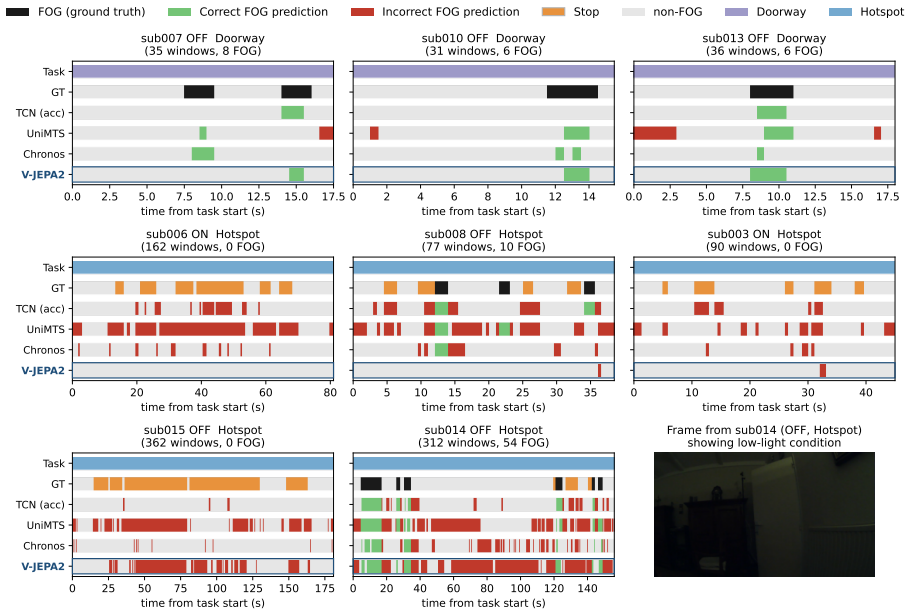


Fig. 5: Predictions over time for the ground-truth (GT), the trained TCN, the IMU FMs, and V-JEPA2. Each timestep represents a model prediction made every 0.5 s using 2 s windows of input signal. The segments are shown as FOG, Stop, or non-FOG. Stop is treated as non-FOG during training and displayed here for comparison. The ego-vision model mostly agrees with the IMU models (top) and in some cases better separates real freezes from voluntary stops (middle). It does not always help (bottom): false FOG predictions for sub015, and a poorly lit sub014 segment (frame on the right) where the visual features fail. Illustrative examples only.

Together, these analyses suggest that ego-vision provides information that is not simply redundant with inertial sensing. At the same time, the errors observed under low-light conditions and in subjects with atypical movement patterns show that visual context alone is insufficient, suggesting further research in multimodal models that can combine contextual and kinematic information.

5 Limitations and Future Work

This study shows that ego-vision enables context-aware clinical motion understanding during home-based ADL through the use case of FOG detection. We provided evidence that ego-vision features from frozen FM representations carry clinically meaningful motion-context information relevant for FOG detection. However, it did not reduce false positives during voluntary stopping relative to the strongest inertial baselines except in a few subjects. Finally, pretrained FM features can reach similar performance to a TCN trained from scratch, supporting their potential for clinical motion applications where labeled datasets

are limited. It remains unclear whether the ego-vision signal reflects motion, environmental context, or both and future work could address this through a deeper interpretability analysis such as applying explainability techniques on the learned feature space.

These findings should be interpreted in light of the scale of the dataset. Collecting synchronized multimodal recordings from people with PD at homes is technically and clinically challenging, particularly when recordings include both medication states, multiple wearable sensors, ego-video, and expert video-based FOG annotation. As a result, the present dataset includes a relatively small cohort and short semi-structured ADLs. Although this setting captures more real-world behavior than laboratory protocols and includes substantial variability in FOG severity and home environments, longer unsupervised free-living recordings are needed to evaluate robustness across broader daily-life conditions. Furthermore, while no participants reported concerns regarding the smart glasses, their long-term usability and acceptability require further investigation, particularly given the potential practical and privacy considerations in everyday use.

Beyond data scale, our framework relies solely on frozen features from pretrained FMs. While this is efficient in a small-data regime and avoids overfitting, it leaves the encoders unadapted to FOG, and fine-tuning strategies such as low-rank adaptation (LoRA) [16] may yield further gains. The results suggest complementary error patterns across modalities, though this remains preliminary since fusion of IMU and ego-vision was not explored here, motivating future work on multimodal fusion and prediction-refinement approaches [44,47]. Finally, the current pipeline performs offline classification over pre-segmented windows and is not directly applicable to real-time use, whereas practical home monitoring and closed-loop interventions ultimately require online, low-latency inference. Bridging this gap will require adapting the approach to streaming, causal prediction and developing lightweight or distilled encoders that can run continuously on wearable or on-device hardware [33].

6 Conclusion

This study presents the first use of ego-vision for FOG detection and demonstrates its potential for context-aware motion understanding during home-based ADLs. Using frozen representations from pretrained video and IMU foundation models with a simple linear probe, we achieved performance comparable to a fully trained TCN baseline, suggesting that clinically relevant predictive information is already present in these representations. Furthermore, ego-vision alone performed competitively with IMU-based models, providing evidence that visual context contains independent cues relevant to FOG. A direct test on annotated stop windows, however, showed that ego-vision FM features alone do not disambiguate voluntary stopping from freezing better than inertial sensing. Together, these findings motivate future work on context-aware multimodal approaches, including FM fine-tuning, fusion strategies, and real-time deployment for home-based clinical motion monitoring.

Acknowledgements

We thank all participants who gave their time and effort to participate in the study. This study was funded, in part, by the AidWear project funded by the Federal Public Service for Policy and Support, the AID-FOG project by the Michael J. Fox Foundation for Parkinson's Research under Grant No.: MJFF-024628, the strategic basic research project RevalExo (S001024N) funded by the Research Foundation Flanders, and the Flemish Government under the Flanders AI Research Program (FAIR). The resources and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by the Research Foundation - Flanders (FWO) and the Flemish Government.

References

1. Adeli, V., Klabučar, I., Rajabi, J., Filtjens, B., Mehraban, S., Wang, D., Seo, H., Hoang, T.H., Do, M.N., Muller, C., de Oliveira, C.N., Coelho, D.B., Ginis, P., Gilat, M., Nieuwboer, A., Spildooren, J., McKay, J.L., Kwon, H., Clifford, G., Esper, C.D., Factor, S.A., Genias, I., Dadashzadeh, A., Shum, L., Whone, A., Mirmehdi, M., Iaboni, A., Taati, B.: Care-pd: A multi-site anonymized clinical dataset for parkinson's disease gait assessment. In: NeurIPS (2025), https://proceedings.neurips.cc/paper_files/paper/2025/hash/bedc73979a95be7727af0c9a99c675ce-Abstract-Datasets_and_Benchmarks_Track.html
2. Ansari, A.F., Shchur, O., Küken, J., Auer, A., Han, B., Mercado, P., Rangapuram, S.S., Shen, H., Stella, L., Zhang, X., Goswami, M., Kapoor, S., Maddix, D.C., Guerron, P., Hu, T., Yin, J., Erickson, N., Desai, P.M., Wang, H., Rangwala, H., Karypis, G., Wang, Y., Bohlke-Schneider, M.: Chronos-2: From univariate to universal forecasting. arXiv preprint arXiv:2510.15821 (2025). https://doi.org/https://doi.org/10.48550/arXiv.2510.15821
3. Assran, M., Bardes, A., Fan, D., Garrido, Q., Howes, R., Komeili, M., Muckley, M., Rizvi, A., Roberts, C., Sinha, K., Zholus, A., Arnaud, S., Gejji, A., Martin, A., Robert Hogan, F., Dugas, D., Bojanowski, P., Khalidov, V., Labatut, P., Massa, F., Szafraniec, M., Krishnakumar, K., Li, Y., Ma, X., Chandar, S., Meier, F., LeCun, Y., Rabbat, M., Ballas, N.: V-jepa 2: Self-supervised video models enable understanding, prediction and planning. arXiv preprint arXiv:2506.09985 (2025). https://doi.org/https://doi.org/10.48550/arXiv.2506.09985
4. Bansal, S.K., Basumatary, B., Bansal, R., Sahani, A.K.: Techniques for the detection and management of freezing of gait in parkinson's disease–a systematic review and future perspectives. MethodsX **10**, 102106 (2023). https://doi.org/https://doi.org/10.1016/j.mex.2023.102106
5. Bloem, B.R., Hausdorff, J.M., Visser, J.E., Giladi, N.: Falls and freezing of gait in parkinson's disease: A review of two interconnected, episodic phenomena. Movement Disorders **19**(8), 871–884 (2004). https://doi.org/https://doi.org/10.1002/mds.20115, https://movementdisorders.onlinelibrary.wiley.com/doi/abs/10.1002/mds.20115
6. Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M.S., Bohg, J., Bosselut, A., Brunskill, E., et al.: On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258 (2021). https://doi.org/https://doi.org/10.48550/arXiv.2108.07258

7. Conde, C.I., Lang, C., Baumann, C.R., Easthope, C.A., Taylor, W.R., Ravi, D.K.: Triggers for freezing of gait in individuals with parkinson's disease: a systematic review. Frontiers in neurology **14**, 1326300 (2023). https://doi.org/https://doi.org/10.3389/fneur.2023.1326300
8. Damen, D., Doughty, H., Farinella, G.M., Furnari, A., Kazakos, E., Ma, J., Moltisanti, D., Munro, J., Perrett, T., Price, W., et al.: Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. International Journal of Computer Vision **130**(1), 33–55 (2022). https://doi.org/https://doi.org/10.1007/s11263-021-01531-2
9. van Dijsseldonk, R.B., van Nes, I.J., Geurts, A.C., Keijsers, N.L.: Exoskeleton home and community use in people with complete spinal cord injury. Scientific reports **10**(1), 15600 (2020). https://doi.org/https://doi.org/10.1038/s41598-020-72397-6
10. Filtjens, B., Ginis, P., Nieuwboer, A., Slaets, P., Vanrumste, B.: Automated freezing of gait assessment with marker-based motion capture and multi-stage spatial-temporal graph convolutional neural networks. Journal of NeuroEngineering and Rehabilitation **19**(1), 48 (2022). https://doi.org/https://doi.org/10.1186/s12984-022-01025-3
11. Filtjens, B., McCrum, C.: Perspectives on interdisciplinary posture and gait research from the ispgr 2025 world congress: Where do we stand and what are the next steps? Gait and Posture **124**, 110058 (2026). https://doi.org/https://doi.org/10.1016/j.gaitpost.2025.110058
12. Gilat, M.: How to annotate freezing of gait from video: a standardized method using open-source software. Journal of Parkinson's disease **9**(4), 821–824 (2019). https://doi.org/https://doi.org/10.3233/JPD-191700
13. Gilat, M., Nonnekes, J., Factor, S.A., Bloem, B.R., Nutt, J.G., Giladi, N., Hallett, M., Nieuwboer, A., Horak, F.B., Weiss, D., et al.: An updated definition of freezing of gait. Nature Reviews Neurology pp. 1–10 (2026). https://doi.org/https://doi.org/10.1038/s41582-025-01179-3
14. Ginis, P., Nackaerts, E., Nieuwboer, A., Heremans, E.: Cueing for people with parkinson's disease with freezing of gait: a narrative review of the state-of-the-art and novel perspectives. Annals of physical and rehabilitation medicine **61**(6), 407–413 (2018). https://doi.org/https://doi.org/10.1016/j.rehab.2017.08.002
15. Grauman, K., Westbury, A., Torresani, L., Kitani, K., Malik, J., Afouras, T., Ashutosh, K., Baiyya, V., Bansal, S., Boote, B., Byrne, E., Chavis, Z., Chen, J., Cheng, F., Chu, F.J., Crane, S., Dasgupta, A., Dong, J., Escobar, M., Forigua, C., Gebreselasie, A., Haresh, S., Huang, J., Islam, M.M., Jain, S., Khirodkar, R., Kukreja, D., Liang, K.J., Liu, J.W., Majumder, S., Mao, Y., Martin, M., Mavroudi, E., Nagarajan, T., Ragusa, F., Ramakrishnan, S.K., Seminara, L., Somayazulu, A., Song, Y., Su, S., Xue, Z., Zhang, E., Zhang, J., Castillo, A., Chen, C., Fu, X., Furuta, R., González, C., Gupta, P., Hu, J., Huang, Y., Huang, Y., Khoo, W., Kumar, A., Kuo, R., Lakhavani, S., Liu, M., Luo, M., Luo, Z., Meredith, B., Miller, A., Oguntola, O., Pan, X., Peng, P., Pramanick, S., Ramazanova, M., Ryan, F., Shan, W., Somasundaram, K., Song, C., Southerland, A., Tateno, M., Wang, H., Wang, Y., Yagi, T., Yan, M., Yang, X., Yu, Z., Zha, S.C., Zhao, C., Zhao, Z., Zhu, Z., Zhuo, J., Arbeláez, P., Bertasius, G., Damen, D., Engel, J., Maria Farinella, G., Furnari, A., Ghanem, B., Hoffman, J., Jawahar, C.V., Newcombe, R., Park, H.S., Rehg, J.M., Sato, Y., Savva, M., Shi, J., Shout, M.Z., Wray, M.: Ego-exo4d: Understanding skilled human activity from first- and third-person perspectives. In: 2024

IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 19383–19400 (2024). https://doi.org/10.1109/CVPR52733.2024.01834
16. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. Iclr **1**(2), 3 (2022). https://doi.org/https://doi.org/10.48550/arXiv.2106.09685
17. Kluge, F., Del Din, S., Cereatti, A., Gaßner, H., Hansen, C., Helbostad, J.L., Klucken, J., Küderle, A., Müller, A., Rochester, L., et al.: Consensus based framework for digital mobility monitoring. PloS one **16**(8), e0256541 (2021). https://doi.org/https://doi.org/10.1371/journal.pone.0256541
18. Koltermann, K., Clapham, J., Blackwell, G., Jung, W., Burnet, E.N., Gao, Y., Shao, H., Cloud, L., Pretzer-Aboff, I., Zhou, G.: Gait-guard: Turn-aware freezing of gait detection for non-intrusive intervention systems. In: 2024 IEEE/ACM Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE). pp. 61–72. IEEE (2024). https://doi.org/https://doi.org/10.1109/CHASE60773.2024.00016
19. Lea, C., Flynn, M.D., Vidal, R., Reiter, A., Hager, G.D.: Temporal convolutional networks for action segmentation and detection. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 156–165 (2017)
20. Liang, Y., Wen, H., Nie, Y., Jiang, Y., Jin, M., Song, D., Pan, S., Wen, Q.: Foundation models for time series analysis: A tutorial and survey. In: Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining. pp. 6555–6565 (2024). https://doi.org/https://doi.org/10.1145/3637528.3671451
21. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: Smpl: A skinned multi-person linear model. In: Seminal Graphics Papers: Pushing the Boundaries, Volume 2, pp. 851–866. Association for Computing Machinery (2023). https://doi.org/https://doi.org/10.1145/2816795.2818013
22. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2019), https://openreview.net/forum?id=Bkg6RiCqY7
23. Mancini, M., Bloem, B.R., Horak, F.B., Lewis, S.J., Nieuwboer, A., Nonnekes, J.: Clinical and methodological challenges for assessing freezing of gait: future perspectives. Movement Disorders **34**(6), 783–790 (2019). https://doi.org/https://doi.org/10.1002/mds.27709
24. Mancini, M., McKay, J.L., Cockx, H., D'Cruz, N., Esper, C.D., Filtjens, B., Heimler, B., MacKinnon, C.D., Palmerini, L., Roerdink, M., et al.: Technology for measuring freezing of gait: Current state of the art and recommendations. Journal of Parkinson's Disease **15**(1), 19–40 (2025). https://doi.org/https://doi.org/10.1177/1877718X241301065
25. Melbourne, J., Kehnemouyi, Y., O'Day, J., Wilkins, K., Gala, A., Petrucci, M., et al.: Kinematic adaptive deep brain stimulation for gait impairment and freezing of gait in parkinson's disease (2023). https://doi.org/https://doi.org/10.1016/j.brs.2023.07.003
26. Mesin, L., Porcu, P., Russu, D., Farina, G., Borzì, L., Zhang, W., Guo, Y., Olmo, G.: A multi-modal analysis of the freezing of gait phenomenon in parkinson's disease. Sensors **22**(7), 2613 (2022). https://doi.org/https://doi.org/10.3390/s22072613
27. Mir, A.N., Nissar, I., Ahmed, M., Masood, S., Rizvi, D.R.: Parkinson's disease diagnosis through deep learning: A novel lstm-based approach for freezing of gait detection. In: International Conference on Advances in Distributed Computing and Machine Learning. pp. 201–215. Springer (2024). https://doi.org/https://doi.org/10.1007/978-981-97-3523-5_16

28. Moore, J., Celik, Y., Stuart, S., McMeekin, P., Walker, R., Hetherington, V., Godfrey, A.: Using video technology and ai within parkinson's disease free-living fall risk assessment. Sensors **24**(15), 4914 (2024). https://doi.org/https://doi.org/10.3390/s24154914
29. Moore, J., Stuart, S., McMeekin, P., Walker, R., Celik, Y., Pointon, M., Godfrey, A.: Enhancing free-living fall risk assessment: contextualizing mobility based imu data. Sensors **23**(2), 891 (2023). https://doi.org/https://doi.org/10.3390/s23020891
30. Nouredanesh, M., Godfrey, A., Powell, D., Tung, J.: Egocentric vision-based detection of surfaces: towards context-aware free-living digital biomarkers for gait and fall risk assessment. Journal of neuroengineering and rehabilitation **19**(1), 79 (2022). https://doi.org/https://doi.org/10.1186/s12984-022-01022-6
31. Pei, B., Chen, G., Xu, J., He, Y., Liu, Y., Pan, K., Huang, Y., Wang, Y., Lu, T., Wang, L., Qiao, Y.: Egovideo: Exploring egocentric foundation model and downstream adaptation. arXiv preprint arXiv:2406.18070 (2024). https://doi.org/https://doi.org/10.48550/arXiv.2406.18070
32. Qian, C., Chi, C., Clapham, J., Qi, J., Zhang, Z., Blackwell, G., Pretzer-Aboff, I., Cloud, L., Ma, M., Zhou, G., et al.: Trigger-finder: A real-time freezing-of-gait trigger detection system using an instruction-tuned multimodal large language model. In: Proceedings of the ACM/IEEE International Conference on Connected Health: Applications, Systems and Engineering Technologies. pp. 1–12 (2025). https://doi.org/https://doi.org/10.1145/3721201.3721387
33. Qu, G., Chen, Q., Wei, W., Lin, Z., Chen, X., Huang, K.: Mobile edge intelligence for large language models: A contemporary survey. IEEE Communications Surveys & Tutorials **27**(6), 3820–3860 (2025). https://doi.org/https://doi.org/10.1109/COMST.2025.3527641
34. Rodríguez-Martín, D., Samà, A., Pérez-López, C., Català, A., Moreno Arostegui, J.M., Cabestany, J., Bayés, À., Alcaine, S., Mestre, B., Prats, A., et al.: Home detection of freezing of gait using support vector machines through a single waist-worn triaxial accelerometer. PloS one **12**(2), e0171764 (2017). https://doi.org/https://doi.org/10.1371/journal.pone.0171764
35. Shi, B., Tay, A., Au, W.L., Tan, D.M., Chia, N.S., Yen, S.C.: Detection of freezing of gait using convolutional neural networks and data from lower limb motion sensors. IEEE Transactions on Biomedical Engineering **69**(7), 2256–2267 (2022). https://doi.org/https://doi.org/10.1109/TBME.2022.3140258
36. Sigcha, L., Borzì, L., Pavón, I., Costa, N., Costa, S., Arezes, P., López, J.M., De Arcas, G.: Improvement of performance in freezing of gait detection in parkinson's disease using transformer networks and a single waist-worn triaxial accelerometer. Engineering Applications of Artificial Intelligence **116**, 105482 (2022). https://doi.org/https://doi.org/10.1016/j.engappai.2022.105482, https://www.sciencedirect.com/science/article/pii/S0952197622004729
37. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S.E., Ramamonjisoa, M., Massa, F., HAZIZA, D., Wehrstedt, L., Wang, J., Darcet, T., Moutakanni, T., Sentana, L., Roberts, C., Vedaldi, A., Tolan, J., Brandt, J., Couprie, C., Mairal, J., Jegou, H., Labatut, P., Bojanowski, P.: DINOv3. Transactions on Machine Learning Research (2026), https://openreview.net/forum?id=2NlGyqNjns, featured Certification
38. Tong, Z., Song, Y., Wang, J., Wang, L.: Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. Advances in neural in-

formation processing systems **35**, 10078–10093 (2022). https://doi.org/https://doi.org/10.48550/arXiv.2203.12602
39. Varghese, R., Sambath, M.: Yolov8: A novel object detection algorithm with enhanced performance and robustness. In: 2024 International conference on advances in data engineering and intelligent computing systems (ADICS). pp. 1–6. IEEE (2024). https://doi.org/https://doi.org/10.1109/ADICS58448.2024.10533619
40. Vinod, A., Pandit, S., Vavre, A., Liu, L.: Egovlm: Policy optimization for egocentric video understanding. arXiv preprint arXiv:2506.03097 (2025). https://doi.org/https://doi.org/10.48550/arXiv.2506.03097
41. Wang, L., Huang, B., Zhao, Z., Tong, Z., He, Y., Wang, Y., Wang, Y., Qiao, Y.: Videomae v2: Scaling video masked autoencoders with dual masking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 14549–14560 (June 2023)
42. Wang, Y., Beuving, F., Nonnekes, J., Cohen, M.X., Long, X., Aarts, R.M., Van Wezel, R.: Freezing of gait detection in parkinson's disease via multimodal analysis of eeg and accelerometer signals. In: 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). pp. 847–850. IEEE (2020). https://doi.org/https://doi.org/10.1109/EMBC44109.2020.9175288
43. Xu, H., Zhou, P., Tan, R., Li, M., Shen, G.: Limu-bert: Unleashing the potential of unlabeled data for imu sensing applications. In: Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems. pp. 220–233 (2021). https://doi.org/https://doi.org/10.1145/3485730.3485937
44. Xu, P., Zhu, X., Clifton, D.A.: Multimodal learning with transformers: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(10), 12113–12132 (2023). https://doi.org/https://doi.org/10.1109/TPAMI.2023.3275156
45. Yang, P.K., Carlon, J., Goris, M., Klaver, E., Nonnekes, J., van Wezel, R.J., Alcock, L., Yarnall, A.J., Rochester, L., Hansen, C., et al.: Deep learning for freezing of gait assessment using inertial measurement units: a multicentre validation study. npj Parkinson's Disease (2026). https://doi.org/https://doi.org/10.1038/s41531-026-01407-7
46. Yang, P.K., Filtjens, B., Ginis, P., Goris, M., Nieuwboer, A., Gilat, M., Slaets, P., Vanrumste, B.: Automatic detection and assessment of freezing of gait manifestations. IEEE Transactions on Neural Systems and Rehabilitation Engineering **32**, 2699–2708 (2024). https://doi.org/https://doi.org/10.1109/TNSRE.2024.3431208
47. Yang, P.K., Filtjens, B., Ginis, P., Goris, M., Nieuwboer, A., Gilat, M., Slaets, P., Vanrumste, B.: Multimodal freezing of gait detection: Analyzing the benefits and limitations of physiological data. IEEE Transactions on Neural Systems and Rehabilitation Engineering (2025). https://doi.org/https://doi.org/10.1109/TNSRE.2025.3545110
48. Yudayev, M., Carlon, J., Lamsal, D., Stefanova, V., Filtjens, B.: Hermes: A unified open-source framework for realtime multimodal physiological sensing, edge ai, and intervention in closed-loop smart healthcare applications (2026), https://arxiv.org/abs/2601.12610
49. Zhang, J., Huang, J., Jin, S., Lu, S.: Vision-language models for vision tasks: A survey. IEEE transactions on pattern analysis and machine intelligence **46**(8), 5625–5644 (2024). https://doi.org/https://doi.org/10.1109/TPAMI.2024.3369699

50. Zhang, W., Yang, Z., Li, H., Huang, D., Wang, L., Wei, Y., Zhang, L., Ma, L., Feng, H., Pan, J., et al.: Multimodal data for the detection of freezing of gait in parkinson's disease. Scientific data **9**(1), 606 (2022). https://doi.org/https://doi.org/10.1038/s41597-022-01713-8
51. Zhang, X., Teng, D., Chowdhury, R.R., Li, S., Hong, D., Gupta, R.K., Shang, J.: Unimts: Unified pre-training for motion time series. Advances in Neural Information Processing Systems **37**, 107469–107493 (2024). https://doi.org/https://doi.org/10.48550/arXiv.2410.19818
52. Zhao, L., Gundavarapu, N.B., Yuan, L., Zhou, H., Yan, S., Sun, J.J., Friedman, L., Qian, R., Weyand, T., Zhao, Y., et al.: Videoprism: A foundational visual encoder for video understanding. arXiv preprint arXiv:2402.13217 (2024). https://doi.org/https://doi.org/10.48550/arXiv.2402.13217
53. Zoetewei, D., Ginis, P., Goris, M., Gilat, M., Herman, T., Brozgol, M., Thumm, P.C., Hausdorff, J.M., Nieuwboer, A., D'Cruz, N.: Which gait tasks produce reliable outcome measures of freezing of gait in parkinson's disease? Journal of Parkinson's Disease **14**(6), 1163–1174 (2024). https://doi.org/https://doi.org/10.3233/JPD-240134

Supplementary Material for "Towards Context-Aware Clinical Motion Understanding in Daily Living at Home: Freezing of Gait Detection with Egocentric Vision"

Vayalet Stefanova[1]*, Diwas Lamsal[1]*, Margot Genbrugge[2], Maxim Yudayev[1], Christian Schlenstedt[3], Moran Gilat[2], Bart Vanrumste[1]†, and Benjamin Filtjens[4,5]†

[1] Department of Electrical Engineering (ESAT), KU Leuven, Leuven, Belgium
[2] Neurorehabilitation Research Group (eNRGy), Department of Rehabilitation Sciences, KU Leuven, Leuven, Belgium
[3] Institute of Interdisciplinary Exercise Science and Sports Medicine, Medical School Hamburg, Hamburg, Germany
[4] Department of Engineering Systems and Services, Delft University of Technology, Delft, The Netherlands
[5] Institute for Health Systems Science, Delft University of Technology, Delft, The Netherlands

1 Complete list of daily life tasks

The data collection protocol for each participant included five activities of daily living: walking through a doorway, two daily-life tasks, and two hotspot tasks. The two daily-life tasks were randomly selected from the following list of 14 predefined tasks:

1. Set the dining table for two people, sit at the table, then clear the table.
2. Walk from the couch/chair to the kitchen, fill a glass with water, get a snack, and return to the couch/chair.
3. (Fictively) water the plants in the living room.
4. Walk to every corner of the living room.
5. Turn on the main lights in all rooms tested.
6. Walk around the coffee table/dining table in both directions.
7. Walk to the front door, open and close it, then return to the living room.
8. Walk from the couch/chair to the table, get a book/newspaper/tablet, walk back to the couch/chair, and sit down.
9. Empty and fill the dishwasher:
   (a) If no dishwasher is present: (fictively) wash some dishes and put the clean tableware in its correct location
10. Close and open the curtains for all windows in all rooms.
11. Tidy up the rooms.
12. Go look at two art pieces or pictures in your house.
13. Clean/dust off some surfaces or furniture in the rooms.
14. Get a pair of shoes, put them next to the couch, sit on the couch, get up, and put the shoes back.

Table 1: Window counts across strides.

Stride	**0.5 s**	**1.0 s**	**1.5 s**
Overlap	75%	50%	25%
Windows	10,249	5,118	3,420
FOG windows	1,190	605	410

Table 2: Window overlap ablation for all models (0.5 s, 1 s, 1.5 s strides; 75%, 50%, 25% overlap). Values are F1/AUROC (%) averaged over the LOSO subjects. **Bold**: best value per column; underline: second-best.

	0.5 s (75%)		1.0 s (50%)		1.5 s (25%)	
Model	F1	AUROC	F1	AUROC	F1	AUROC
IMU foundation models						
UniMTS	29.1	73.5	29.1	73.0	27.4	71.3
Chronos-2	38.7	82.9	36.5	82.2	36.6	82.2
Ego-video foundation models						
DINOv3	17.0	53.6	14.9	52.2	15.6	54.3
VideoMAE-v2	22.0	68.6	22.6	68.5	23.5	68.6
EgoVideo	27.7	72.4	27.9	71.3	29.0	72.2
V-JEPA2	32.6	77.2	31.6	77.4	31.5	78.0
Fully trained IMU Baseline						
TCN (acc only)	**42.3**	83.0	38.4	82.6	**38.7**	**82.8**
TCN (acc+gyro)	35.5	**83.3**	**40.3**	**84.4**	38.4	82.3

2 Effect of window stride

Our main configuration uses a 2 s window with a 0.5 s stride, meaning consecutive windows overlap by 75%. Combined with the 0.5 s minimum-overlap FOG criterion and a median FOG episode duration of 0.9 s, a single FOG episode can generate multiple overlapping windows with varying degrees of temporal purity (i.e., varying proportions of FOG vs. non-FOG content within the window). To assess whether this affects reported performance, we evaluated all models at strides of 0.5 s, 1 s, and 1.5 s, corresponding to 75%, 50%, and 25% window overlap, respectively, while keeping all other hyperparameters fixed at the values optimized for the 0.5 s stride. The resulting number of windows per stride are displayed in Table 1 which shows that the number of FOG windows decreases roughly in proportion with the overall window count as stride increases.

Table 2 shows that performance remains broadly stable across strides for most models, with F1 and AUROC typically varying by only a few points as overlap decreases from 75% to 25%.

3 False alarms during voluntary stopping

A stop window is defined as one containing at least 0.5 s of annotated voluntary stopping and no FOG, mirroring the FOG criterion, giving 2,024 windows across all 13 subjects. Table 3 reports FPR restricted to these windows and to

Table 3: False positive rate on annotated voluntary-stop windows compared with other non-FOG windows, at the 2 s window length. Values are %, computed per subject and averaged over the 13 LOSO subjects. * indicates a significant difference from TCN (acc only) ($p_{\text{corrected}} < 0.05$, Wilcoxon signed-rank test with Holm correction).

Model	**FPR stop↓**	**FPR non-stop↓**
TCN (acc+gyro)	7.1	4.8
Chronos-2	9.7	4.4
TCN (acc only)	11.0	5.4
DINOv3	22.3*	17.6
V-JEPA2	24.9	13.8
VideoMAE-v2	33.3*	15.0
EgoVideo	44.5*	20.5
UniMTS	49.1*	21.7

the remaining non-FOG windows, computed per subject and averaged and the ordering follows each model's overall false-alarm rate. Every model false-alarmed more on stop windows than on other non-FOG windows showing the difficulty in differentiating FOG episodes from voluntary stops. V-JEPA2 (24.9%) did not improve on Chronos-2 (9.7%) or the trained TCN (11.0%), which may partly reflect the ego-video models' generally lower overall performance relative to the IMU-based baselines. Whether combining visual and inertial modalities could yield further improvement remains an open question for future work.