GAITGen: Disentangled Motion-Pathology Impaired Gait Generative Model – Bringing Motion Generation to the Clinical Domain

Supplementary Material

Supplemental Video. Our model generates gait sequences that depict the movement impairments associated with PD. The supplemental video provides qualitative demonstrations that illustrate these motion dynamics and subtle PD-related features in a way that still images cannot. We strongly recommend that readers view the provided supplemental video.

Further clarification on "Motion" and "Pathology" terms. In our work, "Motion" denotes the *general* gait biomechanics (e.g., alternating heel strikes) or the inherent physiological motion constraints (e.g., limited range of motion in knee flexion and extension). "Pathology", on the other hand, describes disease-specific patterns **absent** in healthy gait (e.g., small, shuffling steps, reduced arm swing, festination, etc.). Our model explicitly separates these factors by learning a distinct pathology latent space, ensuring that pathology-related deviations are captured independently of normal motion dynamics. This disentanglement allows for precise control over gait synthesis, enabling pathology-conditioned generation while preserving natural locomotion patterns.

A. More on Clinical Validity

A.1. Elaboration on Clinical Motivation

Parkinsonian gait analysis has direct clinical relevance, particularly in early detection and continuous monitoring. Yet, the scarcity of annotated data for severe PD stages, where motor symptoms become most pronounced, restricts the ability of machine learning models to generalize [1]. Our synthetic generation approach mitigates this data imbalance by creating realistic, pathology-specific gait samples. This enriched training set ultimately supports more robust detection and staging tools, which could improve clinical workflows and inform therapeutic decisions without requiring extensive additional data collection. This approach could also reduce the burden of collecting large amounts of new clinical data.

A.2. Quantitative Validation of SMPL Estimation for Clinical Gait Analysis

To establish the clinical reliability of SMPL-based representations used in PD-GaM, we validated the outputs of WHAM, the mesh extraction backbone used for our dataset, against an external biomechanical benchmark with ground-truth kinematics. The Toronto Older Adults Gait Archive (TOAGA) [14] was selected for this purpose because it uniquely satisfies three key criteria: (i) provides synchronized RGB video and gold-standard full-body 3D motion data from Xsens MVN Analyze IMUs (from 14 participants), (ii) involves straight-walking trials comparable to PD-GaM, and (iii) features an older-adult cohort relevant to parkinsonian studies.

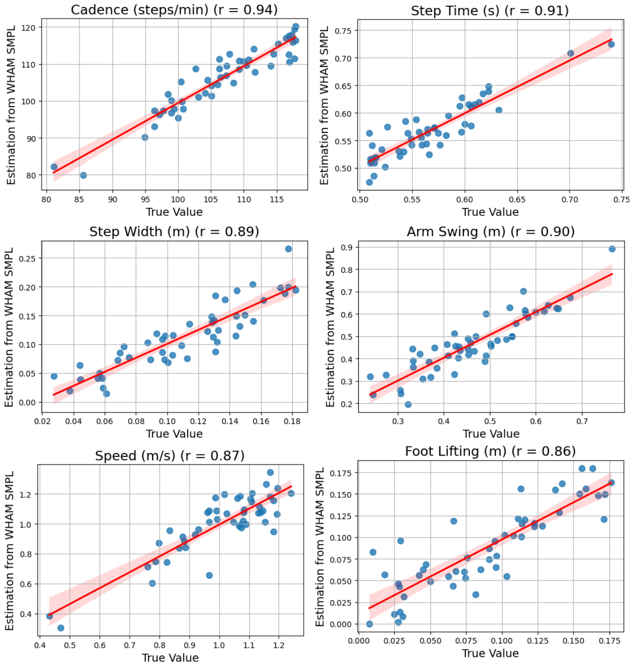

Figure S1. WHAM-derived SMPL estimates vs. Xsens ground truth for six gait features in TOAGA; strong correlations ($r = 0.86 - 0.94$) confirm biomechanical fidelity.

From each TOAGA walk we extracted cadence, walking speed, step time, step width, arm swing amplitude, and foot lifting height from WHAM meshes (features that are clinically meaningful for parkinsonian gait and present in the dataset) and compared them with the same features computed from IMU. Pearson correlations ranged from $r = 0.86$ to $r = 0.94$ (Fig. S1), indicating strong agreement closely matching the high video-vs-IMU correlations reported in the original TOAGA paper; confirming that WHAM preserves clinically relevant gait dynamics.

To assess geometric accuracy, we computed the root-relative MPJPE between WHAM-extracted joints and synchronized Xsens ground truth, obtaining an average error of 39 mm. This falls within the 35–65 mm band reported for multi-view pose estimation systems [3, 5, 10] and the

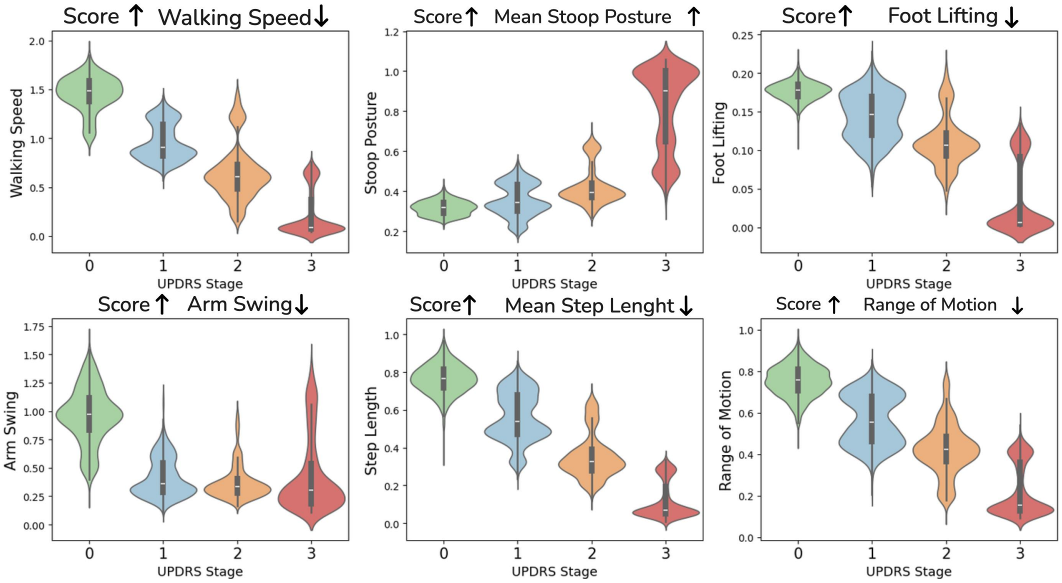

Figure S2. Distributions of gait features from $8k$ synthesized GAITGen sequences across UPDRS stages. Walking speed, step length, foot lifting, arm swing, and ROM decline with severity, with the steepest drop at UPDRS 3. Stoop posture increases sharply at stage 3.

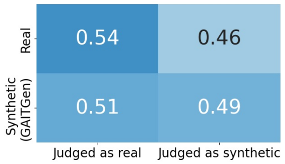

Figure S3. User study (n=6) on distinguising real vs. synthetic data generated by GAITGen .

TULIP dataset for PD gait task [12].

Together, the strong spatiotemporal correlations and low geometric error demonstrate that WHAM-derived SMPL meshes are a reliable surrogate for markerless gait analysis in our study. Nevertheless, while WHAM is used in this study, our method is not tied to it and is expected to benefit from improvements in mesh estimation quality.

A.3. Clinical Gait Features Validation

We extracted six clinically relevant features from $8k$ GAITGen synthetic sequences and plotted their distributions by UPDRS stage (Fig. S2). Walking speed and mean step length consistently decrease with higher UPDRS scores, with a pronounced drop at stage 3, indicating severe motor impairment. Mean stoop posture increases sharply at stage 3, reflecting the stooped posture typical of advanced Parkinsonian gait. Foot lifting and arm swing both decline across stages, again with the steepest reduction at UPDRS 3. Range of motion (ROM) shows a gradual decrease with increasing severity, suggesting reduced joint mobility. The alignment of these trends with established clinical observations [9, 15] supports the validity of GAITGen 's severity-conditioned synthesis. See Sec. J for more details on the gait features calculation.

A.3.1. Cross-dataset Evaluation

To assess the generalizability of GAITGen beyond PD-GaM, we evaluated its impact on two independent datasets: T-SDU-PD [2] and BMClab [21]. Results in Tab. S1 show consistent improvements across three evaluation settings. In the in-domain evaluation (IDE), adding GAITGen samples to the training set improves performance relative to training only on real data, e.g., F_1 scores improve +8 pp on T-SDU-PD and +12 pp on BMClab. In the external validation (EV), training on PD-GaM and testing on the other datasets, GAITGen augmentation raises F_1 by +9 pp on T-SDU-PD and +10 pp on BMClab. Finally, in the target-aware adaptation (TAA), combining real and GAITGen synthetic data across domains yields the best performance overall, achieving F_1 of 0.60 on T-SDU-PD and 0.71 on BMClab. These consistent gains confirm that GAITGen contributes transferable pathological gait patterns that improve severity estimation beyond the source dataset.

Compared to prior SOTA, GAITGen achieves markedly higher F_1 scores (0.60 vs. 0.38 from ST-GCN PD [18] on T-SDU-PD; 0.71 vs. 0.58 from AGIR [24] on BMClab), demonstrating superior performance on both datasets.

A.4. Additional Clinical User Study Results

Our 6 clinical raters completed a tutorial with paired videos/SMPL avatars before blind evaluation. Since clinical evaluation focuses on parkinsonian gait features (not facial or clothing cues) scoring avatar meshes is equivalent to video. For fairness, all real sequences were also rendered

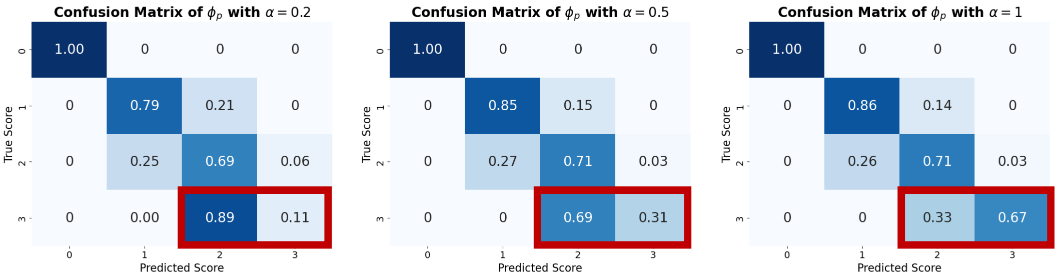

Figure S4. Confusion matrices of the classifier ϕ_p for different values of interference weight α.

Exp.	Train Data	Test Data	Inc. Severe (labels)	F_1	prec.	rec.
IDE	PD-GaM (*no-syn*)	PD-GaM	✓ (0,1,2,3)	0.66	0.66	0.65
IDE	PD-GaM +GAITGen	PD-GaM	✓ (0,1,2,3)	**0.74**	**0.76**	**0.73**
IDE	T-SDU-PD (*no-syn*)	T-SDU-PD	× (0,1,2)	0.49	0.51	0.49
EV	PD-GaM (*no-syn*)	T-SDU-PD	× (0,1,2)	0.44	0.47	0.47
EV	PD-GaM +GAITGen	T-SDU-PD	× (0,1,2)	0.53	0.55	0.52
IDE	T-SDU-PD +GAITGen	T-SDU-PD	× (0,1,2)	0.57	0.56	0.58
TAA	PD-GaM + T-SDU-PD +GAITGen	T-SDU-PD	× (0,1,2)	**0.60**	**0.58**	**0.61**
IDE	BMClab (*no-syn*)	BMClab	× (0,1,2)	0.57	0.59	0.56
EV	PD-GaM (*no-syn*)	BMClab	× (0,1,2)	0.52	0.54	0.52
EV	PD-GaM + GAITGen	BMClab	× (0,1,2)	0.62	0.65	0.61
IDE	BMClab +GAITGen	BMClab	× (0,1,2)	0.69	0.70	0.70
TAA	PD-GaM + BMClab +GAITGen	BMClab	× (0,1,2)	**0.71**	**0.74**	**0.72**

Table S1. Cross-dataset evaluation. In Domain Evaluation (IDE), External Validation(EV), Target-aware Adaptation (TAA). The **best** results per dataset are bold, and second-best are underlined.

to SMPL meshes. The evaluation was fully blind which means raters saw only meshes, without knowing whether a sequence was real or synthetic or how many of each were included. Clinicians easily recognised parkinsonian features from rendered mesh, supported by strong agreement with GT video scores (0.91) and near-chance real vs. synthetic discrimination (precision = 0.52, S3), confirming clinical validity of the mesh avatar format. Fig. S3 shows the confusion matrix for the real vs. synthetic classification task in the clinical user study described in the main manuscript. The near chance-level performance by clinicians highlights the high visual realism of GAITGen 's outputs.

A.5. Interpretation of Interference Weight α

The interference weight α modulates the contribution of pathology-specific features in the reconstructed gait sequences. Experiments with $\alpha = 0.2, 0.5, 1.0$ revealed a trade-off: Lower values of α prioritize motion dynamics over pathology features, slightly improving reconstruction, but reducing classification accuracy of ϕ_p for severe pathology (UPDRS-gait score 3). Fig. S4 presents the confusion matrices of the classifier ϕ_p for different α values. Our final model uses $\alpha = 1.0$ to to achieve better performance on severe cases, while α remains adjustable depending on specific application requirements.

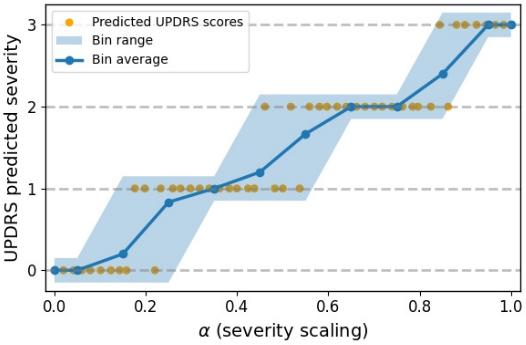

Figure S5. Predicted UPDRS-gait severity as a function of interference weight α. Motion latent ($\mathbf{q}_m$) from a UPDRS 0 sequence were combined with pathology latent ($\mathbf{q}_p$) from a UPDRS 3 sequence. Orange points show discrete classifier predictions for synthetic samples generated by varying α from 0 to 1. Blue shading denotes the min–max prediction range within 0.1-wide bins, and the blue curve indicates the bin-wise mean.

To examine whether pathology is controllably expressed in the latent space, we conducted a mix-and-match experiment in which motion latent ($\mathbf{q}_m$) from a UPDRS 0 sequence were combined with pathology latent ($\mathbf{q}_p$) from a UPDRS 3 sequence and scaled by α. As shown in Fig. S5, the predicted UPDRS-gait severity increases monotonically as α is varied from 0 to 1. At low values of α, predictions remain in the healthy range, while higher values shift decoded motions and predictions toward severe impairment.

To better interpret α, we combined motion latents from ten samples with UPDRS score = 0 sequences with pathology latents from five UPDRS 3 sequences, generating 50 synthetic samples per ranging value of α. Fig. S6 summarizes the resulting gait features. On average, arm swing, step length, and walking speed decrease with increasing α, while stoop posture increases. These feature-level trends align with established Parkinsonian gait characteristics. These results also provide strong evidence that the model achieved disentanglement of motion and pathology, enabling clinically coherent severity modulation.

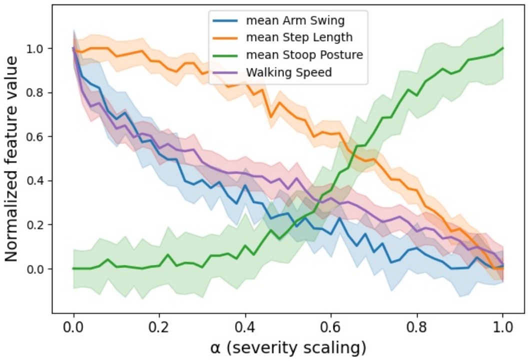

Figure S6. Interpretation of α through four gait features. Lines show mean values and shaded regions denote standard deviations across generated samples.

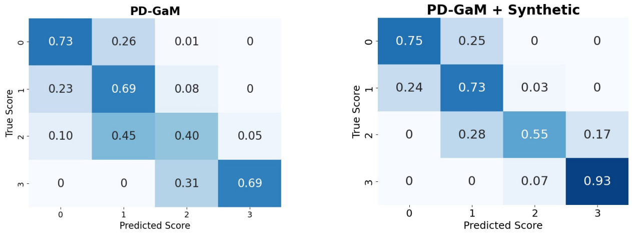

Figure S7. Confusion matrix of downstream classifiers trained on (left) PD-GaM (right) PD-GaM+Synthetic data.

A.6. Downstream Classifier details and Confusion Matrices

For the downstream classification task we used three different pre-trained models and a Random Forest on clinical gait features (Tab. 3 in the paper). Motion embeddings extracted from these encoders are fed into a lightweight trainable classifier head with three fully connected layers (hidden sizes: (128, 64)) to predict UPDRS-gait severity levels. The classifier is trained on the PD-GaM training set along with synthetic data and evaluated *only* on real samples from the PD-GaM test split. The confusion matrices in Fig. S7 compare our primary classifier's (ConvAutoEnc) performance when trained on (left) the PD-GaM training set alone (right) PD-GaM combined with synthetic data, both tested on the same real PD-GaM test set. Adding synthetic data improves UPDRS classification, especially for severe cases (score 3).

B. More on Mix and Match Augmentation.

Mix and Match augmentation is not a naive data augmentation technique. It leverages the decoder's capacity to integrate motion and pathology features. The decoder determines the final output by considering the input motion, the pathology severity level, and their likelihoods, resulting in coherent and realistic gait sequences. For instance, in qualitative results shown in Fig.8 (manuscript), when a severe pathology (score 3) is paired with a motion where the person walks a long distance, the decoder adapts the trajectory to reflect the constraints of severe pathology, which typically limits walking distance. This demonstrates that our method goes beyond naive augmentation by producing contextually appropriate modifications based on the severity level. However, while Mix and Match is effective for augmenting underrepresented classes, it is limited in diversity and control. It serves as a supplementary technique suitable for small-scale augmentation but does not replace the need for training our generative model.

For clarity, GAITGen is purpose-built for *pathology-conditioned* gait generation, isolating pathology indicators from general motion patterns to synthesize gaits that reflect a specific severity level. Modeling subject identity is out of scope, as it shifts focus from pathology modeling to personalization, unnecessary for our data augmentation purpose and could compromise generalization and privacy.

B.1. Alternate Latent Fusion Strategy.

An alternative motion and pathology latent fusion is elementwise multiplication of $\mathbf{q}_m$ and $\mathbf{q}_p$. We choose addition, $\mathcal{D}(\mathbf{q}_m + \alpha \cdot \mathbf{q}_p)$, over elementwise multiplication, $\mathcal{D}(\mathbf{q}_m \odot \mathbf{q}_p)$, for *optimization stability*. Addition behaves as residual learning, where the pathology vector only needs to encode scale-independent corrections. With addition, the zero vector naturally preserves the healthy baseline motion, and small errors from the pathology branch do not distort the underlying motion. In contrast, with multiplication, the network must learn to output $\mathbf{q}_m$ close to one in this case, and even small deviations can disproportionately affect the representation (e.g., if $\mathcal{E}_p$ outputs 0.9 instead of 1, the healthy subcode is reduced by 10% across all dimensions, regardless of whether this change is meaningful). Moreover, gradients under multiplication scale with the magnitude of latents, making optimization unstable and scale-sensitive, whereas in the addition, gradients remain scale independent. This makes addition both more robust and consistent with residual formulations widely used in deep learning architectures.

C. Training with Additional Gait Data.

To test whether healthy gait datasets can compensate for the limited pathological data, we added four external healthy gait datasets [4, 7, 19, 20] (7,971 walks) to the training set. As shown in Tab. S2, this led to only minor gains. Given the relative simplicity of the walking motion, healthy gait is already well-modeled. The main challenge is modeling pathological deviations. Thus, additional normal samples offer limited value and can even bias the model toward UPDRS 0, weakening the pathology latent. To prevent this, we use the extra healthy data only to pretrain the motion encoder $\mathcal{E}_m$, keeping the pathology encoder $\mathcal{E}_p$ trained exclusively on PD gait. These results confirm that future gains will come from more diverse pathological data, not from

	MPJPE ↓	PAMPJPE ↓	ACCL ↓	DS ↑	AVE ↓
GAITGen	28.38	17.91	15.35	1.21	0.19
+ Additional Healthy Gait	28.01	17.43	15.29	1.24	0.18

Table S2. Impact of additional healthy gait.

additional healthy walks.

D. Implementation Details

Model encoders are 1D convolutional ResNet blocks with a temporal downsampling rate of 4. The decoder $\mathcal{D}$ mirrors the encoders' architecture. $\mathcal{E}_m$ is pretrained for 200 epochs (learning rate (lr) $2e^{-6}$) with a reduced lr factor of 0.1 during joint training with $\mathcal{E}_p$ (lr $2e^{-6}$) and the classifiers (lr $2e^{-7}$) for 300 epochs. The loss weights are set as $\lambda_{\text{r}} = 1$, $\lambda_{\text{c}} = \lambda_{\text{adv}} = 0.01$, and $\lambda_{\text{emb}} = 0.02$. We adopt a quantization dropout strategy, with a probability of 0.2. N is set to 6, and both motion and pathology codebook dimensions are 64, with codebook sizes of 512 for motion and 128 for pathology with $\alpha = 1$ for the final model. For the $\mathcal{M}_\theta$ and the $\mathcal{R}_\theta$, we use a lr of $1e^{-3}$ with Adam schedule-free optimizers [6]. The mini-batch size is set to 512 for RVQ-VAE and 256 for the transformers.

VQ-VAE. We employ a 1D convolutional ResNet-based encoder-decoder for gait sequence modeling. The encoder downsamples the input sequence through two convolutional ResNet blocks, reducing the temporal resolution by $4\times$. Initially, a conv layer maps the D-dimensional input motion to $D_m = D_p = 64$ channels, followed by temporal downsampling and quantization into $\#\text{cb}_m = 512$ motion and $\#\text{cb}_p = 128$ pathology codebooks, each with 64-dimensional embeddings. The decoder mirrors the encoder to reconstruct the motion sequence using upsampling ResNet blocks, thereby restoring the original temporal resolution. It also includes one conv layer to recover the original channel size.

Transformers. To ensure diversity in generated sequences, GAITGen integrates residual stochastic sampling techniques, inspired by language modeling approaches such as BERT [11]. During inference, mask tokens are *re-drawn at every sampling step* following a cosine-decay schedule. We apply Top-K filtering and remasking plus Gumbel sampling to enhance diversity. Specifically, in an iterative 10-step process, the top 10% of predicted tokens are retained (based on predicted probabilities), while the remaining 90% are re-masked for refinement. So, the only element that remains constant is the pathology *condition*; neither motion nor pathology tokens are tied to a fixed mask. Therefore, two identical motion seeds are masked differently and converge to distinct completions, ensuring diversity. During training, we introduce additional stochasticity by replacing selected masked tokens with random tokens 8% of the time instead of `<mask>` to enhance robustness. We used CLIP encoding followed by a projection layer for condition, with the four-class c_p functioning similarly to action or style conditioned motion generation [17, 23].

E. Additional Experiments.

Number of Quantization Layer. Tab. S3 shows that increasing the number of quantization layers from 1 to 6 leads to consistent improvements in reconstruction metrics and generation quality (lower AVE). This is because additional layers allow the model to capture finer-grained details of the gait sequences. However, beyond 6 layers, we observe marginal gains in reconstruction and a slight decline in generation performance. This may be due to the increased complexity, making it more challenging for the residual transformer $\mathcal{R}_\theta$ to predict higher-order residuals effectively. Additionally, more layers introduce higher computational costs with diminished disentanglement (lower DS) without significant benefits. Therefore, we selected 6 quantization layers for GAITGen as it offers a good trade-off between reconstruction fidelity and generation quality.

#Quantization	Reconstruction				Generation
layer(s)	MPJPE ↓	PAMPJPE ↓	ACCL ↓	DS ↑	AVE ↓
1	58.42	26.11	25.07	0.77	0.62
2	51.78	24.56	22.43	0.92	0.46
3	43.90	21.87	19.76	1.15	0.47
4	35.77	19.98	17.43	1.02	0.32
5	31.83	18.51	15.67	**1.24**	0.24
6	28.38	17.91	15.35	1.21	**0.19**
7	27.92	17.12	**15.11**	0.95	0.21
8	**27.15**	**16.94**	15.14	0.88	0.25

Table S3. Impact of the number of quantization layers on model performance.

Motion Encoder Reduced Learning Rate Factor. We examine how reducing the learning rate (lr) of the motion encoder $\mathcal{E}_m$ during fine-tuning with the pathology encoder influences disentanglement. As shown in Tab. S4, using an lr reduction factor of 0.1 increases the DS from 0.94 to 1.21. With an lr factor of 1 (no reduction), the motion encoder updates quickly under the influence of the adversarial classifier, causing it to unlearn essential motion features, which degrades reconstruction quality. To compensate, the pathology encoder begins capturing motion information, compromising disentanglement and reducing the DS. Conversely, very small lr factors (0.01 and 0.001) lead to insufficient adjustments by the motion encoder, leaving residual pathology information in the motion latent space and resulting in a lower DS.

Number of Inference Iterations and Masking Scheduler. Here we investigate how the number of inference iterations and the choice of scheduler influence the performance of the Mask Transformer $\mathcal{M}_\theta$. During iterative refinement,

$\mathcal{E}_m$ **lr factor**	**MPJPE** ↓	**PAMPJPE** ↓	**ACCL** ↓	**DS** ↑
1	28.19	17.84	15.31	0.94
0.1	28.38	17.91	15.35	**1.21**
0.01	28.71	18.23	17.01	1.11
0.001	29.94	19.29	17.06	0.86

Table S4. Impact of learning rate reduction factors for the motion encoder $\mathcal{E}_m$ during fine-tuning. Best results are in **bold**, comparable results are underlined.

the masking ratio $\beta(t)$ determines the proportion of tokens to re-mask at each iteration, with t representing the normalized timestep from 0 to 1 over R iterations (i.e., $t = \frac{\text{current iteration}}{R}$). Two schedulers are compared: the linear scheduler ($\beta(t) = 1 - t$) decreases the masking ratio uniformly, while the cosine scheduler ($\beta(t) = \cos\left(\frac{\pi t}{2}\right)$) maintains a higher masking ratio initially, gradually reducing it in later iterations. This higher masking ratio allows the cosine scheduler to focus on easier-to-predict tokens first and progressively tackle harder tokens in later iterations. During training, t is sampled uniformly ($t \sim \mathcal{U}(0, 1)$), and during inference, t is deterministically stepped from 0 to 1 over R iterations. At each iteration, the number of tokens to be masked is calculated as $\beta(t) \times M$, where $M = 2T' + 1$ is the total number of tokens in the sequence. Tab. S5 shows that the cosine scheduler consistently outperforms the linear one, especially at fewer iterations, indicating faster convergence.

#Iterations (Scheduler)	**AVE** ↓	**AAMD** ↓	**ASMD** ↓	**Div** →
5 (linear)	0.367	0.165	0.071	3.744
10 (linear)	0.294	0.142	**0.049**	3.906
20 (linear)	**0.217**	**0.103**	0.051	**3.911**
5 (cosine)	0.313	0.134	0.066	3.824
10 (cosine)	**0.194**	**0.096**	0.048	**3.966**
20 (cosine)	**0.194**	0.099	**0.047**	3.957

Table S5. Impact of inference iterations and scheduler type of the mask transformer model.

VAE Channel Size. The channel size in the VAE determines the number of feature channels in intermediate layers, influencing model's representational capacity. Experiments with varying channel sizes (Tab. S6) show a channel size of 512 achieves the best overall performance, outperforming smaller channel sizes. Increasing the size to 1024 marginally improves ACCL but comes at the cost of increased computational complexity without meaningful benefits to other metrics.

Findings on Codebook Learning. We conducted an experiment to examine the effects of increasing the codebook commitment loss weight (λ_{emb}) and incorporating periodic codebook reset strategy [25] in the quantization process. Raising λ_{emb} from 0.02 to 0.09 resulted in stronger commitment of input vectors to specific codebook entries. However, this came at the cost of worse reconstruction quality and severity prediction performance from pathology latents $\mathbf{q}_p$. A higher λ_{emb} caused the model to commit to a subset of codebook entries early, leaving others unused. We also experimented with periodic codebook resets, where unused codebook entries are reset to random values every 20 iterations instead of at every iteration, giving the codebooks more opportunity to be used before being reset. However, we found that the introduction of new, randomly initialized entries after multiple iterations disrupted the model's established encoding patterns, leading to instability and degraded performance. In contrast, when using a lower λ_{emb} (0.02) with codebook resets at every iteration, the model's commitment to codebook entries was less rigid, allowing the model to adapt more smoothly to the resets without significant disruption, resulting in the most optimal performance.

$\mathcal{E}_m$ **#Channel**	**MPJPE** ↓	**PAMPJPE** ↓	**ACCL** ↓	**DS** ↑
128	58.31	27.65	25.98	0.76
256	39.08	21.78	19.97	0.94
512	**28.38**	**17.91**	15.35	**1.21**
1024	29.12	17.95	**15.23**	1.14

Table S6. Impact of VAE channel size on model performance.

F. Latent space visualization

We present UMAP visualizations of the latent space in Fig. S8. GAITGen achieves well-clustered latent representations aligned with UPDRS-gait scores. Without condition ($\mathcal{E}_p$) or disentanglement, the clusters displays significant overlap among classes.

G. Datasets

G.1. PD-GaM.

PD-GaM is a large PD Gait 3D Mesh dataset derived from the PD4T dataset; but it is anonymized for public release. The original PD4T dataset included 426 gait video recordings from 30 individuals with PD along with per walk UPDRS-scores. Each participant in PD4T was asked to walk forward and backward twice, resulting in 4 walking segments per participant and 1701 walk (≈3 hours) in total. Each recording was reviewed and segmented to retain usable gait segments while excluding frames involving turns. Turns present a unique challenge for individuals with PD, as the gait changes observed during turns differ significantly from those seen during walking. To maintain a consistent representation of gait patterns, only walking sequences are included in this paper. However, we plan to also release the turn segments in PD-GaM to support research on turn-specific gait analysis. SMPL parameters for each subject were extracted at 25 FPS using the WHAM [22]. The camera, positioned at eye level and following the subjects as

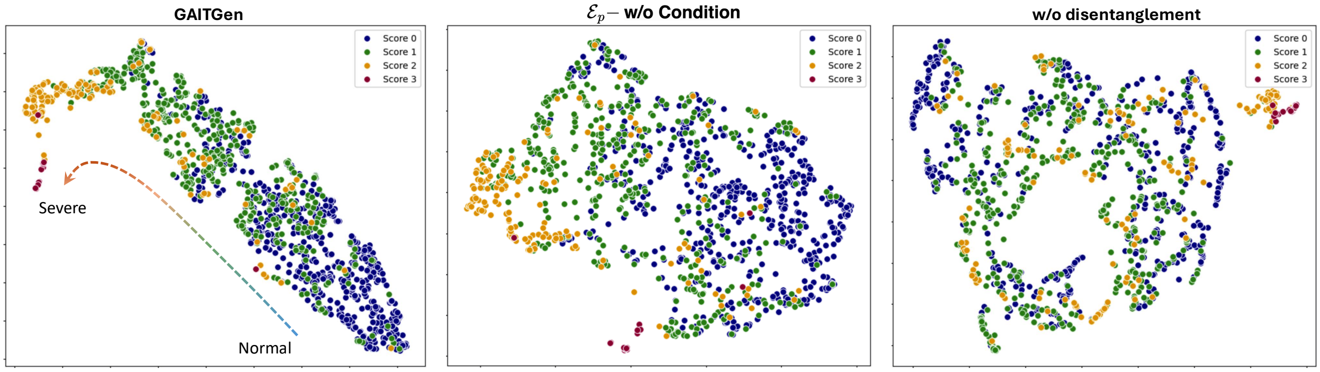

Figure S8. UMAP visualizations of the latent space representations under different settings. The left panel shows GAITGen with well-clustered latent representations aligned with UPDRS-gait scores. The middle panel represents the latent space when $\mathcal{E}_p$ is unconditional, exhibiting overlap among classes. The right panel shows the results without disentanglement, further highlighting increased overlap and less separation of latent clusters.

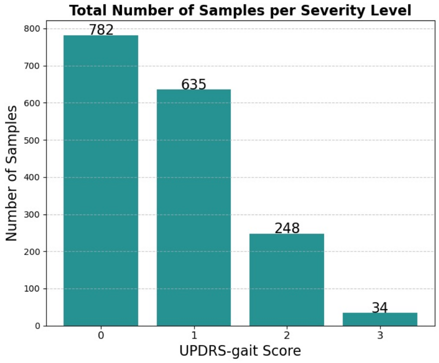

Figure S9. Histogram of UPDRS-gait scores in our PD-GaM dataset.

they walked in original videos, introduced minor distortions in global trajectories. These distortions were corrected during preprocessing to minimize global position artifacts. The dataset is divided into training (1253 samples) and test (448 samples) sets using a participant-based split to ensure unbiased evaluation. A histogram of UPDRS-gait scores in PD-GaM is presented in Fig. S9, highlighting the imbalance in class representation, particularly the limited number of samples in the severe category (class 3). This scarcity emphasizes the importance of our Mix and Match augmentation strategy.

G.2. External Datasets.

The T-SDU-PD dataset [18] contains parkinsonian gait samples from 14 participants, each annotated per walk with UPDRS-gait scores by expert clinicians. Recordings were captured using a ceiling-mounted camera, and 3D meshes were extracted with WHAM [22]. The dataset includes 381 walking trials, totaling approximately 50 minutes.

The BMClab dataset [21] was collected with a Raptor-4 optical motion-capture system (Motion Analysis Corp.) using 44 reflective markers. It contains 781 walking trials (48 minutes) from PD participants, with MDS–UPDRS gait labels assigned at the participant level by expert clinicians. 3D meshes were obtained using SparseFusion optimization [26] to fit SMPL parameters from sparse 3D joints.

H. Input Motion Representation

Our motion representation follows the HumanML3D format [8]. Each frame is encoded as a 263-dimensional vector derived from the SMPL mesh representation of WHAM [22], with joints mapped to 22 AMASS joints [13]. This representation includes root angular velocity ($r^a \in \mathbb{R}$) along the Y-axis, root linear velocities ($r^x, r^z \in \mathbb{R}$) on the XZ-plane, root height ($r^y \in \mathbb{R}$), local joint positions ($j^p \in \mathbb{R}^{3N_j}$), joint velocities ($j^v \in \mathbb{R}^{3N_j}$), joint rotations ($j^r \in \mathbb{R}^{6N_j}$), and binary foot-ground contact features ($c^f \in \mathbb{R}^4$). This redundant representation captures both kinematic and dynamic properties, making it suitable for neural models, particularly generative frameworks. It also allows for the disentanglement of motion dynamics from pathology-specific characteristics.

I. Metric Details

Here we provide details of the AVE, AAMD, and ASMD metrics. All reported generation results are averaged over 10 repetitions.

Average Variance Error (AVE) measures how closely the variance of local joint positions in the generated poses matches the ground-truth variance. For each joint j, the variance $\sigma[j]$ is computed as:

$$\sigma[j] = \frac{1}{T-1} \sum_{t \in T} \left(P_t[j] - \bar{P}[j] \right)^2 \tag{1}$$

where $\bar{P}[j]$ is the mean position for joint j across T frames.

The AVE for each joint is then defined as:

$$\text{AVE}[j] = \frac{1}{N} \sum_{n \in N} \|\sigma[j] - \hat{\sigma}[j]\|_2 \tag{2}$$

where $\sigma[j]$ and $\hat{\sigma}[j]$ are the variances of joint j in the ground-truth and generated poses, respectively. The overall AVE is obtained by averaging the per-joint AVE values.

Absolute Arm Swing Mean Difference (AAMD) measures how closely the generated gait sequences replicate the arm swing ranges of the ground-truth data across different severity classes. For each class c, we calculate the mean arm swing range for the ground-truth ($\overline{AS}_{gt}^{(c)}$) and the generated ($\overline{AS}_{gen}^{(c)}$) sequences. The arm swing range for each sequence is computed by measuring the Euclidean distances between the wrist and shoulder joints at each time step, finding the maximum and minimum distances over the sequence, normalizing by leg length, and selecting the minimum arm swing between the two arms. The AAMD is then defined as the average of the absolute differences between these class-wise mean arm swing ranges:

$$\text{AAMD} = \frac{1}{C} \sum_{c=1}^{C} \left| \overline{AS}_{\text{gen}}^{(c)} - \overline{AS}_{\text{gt}}^{(c)} \right| \tag{3}$$

where C is the total number of classes. A lower AAMD indicates that the generated sequences closely match the real arm swing patterns.

Absolute Stooped Posture Mean Difference (ASMD) quantifies how well the generated gait sequences replicate the stooped posture characteristic across different classes. Similar to AAMD, for each class, we compute the mean stooped posture for the ground-truth ($\overline{SP}_{\text{gt}}^{(c)}$) and generated ($\overline{SP}_{\text{gen}}^{(c)}$) sequences by calculating the vertical distance between the neck and sacrum joints at each time step for each sequence, averaging these distances over all frames, and normalizing by leg length. The ASMD is defined as the average of the absolute differences between these class-wise mean stooped postures:

$$\text{ASMD} = \frac{1}{C} \sum_{c=1}^{C} \left| \overline{SP}_{\text{gen}}^{(c)} - \overline{SP}_{\text{gt}}^{(c)} \right| \tag{4}$$

J. Gait Feature Extraction Details

We derive six clinically relevant features (i.e., Walking Speed, Mean Step Length, Arm Swing, Foot Lifting, Mean Stoop Posture, and Lower-Limb Range of Motion (ROM)) from generated SMPL joints to create an interpretable baseline for UPDRS-gait classification. Following [16], we first detected heelstrike frames by locating alternating peaks in the ankle-to-ankle distance (≤ 8 frames apart, prominence ≤ 0.02). Using these events and the normalized pose by leg length, we compute:

- Walking Speed: total sacrum displacement between first and last heel strike, divided by total time.
- Mean Step Length: Average ankle distance along walking axis between heel strikes.
- Arm Swing: horizontal displacement of the hand joints along the forward axis (sacrum-centered).
- Foot Lifting: vertical range of ankle movement.
- Stoop Posture: measured as the forward-lean distance which is the vertical displacement between neck and sacrum, projected onto the direction of walk.
- Range of Motion: maximum joint displacement, defined as the largest difference between a joint's maximum and minimum positions over time across all joints and axes.

All features are extracted from sequences aligned to a canonical (z-forward) frame.

K. Limitations

The current dataset scoring is based on gait segments that also include turning task, which may influence the UPDRS-gait scores. To address this, we are preparing revised scores focused exclusively on walking segments and will release both formats to accommodate different research objectives. Additionally, the dataset exhibits class imbalance, particularly in severe pathology cases with limited samples. Although our augmentation strategy helped mitigate this issue, integrating additional PD datasets standardized to the PD-GaM format could enhance generalizability and diversity of the generated sequences. Finally, while our model is tailored to PD gait patterns, its applicability to other abnormalities remains an area for future exploration.

References

[1] Vida Adeli, Soroush Mehraban, Irene Ballester, Yasamin Zarghami, Andrea Sabo, Andrea Iaboni, and Babak Taati. Benchmarking skeleton-based motion encoder models for clinical applications: Estimating parkinson's disease severity in walking sequences. In *2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–10. IEEE, 2024. 1

[2] Vida Adeli, Ivan Klabucar, Javad Rajabi, Benjamin Filtjens, Soroush Mehraban, Diwei Wang, Hyewon Seo, Trung-Hieu Hoang, Minh N Do, Candice Muller, et al. CARE-PD: A multi-site anonymized clinical dataset for parkinson's disease gait assessment. In *Advances in Neural Information Processing Systems*, 2025. 2

[3] Andrea Avogaro, Federico Cunico, Bodo Rosenhahn, and Francesco Setti. Markerless human pose estimation for biomedical applications: a survey. *Frontiers in Computer Science*, 5:1153160, 2023. 1

[4] 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. 4

[5] José Carrasco-Plaza and Mauricio Cerda. Evaluation of human pose estimation in 3d with monocular camera for clinical application. In *International Symposium on Intelligent Computing Systems*, pages 121–134. Springer, 2022. 1

[6] Aaron Defazio, Xingyu Alice Yang, Harsh Mehta, Konstantin Mishchenko, Ahmed Khaled, and Ashok Cutkosky. The road less scheduled. *arXiv preprint arXiv:2405.15682*, 2024. 5

[7] 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. 4

[8] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5152–5161, 2022. 7

[9] Joseph Jankovic. Parkinson's disease: clinical features and diagnosis. *Journal of neurology, neurosurgery & psychiatry*, 79(4):368–376, 2008. 2

[10] Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social motion capture. In *Proceedings of the IEEE international conference on computer vision*, pages 3334–3342, 2015. 1

[11] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, page 2. Minneapolis, Minnesota, 2019. 5

[12] Kyungdo Kim, Sihan Lyu, Sneha Mantri, and Timothy W Dunn. TULIP: Multi-camera 3d precision assessment of parkinson's disease. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22551–22562, 2024. 2

[13] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5442–5451, 2019. 7

[14] Sina Mehdizadeh, Hoda Nabavi, Andrea Sabo, Twinkle Arora, Andrea Iaboni, and Babak Taati. The toronto older adults gait archive: video and 3d inertial motion capture data of older adults' walking. *Scientific data*, 9(1):398, 2022. 1

[15] Anat Mirelman, Hagar Bernad-Elazari, Avner Thaler, Eytan Giladi-Yacobi, Tanya Gurevich, Mali Gana-Weisz, Rachel Saunders-Pullman, Deborah Raymond, Nancy Doan, Susan B Bressman, et al. Arm swing as a potential new prodromal marker of parkinson's disease. *Movement Disorders*, 31(10):1527–1534, 2016. 2

[16] Kimberley-Dale Ng, Sina Mehdizadeh, Andrea Iaboni, Avril Mansfield, Alastair Flint, and Babak Taati. Measuring gait variables using computer vision to assess mobility and fall risk in older adults with dementia. *IEEE journal of translational engineering in health and medicine*, 8:1–9, 2020. 8

[17] Mathis Petrovich, Michael J Black, and Gül Varol. Action-conditioned 3d human motion synthesis with transformer vae. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10985–10995, 2021. 5

[18] 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. 2, 7

[19] 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. 4

[20] 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. 4

[21] Thiago Kenzo Fujioka Shida, Thaisy Moraes Costa, Claudia Eunice Neves de Oliveira, Renata de Castro Treza, Sandy Mikie Hondo, Emanuele Los Angeles, Claudionor Bernardo, Luana dos Santos de Oliveira, Margarete de Jesus Carvalho, and Daniel Boari Coelho. A public data set of walking full-body kinematics and kinetics in individuals with parkinson's disease. *Frontiers in Neuroscience*, 17: 992585, 2023. 2, 7

[22] Soyong Shin, Juyong Kim, Eni Halilaj, and Michael J. Black. WHAM: Reconstructing world-grounded humans with accurate 3D motion. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2024. 6, 7

[23] Wenfeng Song, Xingliang Jin, Shuai Li, Chenglizhao Chen, Aimin Hao, Xia Hou, Ning Li, and Hong Qin. Arbitrary motion style transfer with multi-condition motion latent diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 821–830, 2024. 5

[24] Diwei Wang, CÃŠdric Bobenrieth, and Hyewon Seo. Agir: Assessing 3d gait impairment with reasoning based on llms. *arXiv preprint arXiv:2503.18141*, 2025. 2

[25] Will Williams, Sam Ringer, Tom Ash, David MacLeod, Jamie Dougherty, and John Hughes. Hierarchical quantized autoencoders. *Advances in Neural Information Processing Systems*, 33:4524–4535, 2020. 6

[26] Xinxin Zuo, Sen Wang, Jiangbin Zheng, Weiwei Yu, Minglun Gong, Ruigang Yang, and Li Cheng. SparseFusion: Dynamic human avatar modeling from sparse rgbd images. *IEEE Transactions on Multimedia*, 23:1617–1629, 2020. 7