SynthGait-19K: A Physically Grounded Synthetic Video Dataset for Gait Parameter Estimation

Soroush Mehraban[1,2,3] Xin Lei Lin[1,2,3] Vida Adeli[1,2,3]
Majid Mirmehdi[4] Amirhossein Dadashzadeh[4] Clint Hansen[5]
Andrea Iaboni[1,2] Babak Taati[1,2,3]

[1]KITE Research Institute
[2]University of Toronto [3]Vector Institute
[4]University of Bristol [5]Kiel University

Abstract

*Accurate estimation of clinically meaningful gait parameters from monocular video is important for scalable mobility assessment, yet progress is limited by the small scale, restricted viewpoints, and limited visual diversity of existing datasets. We introduce **SynthGait-19K**, a physically grounded synthetic video dataset containing 19,272 walking videos derived from 6,427 MoCap sequences across 437 subjects, with paired SMPL motion and annotations for six gait parameters. To construct the dataset, we develop **Gait2Vid**, which unifies heterogeneous MoCap recordings through SMPL and synthesizes diverse RGB walking videos under controllable viewpoints and scene appearances. We assess the generated videos for consistency with their conditioning gait kinematics and validate extracted gait events against force-platform measurements. Using SynthGait-19K, we benchmark direct RGB, pose-based, biomechanical, and human-mesh-recovery approaches and analyze viewpoint, training-data scale, and synthetic-to-real domain shift. We also introduce **GaitXFormer** as a direct RGB reference model for estimating gait parameters. Synthetic supervision transfers effectively to real videos across both GaitXFormer and a pose-based architecture, demonstrating utility across different representations. We further find that spatial gait parameters are more sensitive to visual domain shift and that improved HMR reconstruction alone does not necessarily translate to improved downstream gait estimation.*

1. Introduction

Accurate estimation of human gait parameters has become an increasingly important goal in computer vision, driven by its broad impact in healthcare and mobility assessment. Gait features such as walking speed, cadence, step length, step width, stooped posture, and arm swing serve as critical biomarkers in diagnosing neurological conditions like Parkinson's disease [13, 24, 31], predicting fall risk in older adults [1, 5, 30], and tracking rehabilitation progress [2]. Traditionally, gait evaluation relies on inertial measurement units (IMUs) [35, 37] or marker-based motion capture (MoCap) systems [8, 15]. Although these approaches offer high accuracy, they require laboratory environments, specialized equipment, and extensive subject preparation, which restricts their use in routine clinical workflows. Furthermore, gait patterns measured in controlled settings often diverge from those observed in daily life [29], highlighting the need for ambient and unobtrusive capture methods for robust assessment in natural, everyday settings.

Vision-based gait assessment offers a promising path toward scalable and non-intrusive analysis, but progress is limited by the availability and scope of existing datasets. Estimating gait parameters from video typically requires supervision from force plates or MoCap systems, making large-scale collection expensive and difficult to share due to the identifiable biometric information contained in gait recordings. Consequently, existing datasets are often collected in a single controlled environment, with restricted camera viewpoints and limited variation in subject appearance. These constraints make it difficult to assess whether a method generalizes beyond the capture setup on which it was developed. They also make it challenging to isolate the effect of individual factors such as camera viewpoint, visual appearance, or scene variation, since these factors typically change together across datasets. As a result, current evaluations provide only limited insight into how different gait-estimation approaches behave under controlled distribution shifts and which aspects of the visual domain are most responsible for performance degradation.

To address these limitations, we introduce **SynthGait-19K**, a large-scale, physically grounded synthetic video dataset for gait parameter estimation, with motion derived

Figure 1. **SynthGait-19K overview.** Source-cohort composition and representative depth-conditioned generations across viewpoints and diverse indoor/outdoor scenes. The dataset contains 19,273 videos from 437 subjects with paired 3D motion and six gait-parameter annotations.

from real MoCap recordings. SynthGait-19K contains 19,273 RGB walking videos derived from 6,427 MoCap sequences across 437 subjects, with paired SMPL motion and annotations for six clinically relevant gait parameters. As summarized in Figure 1, the source data span five independent cohorts, including 231 healthy or asymptomatic participants and 206 participants from clinical populations, providing diversity in both gait characteristics and capture protocols. Each source motion is observed under multiple camera configurations and diverse visual appearances, allowing viewpoint and scene variation to be studied independently of the underlying fitted motion.

To construct SynthGait-19K, we develop **Gait2Vid**, a synthetic video-generation pipeline that converts heterogeneous MoCap recordings into a common SMPL representation, renders them from controllable virtual cameras, and uses depth-conditioned video diffusion to generate diverse RGB walking videos across indoor and outdoor scenes. We validate the generated videos for consistency with their conditioning gait kinematics and independently validate the extracted gait events against force-platform measurements. The resulting dataset enables controlled study of viewpoint, training-data scale, and synthetic-to-real visual shift, while providing a common benchmark for methods with substantially different intermediate representations.

We evaluate human mesh recovery (HMR), biomechanical, pose-based, and direct RGB approaches on a common real-world gait benchmark. As a direct RGB reference model, we introduce **GaitXFormer**, a Video-ViT-based model that predicts gait parameters directly from monocular video. Training both GaitXFormer and a pose-based STT model [19] on SynthGait-19K shows that synthetic supervision transfers effectively to real video across different representations. Our experiments further characterize viewpoint dependence, scaling behavior, and synthetic-to-real domain shift, and show that improved intermediate HMR reconstruction alone does not necessarily translate to improved downstream gait estimation.

Contributions. (1) We introduce **SynthGait-19K**, a physically grounded synthetic gait-video dataset containing 19,272 videos from 6,427 walking sequences across 437 subjects, with paired SMPL motion and six gait-parameter annotations. (2) We develop **Gait2Vid** to construct the dataset and validate the generated motion consistency and gait-event annotations through kinematic-fidelity and force-platform analyses. (3) We establish a benchmark spanning direct RGB, pose-based, biomechanical, and HMR approaches, and use SynthGait-19K to study viewpoint, data scale, and synthetic-to-real transfer, with **GaitXFormer** serving as a

direct RGB reference model.

2. Related Work

2.1. Video-based Gait Parameter Estimation

Vision-based gait analysis methods broadly fall into general-purpose human-motion reconstruction pipelines and task-specific gait-estimation models.

Human mesh recovery–based methods. HMR methods estimate 3D body pose and shape from images [11, 26] or videos [23, 36, 39]. Gait events and spatiotemporal parameters can then be derived from the reconstructed motion, for example using UnderPressure [25]. However, these models are primarily optimized for human reconstruction rather than downstream gait accuracy.

Task-specific gait estimation. Other approaches directly target gait quantities from visual representations. Pose2Gait [22], Kidziński *et al.* [17], and STT [19] estimate gait parameters from 2D pose trajectories, while Transforming Gait [6] operates on estimated 3D joint trajectories. Biomechanical pipelines such as PBL [28] and OpenCap Monocular [10] further combine monocular reconstruction with biomechanical modeling. These methods span substantially different intermediate representations and objectives, motivating comparison based on the resulting gait measurements rather than reconstruction quality alone.

2.2. Gait Datasets and Evaluation

Existing gait datasets with synchronized video and motion or biomechanical measurements remain relatively small and capture-specific. GPJATK [18] provides synchronized MoCap and calibrated multi-view RGB, while prior task-specific studies rely on dedicated dementia [22], cerebral-palsy [17], or instrumented gait-laboratory cohorts [6]. Such datasets provide valuable real-world measurements but make it difficult to vary viewpoint, appearance, or scene independently while preserving the underlying motion. SynthGait-19K complements them with large-scale RGB walking videos whose camera and visual conditions can be controlled independently of the fitted gait motion.

2.3. Synthetic Human Motion and Video Data

Synthetic data has enabled scalable supervision for human-motion understanding. SURREAL [38] generates SMPL-based synthetic humans from MoCap motion, while AGORA [27] and BEDLAM [4] increase diversity in appearance, clothing, scenes, and camera configurations. These datasets primarily target general human reconstruction rather than quantitative gait analysis.

Most closely related, Yamada *et al.* [41] study synthetic musculoskeletal gait data for healthcare applications using projected pose representations from simulated gait. In contrast, Gait2Vid starts from real-world recorded MoCap across multiple cohorts and generates diverse RGB videos while retaining correspondence with fitted 3D motion and gait annotations, enabling controlled analysis of viewpoint and synthetic-to-real visual variation.

3. SynthGait-19K Dataset

Dataset composition. SynthGait-19K contains 19,272 RGB walking videos derived from 6,427 unique MoCap sequences comprising 671 minutes of walking from 437 subjects across five public datasets [3, 12, 32, 34, 40]. As summarized in Figure 1, the source cohorts include both healthy or asymptomatic participants and multiple clinical populations. Each video is paired with its fitted SMPL motion and annotations for six gait parameters: cadence, walking speed, step length, step width, stooped posture, and arm swing. Table 1 summarizes the source-specific subject, sequence, and video counts.

Viewpoint and visual diversity. Each source motion is observed under front, back, sagittal, and oblique camera configurations, with additional variation in camera pitch, subject appearance, scene, and lighting. Because these factors can vary while the fitted motion and gait labels remain fixed, SynthGait-19K enables controlled analysis of viewpoint and visual-domain variation. Figure 3 shows representative samples from the same underlying walking sequence across multiple views and visual conditions. We use an 80/20 subject-level training/validation split so that all videos from a given subject remain within the same partition.

4. Gait2Vid: Dataset Construction Pipeline

We propose Gait2Vid, a synthetic data-generation pipeline for constructing SynthGait-19K from heterogeneous MoCap recordings. As illustrated in Figure 2, Gait2Vid converts source motions into a common SMPL representation, renders them from controllable virtual cameras, synthesizes diverse RGB videos using depth-conditioned video diffusion, and derives gait annotations from the fitted motion.

Unified motion representation. The five source datasets use different marker layouts and joint conventions, preventing their raw trajectories from being combined directly. We therefore convert each dataset into a common SMPL [21] representation, providing a consistent full-body representation for video generation and gait annotation.

Synthetic camera and scene construction. Given a fitted SMPL walking sequence, we render depth videos from front, back, sagittal, and randomly sampled oblique cameras with varying pitch. A planar ground surface is rendered beneath the walking trajectory to provide a stable geometric reference during synthesis.

Depth-conditioned video synthesis. The rendered depth sequence conditions Wan2.1-14B-VACE [14], together with a text prompt describing the subject and environment. Prompts

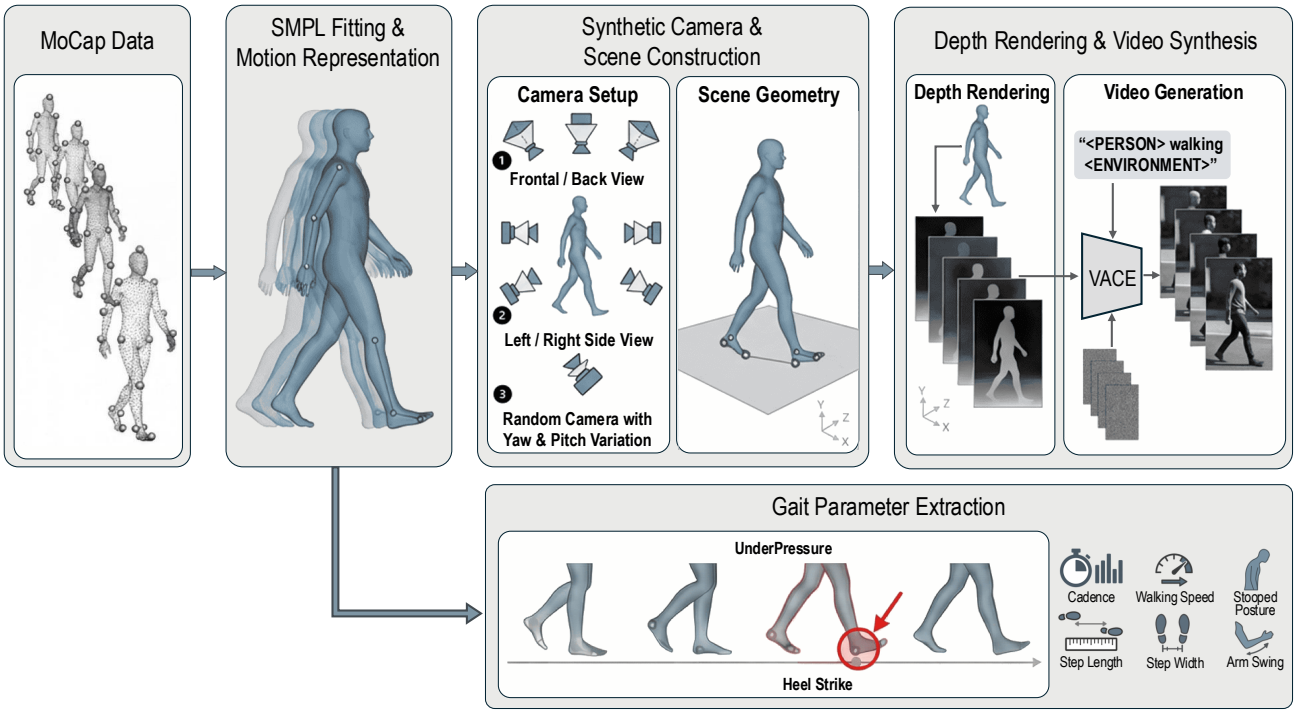


Figure 2. Overview of the synthetic video generation pipeline. 3D walking motions from MoCap data are fitted with SMPL, rendered into depth maps using controllable synthetic cameras and scene geometry, and used to condition a video diffusion model to generate diverse walking videos.

Table 1. **Composition of SynthGait-19K and real-video evaluation data.** SynthGait-19K combines five public MoCap datasets; GPJATK is the independent real-video benchmark. Counts report subjects, sequences, and videos by viewpoint.

Dataset / motion source	# Subjects	# Sequences	# Videos			
			Front/Back	Sagittal	Oblique	Total
SynthGait-19K composition						
Schreiber & Moissenet [34]	49	917	917	918	918	2,752
Santos *et al.* [32]	25	488	488	488	488	1,464
Bertaux *et al.* [3]	186	4,442	4,441	4,442	4,436	13,319
Grouvel *et al.* [12]	10	82	82	82	82	246
Warmerdam *et al.* [40]	167	497	497	497	497	1,491
SynthGait-19K total	**437**	**6,427**	**6,425**	**6,427**	**6,421**	**19,272**
Independent real-video evaluation						
GPJATK [18]	32	152	152	152	304	608

are sampled from 200 curated indoor and outdoor scene templates, producing variation in subject appearance, background, lighting, and recording context while conditioning on the fitted motion.

Gait parameter extraction. Heel strikes are detected from the fitted SMPL motion using UnderPressure [25]. Cadence is computed from their timing; walking speed from pelvis displacement; step length and width from forward and mediolateral foot displacement between successive steps; stooped posture from normalized neck–pelvis forward displacement; and arm swing from normalized wrist range along the walking direction. Each generated video remains paired with gait labels derived from its conditioning motion. Dataset-specific SMPL fitting, video synthesis, and gait-parameter definitions are detailed in the supplementary material.

5. Benchmark Protocol

5.1. Evaluation Dataset and Metrics

Real-world evaluation. We evaluate all methods on GPJATK [18], a synchronized RGB–MoCap gait dataset con-

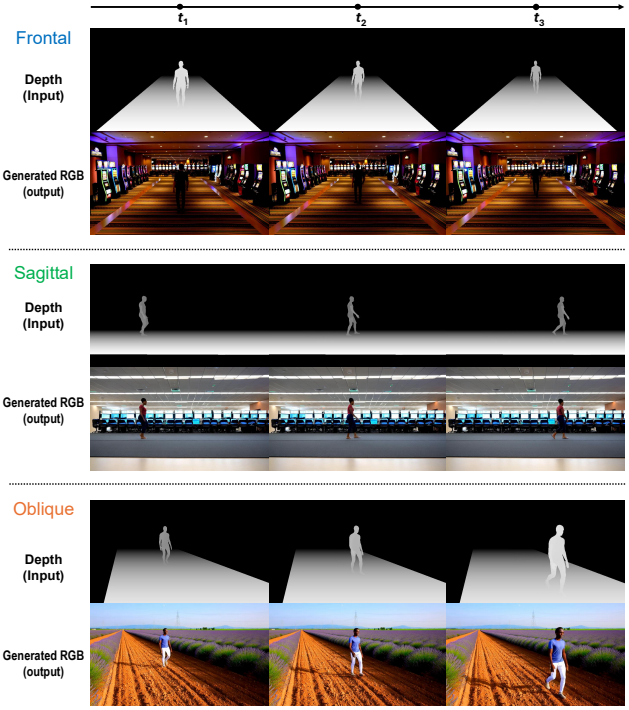


Figure 3. Qualitative samples from SynthGait-19K showing the diversity of generated subjects, environments, viewpoints, and walking motions. Each generated RGB video is paired with the corresponding SMPL motion and gait annotations.

taining 152 walking sequences from 32 subjects. The dataset provides 608 RGB videos across four camera configurations: 152 side, 76 front, 76 back, and 304 oblique views. The RGB videos are used as model input, while the synchronized MoCap recordings are used to derive reference gait parameters using the same gait-annotation definitions as SynthGait-19K.

Preferred-view protocol. Because different gait parameters are most observable from different camera directions, we define a fixed preferred-view protocol used for the main benchmark. Walking speed, step length, stooped posture, and arm swing are evaluated from the side view; step width is evaluated using front and back views; and cadence is evaluated across all views. We additionally report view-wise results to characterize sensitivity to camera viewpoint.

Metrics. We report Pearson correlation (r) to measure how well each method preserves inter-sequence variation in each gait parameter. The overall correlation is obtained by averaging the six parameter-wise correlations using Fisher z-transformation. We additionally report mean absolute error (MAE) in the native units of each gait parameter in the supplementary material to assess absolute prediction accuracy.

5.2. GaitXFormer

We introduce GaitXFormer, a direct RGB model for estimating gait parameters without an intermediate pose or mesh representation. As illustrated in Figure 4, a V-JEPA2-initialized Video ViT encodes the walking video into spatiotemporal tokens. Six learnable gait queries, one for each gait parameter, independently cross-attend to the video representation through a lightweight decoder, and parameter-specific linear heads predict cadence, walking speed, step length, step width, stooped posture, and arm swing. The model is trained end-to-end using mean-squared error over normalized gait parameters. Architecture and training details are provided in the supplementary material.

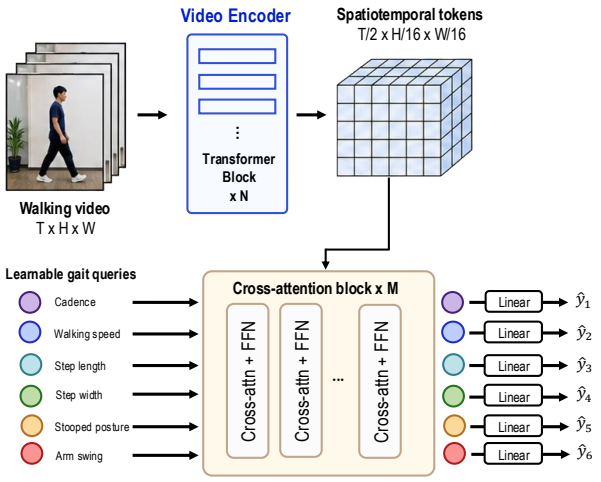


Figure 4. **GaitXFormer overview.** An input walking video is encoded by a Video ViT into spatiotemporal tokens. A set of six learnable gait-query tokens cross-attends, and parameter-specific linear heads produce the final gait estimates.

5.3. Comparison Methods

We benchmark methods spanning different representations and training objectives. For general-purpose human mesh recovery, we evaluate WHAM [36], CameraHMR [26], PromptHMR [39], and FastHMR [23]. For each method, the predicted 3D body motion is converted to the six gait parameters using the same gait-feature extraction procedure, allowing downstream gait accuracy to be compared under a common protocol.

To provide a task-specific comparison with matched synthetic supervision, we additionally train STT [19] on SynthGait-19K. Unlike the HMR baselines, STT operates on pose trajectories and is explicitly optimized for gait-parameter estimation, providing a complementary control for separating the effect of task-specific supervision from the choice of intermediate representation.

We further consider the biomechanical pipelines PBL [28] and OpenCap Monocular [10], which are evaluated using their released formulations. Additional adaptation details and representation-specific constraints are provided in the supplementary material.

Table 2. **Kinematic fidelity of VACE-generated GPJATK videos.** Sapiens2 pose estimates are compared with projected fitted-SMPL motion for paired real and generated videos. Deltas are Generated − Real; 95% CIs are obtained by sequence-level bootstrap.

Metric	Real	Generated	Δ	95% CI
BBox-NMPJPE (%) ↓	3.41	2.32	-1.09	[-1.19, -0.98]
Lower-body NMPJPE (%) ↓	2.90	2.20	-0.70	[-0.79, -0.61]
Lower-body velocity error ↓	0.74	0.76	+0.02	[-0.01, +0.04]
Knee-angle MAE (°) ↓	8.71	4.86	-3.86	[-4.11, -3.62]

Table 3. **UnderPressure heel-strike validation against force-platform measurements.** Timing errors are reported in 30-FPS SMPL frames over force-platform-observed contacts.

Events	MAE (frames) ↓	≤3 frames ↑	≤5 frames ↑
6,092	2.31	82.3%	91.7%

6. Results and Analysis

6.1. Synthetic Video Kinematic Fidelity

We assess whether RGB synthesis preserves the conditioning motion by projecting each fitted SMPL sequence into the calibrated camera and comparing it with Sapiens2 [16] 2D pose estimates from the corresponding RGB video. Confidence intervals are obtained by bootstrapping walking sequences, keeping synchronized views grouped.

As shown in Tab. 2, generated videos have lower pose and knee-angle errors than the corresponding real videos, likely due to cleaner visual conditions and reduced clothing-induced ambiguity, while velocity error is nearly unchanged. We therefore find no evidence that synthesis degrades adherence to the conditioning motion. Since fitted SMPL is not independent marker-level ground truth, this analysis measures motion consistency rather than absolute biomechanical accuracy.

6.2. Gait Annotation Validation

Gait2Vid gait annotations rely on heel strikes detected from fitted SMPL motion using UnderPressure [25]. We validate their timing against force-platform contacts from the Bertaux *et al*. [3] subset. Because the force plates cover only part of the walkway, they provide an independent reference for observed contacts, while UnderPressure remains necessary for the full sequence.

Across 6,092 force-platform-observed heel strikes, UnderPressure achieves a mean absolute timing error of 2.31 30-FPS SMPL frames ($\approx$ 77 ms), with 82.3% and 91.7% localized within 3 and 5 frames, respectively (Tab. 3).

6.3. Real-World Gait Estimation Benchmark

Table 4 compares gait-estimation approaches spanning HMR, biomechanical, 2D pose-based, and direct video-based formulations. The HMR and biomechanical approaches first recover an intermediate body representation from which gait parameters are derived, whereas STT and GaitXFormer are optimized specifically for gait estimation. Across the off-the-shelf HMR and biomechanical approaches, performance varies substantially by gait parameter: OpenCap Monocular performs strongly for walking speed, WHAM for step width, and FastHMR for arm swing.

Table 4. Preferred-view comparison across gait parameters. We report Pearson correlation (r). **Cad**: cadence, **W.Speed**: walking speed, **Step Len**: step length, **Step Wid**: step width, **Stoop Post**: stooped posture, **Arm Swing**: arm swing. **Avg** denotes the Fisher z-transformed average across parameters. † denotes STT trained on SynthGait-19K; – indicates unsupported outputs.

Method	Cad	W. Speed	Step Len	Step Wid	Stoop Post	Arm Swing	Avg
HMR							
WHAM [36]	0.85	0.82	0.40	**0.69**	0.64	0.67	0.70
CameraHMR [26]	0.71	0.65	0.10	0.35	0.39	0.91	0.59
PromptHMR [39]	0.87	0.61	0.13	0.48	0.63	0.90	0.67
FastHMR [23]	0.80	0.63	-0.07	0.18	0.02	**0.93**	0.54
Biomechanical							
PBL [28]	0.47	0.77	0.38	0.50	0.73	0.63	0.60
OpenCap-M [10]	0.76	0.87	0.62	0.47	0.58	0.65	0.68
2D Pose-based							
STT [19]	0.20	0.67	–	–	–	–	–
STT†	0.91	0.85	**0.73**	0.67	0.70	0.92	0.82
Video-based							
GaitXFormer	**0.94**	**0.88**	0.67	0.67	**0.75**	0.91	**0.84**

To evaluate whether the benefit of SynthGait-19K extends beyond GaitXFormer, we additionally train the STT architecture on SynthGait-19K using the same six gait-parameter annotations. The released STT model was trained on side-view videos from a cerebral-palsy cohort and therefore exhibits limited cross-dataset transfer to GPJATK, particularly for cadence ($r = 0.20$), while its walking-speed correlation is 0.67. After training on SynthGait-19K, STT† reaches a Fisher-averaged correlation of 0.82 on real GPJATK, compared with 0.84 for GaitXFormer, and achieves the strongest step-length result among the evaluated methods. This substantial improvement indicates that SynthGait-19K provides useful supervision for a substantially different pose-based architecture, rather than benefiting only GaitXFormer. At the same time, the parameter-wise differences across methods highlight that no single representation is uniformly optimal for all gait quantities.

Inference efficiency. GaitXFormer processes a 5-second walking clip in 0.27 s on a single RTX3090, compared with 1.21–139.10 s for the evaluated baselines, including required preprocessing. As shown in Figure 5, it achieves the most favorable accuracy–runtime trade-off. STT itself requires only 0.32 s, but additionally relies on OpenPose preprocessing.

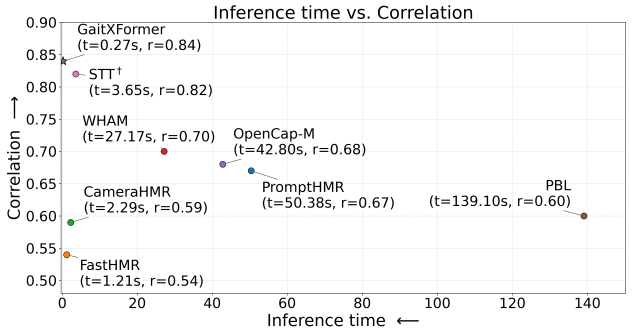


Figure 5. **Accuracy–runtime trade-off.** Pearson correlation versus inference time for a 5-second walking clip on a single NVIDIA RTX3090 GPU.

Table 5. **Synthetic-to-real transfer analysis.** GPJATK-VACE contains VACE-generated videos using GPJATK motions and camera viewpoints. **Syn.** indicates synthetic RGB input, and **Avg** denotes the Fisher z-transformed average across gait parameters.

Eval.	Syn.	Cad	W. Spd	Step Len	Step Wid	Stoop Post	Arm Swing	Avg
GaitXFormer								
Synth. val	✓	0.71	0.96	0.93	0.92	0.96	0.93	0.92
GPJATK-VACE	✓	0.93	0.89	0.74	0.77	0.80	0.93	0.87
GPJATK	×	0.94	0.88	0.67	0.67	0.75	0.91	0.84
WHAM								
GPJATK-VACE	✓	0.82	0.79	0.40	0.56	0.79	0.61	0.68
GPJATK	×	0.85	0.82	0.40	0.69	0.64	0.67	0.70

6.4. Synthetic-to-Real Transfer Analysis

Gait2Vid allows us to examine the effect of visual-domain change while retaining the same source gait motions and camera configurations. We therefore compare performance on GPJATK-VACE, generated from GPJATK motion, with performance on the corresponding real GPJATK domain. We additionally report performance on the held-out SynthGait-19K validation set to distinguish transfer to unseen synthetic motion from transfer to real imagery.

GaitXFormer exhibits a modest decrease in Fisher-averaged correlation from 0.87 on GPJATK-VACE to 0.84 on real GPJATK. Cadence, walking speed, and arm swing remain nearly unchanged, whereas larger reductions occur for step length and step width, suggesting that spatial gait quantities are more sensitive to the synthetic-to-real appearance shift.

WHAM shows a different pattern. Its overall correlation remains similar across generated and real GPJATK (0.68 vs. 0.70), but the effect varies considerably across parameters. Step width improves from 0.56 to 0.69 on real video, while stooped-posture correlation decreases from 0.79 to 0.64. These results indicate that synthetic-to-real sensitivity is not a uniform property of the dataset, but depends on both the estimated gait quantity and the intermediate representation used by the model.

Table 6. **Effect of SynthGait supervision on WHAM.** Reconstruction and gait correlations are evaluated on real GPJATK. Adapt. denotes WHAM with the SynthGait-19K-trained output adapter.

Metric	WHAM	Adapt.	Change
SMPL reconstruction ↓			
Body-pose error (°)	10.54	9.76	−7.43%
Pelvis MPJPE (mm)	46.93	45.63	−2.78%
PA-MPJPE (mm)	33.28	32.44	−2.53%
Gait correlation ↑			
Walking speed	0.826	0.842	+0.017
Step length	0.402	0.442	+0.040
Step width	0.687	0.666	-0.021
Arm swing	0.669	0.651	-0.017
Avg	0.705	0.708	+0.004

6.5. Effect of SynthGait-19K Supervision on HMR

To examine whether SynthGait-19K supervision can improve an HMR-based gait pipeline, we adapt WHAM using the paired SMPL motion available in SynthGait-19K. We keep the released WHAM RGB-to-motion network fixed and train a lightweight temporal output adapter that refines its predicted pose, shape, and trajectory scale. We then evaluate both the resulting motion reconstruction and the gait parameters derived from the adapted predictions.

As shown in Tab. 6, SynthGait-19K supervision improves WHAM reconstruction on real GPJATK, reducing body-pose error by 7.43%, pelvis MPJPE by 2.78%, and PA-MPJPE by 2.53%. These reconstruction gains, however, do not translate uniformly to gait estimation. The Fisher-averaged correlation changes only modestly from 0.7045 to 0.7080 ($\Delta r = 0.0035$), with a participant-bootstrap 95% confidence interval of $[-0.0068, 0.0166]$. Walking speed and step length improve, while step width and arm swing decrease. This result highlights that improved human-motion reconstruction does not necessarily imply improved accuracy for derived gait quantities.

6.6. Controlled Analyses Enabled by SynthGait-19K

Sensitivity to camera viewpoint. SynthGait-19K provides controlled training data across diverse camera configurations while preserving the underlying gait motion. To assess whether the resulting model remains robust to viewpoint changes in real video, we evaluate GaitXFormer on the multi-view GPJATK benchmark. As shown in Tab. 7, cadence is highly stable across camera configurations, whereas other gait parameters exhibit stronger viewpoint dependence. Stooped posture and arm swing are best estimated from the side view, while step width benefits substantially from front and back views. Walking speed and step length remain relatively robust across views, with the highest correlations observed for the oblique cameras. These results

Table 7. Viewpoint-wise Pearson correlation (r, ↑) of GaitXFormer on GPJATK. Numbers in parentheses denote the number of test videos.

Gait Feature	Side (152)	Front (76)	Back (76)	Oblique (304)
Cadence	**0.95**	0.94	0.94	0.92
Walking Speed	0.88	0.87	0.87	**0.91**
Step Length	0.67	0.62	0.64	**0.70**
Step Width	0.50	0.63	**0.70**	0.48
Stooped Posture	**0.75**	0.26	0.24	0.59
Arm Swing	**0.91**	0.77	0.78	**0.91**

Table 8. Effect of training-data scale on gait-estimation performance in terms of Pearson correlation (r). **Avg** denotes the Fisher z-transformed average across gait parameters.

Train	Cad	W. Speed	Step Len	Step Wid	Stoop Post	Arm Swing	Avg
25%	0.87	0.84	0.60	0.59	**0.77**	0.89	0.79
50%	0.92	0.84	0.57	0.62	0.72	0.91	0.80
75%	0.92	0.87	0.66	0.65	0.74	**0.92**	0.83
100%	**0.94**	**0.88**	**0.67**	**0.67**	0.75	0.91	**0.84**

Table 9. **UPDRS classification on PD4T** using frozen video encoders under leave-one-subject-out evaluation.

Encoder	Acc.	Macro Recall	Macro F1
V-JEPA2	71.3	68.3	69.1
GaitXFormer	**76.0**	**71.3**	**73.3**

characterize the remaining viewpoint sensitivity after multi-view synthetic training and motivate the parameter-specific preferred-view protocol used in our main benchmark.

Effect of synthetic training-data scale. SynthGait-19K also allows us to directly measure how increasing the amount of synthetic supervision affects transfer to real video. We train the same GaitXFormer configuration using progressively larger subsets of SynthGait-19K and evaluate each model on real GPJATK. As shown in Tab. 8, the Fisher-averaged correlation increases from 0.79 using 25% of the training data to 0.84 using the full dataset. Most gait parameters improve as additional synthetic data are introduced, with particularly clear gains for cadence, step length, and step width. The continued improvement up to the full training set indicates that the scale of SynthGait-19K contributes directly to downstream real-world performance.

6.7. External Clinical Transfer on PD4T

We further evaluate transfer to PD4T [7], an independent dataset of Parkinson's-disease walking videos with UPDRS severity labels. First, we freeze the GaitXFormer video encoder and train a lightweight classifier for UPDRS prediction, using the same protocol with the pretrained V-JEPA2 encoder as a baseline. Under leave-one-subject-out evaluation, the SynthGait-19K-trained GaitXFormer representation improves accuracy from 71.3% to 76.0% and macro F1 from 69.1% to 73.3% (Tab. 9), indicating that gait supervision produces features that transfer to clinical severity estimation.

Table 10. Patient-aware clinical-anchor analysis on PD4T. Predictions are averaged within each participant and UPDRS-score group before correlation analysis. Dir. denotes within-patient agreement with the predefined expected clinical direction, and SRM denotes standardized response mean.

Parameter	Spearman ρ	95% CI	Dir.	SRM
Walking Speed	-0.83	$[-0.88, -0.76]$	97.9%	1.37
Step Length	-0.83	$[-0.88, -0.78]$	97.9%	1.39
Arm Swing	-0.75	$[-0.84, -0.61]$	91.5%	1.24
Stooped Posture	0.30	$[0.14, 0.47]$	80.9%	0.53

We additionally examine whether GaitXFormer's predicted gait quantities themselves vary consistently with clinical severity. To account for repeated observations, predictions are averaged within each participant and UPDRS-score group, with confidence intervals obtained by bootstrapping participants. As shown in Tab. 10, walking speed, step length, and arm swing show strong negative associations with increasing UPDRS severity ($\rho = -0.83$, -0.83, and -0.75) and change in the expected direction in more than 91% of within-participant comparisons. Stooped posture shows a weaker positive association ($\rho = 0.30$), with 80.9% directional agreement. Together, these results provide complementary evidence that representations and gait quantities learned from synthetic supervision transfer to unseen clinical videos.

7. Conclusion

We introduced SynthGait-19K, a large-scale synthetic video dataset for gait analysis constructed from heterogeneous MoCap recordings and paired with unified SMPL motion and six gait-parameter annotations. Beyond providing training data, SynthGait-19K enables controlled evaluation of viewpoint, visual-domain shift, and training-data scale. Our experiments show that synthetic supervision transfers effectively to real video for several gait parameters, while spatial quantities remain more sensitive to domain shift. Improved HMR reconstruction does not necessarily improve downstream gait estimation, while predicted gait quantities show consistent associations with clinical severity on PD4T.

Limitations include incomplete coverage of real-world variation, particularly severe occlusion, assistive devices, and broader clothing and clinical diversity. The fixed 5-s window also excludes longer-horizon phenomena such as freezing of gait. Evaluation centers on GPJATK, with PD4T validation; broader cohorts would better characterize generalization.

References

[1] Vida Adeli, Navid Korhani, Andrea Sabo, Sina Mehdizadeh, Avril Mansfield, Alastair Flint, Andrea Iaboni, and Babak Taati. Ambient monitoring of gait and machine learning models for dynamic and short-term falls risk assessment in people with dementia. *IEEE journal of biomedical and health informatics*, 27(7):3599–3609, 2023. 1

[2] April M Barthuly, Richard W Bohannon, and Walter Gorack. Gait speed is a responsive measure of physical performance for patients undergoing short-term rehabilitation. *Gait & posture*, 36(1):61–64, 2012. 1

[3] Aurélie Bertaux, Mathieu Gueugnon, Florent Moissenet, Baptiste Orliac, Pierre Martz, Jean-Francis Maillefert, Paul Ornetti, and Davy Laroche. Gait analysis dataset of healthy volunteers and patients before and 6 months after total hip arthroplasty. *Scientific data*, 9(1):399, 2022. 3, 4, 6, 12

[4] Michael J Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. Bedlam: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In *2023 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 8726–8737. IEEE, 2023. 3

[5] ML Callisaya, Leigh Blizzard, Jennifer L McGinley, and VK Srikanth. Risk of falls in older people during fast-walking–the tascog study. *Gait & posture*, 36(3):510–515, 2012. 1

[6] R James Cotton, Emoonah McClerklin, Anthony Cimorelli, Ankit Patel, and Tasos Karakostas. Transforming gait: video-based spatiotemporal gait analysis. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 115–120. IEEE, 2022. 3

[7] Amirhossein Dadashzadeh, Shuchao Duan, Alan Whone, and Majid Mirmehdi. PECoP: Parameter efficient continual pretraining for action quality assessment. In *Proceedings of the IEEE/CVF Winter Conference on applications of computer vision*, pages 42–52, 2024. 8

[8] Roy B Davis III, Sylvia Ounpuu, Dennis Tyburski, and James R Gage. A gait analysis data collection and reduction technique. *Human movement science*, 10(5):575–587, 1991. 1

[9] Saeed Ghorbani, Kimia Mahdaviani, Anne Thaler, Konrad Kording, Douglas James Cook, Gunnar Blohm, and Nikolaus F. Troje. MoVi: A large multi-purpose human motion and video dataset. *PLOS ONE*, 16(6):e0253157, 2021. 13

[10] Selim Gilon, Emily Y Miller, and Scott D Uhlrich. Opencap monocular: 3d human kinematics and musculoskeletal dynamics from a single smartphone video. *arXiv preprint arXiv:2603.24733*, 2026. 3, 5, 6, 14, 16

[11] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4D: Reconstructing and tracking humans with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14783–14794, 2023. 3

[12] Gautier Grouvel, Lena Carcreff, Florent Moissenet, and Stéphane Armand. A dataset of asymptomatic human gait and movements obtained from markers, imus, insoles and force plates. *Scientific Data*, 10(1):180, 2023. 3, 4

[13] Jesse V Jacobs, Diana M Dimitrova, John G Nutt, and Fay B Horak. Can stooped posture explain multidirectional postural instability in patients with parkinson's disease? *Experimental brain research*, 166(1):78–88, 2005. 1

[14] Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. VACE: All-in-one video creation and editing. *arXiv preprint arXiv:2503.07598*, 2025. 3, 11

[15] Mrn P Kadaba, HK Ramakrishnan, and ME Wootten. Measurement of lower extremity kinematics during level walking. *Journal of orthopaedic research*, 8(3):383–392, 1990. 1

[16] Rawal Khirodkar, He Wen, Julieta Martinez, Yuan Dong, Zhaoen Su, and Shunsuke Saito. Sapiens2. In *International Conference on Learning Representations*, pages 58484–58507, 2026. 6

[17] Łukasz Kidziński, Bryan Yang, Jennifer L Hicks, Apoorva Rajagopal, Scott L Delp, and Michael H Schwartz. Deep neural networks enable quantitative movement analysis using single-camera videos. *Nature communications*, 11(1):4054, 2020. 3

[18] Bogdan Kwolek, Agnieszka Michalczuk, Tomasz Krzeszowski, Adam Switonski, Henryk Josinski, and Konrad Wojciechowski. Calibrated and synchronized multi-view video and motion capture dataset for evaluation of gait recognition. *Multimedia Tools and Applications*, 78(22):32437–32465, 2019. 3, 4

[19] Hung Le and Hieu Pham. Learning to estimate critical gait parameters from single-view rgb videos with transformer-based attention network. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2024. 2, 3, 5, 6, 13, 16

[20] Feng Liu, Shiwei Zhang, Xiaofeng Wang, Yujie Wei, Haonan Qiu, Yuzhong Zhao, Yingya Zhang, Qixiang Ye, and Fang Wan. Timestep embedding tells: It's time to cache for video diffusion model. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7353–7363, 2025. 11

[21] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. SMPL: a skinned multi-person linear model. *ACM Transactions on Graphics (TOG)*, 34(6):1–16, 2015. 3

[22] Caroline Malin-Mayor, Vida Adeli, Andrea Sabo, Sergey Noritsyn, Carolina Gorodetsky, Alfonso Fasano, Andrea Iaboni, and Babak Taati. Pose2Gait: Extracting gait features from monocular video of individuals with dementia. In *International Workshop on PRedictive Intelligence In MEdicine*, pages 265–276. Springer, 2023. 3

[23] Soroush Mehraban, Andrea Iaboni, and Babak Taati. FastHMR: Accelerating human mesh recovery via token and layer merging with diffusion decoding. *arXiv preprint arXiv:2510.10868*, 2025. 3, 5, 6

[24] Meg E Morris, Robert Iansek, Thomas A Matyas, and Jeffery J Summers. Stride length regulation in parkinson's disease: normalization strategies and underlying mechanisms. *Brain*, 119(2):551–568, 1996. 1

[25] Lucas Mourot, Ludovic Hoyet, François Le Clerc, and Pierre Hellier. UnderPressure: Deep learning for foot contact detection, ground reaction force estimation and footskate cleanup. In *Computer Graphics Forum*, pages 195–206. Wiley Online Library, 2022. 3, 4, 6, 11, 12

[26] Priyanka Patel and Michael J Black. CameraHMR: Aligning people with perspective. In *2025 International Conference on 3D Vision (3DV)*, pages 1562–1571. IEEE, 2025. 3, 5, 6

[27] Priyanka Patel, Chun-Hao P Huang, Joachim Tesch, David T Hoffmann, Shashank Tripathi, and Michael J Black. Agora: Avatars in geography optimized for regression analysis. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13463–13473. IEEE, 2021. 3

[28] JD Peiffer, Kunal Shah, Irina Djuraskovic, Shawana Anarwala, Kayan Abdou, Rujvee Patel, Prakash Jayabalan, Brenton Pennicooke, and R James Cotton. Portable biomechanics laboratory: clinically accessible movement analysis from a handheld smartphone. *arXiv preprint arXiv:2507.08268*, 2025. 3, 5, 6, 13, 16

[29] David Renggli, Christina Graf, Nikolaos Tachatos, Navrag Singh, Mirko Meboldt, William R Taylor, Lennart Stieglitz, and Marianne Schmid Daners. Wearable inertial measurement units for assessing gait in real-world environments. *Frontiers in physiology*, 11:90, 2020. 1

[30] Alejandro Rodríguez-Molinero, Alexandra Herrero-Larrea, Antonio Miñarro, Leire Narvaiza, César Gálvez-Barrón, Natalia Gonzalo León, Esther Valldosera, Eva de Mingo, Oscar Macho, David Aivar, et al. The spatial parameters of gait and their association with falls, functional decline and death in older adults: a prospective study. *Scientific reports*, 9(1): 8813, 2019. 1

[31] Andrea Sabo, Sina Mehdizadeh, Andrea Iaboni, and Babak Taati. Estimating parkinsonism severity in natural gait videos of older adults with dementia. *IEEE journal of biomedical and health informatics*, 26(5):2288–2298, 2022. 1

[32] Geise Santos, Marcelo Wanderley, Tiago Tavares, and Anderson Rocha. A multi-sensor human gait dataset captured through an optical system and inertial measurement units. *Scientific Data*, 9(1):545, 2022. 3, 4

[33] István Sárándi, Timm Linder, Kai Oliver Arras, and Bastian Leibe. MeTRAbs: metric-scale truncation-robust heatmaps for absolute 3d human pose estimation. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 3(1):16–30, 2020. 13

[34] Céline Schreiber and Florent Moissenet. A multimodal dataset of human gait at different walking speeds established on injury-free adult participants. *Scientific data*, 6(1):111, 2019. 3, 4

[35] Thomas Seel, Jorg Raisch, and Thomas Schauer. Imu-based joint angle measurement for gait analysis. *Sensors*, 14(4): 6891–6909, 2014. 1

[36] Soyong Shin, Juyong Kim, Eni Halilaj, and Michael J Black. WHAM: Reconstructing world-grounded humans with accurate 3d motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2070–2080, 2024. 3, 5, 6, 14

[37] Weijun Tao, Tao Liu, Rencheng Zheng, and Hutian Feng. Gait analysis using wearable sensors. *Sensors*, 12(2):2255–2283, 2012. 1

[38] Gul Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 109–117, 2017. 3

[39] Yufu Wang, Yu Sun, Priyanka Patel, Kostas Daniilidis, Michael J Black, and Muhammed Kocabas. PromptHMR: Promptable human mesh recovery. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1148–1159, 2025. 3, 5, 6

[40] Elke Warmerdam, Clint Hansen, Robbin Romijnders, Markus A Hobert, Julius Welzel, and Walter Maetzler. Full-body mobility data to validate inertial measurement unit algorithms in healthy and neurological cohorts. *Data*, 7(10):136, 2022. 3, 4

[41] Yasunori Yamada, Masatomo Kobayashi, Kaoru Shinkawa, Erhan Bilal, James Liao, Miyuki Nemoto, Miho Ota, Kiyotaka Nemoto, and Tetsuaki Arai. Utility of synthetic musculoskeletal gaits for generalizable healthcare applications. *Nature Communications*, 16(1):6188, 2025. 3

[42] Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. *Advances in neural information processing systems*, 33:5824–5836, 2020. 14

Supplementary Material

SynthGait-19K: A Physically Grounded Synthetic Video Dataset for Gait Parameter Estimation

A. Physically-Grounded Video Generation

A.1. Motion Sources and SMPL Conversion

We obtain walking motion from motion capture (MoCap) datasets. MoCap datasets employ heterogeneous joint placement conventions depending on the capture system. To ensure consistency, each dataset is independently interpolated and converted into a unified joint representation aligned with the SMPL kinematic structure.

A.2. Camera Configuration and Viewpoint Sampling

For each SMPL walking sequence, we render depth videos using three camera configurations. The first configuration places the camera in a frontal or back view relative to the walking direction, where only the camera pitch is varied. The second configuration places the camera in a left or right side view, again varying only the pitch angle. The third configuration samples a random viewpoint by uniformly rotating the camera yaw in the range $[0°, 360°]$.

For all three configurations, the camera pitch is randomly sampled from the range $[-5°, 45°]$, enabling viewpoints ranging from near-horizontal to moderately elevated perspectives.

A.3. Depth Rendering with Planar Ground

Rendering depth maps of the human body alone provides limited environmental context and can lead to undesired zooming and frame-to-frame camera instability during video synthesis. As shown in Figure 6, missing scene geometry can induce spurious changes in camera pitch and scale across frames. To provide a stable geometric reference, we place a synthetic planar ground beneath the walking trajectory during depth rendering. The plane extends along the walking direction with a randomly sampled width.

Since the depth of the planar ground remains static throughout the sequence, this setup implicitly enforces a static camera configuration during RGB video generation. This design significantly improves temporal consistency and prevents camera drift artifacts in the synthesized videos.

A.4. RGB Video Synthesis with VACE

Given a rendered depth video, we synthesize RGB walking videos using the Wan2.1-14B-VACE [14] model. Video generation is conditioned on both the depth sequence and a text prompt describing the subject and environment. We use TeaCache [20] to increase generation speed. The prompt specifies the subject's gender and nationality, along with an environment description sampled from a curated set of indoor and outdoor scenes.

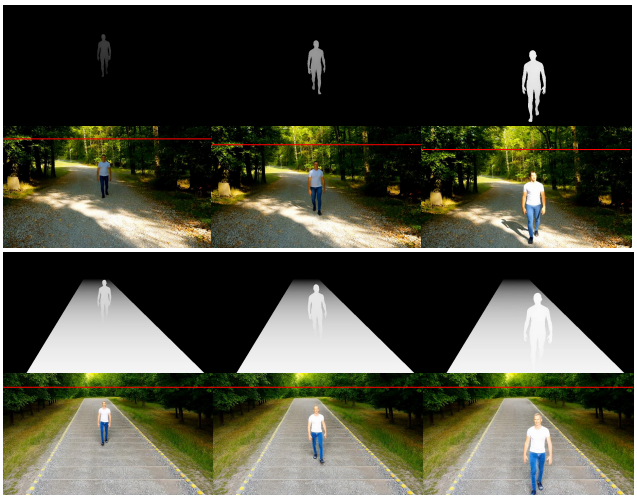


Figure 6. Effect of adding a synthetic ground plane on camera stability, where the red line marks the horizon and shows how missing geometry induces spurious camera pitch and scale changes across frames.

Generation compute. Video synthesis is an offline, one-time dataset construction cost. The production campaign used Wan2.1-VACE-14B at 480×832 resolution with 81 frames, 50 denoising steps, and TeaCache (threshold 0.3) on NVIDIA L40S GPUs. Generation was divided into 48 independent shards and executed in parallel on up to 48 GPUs, requiring approximately 3,131 allocated L40S-hours.

The generation campaign produced a larger candidate pool prior to dataset curation, from which the final 19,273 videos from 6,427 motions comprising SynthGait-19K were retained. Consequently, the reported generation compute also includes samples that were not retained in the final dataset. This cost reflects the particular Wan2.1-VACE-14B configuration used for this release rather than an intrinsic computational requirement of Gait2Vid; the video-generation backbone is modular and can benefit from faster generation models and inference techniques.

A.5. Gait Parameter Extraction

Gait-parameter annotations are computed directly from the fitted SMPL walking sequences. We apply the UnderPressure model [25] to detect heel-strike events from the motion sequence. Using the detected heel strikes together with 3D joint trajectories of the hips, feet, and neck, we compute a set of physically meaningful 3D gait features.

These features serve as supervision signals for the downstream gait estimation task.

Walking Speed. Walking speed measures the average forward velocity of the subject during locomotion. Assuming that walking occurs predominantly along the forward axis, we compute walking speed using the displacement of the pelvis joint between the first and last detected heel strikes, normalized by elapsed time:

$$v = \frac{|z_{\text{pelvis}}(t_{\text{last}}) - z_{\text{pelvis}}(t_{\text{first}})|}{\frac{t_{\text{last}} - t_{\text{first}}}{\text{fps}}}, \tag{1}$$

where $z_{\text{pelvis}}(t)$ denotes the pelvis position along the forward axis at frame t, and fps is the frame rate.

Cadence. Cadence quantifies the temporal rhythm of walking and is defined as the number of steps per minute. Given a sequence of detected heel strikes, cadence is computed as:

$$\text{Cadence} = \frac{(N_{\text{HS}} - 1)}{\frac{t_{\text{last}} - t_{\text{first}}}{\text{fps}}} \times 60, \tag{2}$$

where N_{HS} is the total number of detected heel strikes.

Step Length. Step length represents the forward distance covered between consecutive steps. For each heel strike, we record the forward-axis position of the stepping foot. Step length is computed as:

$$\ell_i = |z_{i+1} - z_i|, \tag{3}$$

where z_i is the forward position of the foot at the i-th heel strike. The reported step length is the mean over all steps:

$$\overline{\ell} = \frac{1}{N-1} \sum_{i=1}^{N-1} \ell_i. \tag{4}$$

Step Width. Step width measures the mediolateral spacing between consecutive steps and reflects balance during walking. At each heel strike, the lateral-axis position of the stepping foot is extracted. Step width is computed as:

$$w_i = |x_{i+1} - x_i|, \tag{5}$$

where x_i denotes the mediolateral foot position at the i-th heel strike. The final step width is given by:

$$\overline{w} = \frac{1}{N-1} \sum_{i=1}^{N-1} w_i. \tag{6}$$

Stooped Posture. Stooped posture characterizes the forward-leaning configuration of the upper body during walking. It is computed as the horizontal distance between the neck and pelvis joints, normalized by leg length:

$$s(t) = \frac{z_{\text{neck}}(t) - z_{\text{pelvis}}(t)}{L_{\text{leg}}}, \tag{7}$$

where L_{leg} denotes the subject's leg length. The reported stooped posture is the temporal mean:

$$\overline{s} = \frac{1}{T} \sum_{t=1}^{T} s(t). \tag{8}$$

Arm Swing. Arm swing captures the amplitude of arm motion along the direction of walking. Joint positions are first centered at the pelvis to remove global translation. For each wrist, the normalized range of forward motion is computed as:

$$\begin{aligned} a_{\text{R}} &= \frac{\max_t z_{\text{rwrist}}(t) - \min_t z_{\text{rwrist}}(t)}{L_{\text{leg}}}, \\ a_{\text{L}} &= \frac{\max_t z_{\text{lwrist}}(t) - \min_t z_{\text{lwrist}}(t)}{L_{\text{leg}}}. \end{aligned} \tag{9}$$

The final arm swing measure is:

$$\overline{a} = \frac{1}{2} \left(a_{\text{R}} + a_{\text{L}} \right). \tag{10}$$

A.6. Force-Platform Validation of Heel-Strike Annotations

We validate the heel-strike events used for gait-parameter extraction against force-platform measurements available in the Bertaux *et al.* [3] dataset. All retained force-platform contacts were recorded at an analog sampling rate of 1,000 Hz. For each vertical-force channel, contact was considered active when $|F_z| \geq 20$ N. Within-plate gaps of at most 0.03 s were bridged, after which contact segments shorter than 0.10 s were discarded. The first above-threshold sample of each retained segment defined the force-platform contact onset. Contacts assigned to the same foot with onset times within 0.25 s were merged, retaining the earliest onset.

All processed SMPL sequences are represented at 30 FPS. Each force-platform onset was therefore mapped to the corresponding processed frame as

$$f_{\text{GRF}} = \text{round}\left(t_{\text{onset}} r_{\text{SMPL}}\right) - f_{\text{offset}}. \tag{11}$$

where t_{onset} is the force-platform onset time, $r_{\text{SMPL}} = 30$ FPS, and f_{offset} denotes the starting-frame offset of the processed sequence.

UnderPressure [25] heel strikes are defined by $0 \to 1$ transitions of its estimated heel-contact signal. To evaluate temporal localization independently of laterality, left- and right-foot events were pooled and each force-platform onset was matched to its nearest UnderPressure event. The signed timing error is defined as

$$e_f = f_{\text{UP}} - f_{\text{GRF}}, \tag{12}$$

such that negative values indicate that UnderPressure detects contact before the force signal crosses the 20 N threshold.

Of 3,245 evaluated trials, 3,053 contain at least one force-platform contact within the processed SMPL interval, yielding 6,092 heel strikes. The resulting mean absolute timing error is 2.305 frames (≈ 76.8 ms), with a median absolute error of 2 frames and an RMSE of 3.045 frames. Overall, 82.3% and 91.7% of heel strikes are localized within 3 and 5 frames, respectively. Because the force platforms instrument only part of the walkway, this evaluation measures timing accuracy for plate-observed contacts rather than recall over the complete walking sequence.

B. Gait Parameter Distribution in SynthGait-19K and GPJATK

Figure 7 compares the distributions of the six gait parameters in SynthGait-19K and GPJATK. Overall, SynthGait-19K provides substantially broader coverage across most gait parameters, while GPJATK forms a smaller and more concentrated evaluation distribution. This is expected since SynthGait-19K is designed to expose the model to diverse walking patterns during training, whereas GPJATK is a real-world evaluation dataset with a limited number of recorded trials.

Importantly, the GPJATK distributions largely fall within regions covered by SynthGait-19K, indicating that the synthetic training set captures gait variations relevant to the real evaluation data. At the same time, SynthGait-19K extends beyond GPJATK in several parameters, such as cadence, step width, stooped posture, and arm swing, which encourages the model to learn a broader representation rather than overfitting to the narrower range of the target evaluation set.

C. Additional GaitXFormer Details

C.1. Architecture and Training

Input processing. Input walking videos are represented by a fixed temporal window of $T = 32$ frames sampled from a 5-second clip. After person-centric cropping, each frame is resized to $H = 256$, $W = 192$. Videos shorter than the required temporal window are zero-padded. All gait parameters are normalized using z-score normalization computed from the SynthGait-19K training set.

Video encoder. Given an input video $X \in \mathbb{R}^{T\times H\times W}$, GaitXFormer first encodes it using a Video Vision Transformer:

$$Z = \mathcal{E}(X) \in \mathbb{R}^{L\times D}, \qquad L = \frac{T}{2}\frac{H}{16}\frac{W}{16}, \tag{13}$$

where D is the embedding dimension and L is the number of spatiotemporal tokens. The encoder is initialized from V-JEPA2 and contains 24 transformer layers with embedding dimension $D = 1024$. It uses a temporal stride of 2 and a spatial patch size of 16×16. All encoder layers are fine-tuned during training on SynthGait-19K.

Gait-query decoder. To obtain parameter-specific representations, we introduce $N = 6$ learnable gait-query tokens,

$$Q = \left\{q^{(i)}\right\}_{i=1}^{N}, \qquad q^{(i)} \in \mathbb{R}^{D}, \tag{14}$$

with one query corresponding to each of cadence, walking speed, step length, step width, stooped posture, and arm swing. The queries attend to the video representation through a three-layer decoder:

$$H = \mathrm{Decoder}(Q, Z) \in \mathbb{R}^{N\times D}. \tag{15}$$

Each decoder layer consists of cross-attention from the gait queries to the encoded video tokens followed by a feed-forward network. We omit self-attention among the gait queries so that each query directly retrieves evidence relevant to its corresponding gait parameter.

Prediction heads. Each decoded representation $h^{(i)}$ is mapped independently to a scalar gait estimate using a parameter-specific linear head:

$$\hat{y}^{(i)} = w_i^{\top} h^{(i)} + b_i, \qquad i = 1, \ldots, N. \tag{16}$$

Stacking the individual predictions gives $\hat{\mathbf{y}} \in \mathbb{R}^N$.

Training objective. We train GaitXFormer end-to-end using mean-squared error over the normalized gait parameters:

$$\mathcal{L}_{\mathrm{MSE}} = \frac{1}{BN}\sum_{b=1}^{B}\sum_{i=1}^{N}\left(\hat{y}_{b,i} - y_{b,i}\right)^2, \tag{17}$$

where B denotes the batch size. We use AdamW with a learning rate of 1×10^{-4} and train for 100 epochs with a global batch size of 16 across four NVIDIA L40S GPUs.

D. Benchmark Method Details

STT. We evaluate both the released STT [19] model and an STT variant trained on SynthGait-19K. The released model was trained on side-view videos from a cerebral-palsy cohort and predicts cadence and walking speed. For STT†, we train the STT architecture from scratch on SynthGait-19K, using six independent regressors corresponding to the six gait parameters. Each model operates on 81-frame, 16-FPS sequences of 25 BODY_25 2D joints extracted using RTMPose. All model parameters are trained using Adam with a learning rate of 6×10^{-4}, a batch size of 128, and normalized mean-squared error, with early stopping based on validation error.

PBL. We evaluate PBL [28] using its released formulation. We do not fine-tune its MeTRAbs pose-estimation component [33] on SynthGait-19K because PBL relies on the dense 87-landmark BML-MoVi convention [9], whereas SynthGait-19K provides SMPL-based motion annotations, and no validated correspondence between these landmark definitions was available.

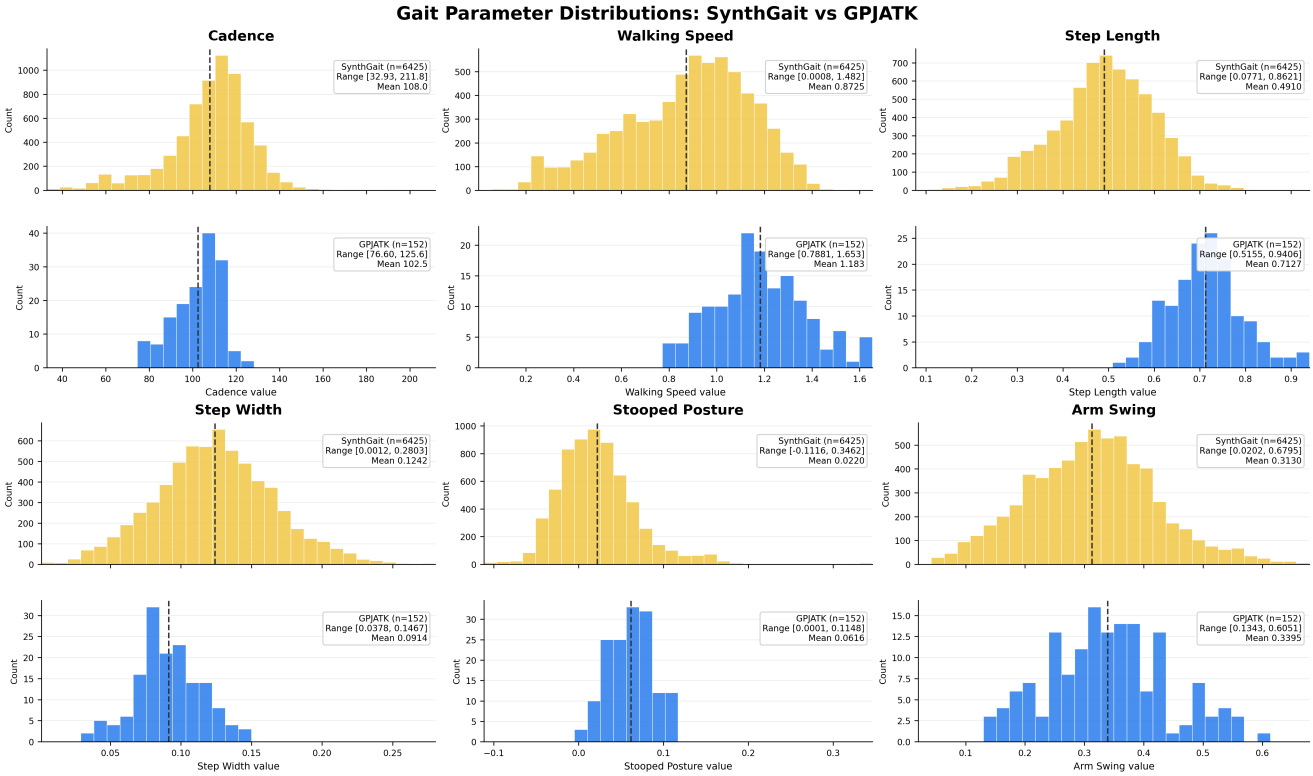


Figure 7. Distribution comparison of gait parameters in SynthGait-19K and GPJATK. For each gait parameter, the top histogram shows the proposed SynthGait-19K training distribution and the bottom histogram shows the GPJATK evaluation distribution. Dashed vertical lines indicate dataset means.

OpenCap Monocular. OpenCap Monocular [10] builds on WHAM [36] as its monocular 3D human-motion estimator and subsequently refines the predicted motion through camera and biomechanical optimization. We evaluate the released OpenCap Monocular pipeline without additional adaptation. In our separate SynthGait-19K adaptation experiment on WHAM, improved SMPL reconstruction produced only a negligible change in downstream gait correlation, suggesting that improving the underlying HMR representation alone does not necessarily translate to improved gait estimation.

WHAM adaptation. To examine whether SynthGait-19K supervision can improve an HMR-based gait pipeline, we adapt the released WHAM [36] model while keeping its RGB feature extractor, motion encoder/decoder, trajectory network, contact prediction, and root-orientation prediction frozen. We attach a lightweight temporal output adapter ($\sim$1.08M trainable parameters) to WHAM's final SMPL predictions. The adapter refines body pose and shape and applies a bounded scale correction to the predicted translation displacement. Training uses released-WHAM predictions from SynthGait-19K RGB videos paired with the corresponding fitted-SMPL motion. Because the constituent motion datasets are highly imbalanced, batches are source-balanced, and PCGrad [42] is used to reduce conflicting pose gradients across motion sources. The training objective supervises SMPL pose, joint and bone geometry, shape, and traveled distance. The adapted model is then evaluated both for SMPL reconstruction on GPJATK-VACE and for downstream gait estimation on real GPJATK using the same gait-extraction protocol as the released WHAM model.

E. Details of the PD4T experiments

E.1. Parkinson's severity estimation

PD4T contains 418 annotated walking trials from 30 participants. Since each trial may contain multiple walking rounds and turns, we extract straight-walking segments, resulting in 1,666 video clips. For this experiment, we remove the GaitXFormer gait-query decoder, freeze the video encoder, and train a lightweight MLP classifier to predict UPDRS severity. As a baseline, we apply the same frozen-encoder and classifier protocol to the pretrained V-JEPA2 encoder. Both representations are evaluated using leave-one-subject-out cross-validation.

Figures 8 and 9 provide additional details on this classification setting. Figure 8 shows that the dataset is fairly bal-

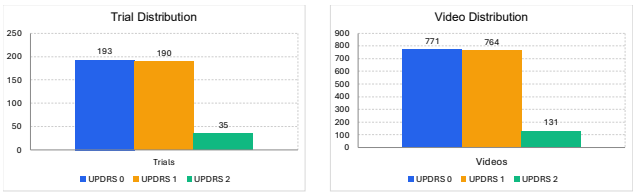


Figure 8. Distribution of PD4T samples across UPDRS severity classes. The left panel shows the number of trials per class, and the right panel shows the number of videos per class. The dataset is relatively balanced for UPDRS 0 and UPDRS 1, while UPDRS 2 is substantially underrepresented.

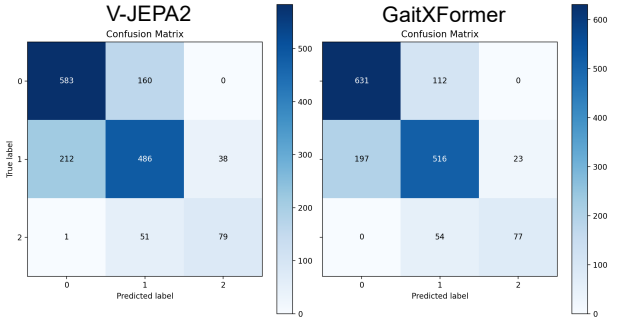


Figure 9. Confusion matrices for PD4T severity classification using frozen V-JEPA2 and GaitXFormer backbones. Rows denote ground-truth labels and columns denote predicted labels. Compared with V-JEPA2, GaitXFormer produces more diagonal predictions, particularly reducing confusion between classes 0 and 1, while most remaining errors occur between adjacent severity classes.

anced between UPDRS 0 and UPDRS 1 at both the trial and video levels, whereas UPDRS 2 has substantially fewer samples. This class imbalance helps explain the trends observed in the confusion matrices in Figure 9, where both models show stronger performance on the more frequent classes and comparatively weaker recognition of the underrepresented UPDRS 2 class. Despite this challenge, GaitXFormer yields a cleaner diagonal structure and reduced confusion between adjacent classes compared with the frozen V-JEPA2 backbone.

E.2. Patient-Aware Clinical-Anchor Evaluation

For each participant, videos with the same clinical score were grouped together before analysis. We refer to each such group as a *patient-score group*; for example, all videos from participant 001 with score 0 form one group, and all videos from participant 001 with score 1 form another group. Gait predictions were averaged within each group before computing correlations with clinical score. This avoids treating multiple videos from the same participant as independent samples.

Confidence intervals were estimated using patient-level cluster bootstrap resampling. In each bootstrap iteration, patients were sampled with replacement, all rows belonging to the sampled patients were retained, and the statistic of interest was recomputed. The reported 95% confidence interval corresponds to the 2.5th and 97.5th percentiles across bootstrap iterations. This preserves the dependency structure among videos and score groups from the same participant.

For the within-patient directional analysis, adjacent clinical-score states from the same participant were compared. For a gait feature f, the change was computed as

$$\Delta f = f_{\text{higher score}} - f_{\text{lower score}}.$$

For features expected to decrease with worsening severity, such as walking speed, step length, and arm swing, the expected sign was set to -1. For stooped posture, which was expected to increase, the expected sign was set to $+1$. A pair was counted as changing in the expected direction when

$$s_f \Delta f > 0,$$

where s_f is the expected sign for feature f. The effect size was computed as the standardized response mean,

$$\text{SRM}_f = \frac{\text{mean}(s_f \Delta f)}{\text{SD}(s_f \Delta f)}.$$

F. Additional Experimental Analysis

F.1. Mean Absolute Error Analysis

Table 11 reports preferred-view MAE in the native units of each gait parameter. The expanded comparison shows that correlation and absolute error capture complementary aspects of performance. STT† achieves particularly low error for cadence and arm swing, while PBL and OpenCap-M obtain the lowest walking-speed and step-length errors, respectively. HMR-based methods remain competitive for stooped posture. Together with the correlation results in the main paper, these results highlight that no single representation is uniformly best across all gait parameters.

F.2. Viewpoint Sensitivity Analysis

To further assess viewpoint sensitivity, we report both a view-averaged comparison across methods and a detailed view-wise breakdown for the HMR-based WHAM pipeline. As shown in Tab. 12, GaitXFormer achieves the strongest overall view-averaged performance, with an average correlation of 0.79 compared with 0.65 for WHAM and 0.60 for PromptHMR. The gains are particularly clear for walking speed and step length, while GaitXFormer also remains competitive for step width. These results suggest that direct RGB-to-gait estimation can be more stable than pipelines that first reconstruct a human mesh for several spatial gait quantities. At the same time, HMR-based methods remain

Table 11. Preferred-view comparison using Mean Absolute Error (MAE). Lower is better. Cadence is reported in steps/min, walking speed in m/s, and step length and step width in meters; stooped posture and arm swing are dimensionless normalized quantities. † denotes STT trained on SynthGait-19K; – indicates unsupported outputs.

Method	Cad	W. Speed	Step Len	Step Wid	Stoop Post	Arm Swing
HMR						
WHAM	7.95	0.24	0.19	0.02	0.05	0.22
CameraHMR	12.14	0.54	0.10	0.03	0.05	0.18
PromptHMR	7.81	0.67	0.11	0.03	**0.03**	0.11
FastHMR	10.41	0.59	0.12	0.03	0.05	0.13
Biomechanical						
PBL [28]	15.66	**0.09**	0.08	0.02	0.04	0.29
OpenCap-M [10]	7.56	0.10	**0.07**	0.04	0.12	0.12
2D Pose-based						
STT [19]	13.99	0.11	–	–	–	–
STT†	**4.21**	0.16	0.14	**0.02**	0.06	**0.05**
Video-based						
GaitXFormer (ours)	9.97	0.22	0.19	0.02	0.08	0.07

Table 12. View-averaged comparison across gait parameters using Pearson correlation (r). For each gait parameter, correlations are first computed separately for each camera view and then averaged across views using Fisher z-transformation. **Avg** denotes the Fisher z-transformed average across gait parameters.

Method	Cad	W. Speed	Step Len	Step Wid	Stoop Post	Arm Swing	Avg
HMR							
WHAM	0.85	0.77	0.34	0.55	0.46	0.72	0.65
CameraHMR	0.71	0.44	-0.16	0.35	0.42	0.88	0.51
PromptHMR	0.87	0.37	-0.20	0.46	**0.64**	0.89	0.60
FastHMR	0.80	0.44	-0.20	0.33	0.12	0.91	0.51
Video-based							
GaitXFormer	**0.94**	**0.88**	**0.66**	**0.58**	0.49	**0.86**	**0.79**

Table 13. Viewpoint-wise Pearson correlation (r, ↑) of WHAM on GPJATK. Numbers in parentheses denote the number of test videos.

Gait Feature	Side (152)	Front (76)	Back (76)	Oblique (304)
Cadence	0.74	0.89	0.89	0.83
Walking Speed	0.83	0.71	0.73	0.78
Step Length	0.40	0.30	0.31	0.35
Step Width	0.29	0.73	0.63	0.45
Stooped Posture	0.64	0.31	0.53	0.29
Arm Swing	0.67	0.46	0.86	0.78

competitive for pose-related quantities such as stooped posture and arm swing, indicating that mesh reconstruction can still provide useful cues for some gait features.

The detailed WHAM results in Tab. 13 further illustrate the parameter-specific effect of viewpoint. Walking speed remains relatively stable across camera directions, whereas step width benefits substantially from frontal and back views. Stooped posture is strongest from the side view, while arm swing performs substantially better from the back and oblique views than from the front. Step length remains comparatively challenging across all viewpoints. These results support the parameter-specific preferred-view protocol used in the main benchmark and show that gait quantities derived through an HMR pipeline can exhibit substantial viewpoint dependence.

Table 14. Effect of decoder design on gait-estimation performance. We report Pearson correlation (r). **Avg** denotes the Fisher z-transformed average.

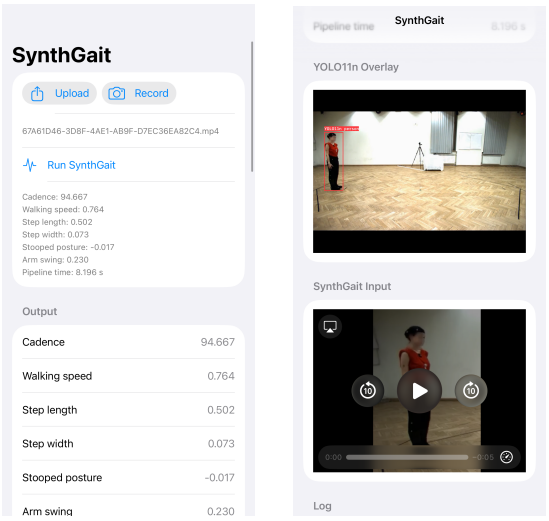
Decoder	Cad	Spd	Len	Wid	Stoop	Arm	Avg
AvgPool	0.92	**0.88**	0.66	0.65	0.74	**0.91**	0.82
Transf. Dec.	0.93	0.87	**0.68**	0.64	0.70	**0.91**	0.82
Cross-Attn.	**0.94**	**0.88**	0.67	**0.67**	**0.75**	**0.91**	**0.84**

Table 15. Effect of encoder training strategy on gait-estimation performance. We report Pearson correlation (r); **Avg** is the Fisher z-transformed average.

Strategy	Cad	Spd	Len	Wid	Stoop	Arm	Avg
Frozen	0.75	0.87	0.68	0.62	0.64	0.88	0.76
LoRA-16	0.88	0.88	**0.72**	0.56	0.68	**0.91**	0.80
LoRA-32	0.88	0.87	0.56	0.53	0.73	0.90	0.81
LoRA-64	0.86	**0.90**	0.69	0.53	0.72	0.89	0.80
Full FT	**0.94**	0.88	0.67	**0.67**	**0.75**	**0.91**	**0.84**

F.3. Additional Ablation Studies

Choice of decoder. Tab. 14 compares different decoder choices for mapping video encoder tokens to gait parameters. For the AvgPool baseline, we average-pool the encoder tokens into a single global representation and use separate linear heads to predict each gait parameter. The transformer decoder uses learnable gait queries with self-attention among queries followed by cross-attention to the encoder tokens, while cross-attention decoder removes query self-attention and directly attends from each gait query to the encoded video representation. Although all three designs achieve similar performance, the cross-attention decoder obtains the best Fisher-averaged correlation and slightly improves several gait parameters, including cadence, step width, and stooped posture. We therefore adopt the cross-attention decoder as the default design, as it provides a modest performance gain while keeping the decoder simple and task-specific.



(a) Prediction interface (b) Tracking and input clip

Figure 10. Smartphone deployment of GaitXFormer. The app tracks the walking subject with YOLO11n and runs GaitXFormer on the cropped clip, completing the full on-device pipeline in 8.2 s.

Table 16. Effect of input length on gait-estimation performance in terms of Pearson correlation (r). Avg denotes the Fisher z-transformed average across gait parameters.

Input	Cad	W. Speed	Step Len	Step Wid	Stoop Post	Arm Swing	Avg
T=16	0.90	**0.91**	**0.72**	0.58	0.70	0.88	0.81
T=32	**0.94**	0.88	0.67	**0.67**	**0.75**	0.91	**0.84**
T=64	0.92	0.88	0.64	0.66	0.71	**0.92**	0.82

Effect of training strategy. Tab. 15 compares different strategies for adapting the V-JEPA2 video encoder to gait estimation. In the frozen setting, the encoder weights are kept fixed and only the decoder and regression heads are trained. For LoRA-based adaptation, we keep the pretrained encoder weights frozen and insert low-rank trainable adapters with ranks 16, 32, and 64. Finally, full fine-tuning updates the entire encoder together with the gait-query decoder and prediction heads. While LoRA improves over the frozen baseline and achieves competitive performance, full fine-tuning obtains the best Fisher-averaged correlation and improves several clinically relevant parameters, including cadence, step width, and stooped posture. We therefore use full fine-tuning as the default training strategy.

Input video length. We additionally study the effect of the number of input frames in Tab. 16. Using $T = 32$ frames achieves the highest overall performance, with a Fisher z-averaged correlation of 0.84. Compared with $T = 16$, the additional temporal context improves several gait parameters, including cadence, step width, and stooped posture. Increasing the input length to $T = 64$ does not provide consistent gains and slightly reduces the overall correlation. We therefore use $T = 32$ frames for the final model.

Table 17. Controlled occlusion robustness on GPJATK. We report Pearson correlation (r) and the Fisher z-transformed average. P denotes persistent occlusion over the full clip, and T denotes transient occlusion over one contiguous 50% segment of the clip. Δ denotes the drop in average correlation relative to clean input. CIs are obtained from 10,000 paired participant-level bootstrap replicates.

Condition	Cad	Spd	Len	Wid	Stoop	Arm	Avg	Δ [95% CI]
Clean	.939	.882	.674	.667	.748	.915	.837	.000
Upper-P	.933	.874	.675	.688	.199	.863	.775	.062 [.038,.089]
Upper-T	.930	.875	.652	.668	.732	.886	.820	.017 [.004,.033]
Lower-P	.918	.874	.600	.269	.782	.898	.789	.049 [.031,.071]
Lower-T	.926	.856	.569	.600	.762	.898	.807	.030 [.019,.047]
Random-P	.925	.863	.578	.487	.572	.781	.748	.089 [.066,.123]
Random-T	.930	.851	.600	.651	.731	.884	.807	.030 [.017,.052]

Figure 11. Controlled spatial occlusion settings. The same GPJATK frame is shown with clean input and persistent upper-body, lower-body, and random person-region occlusion. Each mask covers approximately 25% of the person-centered crop and remains fixed throughout the clip.

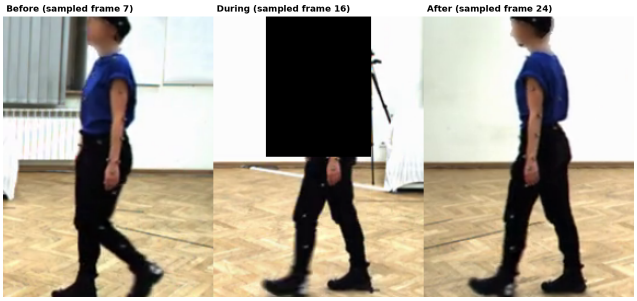


Figure 12. Transient occlusion setting. Example frames from the same GPJATK clip before, during, and after the occluded interval. The mask is applied over one contiguous 50% segment of the sampled frames.

F.4. Controlled Occlusion Robustness

To evaluate robustness to incomplete visual evidence, we apply controlled occlusions to GPJATK videos at inference time without retraining GaitXFormer. Each occlusion covers approximately 25% of the person-centered crop. We consider upper-body, lower-body, and randomly positioned person-region masks, applied either persistently throughout the full clip or transiently over one contiguous 50% temporal

Table 18. Paired participant-bootstrap comparison on GPJATK. Entries report $\Delta r = r_{\mathrm{GXF}} - r_{\mathrm{baseline}}$ with 95% percentile confidence intervals from 10,000 bootstrap replicates. Positive values favor GaitXFormer. Avg denotes the difference in Fisher-z-averaged correlation.

Metric	GXF − STT†	GXF − WHAM
Cadence	$+.025$ $[+.011, +.048]$	$+.094$ $[+.058, +.152]$
Walking speed	$+.033$ $[-.002, +.085]$	$+.056$ $[+.014, +.100]$
Step length	$-.059$ $[-.163, +.019]$	$+.272$ $[+.047, +.454]$
Step width	$+.001$ $[-.098, +.082]$	$-.019$ $[-.145, +.108]$
Stoop posture	$+.044$ $[-.053, +.172]$	$+.104$ $[-.031, +.245]$
Arm swing	$-.007$ $[-.030, +.025]$	$+.247$ $[+.121, +.442]$
Avg	$+.013$ $[-.008, +.035]$	$+.133$ $[+.086, +.183]$

segment. The mask remains spatially fixed within each clip, and all other preprocessing and evaluation settings follow the preferred-view protocol used in the main paper.

As shown in Tab. 17, the effect of occlusion is strongly parameter-dependent. Persistent upper-body occlusion primarily affects stooped posture, whose correlation decreases from 0.75 to 0.20, while persistent lower-body occlusion most strongly affects step width, which decreases from 0.67 to 0.27, and also reduces step-length performance. Random persistent occlusion produces the largest overall degradation, reducing the Fisher z-averaged correlation from 0.84 to 0.75. In all three spatial settings, transient occlusion produces a smaller overall drop than persistent occlusion, indicating that GaitXFormer can partially recover when unoccluded temporal evidence remains available. Paired participant-level bootstrap analysis further shows that the Fisher-averaged degradation is positive for every occlusion condition, with all 95% confidence intervals excluding zero. Figures 11 and 12 illustrate the spatial and temporal masking protocols, respectively.

F.5. Statistical Uncertainty of Headline Comparisons

We additionally quantify uncertainty in the principal GPJATK comparisons using 10,000 paired participant-level bootstrap replicates. Participants are resampled with replacement while all sequences and synchronized views belonging to a participant are kept together. For each gait parameter, methods are compared on their common evaluated samples, while retaining the preferred-view protocol used in the main paper.

Table 18 reports the signed difference in Pearson correlation between GaitXFormer and two representative baselines. Relative to the task-specific STT† model trained on SynthGait-19K, GaitXFormer obtains a slightly higher Fisher-z-averaged correlation ($\Delta = 0.013$), although the 95% confidence interval includes zero, indicating comparable overall performance under matched task supervision. In contrast, GaitXFormer shows a substantially larger average correlation than WHAM ($\Delta = 0.133$, 95% CI $[0.086, 0.183]$). At the individual parameter level, the advantage over WHAM is particularly clear for cadence, walking speed, step length, and arm swing.

G. Mobile Deployment

To evaluate practical deployability, we implement a prototype GaitXFormer mobile application that runs the full RGB-to-gait pipeline on a smartphone. As shown in Fig. 10, the app allows the user to upload or record a walking video, detects the walking subject using YOLO11n, extracts a person-centered input clip, and applies GaitXFormer to estimate cadence, walking speed, step length, step width, stooped posture, and arm swing. The example in Fig. 10, tested on an iPhone 17, demonstrates that the complete pipeline, including detection, preprocessing, and GaitXFormer inference, can be executed on-device in approximately 8 seconds. This result highlights that the proposed direct RGB-to-gait formulation is practical for lightweight mobile deployment.

H. Ethics and Broader Impact

SynthGait-19K is intended to support research on video-based gait analysis and evaluation rather than clinical diagnosis. Although the estimated gait parameters are clinically relevant, predictions from GaitXFormer should not be interpreted as medical assessments without appropriate validation in the target population and acquisition setting. Performance may vary under factors such as severe occlusion, assistive devices, clothing, camera placement, and population characteristics that are not fully represented by the current benchmarks.

Synthetic video provides a way to increase the diversity and scale of gait data without collecting additional identifiable video from human participants. At the same time, synthetic data do not eliminate biases inherited from the underlying motion sources, rendering pipeline, or evaluation datasets. We therefore view SynthGait-19K as a research resource for developing and benchmarking gait-estimation methods, and recommend evaluation on appropriately governed real-world cohorts before deployment in health-related applications.