

Event-specific generalisation of gait-event detectors: a 16-dataset benchmark with real markerless-video validation across two pose-estimation chains

Frédéric Fer

f.fer@institut-myologie.org

Institut de Myologie

Thomas Marques

Institut de Myologie

Jean-Yves Hogrel

Institut de Myologie

Maxime Jacoupy

Institut de Myologie

Romain Feigean

Institut de Myologie

Method Article

Keywords: gait event detection, heel strike, toe-off, Bayesian inference, motion capture, marker-based, markerless

Posted Date: October 1st, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10854607/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Event-specific generalisation of gait-event detectors: a 16-dataset benchmark with real markerless-video validation across two pose-estimation chains

Frédéric Fer^{1*}, Thomas Marques², Jean-Yves Hogrel², Maxime Jacoupy¹, Romain Feigean²

¹ Institut de Myologie, Paris, France, Myodata team

² Neuromuscular Physiology and Evaluation Laboratory, Institute of Myology, Neuromuscular Exploration and Evaluation Center, Paris, France

* Correspondence: f.fer@institut-myologie.org

Abstract

Background. Automated detection of heel strike (HS) and toe-off (TO) underpins clinical gait analysis, yet most are validated on one dataset and rarely tested across unseen populations or markerless chains. It is unclear whether performance is set by algorithmic sophistication or by the transportability of the underlying kinematic features.

Methods. We assembled sixteen marker-based motion-capture datasets spanning healthy adults and seven clinical populations (729 unique individuals), split a priori into ten development and six held-out datasets. BIKE (Bayesian Inference for Kinematic Events) is an interpretable two-stage detector combining limb-crossing segmentation with composite-likelihood classification over biomechanics-informed kinematic features. BIKE and nine published comparators were evaluated at 50-ms tolerance, with macro-averaged F1 as the primary metric. Transfer to real video was tested against marker-based reference through two markerless chains: single-camera MediaPipe (fifteen healthy participants, 200 to 15 fps) and Sapiens 2 (eight participants, 60 to 15 fps); the Sapiens 2 chain also gave a feature-level event-timing analysis.

Results. BIKE achieved the highest macro-averaged F1 (HS 0.87, TO 0.86) through cross-dataset consistency, not per-dataset dominance (first-ranked on 4/16 datasets vs. 7/16 for IntellEvent). On held-out datasets, BIKE and IntellEvent were comparable (combined macro-F1 0.925 vs. 0.924) and both exceeded Zeni (0.82). In real video,

generalisation was event-specific: BIKE had the highest HS F1 in six of seven MediaPipe conditions and, in the Sapiens 2 chain, was the only detector strong on both events at 60 fps (HS 0.83, tied with Zeni; TO 0.87). The marker-to-video transition cost biomechanics-based detectors 6-12 F1 points versus 26-31 for deep-learning comparators. Band-limited velocity features preserved event-timing information (~23 ms), whereas angular accelerations retained amplitude discriminability but not usable timing. Markerless performance was preprocessing-dependent (BIKE HS F1 0.83->0.13 without the video pipeline) and collapsed below 30 fps.

Conclusions. Across marker-based data and two markerless chains, generalisation was event-specific and tracked the transportability of kinematic features more than model architecture: heel-strike cues, carried by band-limited velocity features, transferred reliably across populations, conditions and pose estimators, whereas toe-off stayed sensitive to feature choice and degradation. Pathological gait and markerless acquisition were assessed in disjoint cohorts; their interaction remains untested. Real-video validation, reported per event and chain, should be routine when assessing gait-event detectors for clinical purpose.

Keywords: gait event detection, heel strike, toe-off, Bayesian inference, motion capture, marker-based, markerless

Background

Initial contact (HS, heel or foot strike) and foot rise (TO, toe or foot off) are ones the most used key temporal events of the gait cycle. They delimit stance and swing phases and are required to derive spatiotemporal parameters used in biomechanical and clinical gait analysis [1,2,3]. In laboratory settings, kinetic data from force plates remain the robust gold standard for gait event detection, but kinematic event detection is required whenever force plates are unavailable, including treadmill protocols, long-walkway assessments, and the rapidly expanding field of markerless video-based gait analysis [4,5,6]. As video-based motion analysis moves closer to routine clinical, drug-development, and field use, the need for detectors that remain reliable outside controlled laboratory conditions has become increasingly essential, where pose-estimation noise and reduced frame rates can degrade the signals, these detectors rely on.

A wide range of kinematic gait-event detectors has been proposed. Existing detectors span several families: position-based methods identifying foot-displacement extrema, temporal-based approaches using threshold crossings, multi-cue composite rules, and supervised learning models trained on labelled gait data. In the related context of inertial measurement units, Bayesian inference and hidden Markov models have also been applied to phase detection [7,8],

demonstrating the potential of probabilistic frameworks for sequential event classification. However, despite their methodological diversity, most published detectors have been validated on a single dataset or a narrow set of acquisition conditions, mostly in healthy adults. This makes it difficult to determine whether performance transfers to clinical populations with idiopathic gait patterns, different marker protocols, curved or non-standard walking tasks, or the noisier and temporally less stable signals produced by markerless pose-estimation acquisition chains. This limitation is not trivial: pathology alters joint kinematics, walking speed, segment coordination, and limb-loading patterns, while video-based pipelines add positional noise, temporal jitter, and projection artefacts [6,9]. Features that are highly discriminative in healthy laboratory gait may therefore degrade substantially when gait mechanics or acquisition conditions change.

We postulated that the cross-dataset generalisability of a detector relies less on algorithmic sophistication than on the invariance of the underlying features across a large variability of gait patterns. We define feature transportability as the extent to which a feature retains event-linked discriminative structure across populations, protocols, and acquisition system, a concept distinct from algorithmic generalisation, which may additionally depend on model architecture and training procedure. Features based on absolute positions or amplitudes may be sensitive to speed reduction, altered segment trajectories, or tracking noise, whereas relational or dimensionless features may better preserve useful structure across conditions. We further hypothesised that this transportability is event specific. The biomechanical signatures associated with HS, which reflect ground contact, may remain more stereotyped across conditions than those associated with TO, which depend on underlying propulsive mechanisms and unloading patterns. If so, detectors may generalise differently for HS and TO even when their overall performance appears similar.

The aims of this study were threefold. First, we assembled a benchmark of sixteen open data motion-capture datasets spanning healthy participants and seven clinical populations and partitioned them a priori into 10 development and 6 held-out datasets to evaluate cross-dataset generalisation without dataset-specific parameter retuning. Second, we developed BIKE (Bayesian Inference for Kinematic Events), an interpretable two-stage detector combining limb-crossing intervals with biomechanically motivated kinematic features and literature-informed priors, compared to nine published detectors, including deep-learning approaches. Third, we performed external validation on real markerless video recorded from 15 participants under seven progressively constrained acquisition conditions, from laboratory-grade capture to low-frame-rate hallway recordings, to test whether marker-based detectors transfer to real-world video

conditions without modification. Throughout, HS and TO were evaluated separately in order to characterise the event-specific nature of detector generalisation.

Based on these aims, we tested three specific hypotheses: i) BIKE, an interpretable and training-free detector, achieves the highest macro-averaged benchmark performance, reflecting cross-dataset consistency rather than uniform per-dataset dominance; ii) performance on held-out datasets remains high without dataset-specific retuning, while event-specific differences between HS and TO become more apparent; and iii) in real-video experiments spanning two independent pose-estimation pipelines, HS detection transfers more reliably than TO across increasingly constrained acquisition conditions.

Methods

Study design

The evaluation was structured around three complementary phases of validation (Figure 1). First, the full 16-dataset motion-capture benchmark provided a broad assessment of detector performance across heterogeneous populations, marker protocols, ground-truth sources, and sampling rates. Second, six held-out datasets, not used during feature analysis or prior calibration, tested cross-dataset generalisation without any dataset-specific retuning. Third, real-video markerless evaluation assessed whether detectors originally designed for marker-based input could transfer to actual video-derived joint kinematics under increasingly constrained acquisition conditions. This three-level design separates overall robustness (full benchmark), external validity (held-out), and cross-modal transferability (real markerless video). Dataset-level macro-averaged F1 score (harmonic mean of precision and recall) at 50 ms served as the primary summary metric, alongside pooled micro-averaged F1 to quantify overall event-level performance, thereby preventing the largest datasets from dominating the benchmark summary.

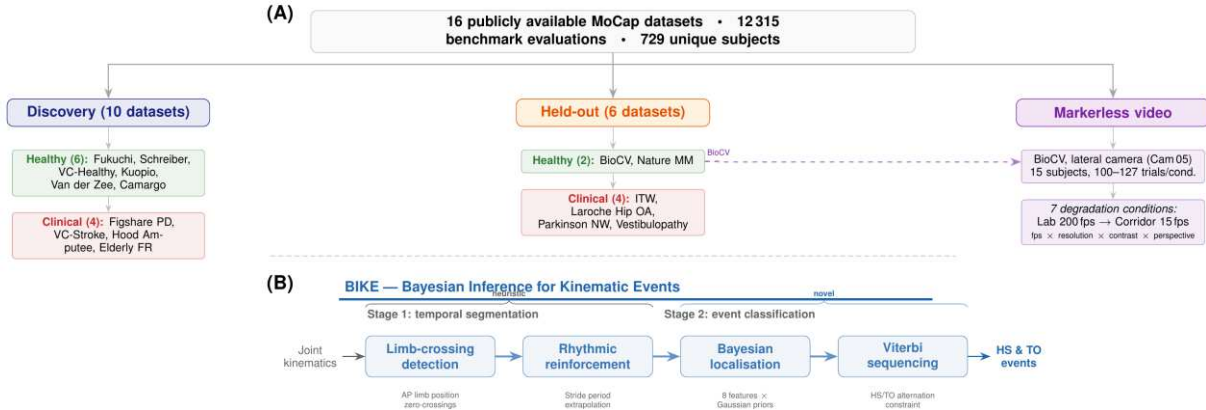


Figure 1. Study design and evaluation framework. (A) The sixteen motion-capture datasets were split into 10 development datasets and 6 held-out datasets, covering healthy adults, neurological, orthopaedic, and paediatric populations. (B) The BIKE two-stage pipeline: Stage 1 identifies half-gait-cycle intervals via limb-crossing detection; Stage 2 localises HS and TO within each interval using eight core biomechanical feature families, instantiated as event-specific priors (9 for HS, 10 for TO) with mixed literature-informed and empirically calibrated parameters. A real-video markerless evaluation was conducted on 15 BioCV-Video participants across seven degradation conditions.

Datasets and cohort accounting

Sixteen publicly available datasets were included (Table 1), selected based on available marker-based motion capture, reference gait events, and sufficient metadata for harmonised processing across diverse clinical populations. Together they encompass walking recordings from healthy young and older adults [10,11,12,13,14,15,16], individuals with Parkinson’s disease [17,18], post-stroke survivors [15], transfemoral amputees [19], elderly adults at risk of falls [20], children with idiopathic toe walking [21], adults with hip osteoarthritis [22,23], and patients with bilateral vestibulopathy [24]. Sampling rates ranged between 100 and 200 Hz. Reference events were derived from force plates (7 datasets, two with partial walkway coverage), expert manual annotation (4 datasets), automated kinematic labelling in Vicon Nexus (2 datasets), or gait-cycle percentage from treadmill software (2 datasets); one dataset (BioCV-motion-capture) used Visual3D-processed joint centres with annotated events.

The datasets were partitioned a priori into a 10-dataset development cohort (8688 benchmark evaluations) used for feature discriminability analysis and prior calibration, and a 6-dataset held-out cohort (3627 evaluations) reserved exclusively for external validation. The held-out cohort introduced several new clinical settings, including idiopathic toe walking (ITW), hip osteoarthritis, vestibulopathy, and a Parkinson cohort collected under a different protocol. Across all datasets, the benchmark comprised 778 dataset–subject entries (dataset-specific participant records), 729 unique individuals after overlap resolution (i.e. removal of duplicate participants across datasets), and 12315

benchmark evaluations corresponding to 9808 unique recordings. The only cross-dataset overlap involved the Schreiber and Nature Multimodal entries, which draw on the same underlying recordings from the Schreiber and Moissenet cohort: Schreiber is the annotated subset used for development, and Nature Multimodal is the fuller release retained as a same-cohort held-out check. The two are therefore not independent, and all pooled, subject-level and bootstrap analyses deduplicate this overlap so that no recording is counted twice.

One dataset (Camargo) included both treadmill and curved-path overground conditions; only treadmill recordings were retained because the curved trajectory violates the straight-line walking assumption underlying anteroposterior-based detectors. Curved-path results are reported in the Additional file 1. Two datasets (Kuopio and Van der Zee) used force plates covering only part of the walkway; for these datasets, a zone-aware scoring strategy was applied (see “Outcomes and statistical analysis”).

Table 1. Datasets included in the benchmark. Note: Rec. = unique recordings; benchmark evaluations = detector-recording instances after exclusions.

Dataset	Role	Population	GT	Hz	Subj.	Rec.
BioCV	Held-out	Healthy	Visual3D	200	15	139
Camargo	Dev.	Healthy + treadmill	GC%	200	22	157
Elderly Fall Risk	Dev.	Elderly fall risk	Ann	100	40	150
Figshare PD	Dev.	Parkinson’s	FP	150	26	506
Fukuchi	Dev.	Healthy lifespan	GC%	150	42	1683
Hood Amputee	Dev.	Lower-limb amputee	Ann	200	13	323
ITW	Held-out	Children (ITW+ctrl)	FP	100	29	300
Kuopio	Dev.	Healthy	FP*	100	51	3173
Laroche Hip OA	Held-out	Hip osteoarthritis	FP	100	189	992
Nature Multimodal	Held-out*	Healthy lifespan	Ann	100	50	251
Parkinson NW	Held-out	Parkinson’s (NW)	FP	100	24	72
Schreiber	Dev.	Healthy	Ann	100	49	1105
VanCrieke Healthy	Dev.	Healthy lifespan	Nexus	100	138	449
VanCrieke Stroke	Dev.	Stroke (subacute)	Nexus	100	50	121
Van der Zee	Dev.	Healthy	FP	120	10	327

Dataset	Role	Population	GT	Hz	Subj.	Rec.
Vestibulopathy	Held-out	Vestibular disorders	FP	100	30	60

Abbreviations: HS = heel strike; TO = toe-off; F1 = F1-score; MAE = mean absolute error (ms); Dev. = development; Rec. = recordings. *Nature Multimodal comprises the same recordings as the development dataset Schreiber (the fuller release of the same Schreiber and Moissenet cohort) and is therefore not an independent dataset; it is retained only as a same-cohort robustness check, and the overlap is deduplicated in all pooled and subject-level statistics.

Outcomes and statistical analysis

HS and TO were evaluated as separate primary outcomes throughout. For the motion-capture benchmark (full 16-dataset and held-out analyses), the primary summary metric was dataset-level macro-averaged F1 at 50-ms tolerance: within each dataset, micro-F1 was computed by pooling true positives (TP), false positives (FP), and false negatives (FN) across all trials, and the macro-average was then obtained as the unweighted mean of per-dataset micro-F1 values. This approach reduces domination by the largest datasets and ensures equal contribution of each dataset to the primary benchmark summary. A combined F1 score was defined as the arithmetic mean of HS F1 and TO F1: Combined F1 = $(\text{HS F1} + \text{TO F1}) / 2$. Pooled micro-averaged F1 across all trials was reported as a secondary metric to quantify overall event-level performance.

$$F_1 = \frac{2 \text{ TP}}{2 \text{ TP} + \text{FP} + \text{FN}}$$

For the real-video markerless evaluation, pooled micro-F1 was computed separately for each acquisition condition and served as the primary condition-level performance metric, since all trials originated from a single dataset and the main question concerned robustness across increasingly constrained acquisition conditions. Secondary outcomes included mean absolute error (MAE) of matched events (computed on matched true positives only), signed timing error, detector failure rate (proportion of trials with zero detected events despite reference events being present), and F1 at alternative tolerances of 25 ms and 75 ms.

For each trial, detected events were matched to reference events using nearest-neighbour matching within the tolerance window, with each reference event matched to its closest unmatched detection. Matched events contributed to TP; unmatched detections to FP; unmatched reference events to FN. This procedure was preferred over optimal Hungarian matching because the minimum same-type inter-event interval across all 11248 included trials was 232 ms, making the two procedures effectively equivalent at 50 ms in the present datasets.

Two datasets (Kuopio and Van der Zee) used force plates covering only part of the walkway. For these datasets, a valid frame range was computed from the temporal extent of reference events with a 0.5-s margin; events outside this

range were excluded from both reference and detections before scoring, thereby avoiding systematic false positives from detections outside the instrumented zone.

Detector rankings were assessed using Friedman tests on per-dataset summary values ($N = 16$ for the full benchmark, $N = 6$ for held-out), followed by pairwise Wilcoxon signed-rank tests with Bonferroni correction. Because trials from the same participant are not independent, a subject-level analysis was also performed by averaging per-trial F1 within each participant and repeating rank-based comparisons on participant-level summaries. For the markerless evaluation, condition effects were additionally examined using subject-level repeated-measures summaries across acquisition conditions. Because DeepEvent could not be evaluated on all datasets due to marker-protocol incompatibility, it was excluded from omnibus paired comparisons and included only in shared-coverage pairwise analyses. Because force plates provide the most reliable event timing, analyses stratified by ground-truth type, including the force-plate subset, were reported as a secondary temporal-accuracy analysis.

The BIKE detector

BIKE (Bayesian Inference for Kinematic Events) decomposes gait event detection into two sequential inference stages: temporal segmentation into half-cycle intervals (Stage 1) and prior-constrained probabilistic event classification within each interval (Stage 2).

Stage 1 - Limb-crossing detection. In gait, the lower limbs alternate in anteroposterior (AP) position. A crossing signal $d(t)$ is computed as the difference in mean AP position between the left and right knee–ankle complexes, relative to the pelvis midpoint. Crossings occur when $d(t)$ approaches zero with high velocity. A composite crossing score combines three components: a proximity term (Gaussian closeness to zero), a normalised velocity magnitude, and a rhythmic regularisation term predicting future crossings from observed inter-crossing intervals. Crossings are detected as peaks of this score using a hierarchical threshold strategy. Consecutive crossings define half-cycle intervals, each expected to contain one HS and one TO from the other limb.

Stage 2 - Probabilistic event classification. Within each half-cycle interval, BIKE localises HS and TO by maximising a log-posterior over candidate frames. At each frame, an event-specific feature set is computed from pre-specified kinematic cues: (1) normalised heel-to-knee AP distance, (2) pelvis-relative ankle AP velocity (normalised by thigh length), (3) normalised heel vertical position (or toe for TO), (4) heel vertical velocity (or toe for TO), (5) knee flexion–extension angle, (6) knee angular velocity, (7) toe velocity ratio (R_{toe}), a dimensionless fraction of total

lower-limb velocity carried by the toe marker, and (8) differential angular velocity between knee and ankle ($\Delta\omega$). Features 1 to 6 are common biomechanical quantities whose expected values at HS and TO are grounded in normative gait data and follow ISB joint-angle conventions. Features 7 and 8 were designed to improve transportability across heterogeneous acquisition conditions: R_{toe} is dimensionless by construction and less sensitive to absolute scale, whereas $\Delta\omega$ was intended to attenuate offsets common to the two joint-angle estimates, for example from markerless pose-estimation bias; the extent to which thigh- and shank-level soft-tissue artefacts are genuinely shared is more limited [25,26], consistent with the modest empirical contribution of $\Delta\omega$ to detector performance (see Results, “Feature transportability and mechanistic interpretation”). In addition to these eight core features, the active feature set includes R_{knee} (ratio of knee angular velocity to total lower-limb angular velocity) and, for TO only, ankle_toe_ap_vel (anteroposterior velocity difference between ankle and toe), yielding 9 HS features and 10 TO features.

The likelihood of the feature vector given event type E is modelled as a weighted product of marginal Gaussian likelihoods (composite likelihood), allowing feature-specific contributions without estimating a full covariance matrix. Gaussian parameters encode expected event-specific kinematic configurations derived from normative gait biomechanics. A temporal phase prior further constrains HS to the early portion and TO to the late portion of each half-cycle. A Viterbi-style dynamic programming algorithm jointly optimises event assignments across all intervals, enforcing strict HS–TO alternation for each limb. Detected events are then locally refined using nearby AP-velocity zero-crossings and event-specific vertical extrema.

$$\log P(E | t) = \sum_k w_k \cdot \log N(x_k(t); \mu_k, \sigma_k) + \log P_{\text{phase}}(t)$$

Parameters. BIKE uses 99 parameters in total: 57 active Stage-2 feature parameters (9 feature instances for HS and 10 for TO, drawn from eight core feature families, each defined by a mean, spread, and weight), 6 phase-prior parameters, 19 signal-processing constants for Stage 1, and 17 post-processing refinement constants. All parameters have direct physical interpretation. Six classical features use literature-informed priors; three coordination features (R_{toe} , R_{knee} , and the differential angular velocity) and the TO-specific ankle-toe AP velocity use empirically calibrated priors derived from development data only. The Gaussian priors were not re-optimised on any held-out or markerless data. A general, frame-rate-dependent timing correction (a curvature-modulated compensation of the Savitzky-Golay group delay) was introduced during development to counter a low-frame-rate bias; it is scale-general and leaves the marker-based results at 100 Hz and above unchanged. The complete parameter listing is provided in the Additional file 1.

Comparator methods

Nine published detectors were evaluated alongside BIKE, spanning rule-based, composite-rule, and learned approaches (Table 2).

Primary comparators. Zeni is a position-based detector that identifies HS from the maximum anterior ankle displacement relative to the pelvis and TO from the posterior extremum. It requires no training and relies on a single positional cue per event. IntellEvent is a bidirectional long short-term memory (BiLSTM) network-based detector trained on a large proprietary clinical dataset using an extended Cleveland Clinic marker protocol. These two methods were selected as primary comparators because they represent, respectively, a widely used rule-based baseline and a clinically validated deep-learning approach for gait event detection.

Secondary comparators. Five classical rule-based methods were included: O'Connor, a velocity-based method using heel and toe vertical marker velocity; Hreljac and Mickelborough, both acceleration-based; Ghousayni, a velocity zero-crossing method; and Vancanneyt, a multi-axis velocity-threshold algorithm. Two additional data-driven methods were evaluated: DGEI (Dynamic Gait Event Identification), a temporal convolutional network (TCN), originally designed for inertial-measurement-unit-derived signals and included here as a cross-modality stress comparator, and DeepEvent, a BiLSTM network trained on paediatric gait data. DeepEvent could be evaluated on 13 of 16 datasets owing to marker-protocol incompatibility.

All detectors were evaluated from a common harmonised kinematic extraction pipeline derived from the same 3D marker trajectories, with dataset-specific marker names mapped to a canonical landmark set. Published decision rules were implemented as faithfully as possible within this harmonised framework. A positive-control analysis confirming comparator implementation fidelity is provided in Additional file 1. Because the deep-learning methods were trained outside the benchmark, the comparison probes cross-domain generalisation under fixed detector parameters: BIKE with biomechanics informed priors and limited calibration on development data, versus IntellEvent and DeepEvent with learned representations trained on external datasets.

Table 2. Detectors evaluated in the benchmark.

Detector	Type	Key cue	Training	Reference
BIKE	Bayesian	AP velocity + ratios	No	Present study
IntellEvent	BiLSTM	Raw kinematics	Yes	[27]

Detector	Type	Key cue	Training	Reference
Zeni	Rule-based	Sacrum–foot AP	No	[28]
DeepEvent	BiLSTM	Raw kinematics (paed.)	Yes	[29]
DGEI	TCN	IMU-derived signals	Yes	[30]
Vancanneyt	Velocity- threshold	Knee angular velocity	No	[31]
Ghoussayni	Rule-based	Foot vertical velocity	No	[32]
Hreljac	Rule-based	Ankle acceleration	No	[33]
Mickelb.	Rule-based	Foot vert. acceleration	No	[34]
O'Connor	Rule-based	Foot vertical velocity	No	[35]

Training = supervised model training on labelled gait events; BIKE requires no dataset-specific retraining at evaluation time.

Real-video markerless evaluation

To test whether marker-based detectors transfer to real-world video conditions, we used recordings from the BioCV dataset, which includes synchronised motion-capture and multi-camera video of 15 healthy adults walking overground. A single lateral camera (Camera #5, right lateral view) was selected for pose estimation, as sagittal-plane views best capture the AP and vertical kinematics used by most detectors. Recordings were collected under seven progressively constrained acquisition conditions, ranging from laboratory grade (video acquisition at 200 fps) to low frame rate recordings at 15 fps. This experiment provided an external test of whether detectors developed for marker-based kinematics could transfer to real-video-based pose estimation without dataset-specific modification. Full condition-wise per-detector markerless results are provided in Additional file 1, Tables S8a–S8c.

Seven conditions were defined to represent increasingly constrained recording settings, from laboratory-grade capture to low-resource clinical use (Table 3). The first three conditions altered frame rate only (200, 100, and 60 fps at native 1080p resolution). The remaining four combined frame-rate reduction with resolution downscaling, contrast attenuation, and geometric perspective distortion, representing handheld smartphone recording (30 fps, 1080p, mild perspective distortion), basic smartphone recording (30 fps, 720p, reduced contrast), teleconsultation (30 fps, 480p), and hospital-corridor recording (15 fps, 480p, stronger contrast and perspective degradation). Importantly, all degradations were applied to the video before running the pose-estimation algorithm, so that the pose estimator

received degraded visual input rather than synthetically perturbed landmark coordinates. This design captures the full chain of error propagation from image quality through pose estimation to gait-event detection.

Two-dimensional pose estimation was performed independently on each degraded video using MediaPipe Pose [36] (heavy model, 33 COCO landmark from BlazePose topology). Landmark trajectories were low-pass filtered at 4 Hz with a fourth order Butterworth filter [37] and converted to sagittal-plane kinematics and joint angles. All detectors were then run from the harmonised kinematic signals derived from this common pose-estimation pipeline, implemented in the open-source myogait toolkit [38], without re-optimising the biomechanical priors (the general low-frame-rate timing correction described above was applied). Reference gait events were obtained from the synchronised marker-based motion-capture data processed in Visual3D. MediaPipe Pose was selected as the pose estimator because it is freely available, runs on consumer hardware without GPU, and is increasingly used in clinical and field-based gait studies. However, the present markerless evaluation is limited to this single pose-estimation pipeline applied to a single lateral camera view. Different pose estimators (e.g., OpenPose, MMPose, RTMPose, Sapiens) may produce different landmark noise profiles, systematic biases, and failure modes, which could affect detector performance in ways not captured here [39]. The results should therefore be interpreted as characterising detector transfer through one specific markerless acquisition chain rather than markerless conditions in general. However, because the degradations were applied synthetically to originally high-quality video rather than recorded natively under each condition, the results may not fully reflect the imaging artefacts of actual consumer-grade devices (e.g., rolling shutter, autofocus hunting, motion blur).

Table 3. Real-video markerless acquisition conditions.

Condition	FPS	Resolution	Contrast	Persp.	Trials	Subj.
Lab reference	200	1080p	1.0	0.0	127	15
Clinical pro	100	1080p	1.0	0.0	101	13
Smartphone tripod	60	1080p	1.0	0.0	101	13
Smartphone handheld	30	1080p	1.0	0.10	101	13
Smartphone basic	30	720p	0.8	0.10	101	13
Teleconsultation	30	480p	0.7	0.15	101	13
Hospital corridor	15	480p	0.6	0.20	100	13

FPS : Frame Per Second, Persp : Geometric Perspective distortion, Subj : number of participants.

Second pose-estimation chain (Sapiens 2). To test whether these findings depend on the specific pose estimator, we repeated the evaluation through a second, independent markerless chain. Videos from two sagittal cameras (Cameras 01 and 05) were processed with the Sapiens 2 pose estimator [40] and converted to joint kinematics through the myogait video pipeline (v0.6.0, open source; 4-Hz zero-phase Butterworth low-pass, sagittal perspective correction, linear detrending) [38]. This pipeline was developed on this cohort prior to and independently of the present analyses; no BIKE parameter was tuned on these data, and every detector was run through the identical pipeline. Pose extractions were available for ten of the 15 BioCV participants; one (P09), whose Visual3D reference contained corrupted right heel-strike annotations on seven of ten trials (verified against raw marker trajectories), was excluded, and one yielded no usable extractions, leaving eight participants (79 walking trials, 158 trial–view pairs). Native 60 fps kinematics were evaluated together with 30- and 15-fps decimated versions and scored against the marker-based reference events (micro-F1 at 50 ms, subject-level cluster-bootstrap 95% CIs). Because monocular pose estimation can mirror anatomical left and right, primary scores pool both limbs.

This second chain additionally supported a feature-level analysis of event-timing information: for each of 219 kinematic features (seven families), we measured the offset between the feature's event-locked local extremum and the reference event (search window ± 150 ms, sub-frame parabolic interpolation), summarised as capture rate, median bias, and jitter (robust SD) with subject-level cluster-bootstrap 95% confidence intervals, and repeated the analysis across Butterworth cut-offs (2–25 Hz) to quantify the smoothing–timing trade-off. Full protocol and results are provided in the dedicated additional file.

Feature transportability analysis

To evaluate feature transportability independently from full-detector performance, we performed a cross-transfer analysis on the eight shared core features. For each feature and event, Gaussian discriminators were calibrated on healthy-only data and evaluated on clinical data, and vice versa. This yielded AUC values for four transfer conditions: $H \rightarrow H$, $H \rightarrow C$, $C \rightarrow C$, and $C \rightarrow H$. We defined the transfer gap as the difference between within-group and cross-group AUC, such that values closer to zero indicate better transportability across populations (Additional file 1, Figures S5 and S6). To assess whether feature-level transportability translated into detector-level robustness, BIKE was recalibrated under three prior configurations: pooled priors, healthy-only priors, and clinical-only priors. These configurations were evaluated on the held-out datasets without markerless-data calibration.

Results

Cohort and benchmark overview

All ten detectors returned results on all 16 datasets except DeepEvent, which was evaluated on 13 datasets because of marker-protocol incompatibility. Detector failure rates were low overall: BIKE produced no events on 2 of 12315 evaluated detector–recording instances (0.016 %), both involving short recordings with fewer than three strides (Table S4d). The complete per-dataset results for all 10 detectors are provided in the full 10-detector benchmark (Figure 2 and Tables S1a–S1b).

Full benchmark performance

Across all datasets at 50-ms tolerance, BIKE achieved the highest dataset-level macro-averaged F1 for both HS (0.870) and TO (0.856), yielding the highest combined score (0.863; Table 4, Figure 2). This advantage reflected cross-dataset consistency rather than frequent per-dataset dominance: BIKE ranked first on 4 of 16 datasets, second on 9, and third on 3, and was never ranked below third. By comparison, IntelliEvent achieved the highest per-dataset combined score on 7 of 16 datasets but showed greater between-dataset variability, particularly for TO (combined macro-F1 0.823). Thus, the two leading detectors exhibited complementary strength profiles: IntelliEvent more frequently achieved top performance on individual datasets, whereas BIKE’s narrower between-dataset spread produced the highest macro-average. Zeni ranked third (combined 0.720). The remaining comparators ranged from 0.649 (DeepEvent, evaluated on 13 datasets) to 0.192 (Mickelborough). The near-zero HS score of DGEI illustrates a cross-modality caveat: it does not reflect a failure to detect gait cadence (its HS detection count matched the reference), but an event-timing rule inherited from its native inertial-sensor formulation, namely the peak of foot-descent velocity, which does not coincide with the contact instant on kinematic vertical-position signals; its toe-off, defined from foot rise, transferred normally. DGEI HS should therefore be interpreted as a cross-modality event-definition mismatch rather than an inability to detect gait events.

Subject-level bootstrap 95% confidence intervals confirmed BIKE’s advantage and, at the subject level, placed Zeni above IntelliEvent for toe-off: BIKE HS [0.742, 0.791], IntelliEvent HS [0.659, 0.730], Zeni HS [0.605, 0.669]; for TO, BIKE [0.680, 0.759], Zeni [0.592, 0.695], IntelliEvent [0.379, 0.513] (Tables S5e–S5f). Overall differences among the nine full-coverage detectors were significant at the dataset level (Friedman on per-dataset micro-F1, $N = 16$: $\chi^2 = 100.7$

for HS and 90.4 for TO, both $p < 0.001$); dataset-level pairwise Wilcoxon signed-rank tests confirmed BIKE's advantage over Zeni for HS (adjusted $p = 0.003$) and over the classical rule-based detectors, while BIKE and IntellEvent did not differ significantly for either event. On the force-plate subset, the most reliable timing reference, IntellEvent achieved marginally higher HS F1 and BIKE higher TO F1 (Additional file 1, Section S4), so the two leading detectors are best described as complementary rather than one being uniformly superior. Trial-level secondary comparisons are provided in Additional file 1, Tables S5b and S5d.

BIKE's weakest performance occurred on Kuopio (HS 0.570, TO 0.631), a large and challenging dataset characterised by partial force-plate coverage and many short walkway passages, conditions associated with elevated false-positive rates across detectors. On a secondary micro-F1 basis, Kuopio remained the most challenging dataset for all detectors.

Table 4. Full 16-dataset benchmark: macro-averaged F1 at 50-ms tolerance.

Combined = mean of HS and TO macro-F1; MAE computed on matched true-positive events only. Subject-level bootstrap 95% confidence intervals are reported in Additional file 1, Tables S5e–S5f.

Detector	Macro-averaged F1			MAE (ms)	
	HS F1	TO F1	Combined	HS MAE	TO MAE
BIKE	0.870	0.856	0.863	24.2	15.5
IntellEvent	0.843	0.804	0.823	18.3	19.7
Zeni	0.740	0.700	0.720	30.1	21.1
DeepEvent	0.614	0.685	0.649	26.4	16.4
Vancanneyt	0.638	0.350	0.494	15.3	22.4
DGEI	0.045	0.747	0.396	26.2	25.0
Ghoussayni	0.504	0.241	0.372	22.6	25.8
O'Connor	0.180	0.466	0.323	24.1	26.7
Hreljac	0.384	0.177	0.281	23.3	30.5
Mickelb.	0.323	0.061	0.192	24.2	31.5

HS : Heel Strike, TO : Toe Off, MAE : Mean Absolute Error

Held-out validation

On the six held-out datasets (3627 evaluations), BIKE and IntellEvent achieved closely comparable macro-averaged performance (BIKE: HS 0.935, TO 0.915, combined 0.925; IntellEvent: HS 0.930, TO 0.918, combined 0.924), both

exceeding all other detectors (the six-dataset held-out comparison is descriptive and statistically underpowered for pairwise inference) (Table 5, Additional file 1, Tables S1a–S1b). This near-equivalence indicates that both an interpretable probabilistic framework and a deep-learning model can generalise well to unseen clinical datasets under marker-based acquisition, and that the marker-based held-out tier alone does not discriminate between these two modelling strategies. Zeni was clearly lower on both events (combined 0.820). In a sensitivity analysis excluding Nature Multimodal (the only held-out dataset sharing participants with the development cohort, 49 of 50 subjects), the held-out estimates for BIKE and IntelliEvent changed by ≤ 0.01 and remained near-identical, both clearly exceeding Zeni, confirming that the held-out conclusion does not depend on this partially overlapping dataset.

These findings show that held-out performance remained high, but that the relative advantage of BIKE was event-specific rather than uniform. At the dataset level, HS leadership was distributed across methods: BIKE ranked first on two held-out datasets, IntelliEvent on two, and Zeni and DeepEvent on one each. TO performance was more variable, a nuance that tempers interpretation of aggregate held-out metrics, with BIKE ranking first on two of six datasets and showing greater between-dataset fluctuation than for HS.

Timing accuracy on the held-out datasets is detailed below (“Timing accuracy and degradation boundary”). Briefly, BIKE achieved the lowest HS MAE (15.1 ms) while DeepEvent achieved the lowest TO MAE (14.1 ms).

Table 5. Held-out validation (6 datasets): macro-averaged F1 at 50-ms tolerance.

Detector	Macro-averaged F1			MAE (ms)	
	HS F1	TO F1	Combined	HS MAE	TO MAE
BIKE	0.935	0.915	0.925	15.1	17.2
IntelliEvent	0.930	0.918	0.924	18.9	14.3
Zeni	0.850	0.790	0.820	21.8	26.4
DeepEvent	0.777	0.903	0.840	25.2	14.1
Vancanneyt	0.829	0.483	0.656	15.7	26.9
DGEI	0.007	0.916	0.462	29.2	15.3
Ghoussayni	0.580	0.236	0.408	25.3	30.9
O’Connor	0.141	0.549	0.345	24.0	26.6
Hreljac	0.405	0.168	0.286	26.8	33.1
Mickelb.	0.358	0.053	0.205	27.5	30.7

Combined = mean of HS and TO macro-F1; MAE computed on matched true-positive events only. Bootstrap CIs for held-out datasets are reported in Additional file 1, Section S5.

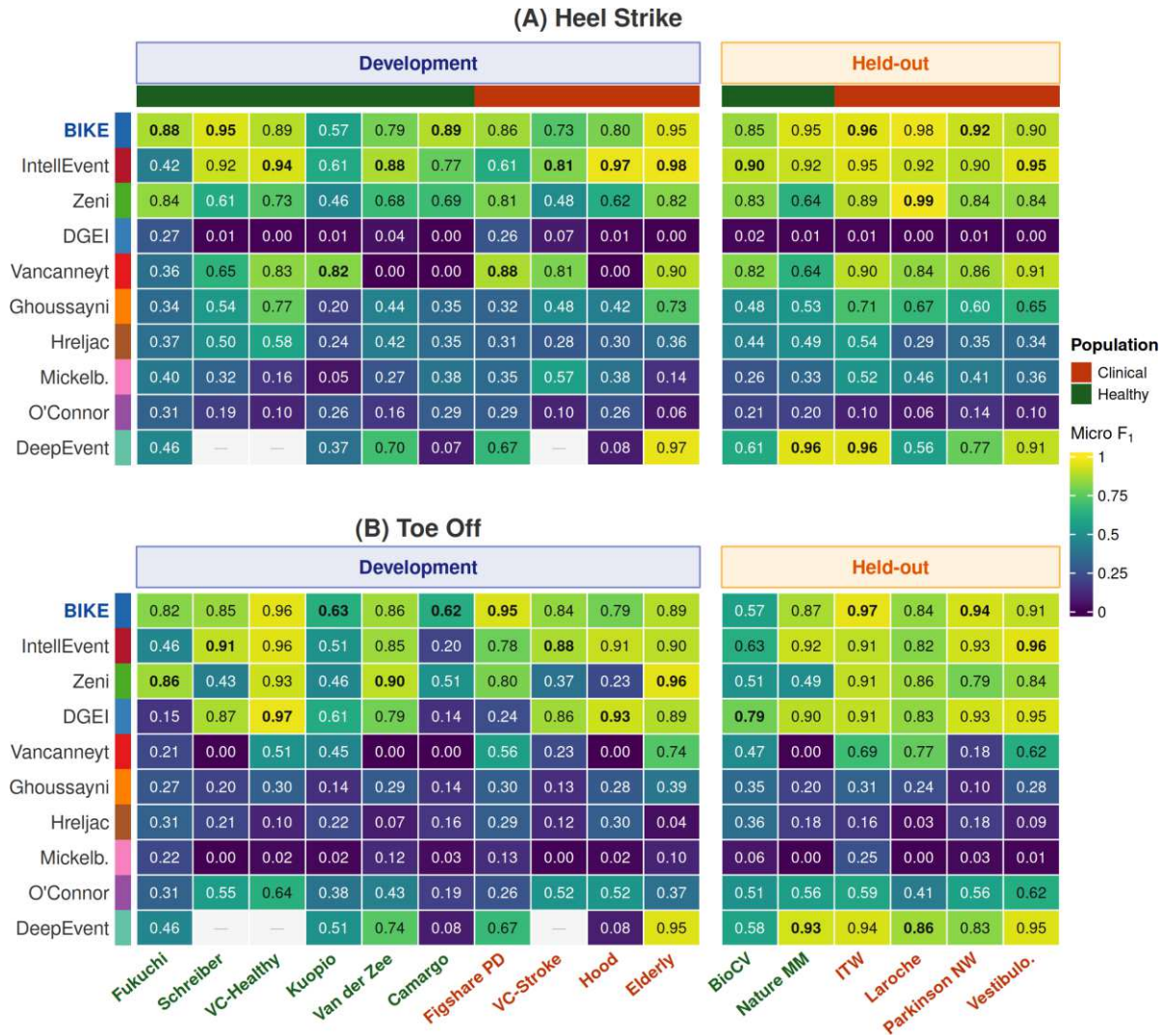


Figure 2. Complete benchmark heatmap showing micro-F1 scores for all 10 detectors across 16 datasets at 50 ms tolerance. Development datasets (top) and held-out datasets (bottom) are separated. BIKE achieved the highest combined macro-averaged F1 (0.863). The full 10-detector benchmark is presented here; detailed per-dataset values are reported in Additional file 1, Tables S1a–S1b.

Real-video markerless evaluation

After quality control, the markerless dataset comprised 127 trials in the reference laboratory condition and 100–101 trials in each degraded condition. Participants P27 and P28 were available only in the laboratory-reference condition. A sensitivity analysis on the marker-based BioCV benchmark, excluding seven trials from one participant (P09) with corrupted right heel-strike reference annotations, raised every detector’s BioCV heel-strike F1 uniformly by about

0.015 without altering any ranking (Additional file 1). At 15 fps, all detectors showed marked performance loss compared with conditions at 30 fps or above.

For HS detection, BIKE achieved the highest pooled F1 at 50-ms tolerance in six of seven conditions (Figure 3). Performance remained high from the laboratory condition (0.883 at 200 fps) through the clinically realistic 30-fps settings, including handheld smartphone recording (0.901), basic smartphone recording (0.876), and teleconsultation-like video (0.858). Zeni ranked second overall for HS and remained competitive in several conditions but showed a marked drop in the most degraded hallway setting (0.178 at 15 fps), whereas BIKE retained moderate performance (0.655). In contrast, the deep-learning detectors transferred poorly to markerless video: IntelliEvent fell from 0.493 at 200 fps to ≤ 0.090 in all 30-fps conditions, and DeepEvent remained near floor level throughout.

The pattern differed for TO. Zeni achieved the highest pooled F1 in five of seven conditions, representing the strongest and most stable overall TO performance across markerless conditions, whereas BIKE remained competitive from 200 to 30 fps (0.798, 0.792, 0.799, 0.787, 0.793, and 0.763 across the first six conditions) but deteriorated sharply at 15 fps (0.086). IntelliEvent and DGEI showed good TO performance in the laboratory condition and at 100 fps but degraded markedly in the 30-fps video settings.

Timing accuracy showed a complementary pattern. For HS, BIKE achieved the lowest MAE in four of the seven conditions and remained below 16 ms in all 30-fps settings. For TO, BIKE showed moderate error from 200 to 30 fps (14.2 to 23.8 ms depending on condition) but no longer outperformed Zeni or IntelliEvent consistently. All detectors deteriorated at 15 fps, indicating a lower boundary condition for reliable event timing under sparse temporal sampling.

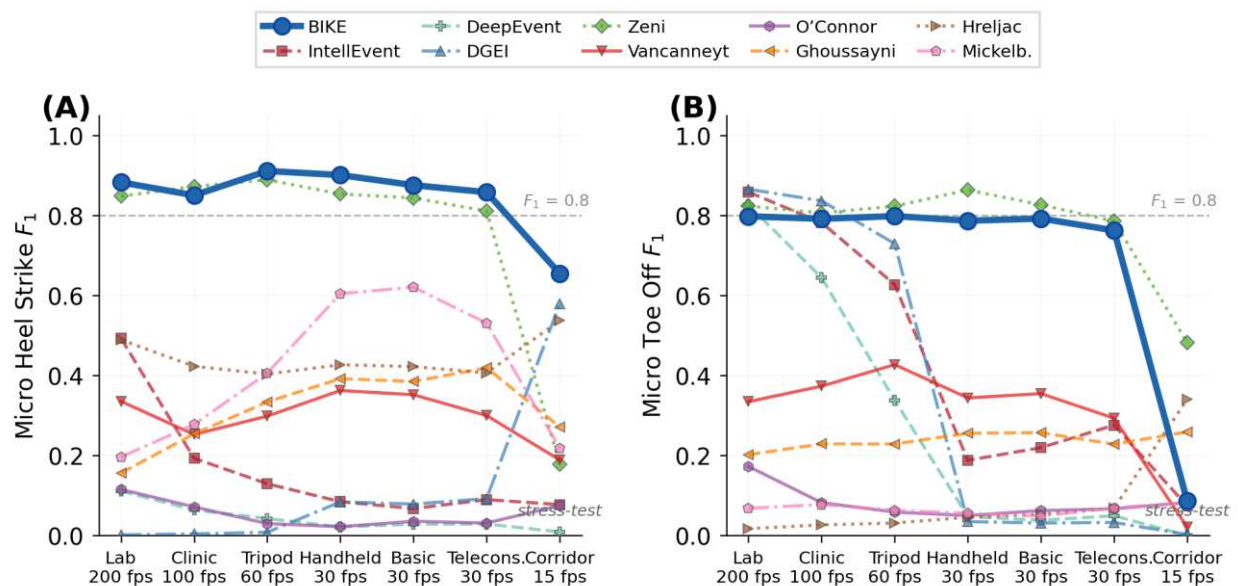


Figure 3. Real-video markerless evaluation. Pooled micro-F1 at 50-ms tolerance across seven acquisition conditions of increasing degradation. (A) Heel strike: BIKE achieved the highest F1 in six of seven conditions (0.655–0.912). (B) Toe off: Zeni achieved the highest F1 in five of seven conditions. At 15 fps (corridor condition), all detectors showed substantial performance loss, supporting 15 fps as a practical lower boundary in the present markerless dataset. Per-detector values are provided in Additional file 1, Tables S8a–S8c.

Replication through a second pose-estimation chain (Sapiens 2). All main findings replicated through an independent Sapiens 2 chain on the eight participants with concurrent marker-based ground truth (two lateral cameras, 79 trials, 158 trial–view pairs; Methods). At native 60 fps, BIKE was the only detector performing strongly on both events: HS micro-F1 0.828 (subject-level bootstrap 95% CI 0.749–0.890), statistically tied with Zeni (0.832, CI 0.758–0.904), and TO 0.867 (CI 0.788–0.940), the highest overall; IntellEvent reached 0.661/0.694 and DeepEvent could not run through the video pipeline. A within-cohort comparison against the marker-based benchmark on the same participants quantified the cost of the video modality itself: the marker-to-video transition reduced HS F1 by 0.119 for BIKE (0.947→0.828) and 0.113 for Zeni, but by 0.314 for IntellEvent (0.975→0.661); for TO the pattern was similar (BIKE –0.063, Zeni –0.093, IntellEvent –0.257). The deep-learning detector lost 2.6–4.1 times more performance in the marker-to-video transition than the biomechanics-based detectors (values recomputed from the frozen benchmark source on the same 79 trials). Frame-rate decimation reproduced the boundary observed with MediaPipe: at 15 fps all derivative-based detectors collapsed (BIKE HS 0.298) whereas position-based detection was largely preserved (Zeni 0.819), and relaxing the tolerance to 100 ms did not rescue the derivative-based detectors.

One result proved chain-dependent: in the Sapiens 2 chain BIKE provided the strongest TO detection (0.867 vs. 0.682 for Zeni), whereas Zeni led TO in the MediaPipe chain. This sensitivity to TO but not to HS to the pose-estimation pipeline reinforces the event-specific picture: heel-strike detection was robust across populations, acquisition conditions, and pose estimators, whereas toe-off remained the fragile event under every domain shift tested.

Timing accuracy and degradation boundary

Timing accuracy showed a complementary pattern between the two leading detectors (Figure 4). On the six held-out datasets, BIKE achieved the lowest HS mean absolute error (15.1 ms, versus 18.9 ms for IntellEvent and 21.8 ms for Zeni), whereas DeepEvent achieved the lowest TO MAE (14.1 ms, versus 14.3 ms for IntellEvent and 17.2 ms for BIKE). Across all datasets, the MAE ranking was less event-consistent: IntellEvent achieved lower HS MAE than BIKE (18.3 vs. 24.2 ms), whereas BIKE achieved lower TO MAE than IntellEvent (15.5 vs. 19.7 ms). Illustrative detections on two challenging gait patterns are shown in Figure 5.

Tolerance sensitivity analysis further clarified the relationship between timing precision and detection performance. On the full benchmark, BIKE macro-averaged F1 rose from 0.542 to 0.870 to 0.918 for HS and from 0.636 to 0.856 to 0.895 for TO as tolerance increased from 25 to 50 to 75 ms; parallel development-set micro-F1 curves are shown in Additional file 1, Figure S1.

In the real-video markerless experiment, BIKE HS's MAE remained stable across all 30 fps conditions: 14.5 ms for handheld smartphone recording, 14.8 ms for basic smartphone recording, and 15.2 ms for teleconsultation, all close to the held-out motion-capture reference (15.1 ms). TO's MAE was more condition-dependent, ranging from 14.2 ms in teleconsultation to 23.8 ms in the laboratory reference condition. Across the six conditions at 30 fps or above, BIKE maintained HS micro-F1 at least 0.851 and TO micro-F1 at least 0.763.

The 15 fps hospital-corridor condition represented the most degraded markerless setting. In this condition, BIKE HS micro-F1 dropped to 0.655 and TO micro-F1 to 0.086, whereas Zeni reached HS F1 = 0.178 and TO F1 = 0.482. BIKE HS MAE increased to 21.4 ms, while TO detection deteriorated sharply.

Feature-level timing information (Sapiens 2 chain). The second markerless chain supported timing analysis at the level of individual features rather than detectors. Band-limited knee angular velocity carried the best usable timing for TO (jitter 23–24 ms, bootstrap CI upper bounds 32–33 ms; bias –29 ms; capture rate 0.86–0.89), whereas knee angular acceleration showed only marginal timing at HS (bias –44 ms, jitter CI up to 54 ms). A Butterworth cut-off ablation (2–25 Hz) identified this as an irreducible trade-off: raising the cut-off reduced the bias (48 to 24 ms) while inflating the jitter (19 to 66 ms), with no setting bringing both below ~35 ms. The optimal cut-off decreased with derivative order (accelerations ≤ 2 Hz, velocities ≈ 4 Hz, joint angles exhibiting no sharp temporal landmark at any cut-off), yielding a design principle: filter to the bandwidth of the event, not of the gait cycle, and prefer band-limited velocity features when event timing is required.

Feature transportability and mechanistic interpretation

Feature-level transferability differed across events and kinematic cues. For HS, knee–heel anteroposterior distance, heel vertical position, and the toe-velocity ratio R_{toe} showed high cross-transfer performance with small transfer gaps. For TO, knee angular velocity and toe vertical velocity were among the most transferable features. In contrast, ankle anteroposterior velocity and $\Delta\omega$ showed larger transfer gaps for HS, whereas TO-related features showed more heterogeneous transfer behaviour overall.

Cross-dataset stability analyses showed that knee–heel anteroposterior distance remained strongly discriminative across datasets with low between-dataset variability, whereas toe vertical position showed greater dispersion across populations.

Prior-configuration analyses produced similar held-out HS performance across pooled, healthy-only, and clinical-only priors. Across the five marker-based held-out datasets (excluding the markerless BioCV set), macro HS F1 was 0.940 for BIKE-P, 0.927 for BIKE-H, and 0.938 for BIKE-C. For TO, BIKE-H showed a selective deficit on Nature Multimodal, whereas BIKE-C closely matched BIKE-P across datasets. Additional feature-level statistics, transfer matrices, and sensitivity analyses are provided in Additional file 1, Section S7 (Table S7 and Figures S5-S7).

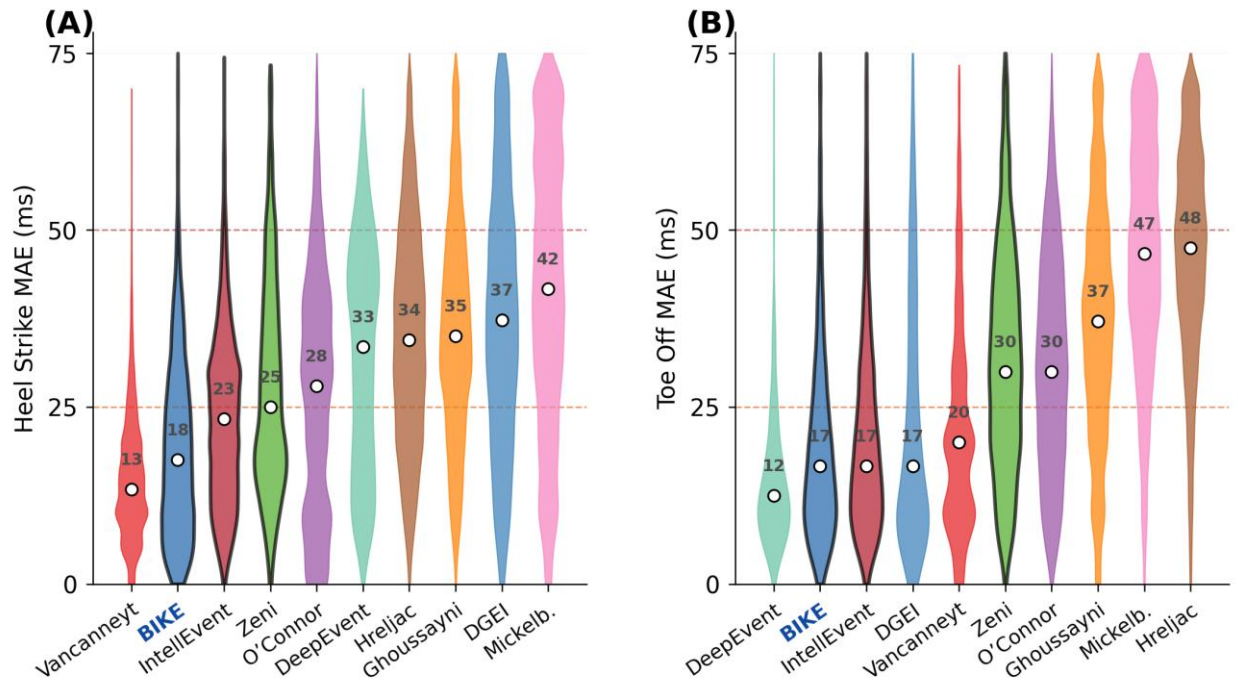


Figure 4. Timing accuracy. Distribution of mean absolute error (MAE) across correctly matched events (true positives at 75-ms tolerance) on the full 16-dataset benchmark. (A) Heel strike, (B) Toe off. Detectors are ordered by median MAE; white circles indicate medians. Dashed lines mark the 25-ms and 50-ms tolerance thresholds.

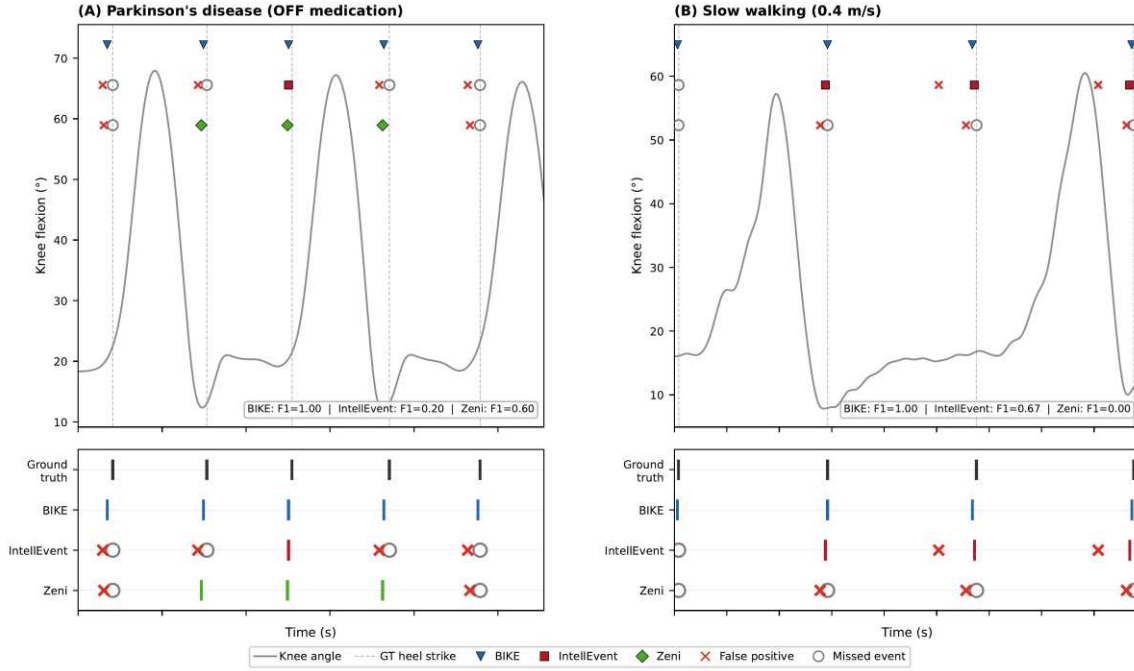


Figure 5. Illustrative detection on two challenging gait patterns. Knee flexion angle (top) and event raster (bottom) for two representative trials: (A) Parkinson’s disease patient (OFF medication), (B) slow walking (0.4 m/s). BIKE correctly detects all heel strikes in both conditions ($F1 = 1.00$), while IntellEvent and Zeni produce false positives ($\times$) and missed events ($\circ$). These selected cases illustrate contrasting detector behaviour under reduced stride amplitude and atypical rhythm, consistent with BIKE’s multi-feature Bayesian framework.

Feature transportability to video (Sapiens 2 chain). Amplitude discriminability recomputed for 219 kinematic features on the real markerless signals ranked angular accelerations (median $|d| = 0.47$) above joint angles (0.34) and velocities (0.28; this family median is pulled down by many weak features and the specific knee angular velocities used by detectors reach $|d| \geq 1.2$), with relative positions (0.15), inter-segment distances (0.12), and left–right asymmetry features (0.06) strongly degraded.

Budget ratios under real markerless conditions (Sapiens 2 chain). The second markerless chain (below and Methods) also tested the ratio features outside marker-based acquisition. $\Delta\omega$ retained both large amplitude discriminability and usable timing at TO ($|d| = 1.20$; landmark capture rate 0.94, jitter 23 ms; bias -40 ms, correctable).

Notably, R_{toe} — among the most population-transferable HS features above — collapsed under video-based estimation ($|d| = 0.05$, versus ≥ 1.30 on pathological marker-based data), because it depends on toe vertical velocity, which monocular pose estimation resolves poorly. Population transportability and acquisition robustness are therefore distinct axes that must be tested separately.

Sensitivity to the calibrated ratio features (preliminary). To quantify whether the empirically calibrated ratio features drive the benchmark results, we re-ran BIKE with all ratio weights set to zero (classical, literature-informed features only) in a paired design on identical trials. In an interim analysis covering the six datasets completed at the time of writing — including two clinical populations (Parkinson's disease, post-stroke; 4,191 trials in total) — the classical-only variant left pooled HS performance strictly unchanged (micro-F1 $0.838 \rightarrow 0.838$) and slightly improved TO ($0.854 \rightarrow 0.870$); per-dataset HS deltas ranged from -0.011 to $+0.002$. These preliminary results indicate that the calibrated ratio features do not drive BIKE's benchmark performance; the complete paired run will be reported in the additional file.

Discussion

This study evaluated the generalisation of kinematic gait-event detection across heterogeneous marker-based datasets and real-video markerless acquisition conditions. Three main findings emerge. First, across the 16-dataset motion-capture benchmark, BIKE achieved the highest macro-averaged performance, not by dominating every dataset, but by maintaining consistent performance across datasets. IntelEvent [27] more frequently achieved the best score on individual datasets but showed greater between-dataset variability. Second, on the six held-out datasets, BIKE and IntelEvent performed almost identically, indicating that both biomechanics-informed probabilistic modelling and deep-learning-based sequence modelling can generalise well to unseen populations when the input remains marker-based. Third, the markerless evaluation revealed a stronger domain shift and showed that generalisation was event-specific: HS detection transferred more robustly than TO detection, BIKE was strongest for HS across most acquisition conditions, Zeni [28] was strongest for TO, and deep-learning comparators showed substantial degradation under video-derived kinematics. The two leading detectors diverged by acquisition modality rather than by population: BIKE and IntelEvent were comparable on marker-based data, including pathological gait, but only BIKE retained reliable heel-strike detection under real-video markerless conditions, where the deep-learning detectors collapsed. This degradation propagated to the reconstructed joint kinematics: robust detectors preserved the joint-angle waveforms

and recovered most gait cycles, whereas detectors that lost accuracy under video-derived input, including the deep-learning models and several classical heuristics, distorted or lost them (Additional file 1, Figure S9 and Table S11), consistent with the dependence of phase-normalised kinematics and derived spatiotemporal measures on accurate event identification [1,2].

A central interpretation of these results is that detector generalisation is constrained by the transportability of the underlying kinematic features. In the present context, feature transportability refers to the ability of a feature to preserve event-linked discriminative structure across populations, protocols, and acquisition modalities, the last of which introduces systematic differences between marker-based and video-derived joint kinematics [4,6,9]. The feature-level analyses showed that high within-group discriminability alone was not sufficient. The most useful features were those that remained informative when transferred between healthy and clinical populations. For HS, knee–heel anteroposterior distance, heel vertical position, and the toe-velocity ratio showed strong cross-transfer behaviour; these are the quantities underlying widely used position- and velocity-threshold detectors [28,32,35]. For TO, transferable features were more heterogeneous, with knee angular velocity and toe vertical velocity among the most stable cues [31,35]. This suggests that the stability of the biomechanical information carried by a feature, grounded in the stereotyped kinematics of initial contact and pre-swing [2,37], may be as important as the classification framework used to exploit it.

This feature-level interpretation was consistent with model-level behaviour. When BIKE was recalibrated using pooled, healthy-only, and clinical-only priors, HS performance remained similar across configurations, suggesting that HS detection was relatively insensitive to the population used for calibration. By contrast, TO showed more population-dependent behaviour, including a selective deficit for the healthy-only prior configuration on Nature Multimodal. This supports the interpretation that HS-related cues are more stable across populations, whereas TO-related cues are more sensitive to gait strategy, pathology, and acquisition context, consistent with the marked alterations of push-off, pre-swing timing and distal-segment control reported in Parkinson's disease [17,18], post-stroke gait [15], lower-limb amputation [19], hip osteoarthritis [22,23] and idiopathic toe walking, where the heel rocker is absent altogether [21]. These findings also explain why aggregate performance metrics can obscure important event-specific behaviour: two detectors may show similar combined F1 scores while relying on features that differ substantially in their robustness across events and domains.

The event-specific pattern observed in the markerless experiment is biomechanically plausible. HS is associated with relatively abrupt and temporally localised changes in sagittal-plane kinematics, including limb deceleration and ground contact rapidly arrests distal-segment motion during the loading response [2,37]. Because these changes are carried by geometric and velocity relationships in the sagittal plane, the plane best preserved by monocular pose estimation [4,9], they can remain detectable even when joint localisation is noisy, provided that the relevant geometric or velocity-based relationships are preserved. By contrast, TO coincides with the end of a smoother, lower-acceleration foot-rise governed by pre-swing limb unloading, ankle plantar-flexor output and ankle-foot coordination, and is further modulated by walking speed, and pathology-specific adaptations [2,12,16,37]. These mechanisms produce a shallower kinematic landmark and are more sensitive to frame rate, landmark noise, and degraded foot tracking [4]. This likely explains why BIKE maintained robust HS performance down to clinically realistic 30-fps video, whereas TO transfer was more variable and position-based approaches such as Zeni [28] remained advantageous in several markerless conditions; acceleration-based landmarks, most exposed to the broadband noise that differentiation amplifies and that already characterises the impact transient at contact, were the least transportable [33,34,41,42].

The 15 fps hospital-corridor condition further illustrates the interaction between detector design and acquisition constraints. In this condition, all detectors deteriorated substantially, indicating that very low frame rate markerless acquisition may impose a practical lower boundary for reliable gait event timing [6]. This degradation is not only algorithmic. At 15 fps, the inter-frame interval is approximately 66.7 ms, which exceeds the 50 ms scoring tolerance used in the benchmark, and differentiated features suffer additional broadband noise amplification as temporal sampling becomes sparse relative to the frequency content of the events [37,41]. Therefore, even an otherwise accurate detector may fail to localise events within the tolerance window when temporal sampling is too sparse. This issue is particularly relevant for TO, where the temporal signature appears less robust and more dependent on the quality of foot and distal-segment tracking [4]. These findings support the need to report acquisition constraints explicitly when evaluating markerless gait-event detection [5,39].

The comparison between BIKE and IntelliEvent refines the relationship between model complexity and generalisation. Under marker-based conditions, both approaches generalised well to held-out datasets, suggesting that complex sequence learning [27,29] and interpretable probabilistic inference [7,8] can both be effective when the input domain remains close to the expected kinematic structure. However, the markerless evaluation showed that strong marker-based held-out performance does not necessarily imply robustness to changes in signal source, frame rate, camera

quality, or pose-estimation noise [4,9]. In this setting, the decisive issue appeared to be whether the detector relied on cues that survived the acquisition-chain shift. This does not argue against deep learning; rather, it suggests that learned models may require explicit exposure to markerless variability, pose-estimation artefacts, and low-frame-rate conditions during training to achieve similar cross-modal robustness, as achieved by models trained directly on single-camera video [43]. Conversely, interpretable feature-based models offer the advantage that their failure modes can be traced back to specific biomechanical cues [28,31,35], which may facilitate targeted event-specific refinement.

These findings have practical implications for markerless gait analysis. First, markerless validation should not be inferred from marker-based performance alone. Even detectors that generalise well across marker-based datasets can degrade when applied to video-derived kinematics, which differ systematically from marker-based estimates [6,9,39]. Second, HS and TO should be evaluated and optimised separately, because their biomechanical signatures and sensitivities to acquisition degradation differ. Third, detector benchmarking should include not only global F1 scores but also event-specific performance, timing error, failure rates, and sensitivity to frame rate and pose-estimation quality, together with the downstream effect on the phase-normalised kinematics and summary indices used clinically [1,3]. Finally, the results suggest that future clinical tools should define minimum acquisition requirements, particularly for TO detection, before deploying markerless gait-event detection in low-resource or teleconsultation settings [5,39]. Although this study focused on marker-derived and video-derived joint trajectories, the concept of feature transportability may extend to other sensing modalities. IMU-based gait-event detection methods, including hidden Markov models and Bayesian inference frameworks, have shown effective gait phase classification from accelerometer and gyroscope signals [7,8], and temporal convolutional networks have been applied to the same problem [30]. These approaches rely on different signal features, such as angular velocity, acceleration peaks, and temporal phase structure. The present results raise the question of whether IMU-derived cues exhibit similar event-specific transportability, particularly under pathological gait patterns, altered walking speeds, or sensor-placement variability. Datasets that now provide concurrent marker-based, IMU and video recordings [13] make such cross-modality comparisons feasible; future work comparing marker-based, markerless, and IMU-derived features within a common transportability framework could help identify which biomechanical cues are most robust across sensing modalities.

Several limitations should be considered. First, the real-video markerless experiment was performed on a relatively small sample of healthy participants from a single site [11], using two pose-estimation pipelines (MediaPipe Pose on one lateral camera [36]; Sapiens 2 on two lateral cameras [40] with the myogait preprocessing pipeline [38]). The

findings therefore characterise detector transfer through these specific acquisition chains rather than markerless gait-event detection in general; notably, video performance is contingent on preprocessing (running BIKE on the Sapiens 2 kinematics without the myogait pipeline collapses HS F1 from 0.83 to 0.13, whereas Zeni is barely affected), so markerless results characterise detector-plus-pipeline systems. Different pose estimators, camera viewpoints, clothing conditions, lighting environments, and clinical gait patterns may produce different landmark errors and failure modes [4,39]. Second, TO detection remained imperfect in markerless conditions, and no tested detector maintained uniformly high performance for both HS and TO across all acquisition settings. This limits immediate clinical deployment when both events are required to compute stance, swing, or double-support parameters. Third, BIKE currently assumes approximately straight-line walking during the crossing-segmentation stage, an assumption shared with other anteroposterior-based detectors [28], which limits its applicability to curved walking, turning, and free-form locomotion [10]. Fourth, the comparison between BIKE and deep-learning models was asymmetric: BIKE used literature-informed and development-calibrated priors, whereas IntelEvent and DeepEvent were applied as externally trained models without adaptation to the benchmark datasets [27,29]. Fifth, ground-truth heterogeneity across datasets increased ecological validity but introduced variation in temporal precision between reference standards, including force-plate, manual, Nexus-derived, and treadmill-software annotations.

Future work should therefore pursue four priorities. First, real-video validation should be extended to pathological cohorts, especially populations with altered propulsion, foot clearance, reduced walking speed, or atypical limb coordination, such as post-stroke [15], parkinsonian [17,18], amputee [19], fall-prone older adults [20] and toe-walking gait [21]. Second, TO detection should be treated as a dedicated methodological problem, with features designed specifically to remain stable under degraded foot tracking and low-frame-rate acquisition. Third, detector robustness should be evaluated across multiple pose-estimation pipelines and camera viewpoints to separate detector limitations from pose-estimator-specific artefacts [4,39]. Fourth, feature transportability should be tested beyond level straight-line walking, including curved walking, turning, running, and non-level locomotion such as stairs and ramps [10,24]. Together, these developments would support more clinically reliable gait-event detection and help define the acquisition conditions under which markerless gait analysis can be used for robust spatiotemporal assessment.

Conclusions

Across sixteen marker-based datasets and two independent markerless-video chains, gait-event detector generalisation depended less on model architecture than on the kinematic quantities a detector uses, and it was consistently event-specific. On marker-based data, the interpretable Bayesian detector BIKE and the deep-learning detector IntellEvent generalised equally well to unseen clinical populations; the two approaches separated only under real video. Heel-strike detection, carried by band-limited velocity and sagittal-plane geometric features, transferred reliably across populations, across frame rates down to 30 fps, and across both pose estimators, whereas TO remained sensitive to feature choice, pose-estimation chain and temporal sampling, and no detector, BIKE included, was uniformly strong on both events. Moving from markers to video cost the biomechanically grounded detectors 6–12 F1 points, against 26–31 for the deep-learning comparators trained on marker data. Feature-level analysis showed that population transportability and acquisition robustness are distinct axes, a cue that is stable across pathologies can still collapse under monocular pose estimation, and that usable event timing was confined to features filtered to the bandwidth of the event rather than of the gait cycle. BIKE reached the highest macro-averaged benchmark F1 through cross-dataset consistency rather than per-dataset dominance and retained the most reliable heel-strike detection under real video; a paired ablation indicated that this robustness rests more on its band-limited velocity features and probabilistic fusion than on its empirically calibrated coordination ratios. Markerless performance was therefore a property of the detector-and-preprocessing system, not of the detector alone. These findings argue for evaluating gait-event detectors per event and per acquisition chain, for treating heel strike and toe-off as distinct design problems, and for making real-video validation routine rather than inferring it from marker-based benchmarks. All benchmark outputs, analysis code, and detector implementations are released in the open-source gaitkit framework [44]; the source datasets remain available from their original repositories.

Abbreviations

AP: anteroposterior; AUC: area under the curve; BIKE: Bayesian Inference for Kinematic Events; BiLSTM: bidirectional long short-term memory; CRediT: Contributor Roles Taxonomy; DL: deep learning; F1: F1-score; FN: false negative; FP: false positive; fps: frames per second; GT: ground truth; HS: heel strike; IMU: inertial measurement

unit; ISB: International Society of Biomechanics; ITW: idiopathic toe walking; MAE: mean absolute error; OA: osteoarthritis; TCN: temporal convolutional network; TO: toe off; TP: true positive.

Declarations

Ethics approval and consent to participate

Not applicable. All datasets analysed are publicly available and were collected under ethics approvals granted to the original investigators.

Consent for publication

Not applicable.

Availability of data and materials

The benchmark datasets used in this study are available from their original repositories as cited in the text. The benchmark outputs, analysis code, and detector implementations are available at <https://github.com/IDMDataHub/gaitkit> (gaitkit version 1.4.9, tag v1.4.9-article, archived at <https://doi.org/10.5281/zenodo.18653109>; this exact release: <https://doi.org/10.5281/zenodo.21323295>). Detailed installation instructions and documentation are provided in the package.

Competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. We note that the benchmark, the BIKE detector, and the choice of primary metric (macro-averaged F1) and tolerance (50 ms) originate from the same group; the reported ranking is dependent on these choices (for example, at the stricter 25-ms tolerance IntellEvent leads for heel strike).

Funding

Crédit Agricole CIB-LCL, AFM-Téléthon

Authors' contributions

F. Fer: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Writing - original draft, Visualization. T. Marques: Formal analysis - review & editing. J.-Y. Hogrel: Writing - review & editing. M. Jacoupy: Formal analysis, Writing - review & editing. R. Feigean: Supervision, Writing - review & editing, Funding acquisition.

Acknowledgements

The authors thank the researchers who made their gait datasets publicly available, enabling the large-scale benchmark presented in this work. We are especially grateful to the University of Bath group (L. Needham, L. Wade, M. Evans, S. Colyer, P. McGuigan and colleagues) for access to the BioCV synchronised video, motion-capture and force-plate dataset, which made the markerless proof-of-concept evaluation possible. We also thank Ersin Metin for setting up the computational tools and infrastructure.

Use of large language models

Large language models (Claude, Anthropic; and GitHub Codex, OpenAI) were used as development tools for the benchmark software infrastructure and during manuscript preparation. The study design, data analysis, interpretation of results, and the final content of the manuscript are entirely the responsibility of the authors. No large language model is listed as an author or meets authorship criteria.

References

- [1] Baker R, McGinley JL, Schwartz MH, Beynon S, Rozumalski A, Graham HK, et al. The Gait Profile Score and Movement Analysis Profile. *Gait Posture*. 2009;30(3):265–269. doi:10.1016/j.gaitpost.2009.05.020
- [2] Perry J, Burnfield JM. *Gait Analysis: Normal and Pathological Function*. 2nd ed. SLACK Incorporated; 2010.

- [3] Schwartz MH, Rozumalski A. The Gait Deviation Index: a new comprehensive index of gait pathology. *Gait Posture*. 2008;28(3):351–357. doi:10.1016/j.gaitpost.2008.05.001
- [4] Needham L, Evans M, Cosker DP, Wade L, McGuigan PM, Bilzon JL, et al. The accuracy of several pose estimation methods for 3D joint centre localisation. *Sci Rep*. 2021;11:20673. doi:10.1038/s41598-021-00212-x
- [5] Uhlich SD, Falisse A, Kivelö J, Muccini J, Ko M, Chaudhari AS, et al. OpenCap: human movement dynamics from smartphone videos. *PLoS Comput Biol*. 2023;19(10):e1011462. doi:10.1371/journal.pcbi.1011462
- [6] Wade L, Needham L, McGuigan P, Bilzon J. Applications and limitations of current markerless motion capture methods for clinical gait biomechanics. *PeerJ*. 2022;10:e12995. doi:10.7717/peerj.12995
- [7] Mannini A, Sabatini AM. Gait phase detection and discrimination between walking–jogging activities using hidden Markov models applied to foot motion data from a gyroscope. *Gait Posture*. 2012;36(4):657–661. doi:10.1016/j.gaitpost.2012.06.017
- [8] Martinez-Hernandez U, Dehghani-Sani AA. Adaptive Bayesian inference system for recognition of walking activities and prediction of gait events using wearable sensors. *Neural Netw*. 2018;102:107–119. doi:10.1016/j.neunet.2018.02.017
- [9] Kanko RM, Laende EK, Davis EM, Selbie WS, Deluzio KJ. Concurrent assessment of gait kinematics using marker-based and markerless motion capture. *J Biomech*. 2021;127:110665. doi:10.1016/j.jbiomech.2021.110665
- [10] Camargo J, Ramanathan A, Flanagan W, Young AJ. A comprehensive, open-source dataset of lower limb biomechanics in multiple conditions of stairs, ramps, and level-ground ambulation and transitions. *J Biomech*. 2021;119:110320. doi:10.1016/j.jbiomech.2021.110320
- [11] Evans M, Needham L, Wade L, Parsons M, Colyer S, McGuigan P, et al. Synchronised video, motion capture and force plate dataset for validating markerless human movement analysis. University of Bath Research Data Archive; 2024. doi:10.15125/BATH-01258
- [12] Fukuchi CA, Fukuchi RK, Duarte M. A public dataset of overground and treadmill walking kinematics and kinetics in healthy individuals. *PeerJ*. 2018;6:e4640. doi:10.7717/peerj.4640
- [13] Lavikainen J, Vartiainen P, Stenroth L, Karjalainen PA, Korhonen RK, Liukkonen MK, et al. Gait data from 51 healthy participants with motion capture, inertial measurement units, and computer vision. *Data Brief*. 2024;56:110841. doi:10.1016/j.dib.2024.110841

- [14] Schreiber C, Moissenet F. A multimodal dataset of human gait at different walking speeds established on injury-free adult participants. *Sci Data*. 2019;6:111. doi:10.1038/s41597-019-0124-4
- [15] Van Crielinge T, Saeys W, Truijen S, Vereeck L, Sloot LH, Hallemans A. A full-body motion capture gait dataset of 138 able-bodied adults across the life span and 50 stroke survivors. *Sci Data*. 2023;10:852. doi:10.1038/s41597-023-02767-y
- [16] van der Zee TJ, Munding EM, Kuo AD. A biomechanics dataset of healthy human walking at various speeds, step lengths and step widths. *Sci Data*. 2022;9:704. doi:10.1038/s41597-022-01817-1
- [17] Viglialoro RM, Centrone A, Mattei L, Dalise S, Ferrari NA, Azzollini V, et al. Dataset on gait analysis of parkinsonian subjects: effect of Nordic walking. *Sci Data*. 2025;12:1937. doi:10.1038/s41597-025-06209-9
- [18] Shida TKF, Costa TM, de Oliveira CEN, Leite MFV, Huber P, Machado L, et al. A public data set of walking full-body kinematics and kinetics in individuals with Parkinson's disease. *Front Neurosci*. 2023;17:992585. doi:10.3389/fnins.2023.992585
- [19] Hood S, Ishmael MK, Gunnell A, Foreman KB, Lenzi T. A kinematic and kinetic dataset of 18 above-knee amputees walking at various speeds. *Sci Data*. 2020;7:150. doi:10.1038/s41597-020-0494-7
- [20] Caicedo PE, Rengifo CF, Rodriguez LE, Sierra WA, Gomez MC. Dataset for gait analysis and assessment of fall risk for older adults. *Data Brief*. 2020;33:106550. doi:10.1016/j.dib.2020.106550
- [21] Bonnefoy-Mazure A, Franco-Carvalho M, Leroquais C, De Coulon G, Lascombes P, Armand S. Gait data of idiopathic toe walking patients with pre- and post-treatment sessions and healthy participants, including ankle kinematics and kinetics. *Zenodo*; 2024. doi:10.5281/zenodo.10796424
- [22] Bertaux A, Gueugnon M, Moissenet F, Orliac B, Martz P, Maillefert J-F, et al. Gait analysis dataset of healthy volunteers and patients before and 6 months after total hip arthroplasty (part 1). *figshare*; 2021. doi:10.6084/m9.figshare.14420645
- [23] Bertaux A, Gueugnon M, Moissenet F, Orliac B, Martz P, Maillefert J-F, et al. Gait analysis dataset of healthy volunteers and patients before and 6 months after total hip arthroplasty (part 2). *figshare*; 2021. doi:10.6084/m9.figshare.14420756
- [24] Grouvel G, Boutabla A, Corre J, Revol R, Revaz M, Cavuscens S, et al. Multiple motor tasks while walking in patients with chronic bilateral and unilateral vestibulopathy. *Zenodo*; 2024. doi:10.5281/zenodo.14179651

- [25] Leardini A, Chiari L, Della Croce U, Cappozzo A. Human movement analysis using stereophotogrammetry. Part 3. Soft tissue artifact assessment and compensation. *Gait Posture*. 2005;21(2):212–225. doi:10.1016/j.gaitpost.2004.05.002
- [26] Peters A, Galna B, Sangeux M, Morris M, Baker R. Quantification of soft tissue artifact in lower limb human motion analysis: a systematic review. *Gait Posture*. 2010;31(1):1–8. doi:10.1016/j.gaitpost.2009.09.004
- [27] Dumphart B, Slijepcevic D, Zeppelzauer M, Kranzl A, Unglaube F, Baca A, et al. Robust deep learning-based gait event detection across various pathologies. *PLoS ONE*. 2023;18(8):e0288555. doi:10.1371/journal.pone.0288555
- [28] Zeni JAJ, Richards JG, Higginson JS. Two simple methods for determining gait events during treadmill and overground walking using kinematic data. *Gait Posture*. 2008;27(4):710–714. doi:10.1016/j.gaitpost.2007.07.007
- [29] Lempereur M, Rousseau F, Rémy-Néris O, Pons C, Houx L, Quellec G, et al. A new deep learning-based method for the detection of gait events in children with gait disorders: proof-of-concept and concurrent validity. *J Biomech*. 2020;98:109490. doi:10.1016/j.jbiomech.2019.109490
- [30] Liu Y, Liu X, Zhu Q, Chen Y, Yang Y, Xie H, et al. Adaptive detection in real-time gait analysis through the Dynamic Gait Event Identifier. *Bioengineering (Basel)*. 2024;11(8):806. doi:10.3390/bioengineering11080806
- [31] Vancanneyt T, Le Moal C, Blard M, Lenoir J, Roche N, Bonnyaud C, et al. Automatic detection of human gait events: a simple but versatile 3D algorithm. *J Neuroeng Rehabil*. 2025;22(1):110. doi:10.1186/s12984-025-01544-9
- [32] Ghousayni S, Stevens C, Durham S, Ewins D. Assessment and validation of a simple automated method for the detection of gait events and intervals. *Gait Posture*. 2004;20(3):266–272. doi:10.1016/j.gaitpost.2003.10.001
- [33] Hreljac A, Marshall RN. Algorithms to determine event timing during normal walking using kinematic data. *J Biomech*. 2000;33(6):783–786. doi:10.1016/S0021-9290(00)00014-2
- [34] Mickelborough J, van der Linden ML, Richards J, Ennos AR. Validity and reliability of a kinematic protocol for determining foot contact events. *Gait Posture*. 2000;11(1):32–37. doi:10.1016/S0966-6362(99)00050-8
- [35] O'Connor CM, Thorpe SK, O'Malley MJ, Vaughan CL. Automatic detection of gait events using kinematic data. *Gait Posture*. 2007;25(3):469–474. doi:10.1016/j.gaitpost.2006.05.016
- [36] Lugaresi C, Tang J, Nash H, McClanahan C, Uboweja E, Hays M, et al. MediaPipe: a framework for building perception pipelines. *arXiv [Preprint]*. 2019. doi:10.48550/arXiv.1906.08172
- [37] Winter DA. *Biomechanics and Motor Control of Human Movement*. 4th ed. Hoboken (NJ): John Wiley & Sons; 2009.

- [38] Fer F. myogait: a markerless video-based gait analysis toolkit. Version 0.6.0. Zenodo; 2026. doi:10.5281/zenodo.20050937
- [39] Domènech-Garcia G, Marimon X, Carrasco-Urribarren A, Portela AE, Bagur-Calafat C. Markerless motion capture for human movement estimation using artificial intelligence: a systematic review. *Pediatr Rep.* 2026;18(4):83. doi:10.3390/pediatric18040083
- [40] Khirodkar R, Wen H, Martinez J, Dong Y, Su Z, Saito S. Sapiens2. In: *International Conference on Learning Representations*. 2026. p. 58484–58507.
- [41] Antonsson EK, Mann RW. The frequency content of gait. *J Biomech.* 1985;18(1):39–47. doi:10.1016/0021-9290(85)90043-0
- [42] Whittle MW. Generation and attenuation of transient impulsive forces beneath the foot: a review. *Gait Posture.* 1999;10(3):264–275. doi:10.1016/S0966-6362(99)00041-7
- [43] Kidziński Ł, Yang B, Hicks JL, Rajagopal A, Delp SL, Schwartz MH. Deep neural networks enable quantitative movement analysis using single-camera videos. *Nat Commun.* 2020;11:4054. doi:10.1038/s41467-020-17807-z
- [44] Fer F. gaitkit: a unified toolkit for kinematic gait event detection. Version 1.4.9. Zenodo; 2026. doi:10.5281/zenodo.18653109

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Additionalfile1JNER20260827RF.docx](#)