DOI: 10.1111/cgf.70334

EUROGRAPHICS 2026 / B. Masia and J. Thies
(Guest Editors)

COMPUTER GRAPHICS *forum*
Volume 45 (2026), Number 2

ContactVision: Learning Foot Contact from Video for Physically Plausible Gait Animation

Daeyong Kim[1], Gyuseok Yi[2], Ri Yu[†1,2]

[1]Dept. of Artificial Intelligence, Ajou University, South Korea
[2]Dept. of Software and Computer Engineering, Ajou University, South Korea

Abstract

Foot-ground contact information plays a crucial role in character animation and gait analysis, as it helps accurately simulating realistic movement patterns and understanding the biomechanics of walking. Existing motion datasets do not explicitly include foot-ground contact information, requiring separate computation or manual annotation. Obtaining accurate foot-ground contact information typically requires additional sensors such as pressure mats or force plates. Without such devices, estimating contact becomes a highly challenging task. We propose ContactVision, a deep learning framework that detects heel and toe contact states directly from video. Our network is trained in a supervised manner using contact labels derived from motion capture data via ground reaction force estimation. This enables training on existing datasets without the need for additional hardware. We demonstrate the utility of our contact detection network in two downstream tasks: gait motion reconstruction and gait analysis. For animation, we incorporate predicted contact labels into a reinforcement learning framework with a two-segment foot model, enabling realistic foot articulation behavior. For analysis, we estimate clinically relevant gait parameters such as double and single support times, and validate the accuracy against pressure sensor mat data and prior video-based methods. Our results show competitive performance in both animation and analysis settings. The code is publicly available at github.com/DaeeYong/ContactVision.

Keywords: foot-ground contact detection, video processing, motion reconstruction, gait analysis

CCS Concepts
• **Computing methodologies** → **Motion processing; Animation;**

1. Introduction

Foot-ground contact detection plays a crucial role in various aspects of motion understanding, including character animation and gait analysis. Accurate foot-ground contact information is essential for reconstructing realistic human motion and preventing motion artifacts, such as foot skating or ground penetration. Gait analysis is an important method for assessing movement disorders commonly found in neurological conditions such as Parkinson's disease, cerebral palsy, and stroke, helping to evaluate walking patterns and assist in developing treatment plans. Normally, to determine the contact state, clinical facilities use pressure sensor mats, force plates, or insole shoe sensors. Although the contact data obtained from these sensors is accurate, a drawback is that it can only be collected in controlled environments with embedded sensor equipment and under the supervision of trained professionals. This makes it difficult to obtain gait information in everyday settings. Most existing human motion datasets have no explicit foot-ground contact information. Therefore, researchers have had to resort to heuristic methods, such as calculating velocity from the foot joint position in kinematic motion data to determine contact information, or alternatively, they have had to endure the laborious task of manual annotation. These methods are not only time consuming and costly, but also unlikely to achieve high accuracy.

Recently, efforts have been made to estimate ground reaction force (GRF) from existing kinematics-only motion datasets. These studies have proposed deep learning-based GRF estimation networks using proprietary datasets that simultaneously collect motion capture and GRF data. To collect GRF information, pressure insole systems or force plates have been used. GroundLink [HST*23] is a public dataset that collects both GRF and motion capture data. This paper trains GroundLinkNet using the GroundLink dataset. GroundLinkNet is a network designed to estimate the GRF from kinematic motion. However, a limitation of this network is that it predicts only one GRF per foot, meaning that it cannot provide separate contact information for the heel and toe. Mourot et al. [MHCH22] proposed a heuristic method for obtaining foot contact labels from the estimated GRF, which is predicted by a network

† Corresponding author

trained on UnderPressure, a dataset collected by the authors. This method has the drawback of not directly inferring contact labels from the input, but instead requires additional calculations through the estimated GRF, making it less efficient.

While there have been studies aimed at contact estimation from kinematic motion capture (mocap) data, research has also been conducted to obtain contact information directly from video inputs. Rempe et al. [RGH*20] focused on detecting foot-ground contact from video footage, particularly videos capturing dynamic movements such as dance or sports. Although the study successfully derived contact information from dynamic movement videos, its performance degraded when used for common locomotion tasks like walking or running. As a result, it was found to be inadequate for medical applications such as gait analysis in patients with movement disorders.

In this paper, we propose ContactVision, a novel framework for detecting foot-ground contact from monocular video of daily movements, particularly gait, that can serve as an effective tool for gait analysis.

For training, we use an existing human motion dataset [ENW*24], which contains various synchronized and calibrated data for the same motions, including 9-camera color videos, 3D motion capture data, force plate measurements, and photogrammetry scans. Among these, we utilize only the RGB video data and the corresponding 3D motion capture data. Because the dataset contains recordings from multiple viewpoints, our model is trained on a wide variety of perspectives. This multi-view training enables our model to robustly detect foot contact states even in gait videos captured by dynamic cameras with continuously changing viewpoints. The construction of the training dataset is divided into two parts: pose extraction and contact label extraction. To estimate poses of the human skeleton, we use OpenPose [CHS*19], the pre-trained human pose estimation (HPE) model, which takes each video frame as input. We extract contact labels from motion capture data. These contact labels are used as ground truth when training the contact detection network. The motion capture data comprises 3D marker data that represent sequential human movement. Using an existing framework [MHCH22], we extract foot-ground contact information (heel and toe) from the motion capture data. Once the training data is prepared, we train the contact detection network, a deep learning model capable of detecting the contact state. We adopt a network based on the transformer [VSP*17] architecture.

To validate the effectiveness of our contact detection model, we apply it to two downstream tasks: gait motion reconstruction and gait parameter estimation. These tasks demonstrate the broad applicability of our method in both animation and clinical analysis. For gait motion reconstruction, we employ a two-segment foot model and generate appropriate reference motions for the toe joint between the segments using our ToeRefEstimator, enabling realistic foot articulation. Specifically, we show how foot-ground contact information can be used to (1) reconstruct more realistic walking motions from video, and (2) identify gait cycles and estimate gait parameters for gait analysis purposes. We further validate our framework's accuracy by estimating critical gait parameters—such as stance phase ratio, swing phase ratio, single support ratio, and double support ratio—and comparing them against sensor-based measurements. The high accuracy achieved highlights the potential of our method for applications in medical gait analysis and movement disorder assessment.

The main contribution of this paper can be stated as follows:

- We propose an end-to-end contact detection network from video input, which can accurately detect the contact labels for both the toe and heel.
- Our model demonstrates robustness to viewpoint changes in gait videos captured with dynamic cameras, achieved by a training strategy involving multi-view data and random view sampling.
- We introduce a two-segment foot model and the ToeRefEstimator to reconstruct realistic foot articulation, improving motion realism in gait reconstruction.
- Our model allows more accurate gait parameter estimation than existing video-based approaches [RGH*20], with results closer to sensor mat measurements.

2. RELATED WORK

Foot-ground contact detection plays a critical role in character animation and gait analysis by enhancing realism and supporting structured motion understanding. This section first reviews existing methodologies for acquiring contact information, and then examines their applications in the domains of character animation and gait analysis.

2.1. Foot Contact Acquisition

Heuristic Approach The high cost and environmental constraints of kinetic sensor-based data collection methods are significant. As an alternative, foot-ground contact is often inferred from motion capture using several common approaches. One major approach is heuristic thresholding. Various methods are employed to obtain foot-ground contact information from public datasets [KPLK15, IPOS13, Lab03, MRC*17]. Typically, it is inferred from motion capture by using threshold-based heuristics applied to joint positions, velocities, and heights. Manual annotation is also commonly performed. Threshold-based approaches determine foot-ground contact by classifying it when the foot's velocity or the distance between the foot and the ground falls below a predefined threshold [ZYC*20a, LCR*02], or by assuming a zero-velocity condition at foot-ground contact [ZYC*20b]. Manual frame-by-frame labeling of foot-ground contact is also commonly employed in character animation research [LS99, KG03, KSG02a, LSC*19]. Heuristic threshold-based methods are highly sensitive to subtle noise and measurement errors in motion capture data, and slight variations in threshold settings can significantly affect the quality of contact labels. Although manual annotation can yield relatively high accuracy, it requires substantial cost, labor, and time, and relies on subjective judgment, which can lead to inconsistent label quality across experts.

Deep Learning-based Prediction While heuristic approaches are cost-effective compared to building new datasets, they suffer

from a lack of consistency in foot-ground contact labeling. This is because the thresholds are set manually, leading to significant variability between operators. Consequently, recent studies have proposed methods that directly learn contact patterns from observational data using deep learning, thereby enabling the consistent and reproducible generation of foot-ground contact labels. Using pressure insoles aligned with motion capture, Mourot et al. [MHCH22] train a supervised deep learning model for foot-ground contact prediction, while Han et al. [HST*23] achieve the same goal using foot-ground contact information derived from force plate measurements. Advances in computer vision have enabled video-based human pose estimation [CHS*19, SGX*21, SAA*20, KSW*25, LXC*21] and several works attempt to infer contact directly from video. However, studies such as Zou et al. [ZYC*20a] and Rempe et al. [RGH*20] still rely on threshold-based heuristics to construct training labels, which leads to label noise and operator dependence.

In this work, we propose a deep learning-based framework that requires only a single RGB monocular video as input. Following the procedure of Mourot et al. [MHCH22], we first derive reliable ground-truth foot-ground contact labels from the motion capture data, and then train a deep learning model for foot-ground contact by combining these labels with 2D poses estimated from video. This approach reduces the heuristic-label dependence common to purely video-based methods and yields more consistent training labels. Our video-based approach can be uniformly applied to public datasets to extract consistent, high-quality foot-ground contact information without expensive equipment. This enables large-scale foot-ground contact data acquisition and extends applicability to various fields such as sports science, medicine, and character animation.

2.2. Applications of Foot Contact Information

Character Animation Foot-ground contact information, which defines the physical interaction between a character and the ground, serves as one of the most fundamental cues in character animation. One of its most common uses is the correction of motion artifacts such as foot-skating, where the feet slide across the ground instead of remaining fixed during contact. Many studies [ZYC*20b, KSG02b, LDZ*24] leverage this information to fix foot positions and thereby prevent sliding. Song et al. [SJL*24] proposed a framework for changing motion style while preserving motion content, in which foot contact information is explicitly used to prevent foot-skating in the generated motion. Mourot et al. [MHCH22] adopted a post-processing approach to correct already generated motion, enhancing physical realism by using vertical Ground Reaction Force (vGRF) as a constraint in their Inverse Kinematics (IK) technique. Additionally, research in areas such as motion generation [TPXL24], character control [CSH*24], motion synthesis [ZLHA24], and motion reconstruction [ZYC*20b] also utilizes foot-ground contact information to address the foot-skating issue. In addition to correcting artifacts, foot-ground contact information provides essential temporal structure that supports coherent and consistent motion representation. Kim et al. [KEY*24] demonstrated that gait cycle features derived from foot contact can be effectively used for motion retrieval and interactive locomotion style control. Their work highlights that foot contact information not only facilitates the search for suitable motions in large databases but also enables intuitive manipulation of locomotion styles. Foot-ground contact information actively contributes to the creation of realistic motion by ensuring stability in physics-based simulation [KLVDP20, XLKVDP20, WL19, YTL18, ALX*19]. For instance, Yu et al. [YPL21] proposed a DRL-based framework that reconstructs human motion from monocular video, where foot contact information is incorporated as a reward term. Similarly, Shimada et al. [SGX*21] reconstruct 3D motion from a monocular video and utilize foot-ground contact labels as a clue to estimate an accurate pose.

Gait Analysis Research on gait analysis has been actively conducted in diverse fields such as healthcare and computer vision, playing an important role in applications including clinical diagnosis, biometric recognition, and motion analysis. Previous studies have utilized equipment such as motion capture systems, force plates, and pressure sensors to perform gait analysis [JJLK14, MKD*20, RFCA10, EKA*17, BBS*08]. While these sensor-based approaches provide high accuracy, they are limited by high costs and dependency on specialized equipment.

Recently, with advances in computer vision techniques for human pose estimation, various video-based gait analysis studies have been conducted [LZD*22, JP20, HGZ*23]. Åberg et al. [ÅOÅ*21] utilized heel keypoints obtained from OpenPose as auxiliary information and performed visual inspection to analyze gait cycles, while Cimorelli et al. [CPKC24] validated a video-based gait analysis system for prosthesis users. These studies demonstrate that foot-ground contact information plays a crucial role in video-based gait analysis. Therefore, accurate estimation of foot-ground contact is critical for reliable video-based gait analysis. To address this challenge, we propose a robust approach that operates under challenging real-world conditions, including dynamic viewpoints.

3. Overview

We present a framework to detect foot-ground contact states from RGB monocular videos, which enables comprehensive and accurate motion analysis. Our framework consists of two main components: training data preparation and network training. To train a network that estimates foot contact states from videos, a dataset consisting of frame-wise 2D human joint positions in the image paired with corresponding toe and heel contact state labels is required. First, we constructed a dataset by extracting the necessary information from an existing human motion dataset [ENW*24]. To obtain accurate foot-ground contact state labels, we leveraged a previous study [MHCH22] that estimates ground reaction forces and contact labels from motion capture data. 2D human joint positions are extracted using the pose estimation technique (OpenPose [CHS*19]). Next, we train a foot-ground contact detection network using a transformer-based encoder. We use lower body joint positions obtained by applying pose estimation to video, along with the corresponding contact labels, as training data to perform supervised learning. The trained model then estimates the contact states of heels and toes for each frame of the video.

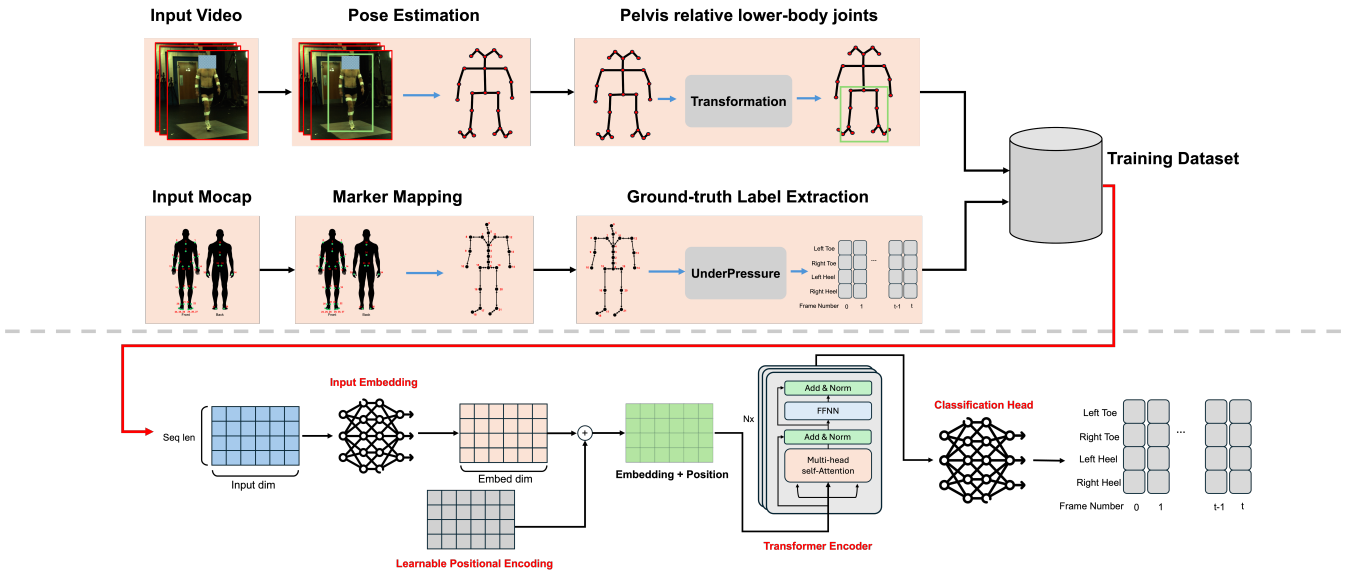

Figure 1: *Overview of our framework, including dataset preparation and model architecture. We use the BioCV dataset [ENW*24], which provides synchronized video and motion capture data, to construct the training dataset. From the video, 2D joint positions are extracted using OpenPose [CHS*19], and 13 lower-body joints are converted into pelvis-relative coordinates to serve as input features. From the motion capture data, foot-ground contact labels are obtained via marker mapping and the UnderPressure framework [MHCH22], and used as ground-truth supervision during training. Given a motion sequence of shape* (seq_len × input_dim)*, the network first projects it into an embedding space through an input embedding layer, followed by the addition of learnable positional encodings. The resulting representation is processed by a Transformer encoder, and the output is passed to a classification head, which predicts binary foot-ground contact labels for four keypoints: the left toe, right toe, left heel, and right heel.*

4. Training Data Preparation

Since our goal is to accurately detect foot-ground contact from videos of daily activities such as walking, we needed motion data that included synchronized video and accurate foot-ground contact information. However, such datasets are extremely rare. To address this, we utilized the BioCV dataset [ENW*24], which contains synchronized video and motion capture data of daily human activities, and estimated the foot-ground contact labels from the motion capture data to construct the training dataset. In the following subsections, we first describe the specifications of the selected dataset [ENW*24], and then explain how we constructed the training data by extracting the subject's skeletal keypoints as input features and generating foot-ground contact labels as ground truth. See Figure 1 (top) for the dataset preparation pipeline.

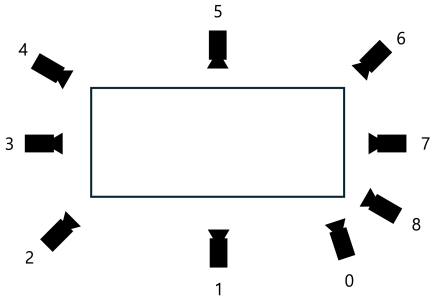

Figure 2: *Illustration of the camera viewpoints used in the BioCV dataset [ENW*24] employed for training.*

4.1. Data Specification

The original dataset [ENW*24] comprises recordings of 15 healthy adults (8 males and 7 females), each performing five distinct movement tasks: walking, running, countermovement jumps (both maximal and sub-maximal effort), and hopping, with up to ten repetitions per activity. Each trial is available in multiple synchronized modalities, including nine calibrated camera views (Figure 2), motion capture data, and photogrammetry scan data. In this study, we utilize only the video and motion capture data. Video streams and motion capture sequences were originally recorded at 200 fps and subsequently downsampled to 100 fps by selecting every even-indexed frame to meet the requirements of our processing pipeline. Camera 2 was excluded due to poor video quality, resulting in eight usable camera views. After preprocessing and applying quality-control filtering, data from 13 subjects were retained for experiments. The processed dataset contains a total of 1,277,093 frames, which were partitioned into training, validation, and test splits following a 70%/15%/15% ratio: 887,847 frames (148.0 min) for training, 196,833 frames (32.8 min) for validation, and 192,413 frames (32.1 min) for testing.

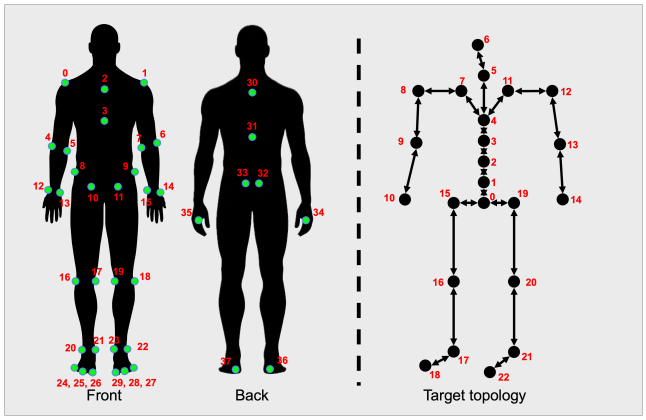

Figure 3: *Mapping of motion capture markers from BioCV [ENW*24] (left) to UnderPressure [MHCH22] (right). Marker indices are shown in the figure; full index-to-name mapping is in Table 4, Appendix A.*

4.2. Keypoint Extraction

To extract the 2D keypoint positions forming the human skeleton from video, we used OpenPose [CHS*19], an off-the-shelf human pose estimation model. Given a gait video, OpenPose estimates human poses for each frame, producing 25 body keypoints per frame based on the BODY_25 model. Each keypoint is represented by 2D pixel coordinates along with a confidence score, forming a pose vector $X_t \in \mathbb{R}^{J\times 3}$, where J denotes the number of keypoints. For our task, we focused on the 13 lower-body joints ($J = 13$) that are related to foot-ground contact, including the pelvis, hips, knees, ankles, heels, and toes.

The coordinates of estimated keypoints vary depending on the subject's position within the video frame, even for identical movements. Additionally, differences in video resolution or variations in the distance between the subject and the camera further affect these coordinate values. Directly inputting raw keypoint data without positional adjustments introduces inconsistencies, hindering the model's ability to generalize across varying subject positions, video resolutions, and camera distances. To address this issue, we converted the keypoint coordinates into relative coordinates based on the pelvis joint. By using the pelvis as a reference point, this transformation preserves the spatial relationships among keypoints, ensuring consistency regardless of changes in the subject's position. Consequently, the input data maintains robustness against the positional shifts during training. The Equation 1 defines how each joint coordinate is transformed relative to the pelvis coordinate. This relative coordinate transformation ensures that, regardless of the subject's position within the frame, the spatial relationships among keypoints are preserved, thereby improving the network's robustness to scale and positional variations during training.

4.3. Contact Labeling

Human motion datasets typically encompass a variety of types such as video sequences and 3D marker trajectories. Most of these datasets do not include explicit labels indicating physical contact between the feet and the ground. Consequently, many approaches have relied on threshold-based heuristics for contact detection by comparing marker displacement or velocity against threshold values or on manual frame-by-frame annotation. However, such methods are fundamentally limited by empirically defined decision criteria that do not generalize well and by the extensive time and effort required for manual annotation. To ensure consistent label quality and reduce the cost associated with manual annotation, this study employs the existing framework UnderPressure. This framework estimates binary foot-ground contact labels from 3D human motion capture data. While various studies have explored inferring foot-ground contact from motion capture data, most provide only indirect indicators such as ground reaction forces (GRFs), or if contact labels are provided directly, they typically offer a single contact state per foot. In contrast, UnderPressure generates more detailed information by providing independent contact states for both the heel and toe of each foot. Since fine-grained contact information is essential for analyzing diverse gait parameters, we chose to adopt the UnderPressure framework in our study. However, the topology of the motion dataset [ENW*24] differs from that required by UnderPressure [MHCH22], as shown in Figure 3. Therefore, to apply the motion data to UnderPressure, we first performed marker mapping to align the topology of the BioCV [ENW*24] to match the one used in UnderPressure [MHCH22]. A detailed description of the marker mapping process is provided in the appendix A. After marker mapping, the transformed motion capture data are fed into the GRF estimation model proposed by UnderPressure [MHCH22] to obtain the estimated GRF. The method for deriving heel and toe contact labels from these GRF estimates is described in the referenced paper [MHCH22], and we used this method to generate the final contact labels serving as ground truth for our contact estimation network.

5. Contact Detection Network

Using the training dataset constructed in the previous chapter, this chapter describes the training procedure for the contact detection network. Our proposed network takes poses estimated from each frame of the video as input and outputs binary contact labels for both the heels and toes in every frame.

5.1. Network Architecture

Figure 1 illustrates the architecture of our contact detection network. Our model comprises four main components: an input embedding, a learnable positional encoding, a Transformer encoder, and a classification head. The input $X \in \mathbb{R}^{T\times J\times 3}$ is a sequence of per-frame joint features, where T is the number of frames and $J = 13$ is the number of joints we used. We employ OpenPose [CHS*19] to extract 2D joint predictions for each video frame. Specifically, OpenPose outputs its pixel coordinates (x, y) and a confidence score $c \in [0, 1]$ for each joint. Concatenating these triplets for 13 lower-body joints (Mid hip, Right hip, Right knee, Right ankle, Left hip, Left knee, Left ankle, Left big-toe, Left small-toe, Left heel, Right big-toe, Right small-toe, Right heel) yields a per-frame feature vector of length 39, producing an input tensor X for the network. We then embed this vector ($input\_dim = 13 \times 3$) at each time step into a d-dimensional representation via

a fully connected layer. This allows the model to learn an optimal representation of the raw joint data before temporal processing. To inject explicit temporal order information, we add a learnable positional encoding. This allows the network to adaptively capture both absolute and relative timing signals suited to our foot-contact estimation. The resulting embeddings are passed through 7 stacked Transformer encoder blocks. Finally, a classification head implemented as a fully connected layer maps each encoder output to four logits (Left Toe, Right Toe, Left Heel, Right Heel). These logits are passed through a sigmoid activation function to produce contact probabilities, and a threshold of 0.5 is applied to classify each probability as either 1 (contact) or 0 (no contact).

5.2. Network Training

We randomly sample training subsequences from the motion sequences obtained with OpenPose [CHS*19]. Each frame is represented by a feature vector $x_t^{(v)} \in \mathbb{R}^{39}$, where t indexes the time step and v denotes the camera view. The 39 dimensions correspond to the pelvis-relative 2D coordinates of 13 lower-body joints expressed as:

$$x_{t,j}^{(v)} = p_{t,j}^{(v)} - p_{t,\text{pelvis}}^{(v)}, \quad j \in J \tag{1}$$

with J denoting the set of 13 lower-body joints (see Section 4.2) and $p_{t,j}^{(v)}$ the original 2D position of joint j at time t in view v. During the creation of the train/validation/test splits, samples from the available views are randomly assigned to each subset and remain fixed throughout training. From a selected view, a random starting frame s is chosen, and a contiguous subsequence of length L is extracted:

$$\tilde{X}^{(v)} = \{x_s^{(v)}, x_{s+1}^{(v)}, \dots x_{s+L-1}^{(v)}\} \tag{2}$$

If the number of frames after s is insufficient, zero-padding is applied so that the input sequence always has length L.

The model outputs frame-wise contact probabilities for four contact keypoints (Left toe, Right toe, Left heel, Right heel), and the training is formulated as an independent multi-label binary classification problem. The objective function is the binary cross-entropy loss, which minimizes the discrepancy between the predicted probabilities $\hat{y}_{t,c}$ and the ground-truth labels $y_{t,c} \in \{0,1\}$, averaged over all frames and contact channels:

$$\mathcal{L}_{\text{BCE}} = \frac{1}{L \cdot 4} \sum_{t=1}^{L} \sum_{c=1}^{4} \Big[-y_{t,c} \log \hat{y}_{t,c} - (1 - y_{t,c}) \log(1 - \hat{y}_{t,c}) \Big], \tag{3}$$

where L denotes the sequence length (we set $L = 128$ in our experiments), $t \in \{1, \dots, L\}$ indexes the time steps, and $c \in \{1, \dots, 4\}$ indexes the classes, with $c = 1$, 2, 3, and 4 corresponding to the left toe, right toe, left heel, and right heel, respectively.

Model parameters are optimized using the Adam optimizer, and early stopping is applied when the validation performance does not improve for a fixed number of epochs, in order to prevent overfitting. In addition, the Optuna framework [ASY*19] is employed to efficiently search for the optimal hyperparameters such as learning rate, embedding dimension, and dropout rate.

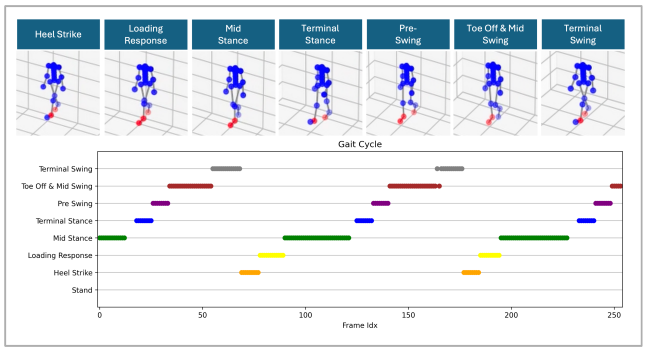

Figure 4: *Prediction results on the validation set of the BioCV dataset. The top row illustrates foot-ground contact during a gait cycle, where red indicates foot-ground contact. The bottom row shows the predicted cycle phase for each frame.*

6. Results

6.1. Implementation Details

Architecture The proposed model is based on a Transformer encoder architecture. Each input frame is represented as a 39-dimensional feature vector, which is projected into a 128-dimensional embedding space through a fully connected layer. Learnable positional embeddings are then added to form the input sequence. The encoder consists of seven stacked layers, each comprising a multi-head self-attention mechanism with four heads and a feed-forward network with a hidden dimension of 512.

Training Strategy The model was implemented in PyTorch and trained, validated, and tested on an NVIDIA GeForce RTX 3080 GPU with 12 GB of VRAM. We employed the Adam optimizer with a batch size of 8 and binary cross-entropy as the loss function. To identify optimal hyperparameters, we conducted 500 hyperparameter optimization trials using the Optuna framework [ASY*19], ultimately selecting the highest F1 score on the validation set. The selected hyperparameters were fixed as follows: a learning rate of 4×10^{-4}, a dropout rate of 9×10^{-3}, seven Transformer layers, an embedding dimension of 128, four attention heads, a feed-forward hidden dimension of 512, a batch size of 8, and early stopping with a patience of 10 epochs was applied. The full hyperparameter search process required approximately two days to complete.

6.2. Performance Evaluation

We evaluated our model on the held-out test split. It achieved an F1 score of 0.93, with a precision of 0.89 and a recall of 0.94, demonstrating reliable discrimination of foot–ground contact events across diverse movement types. Figure 4 illustrates the validation results of foot-ground contact detection during gait.

Our model also performs robustly on gait videos captured with dynamic cameras, accurately detecting foot contact states despite continuous viewpoint changes. This robustness is attributed to our training strategy. First, each motion in the training dataset was recorded simultaneously from 8 cameras placed at diverse angles

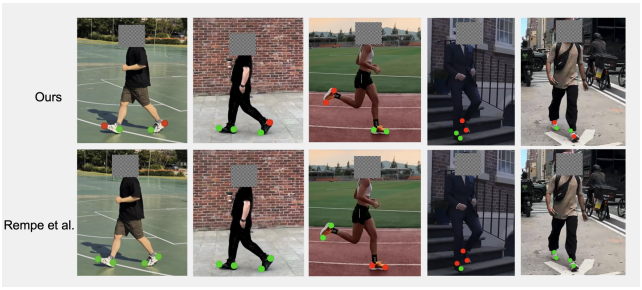

Figure 5: *Qualitative comparison of foot contact estimation under dynamic viewpoints. Baseline results (bottom) show unstable and noisy contact labels when the camera is shaking, while our method (top) produces temporally consistent and physically plausible contacts. Green indicates foot-ground contact and red indicates no contact.*

around the subject (see Figure 2). Second, during training, views were randomly sampled at each iteration, forcing the model to observe the same motion from various perspectives. This setup encourages the model to learn view-invariant features, resulting in strong generalization to unconstrained video conditions. While most existing studies assume static camera viewpoints, real-world gait videos are often captured with handheld devices, introducing camera shake and dynamic perspectives. To evaluate performance under such realistic conditions, we conducted qualitative experiments on walking sequences recorded with moving cameras. Predicted contact states were overlaid on video frames and compared with those from a baseline model. As shown in Figure 5, Rempe et al. [RGH*20] produce unstable and inconsistent labels under camera motion, whereas our model maintains generally accurate and temporally coherent predictions. While the predicted contact occasionally occurs a few frames earlier than the actual contact, likely due to temporal quantization resulting from the lower frame rate of the input video, overall, our model demonstrates strong temporal consistency. These results demonstrate that our method generalizes well to real-world video settings and is robust to viewpoint variation. Additional comparisons are provided in the accompanying demo video.

In the following sections, we demonstrate the applicability of the contact detection network in two downstream tasks: walking motion reconstruction (Section 6.4) and gait analysis (Section 6.5).

6.3. Ablation Study

Table 1: *Ablation study results on positional encoding (PE), joint selection for model input, and subsequence sampling.*

Ablation variant	F1 Score	Precision	Recall
Sinusoidal PE	0.87	0.87	0.89
Full body joints	0.86	0.86	0.88
Sliding window	0.88	0.89	0.89
Ours	**0.93**	**0.89**	**0.94**

We conduct ablation studies to analyze the design choices in our framework, namely positional encoding, joint selection for model input, and subsequence sampling during training. For all ablation experiments, the training, validation, and test splits, as well as the training protocols and hyperparameters, are kept identical to those of the proposed model.

Positional Encoding We evaluate the impact of positional encoding (PE) by replacing the learnable PEs in the proposed model with fixed sinusoidal PEs as introduced in the original Transformer architecture [VSP*17]. As shown in Table 1, this modification leads to a noticeable performance degradation, reducing the F1 score from 0.93 to 0.87, with corresponding decreases in precision and recall. This result indicates that fixed PEs are insufficient for capturing precise contact timing of foot–ground contact events. In contrast, learnable PEs allow the model to adaptively encode temporal structure, leading to more accurate contact detection.

Joint Selection We study the effect of input joint selection by replacing the proposed lower-body joint input with a full-body representation that includes all OpenPose joints [CHS*19]. As shown in Table 1, using the full-body joint representation degrades performance, reducing the F1 score from 0.93 to 0.86. This result indicates that focusing on lower body joints is more effective for foot-ground contact detection than using all available joints.

Subsequence Sampling To examine the impact of the subsequence sampling strategy, we replace the random subsequence sampling described in Section 5.2 with a sliding window approach using a stride of 1. In this setting, training sequences are extracted densely by sliding a fixed length window over the motion sequence, instead of sampling subsequences at random starting frames. As reported in Table 1, this window-based training strategy results in a lower F1 score of 0.88 compared to 0.93 achieved by the proposed training scheme. This result indicates that the random subsequence sampling strategy used in our method is more effective.

6.4. Application 1: Motion Reconstruction

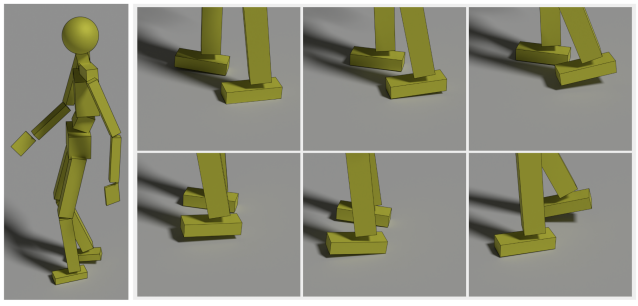

Figure 6: *Walking motion reconstructed using a one-segment foot model. This results in an unnatural stomping gait, where the foot lands flatly on the ground without the gradual heel-to-toe sequence.*

Contact information between feet and ground is important for realistic human motion reconstruction. To showcase the applicability

of our framework, we apply our foot contact labels extracted from our framework to the motion reconstruction framework. Previous motion reconstruction researches normally use single contact information for each foot because the skeleton has a one-segment foot model. In this case, when the character moves, stomping is observed instead of smooth and articulated foot motion (see Figure 6). To achieve more realistic walking motion reconstruction, we employ a two-segment foot model that divides the foot into toe and heel segments, as opposed to the conventional one-segment foot model commonly used in previous motion reconstruction studies. Our walking motion is reconstructed on this model by integrating physics-based simulation with deep reinforcement learning. More specifically, we slightly revise the motion reconstruction framework proposed by Yu et al. [YPL21] to apply our foot contact labels to the two segment foot model.

We first introduce our proposed two-segment foot model. Then, we describe ToeRefEstimator, an algorithm designed to estimate the reference motion of the joint connecting the toe and heel segments, based on the foot contact labels predicted by our contact network. Next, we explain how we leverage deep reinforcement learning to imitate the reference motion and contact patterns extracted from gait videos, and detail the reward design used during training.

Two-segment foot model We designed the two-segment model to mimic actual human foot motion by dividing the foot into heel and toe segments as shown in Figure 7. The key design decision is the location of the joint separating these segments, which should correspond to the bending point of the foot during walking. Based on anthropometric analysis of over 1.2 million feet across North America, Europe, and Asia [JŽD19], this joint is set at approximately 65–80% of the foot length from the heel. Accordingly, we set the heel-to-toe length ratio to 8:2. The joint connecting the two segments is modeled as a single-degree-of-freedom revolute joint.

ToeRefEstimator: Target Angle Generation from Contact Labels Our foot contact detection model provides contact information (contact labels) indicating the contact states of the heel and toe on both feet. However, this information alone does not provide explicit target angles necessary for controlling the two-segment foot model. Without such explicit target angles, the reinforcement learning agent must rely solely on indirect rewards, making it challenging to learn complex foot behaviors efficiently and stably. Empirically, when pose estimation does not supply explicit reference motions for the toe joint—often represented as zero throughout the motion sequence—training with only pose and contact label rewards results in foot motions resembling a one-segment foot. In particular, the revolute joint connecting the toe and heel fails to bend and roll naturally, leading to unrealistic foot dynamics. These theoretical and empirical observations underscore the necessity of providing explicit reference angles for the toe joint. To address this, we propose the *ToeRefEstimator*, a module that generates appropriate target angles from the contact labels. This intermediate step enables the reinforcement learning agent to follow more direct targets, facilitating more stable and efficient learning of natural foot articulation.

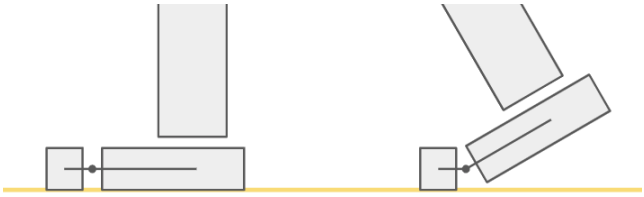

Figure 7: *Two-segment foot model, consisting of toe and heel segments connected by a revolute joint.*

Algorithm 1: Toe Joint Target Angle Generation

1. Detect heel-off time $t_{\text{heel\_off}}$ and toe-off time $t_{\text{toe\_off}}$.
2. For each frame t:
   - If $t < t_{\text{heel\_off}}$: set $\theta(t) = 0$.
   - If $t_{\text{heel\_off}} \leq t \leq t_{\text{toe\_off}}$:
     - Let $T = t_{\text{toe\_off}} - t_{\text{heel\_off}}$.
     - If $T \geq 5$:
       - If $t - t_{\text{heel\_off}} < 5$: linearly increase $\theta(t)$ from 0 to 42°.
       - Else: $\theta(t) = 42^\circ$.
     - Else: linearly increase $\theta(t)$ from 0 to 42° over T frames.
   - If $t_{\text{toe\_off}} < t \leq t_{\text{toe\_off}} + 2$: linearly decrease $\theta(t)$ to 0 over 2 frames.
   - Else: $\theta(t) = 0$.

The ToeRefEstimator generates the reference toe joint angle based on foot contact timings predicted by our foot contact detection network as detailed in Algorithm 1. Specifically, we identify two key events, *heel-off* (denoted as $t_{\text{heel\_off}}$) and *toe-off* (denoted as $t_{\text{toe\_off}}$), and use these to define a time interval where the target angle increases to simulate toe push-off. Using the identified interval $[t_{\text{heel\_off}}, t_{\text{toe\_off}}]$, we define a target angle trajectory for the toe joint. Based on findings from gait studies on metatarsophalangeal joint motion during walking [NBU99], the target angle increases linearly from 0 to 42 degrees starting at $t_{\text{heel\_off}}$. Empirically, we increase the angle to its maximum over 5 frames. If the interval $t_{\text{toe\_off}} - t_{\text{heel\_off}}$ exceeds 5 frames, the target angle remains at 42 degrees until $t_{\text{toe\_off}}$; if it is shorter, the peak angle is scaled proportionally. After $t_{\text{toe\_off}}$, the target angle rapidly decreases back to 0 over 2 frames to simulate the natural return motion of the toe joint. This angle trajectory effectively emulates the toe-pushing action during push-off and empirically contributes to more stable and realistic gait pattern learning. See Figure 8 for an illustration of the resulting angle curve.

Rewards We employed deep reinforcement learning to learn a control policy that mimics reference motion obtained from pose estimation methods developed by prior work [LXC*21], while following contact information extracted from video data using our framework. We primarily adopt the reward functions introduced in prior work [YPL21]. In this section, we focus on the contact reward, which utilizes the contact labels extracted by our method. A detailed description of the full set of reward functions used in this work, beyond the contact reward, is provided in Appendix C.

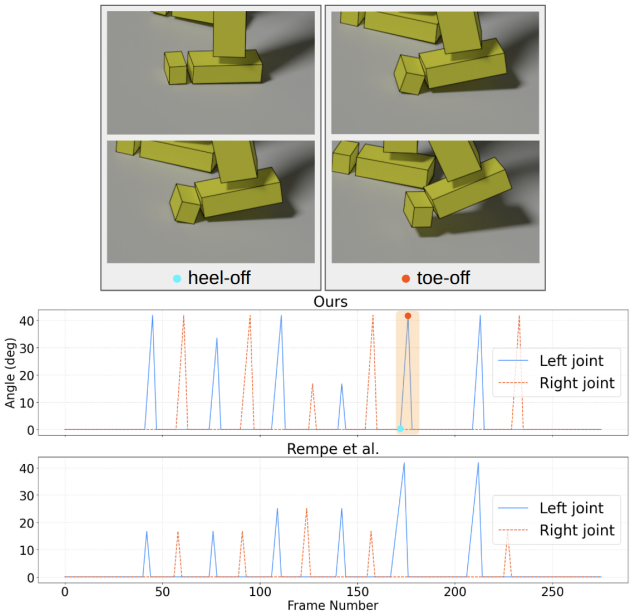

Figure 8: *Target toe joint angles generated by the ToeRefEstimator based on predicted contact labels. The angle increases from zero at the heel-off event (blue dot) to a predefined maximum value, and then decreases rapidly after the toe-off event (red dot), mimicking the natural foot motion during gait.*

Foot-ground contact reward, r_{fc}, encourages the character to follow the foot contact labels, $\hat{c}$, extracted from the video using our contact detection network.

$$r_{\text{fc}} = \exp\left(-\alpha_{\text{fc}} \sum_{l \in \{\text{LT,RT,LH,RH}\}} \|\text{xor}(\hat{c}_l, c_l)\|^2\right), \quad (4)$$

where c and $\hat{c}$ denote the simulated contact states and estimated contact labels, respectively, and both are binary variables, where 0 represents a non-contact state and 1 represents a contact state. The exclusive OR function $\text{xor}(\hat{c}, c)$ returns 1 when $\hat{c} \neq c$, and 0 otherwise.

We reconstructed walking motion using the two-segment foot model with contact labels obtained from our method. For comparison, we also conducted the same experiment using contact labels generated by the method proposed by Rempe et al. [RGH*20]. Figure 9 shows the reconstructed motion results for each contact detection approach. The results from Rempe et al. [RGH*20] exhibited a shuffling gait without proper forward progression. In contrast, our results demonstrated smooth foot-ground contact without heavy stomping, enabling natural forward movement. Figure 10 provides a visualization of the contact labels used during training for both cases. To aid interpretation, the interval $[t_{\text{heel\_off}}, t_{\text{toe\_off}}]$ in each graph of Figure 10 is highlighted in light pink. As seen in the bottom two graphs of Figure 10, the light pink region is barely visible in the results from Rempe et al. [RGH*20], indicating that the contact timings of the toe and heel segments are nearly identical.

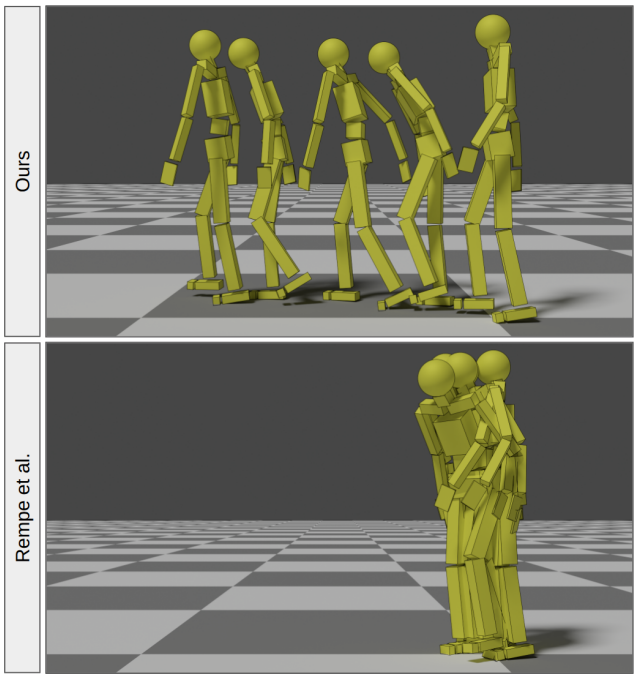

Figure 9: *Comparison of walking motion reconstruction results. Multiple frames over time are overlaid in a single image to visualize motion, which progresses from right to left. The top row shows the result of our method, where the character walks forward naturally. The bottom row shows the result using the same experimental setup, but with contact labels extracted by the method of Rempe et al. [RGH*20], resulting in in-place stepping.*

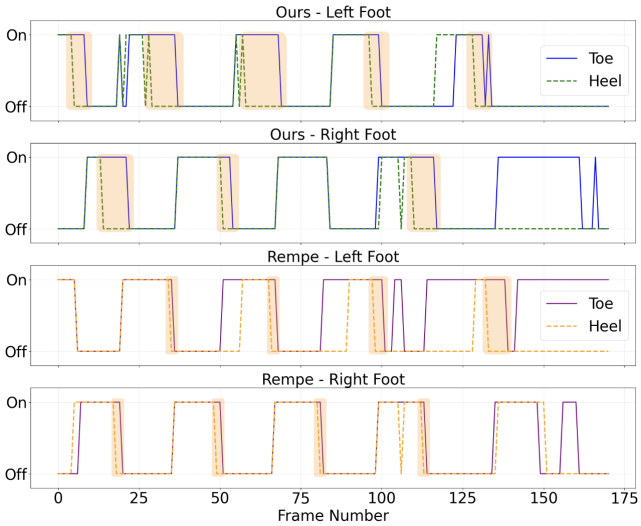

Figure 10: *Comparison of foot-ground contact labels predicted by our method (top) and by Rempe et al. [RGH*20] (bottom). The top two plots show the predicted contact labels for the left and right foot using our method, while the bottom two show the corresponding predictions by Rempe et al. A clearer separation (highlighted in light pink) can be observed in our results, indicating higher temporal accuracy.*

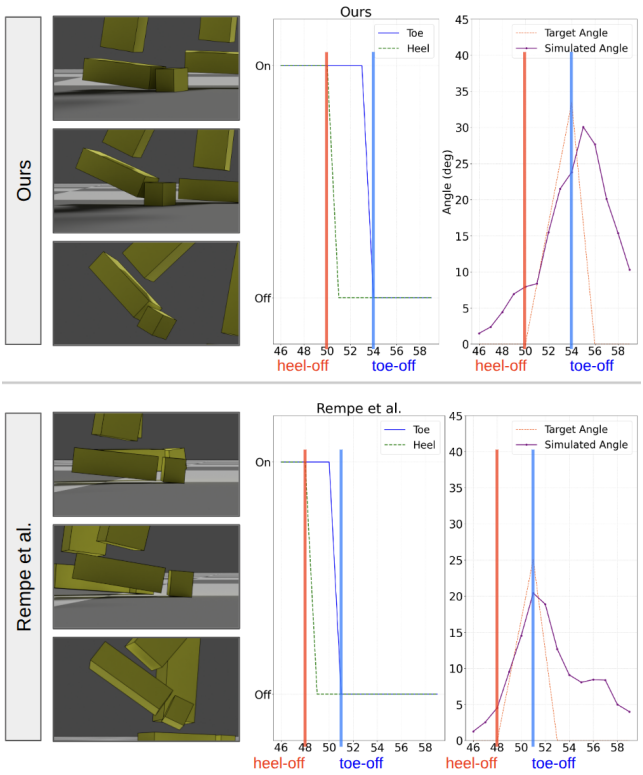

Figure 11: *The top row shows our results and the bottom row shows the results of Rempe et al [RGH*20]. From left to right, each column displays a close-up of the character's foot, the contact labels during the heel-off and toe-off phases, and the target versus simulated angles during those same phases.*

As a result, although using a two-segment foot model, it effectively behaves like a one-segment foot, which cannot generate the propulsive force during toe-off necessary for forward motion, explaining the lack of progression. In contrast, the top two graphs in our results (Figure 10) clearly show the light pink regions, representing the interval $[t_{\text{heel\_off}}, t_{\text{toe\_off}}]$. This indicates a temporal gap between the heel-off and toe-off events, allowing time for toe flexion, which contributes to a natural forward walking motion.

Figure 11 illustrates the simulation results (left column). The top row corresponds to our method, and the bottom row to that of Rempe et al. [RGH*20]. The middle column shows the estimated contact labels, and the right column shows the simulated joint angles tracking the target angles. In the case of Rempe et al., the interval $[t_{\text{heel\_off}}, t_{\text{toe\_off}}]$ was estimated to be shorter, resulting in a lower peak value for the toe joint's target angle. As a consequence, the toe joint fails to achieve sufficient toe flexion, which limits forward propulsion.

6.4.1. Quantitative Evaluation

To quantitatively evaluate the quality of motion reconstruction, we compare the joint angles of the reconstructed motion against those of the reference motion obtained from the input video. Specifically, we measure the root mean square error (RMSE) of joint angles over the entire reconstructed sequence. Our method achieves a joint angle RMSE of 3.27 degrees, indicating a reasonable level of agreement between the reconstructed motion and the IK-based reference. To evaluate how well the reconstructed humanoid motion follows the reference foot–ground contact labels, we measure the F1 score, precision, and recall. The learned controller achieves an F1 score of 0.78 (precision 0.80, recall 0.77), compared to the original estimator's F1 score of 0.93. This reduction reflects the inherent challenges of reproducing precise contact states through control, yet the results indicate that the controller successfully captures the essential contact patterns.

6.4.2. Ablation Study

We conducted ablation studies to evaluate the impact of the contact reward and the ToeRefEstimator by removing each component individually. As illustrated in the first row of Figure 12, when the contact reward was removed, the character attempted to perform forward locomotion, but the actual displacement was minimal. In the second row of Figure 12, in the absence of the ToeRefEstimator, the foot model operated as a one-segment structure, failing to produce natural toe flexion. Consequently, the character exhibited in-place stepping behavior and ultimately lost balance and collapsed. In contrast, our method, which incorporates both components, produces natural toe flexion while the character walks forward (see the third row of Figure 12). These results demonstrate that both components significantly contribute to producing more stable and realistic foot motions.

6.5. Application 2: Gait Analysis

Our framework enables accurate gait analysis directly from video without the need for any sensors. First, we perform per-frame phase labeling by classifying each pose into one of the main gait phases within the gait cycle. Based on these labels, we extract key gait parameters, stance/swing phase ratios, single-support ratio, and double-support ratio, which are widely recognized spatio-temporal indicators in gait assessment. The stance and swing ratios reflect the proportion of time a foot spends on the ground versus in the air, while single- and double-support ratios quantify balance and weight transfer dynamics. Finally, we compare these vision-based measurements against ground-truth obtained via sensor mats to validate our framework's accuracy and clinical utility.

6.5.1. Gait Phase Classification

The gait cycle refers to the sequence of movements that occur during walking, beginning when one foot makes contact with the ground and ending when the same foot contacts the ground again. This cycle is fundamental in gait analysis as it provides a structured framework to assess and understand human locomotion. By examining each phase of the gait cycle, clinicians and researchers can identify abnormalities, design rehabilitation strategies, and enhance overall movement efficiency. The trained network predicts four contact labels for each frame: $[LT, RT, LH, RH] \in \{0,1\}^4$. These labels indicate whether each of the four regions, Left Toe (LT), Right Toe (RT), Left Heel (LH), and Right Heel (RH), is in contact with the ground. Based on these contact labels, we classify each frame into one of the following nine gait phases: Stand, Heel Strike (HS),

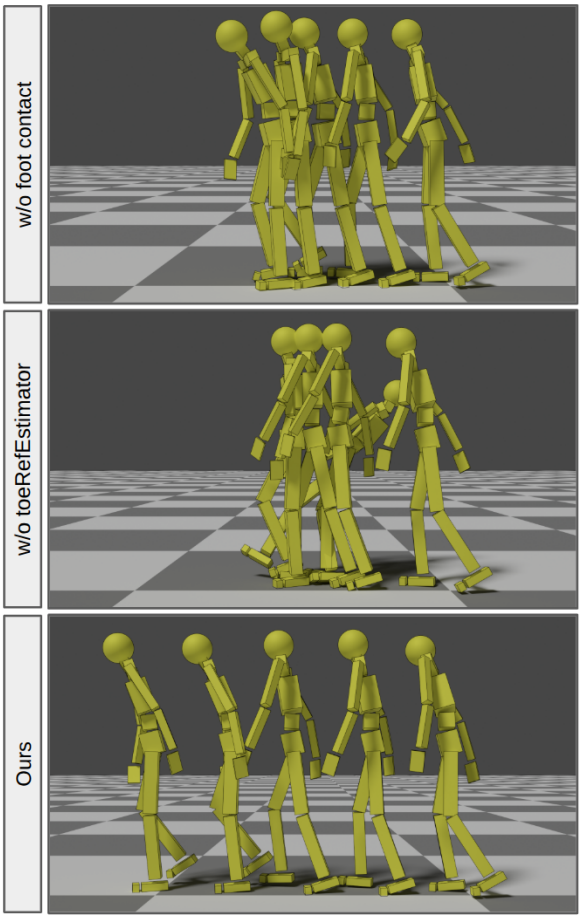

Figure 12: *Ablation results. Without the contact reward (top), the character fails to move forward properly. Without the ToeRefEstimator (middle), the foot motion is unnatural and the character falls over. With both (bottom), the motion is stable and natural.*

Table 2: *Gait phase classification based on four foot contact labels (LT: left toe, RT: right toe, LH: left heel, RH: right heel), defined with respect to the right foot.*

Contact Labels (LT, RT, LH, RH)	Phase	Phase Category	Support Type
(1, 1, 1, 1)	Stand	-	-
(1, 0, 0, 1)	Heel Strike	Stance	Double
(1, 1, 0, 1)	Loading Response	Stance	Single
(0, 1, 0, 1)	Mid Stance	Stance	Single
(0, 1, 1, 0)	Terminal Stance	Stance	Single
(1, 1, 1, 0)	Pre-Swing	Stance	Double
(1, 0, 1, 0)	Toe-Off / Mid Swing	Swing	Single
(1, 0, 0, 0)	Terminal Swing	Swing	Single
Otherwise	Undefined	-	-

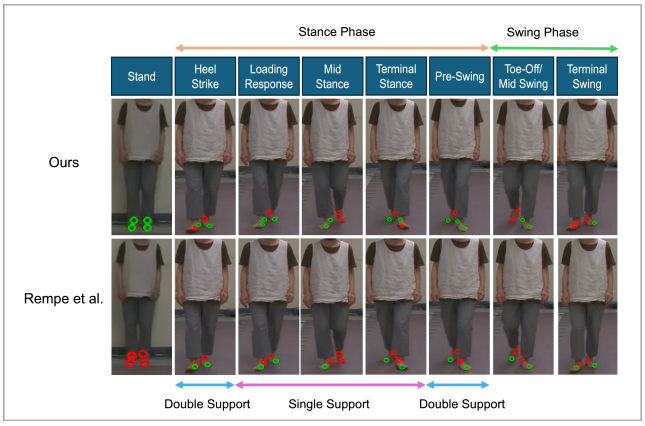

Figure 13: *The visualization compares the contact labels produced by our model and the baseline. Green circles denote contact, while red circles denote no contact. As Shown, our model infers heel contact more accurately than the baseline [RGH*20].*

Loading Response (LR), Mid Stance (MS), Terminal Stance (TS), Pre-Swing (PS), Toe-Off/Mid-sWing (TO/MW), Terminal sWing (TW), and Undefined. The contact label-to-phase mapping is detailed in Table 2. In cases where the contact labels do not correspond to any of the defined phases, the gait phase is labeled as *Undefined*. In Figure 13, the top row presents the per-frame gait phase labeling results derived from our video-based contact detection network. When compared to the second row, which shows results from a previous study (baseline method) [RGH*20], it is evident that our approach achieves significantly higher accuracy in identifying each phase of the gait cycle.

This mapping allows for detailed analysis of gait phases, which is essential for applications such as gait rehabilitation, prosthetic design, and athletic performance assessment. By accurately identifying these phases, clinicians and researchers can assess gait abnormalities and design targeted interventions.

6.5.2. Gait Parameters Extraction

In this section, we demonstrate how frame-by-frame gait phase labels are used to measure key gait parameters. From the per-frame gait phase labels obtained in the previous section, we extracted key spatiotemporal gait parameters—specifically, the stance and swing phase ratios, single-support ratio, and double-support ratio—which are essential for quantifying gait timing and support patterns.

Stance/Swing Phase Ratios Each gait phase falls into one of two primary categories: stance and swing phases. As shown in the second and third columns of Table 2, HS, LR, MS, TS, and PS are classified as stance phases, while TO/MW and TW belong to swing phases.

We define the stance and swing phase ratios as follows:

$$\text{Stance phase ratio} = \frac{N_{\text{stance}}}{N_{\text{valid}}}, \qquad \text{Swing phase ratio} = \frac{N_{\text{swing}}}{N_{\text{valid}}}.$$

Here, N_{stance} and N_{swing} represent the number of frames labeled

stance and swing phases, respectively. We define the number of valid frames, N_{valid}, as the total number of frames excluding those labeled "stand" or "undefined." This ensures that only frames corresponding to active gait phases are considered in our analysis.

$$N_{\text{valid}} = N_{\text{HS}} + N_{\text{LR}} + N_{\text{MS}} + N_{\text{TS}} + N_{\text{PS}} + N_{\text{TO}} + N_{\text{TW}} \quad (5)$$

Single/Double-Support Ratios Single support and double support phases characterize whether one or both feet are in contact with the ground during a gait cycle. The single-support occurs when only one foot is in contact, comprising the Loading Response (LR), Mid Stance (MS), Terminal Stance (TS), Toe-Off/Mid-sWing (TO/MW), and Terminal sWing (TW) phases. Double-support occurs when both feet are simultaneously in contact with the ground, corresponding to the Heel Strike (HS) and Pre-Swing (PS) phases. As shown in the second and fourth columns of Table 2, the support type for each gait phase is presented alongside its corresponding phase.

We can derive single- and double-support ratios:

$$\text{Single support Ratio} = \frac{N_{\text{single}}}{N_{\text{valid}}}, \qquad \text{Double support Ratio} = \frac{N_{\text{double}}}{N_{\text{valid}}}.$$

Here, N_{single} and N_{double} are the number of frames classified as single-support and double-support, respectively, and N_{valid} is the total number of valid gait-phase frames (excluding stand and undefined). These ratios reflect the proportion of the gait cycle during which one or both feet are in contact with the ground—critical metrics widely used in spatiotemporal gait analysis for assessing balance and stability.

6.5.3. Evaluation

To demonstrate the effectiveness of our framework as a sensor-free gait analysis tool, we compared gait parameters derived from our framework with ground-truth measurements from a sensor mat. To this end, we used a dataset containing video and pressure-sensor recordings captured simultaneously during walking trials. From the per-frame gait phase labels, we extracted key spatio-temporal gait parameters: stance and swing phase ratios, single-support ratio, and double-support ratio. We then quantified the accuracy of our method by computing the Mean Absolute Error (MAE) between our video-derived parameters and the corresponding values from the sensor mat.

Dataset The dataset used for the evaluation was collected under controlled conditions: a GaitRite® pressure mat was installed in a straight corridor. Simultaneously, video recordings were made from two perspectives: Frontal view, where subjects walking toward the camera, and rear view where subjects walking away from the camera. Subjects were instructed to walk across the mat at a comfortable pace. The original dataset was collected from a total of 26 subjects by the Department of Neurology at Seoul National University Hospital, comprising 13 healthy individuals and 13 patients diagnosed with Parkinson's disease. All participants provided informed consent prior to data collection, and the study was approved by the Institutional Review Board (IRB) of Seoul National University Hospital under approval number (1908–175–1059). From the original dataset, we used only data from healthy adults, and excluded from the test dataset any videos where subjects were occluded by others or where clothing interfered with accurate joint position estimation by the pose estimator. Consequently, the experiments were conducted on 9 videos from 8 healthy subjects.

Table 3: *Error comparison of gait parameters (unit: %): Our framework vs. Baseline [RGH*20].*

Gait Parameter	Ours (MAE ± SD)	Rempe et al. (MAE ± SD)
Stance phase ratio	**9.3** ± 3.4	37.7 ± 1.3
Swing phase ratio	**9.3** ± 3.4	37.7 ± 1.3
Single-support ratio	**5.6** ± 5.5	59.3 ± 7.3
Double-support ratio	**12.9** ± 4.7	21.3 ± 7.1

The table 3 summarizes the mean absolute error (MAE) and standard deviation (SD), between sensor-based measurements and values estimated by our framework and a previous method for each gait parameter. For the stance and swing phase ratios, the baseline method exhibits significantly large errors, $37.7 \pm 1.3\%$ and $37.7 \pm 1.3\%$, whereas our model achieves an order-of-magnitude reduction with $9.3 \pm 3.4\%$ and $9.3 \pm 3.4\%$, demonstrating substantially improved fidelity in capturing gait cycle phase proportions. When examining the single- and double-support ratios, the baseline approach again shows large discrepancies: $59.3 \pm 7.3\%$ (single); $21.3 \pm 7.1\%$ (double). Our framework dramatically lowers these to $5.6 \pm 5.5\%$ for single-support, and $12.9 \pm 4.7\%$ for double-support—highlighting significantly improved accuracy across all support-phase metrics. Individual values for each gait parameter are provided in Appendix B.

7. Discussion

We proposed ContactVision, a gait-specialized deep learning framework that estimates foot–ground contact information from monocular video. The estimated foot-ground contact information provides valuable cues for 3D motion reconstruction, and its effectiveness was validated through reinforcement learning-based experiments. In particular, our framework extracts contact labels for two foot parts, the heel and the toe, and we introduce a two-segment foot model to effectively utilize this information in motion reconstruction. Furthermore, by implementing the ToeRefEstimator, we successfully reconstructed natural foot articulation during gait, significantly enhancing the realism and accuracy of the walking motion.

Beyond computer graphics, the same information was also validated in gait cycle analysis, showing their potential in medical contexts that require high accuracy. While our primary focus is on character animation and motion reconstruction, the additional evaluation in gait analysis highlights the broader applicability of our framework. Furthermore, since our method estimates gait indicators directly from video without relying on sensors such as force

plates, pressure sensors, or IMUs, it offers advantages in terms of cost efficiency, portability, and accessibility. Unlike prior studies that largely assumed fixed camera viewpoints, our framework has been experimentally shown to operate robustly under dynamic camera conditions, such as changes in viewpoint during recording or handheld scenarios. This demonstrates a significant distinction, highlighting its potential applicability in real-world environments.

Nevertheless, several limitations remain. First, the framework relies solely on the output of video-based pose estimators, making its performance highly dependent on the accuracy of the underlying pose estimator. Consequently, errors in pose estimation can directly affect the accuracy of foot–ground contact detection. To address these limitations, future work will focus on incorporating additional cues such as physical constraints or multi-modal signals beyond pose estimator outputs, thereby reducing dependency on a single input and improving robustness to inaccurate pose estimations. In addition, the ToeRefEstimator relies on simple heuristics tailored to typical forward walking and may not generalize to non-standard or pathological gait patterns, motivating future work toward more adaptive formulations. Second, because the framework is specialized for gait, its generalizability to other types of human motion is limited. By training on more diverse human motion datasets, we aim to extend the framework to reliably estimate foot–ground contact information for a broader range of motions beyond gait.

Acknowledgements

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) under the Artificial Intelligence Convergence Innovation Human Resources Development Program (IITP-2026-RS-2023-00255968), funded by the Korean government (MSIT), and by the 2025 Digital Therapeutic Device Development and Demonstration Support Program, funded by the Ministry of Science and ICT (MSIT) and the National IT Industry Promotion Agency (NIPA), Republic of Korea.

Appendix A: Marker Mapping

The marker topology of our motion capture dataset differs from the one required by the UnderPressure framework. To bridge this discrepancy, we define a marker mapping procedure that transforms the original marker set into the format expected by UnderPressure.

Four types of mapping strategies are used: **(1) direct mapping**, a one-to-one correspondence between source and target markers; **(2) average mapping**, where the target marker is computed as the average of two or more source markers; **(3) interpolated mapping**, which places the target marker at a specified ratio along the line segment between two source markers; and **(4) offset mapping**, where the target marker is placed by offsetting a source marker along a specified direction vector by a fixed ratio. We apply these rules to convert our dataset markers into the required set of 23 input markers. The full mapping specification is provided in Table 5.

Table 4: *Mapping between BioCV [ENW*24] mocap markers (source) and UnderPressure [MHCH22] mocap markers (target). Each marker index corresponds to the label shown in Figure 3.*

Index	Target Marker	Source Marker
0	Pelvis	ACROM_R
1	L5	ACROM_L
2	L3	CLAV
3	T12	XIP_PROC
4	T8	ELB_LAT_R
5	Neck	ELB_MID_R
6	Head	ELB_LAT_L
7	Right Clavicle	ELB_MID_L
8	Right Shoulder	ILCREST_R
9	Right Elbow	ILCREST_L
10	Right Hand	ASIS_R
11	Left Clavicle	ASIS_L
12	Left shoulder	WRI_LAT_R
13	Left Elbow	WRI_MID_R
14	Left Hand	WRI_LAT_L
15	Right Hip	WRI_MID_L
16	Right Knee	KNEE_LAT_R
17	Right Foot	KNEE_MID_R
18	Right Toe	KNEE_LAT_L
19	Left Hip	KNEE_MID_L
20	Left Knee	MAL_LAT_R
21	Left Foot	MAL_MID_R
22	Left Toe	MAL_LAT_L
23	-	MAL_MID_L
24	-	MTP5_R
25	-	TOE_R
26	-	MTP1_R
27	-	MTP5_L
28	-	TOE_L
29	-	MTP1_L
30	-	C7
31	-	T10
32	-	PSIS_R
33	-	PSIS_L
34	-	HAND_R
35	-	HAND_L
36	-	HEEL_R
37	-	HEEL_L

Appendix B: Detailed Gait Parameter Measurements

In normal walking, the stance and swing phase ratios average approximately 60% and 40%, respectively, of the gait cycle. Within the stance phase, single-support ratio (only one foot on the ground) is about 40%, while double-support ratio (both feet in contact) comprises about 20–25% of the cycle. Typical spatio-temporal values in healthy adults walking at a comfortable speed are well-documented [LJFN21,NCCV21]. These benchmarks provide valuable reference points for interpreting deviations in gait parameters measured by our frame-by-frame video-based framework. To facilitate an easy comparison of the actual values, we plotted graphs (Figure 14, 15, 16, and 17) of the ground truth values of four gait parameters (stance phase ratios, swing phase ratios, single-support ratios, and double-support ratios) from the GaitRite sensor (GT), the values estimated by our method (Our), and the values estimated by

Table 5: *Marker mapping Table used for conversion to UnderPressure input format*

Target Marker	Type	Source Marker(s)
Pelvis	average	ILCREST_L, ILCREST_R
L5	interp	Pelvis, Neck, t=1/5
L3	interp	Pelvis, Neck, t=2/5
T12	interp	Pelvis, Neck, t=3/5
T8	interp	Pelvis, Neck, t=4/5
Neck	average	C7, CLAV
Head	offset	Neck, T8→Neck, ratio=0.8
Right Clavicle	interp	ACROM_R, ACROM_L, t=1/3
Right Shoulder	direct	ACROM_R
Right Elbow	average	ELB_LAT_R, ELB_MED_R
Right Hand	direct	HAND_R
Left Clavicle	interp	ACROM_R, ACROM_L, t=2/3
Left Shoulder	direct	ACROM_L
Left Elbow	average	ELB_LAT_L, ELB_MED_L
Left Hand	direct	HAND_L
Right Hip	direct	ASIS_R
Right Knee	average	KNEE_MED_R, KNEE_LAT_R
Right Foot	average	MAL_MED_R, MAL_LAT_R
Right Toe	direct	TOE_R
Left Hip	direct	ILCREST_L
Left Knee	average	KNEE_MED_L, KNEE_LAT_L
Left Foot	average	MAL_MED_L, MAL_LAT_L
Left Toe	direct	TOE_L

the baseline [RGH*20] (Rempe et al.) across the 9 walking samples. Each sample follows the format `{subject_id}_{view}`, where view can be either `F` (forward-facing) or `B` (backward-facing), representing the walking direction relative to the camera.

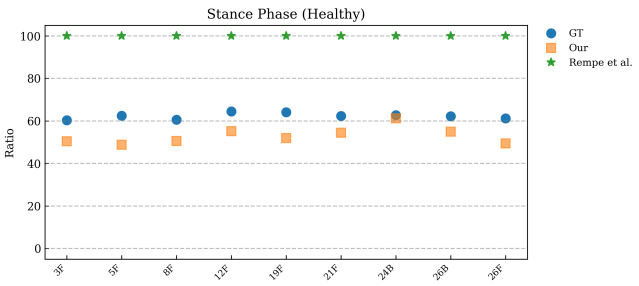

Figure 14: *Quantitative comparison of the stance phase ratio based on the right foot during a single gait cycle in healthy adults.*

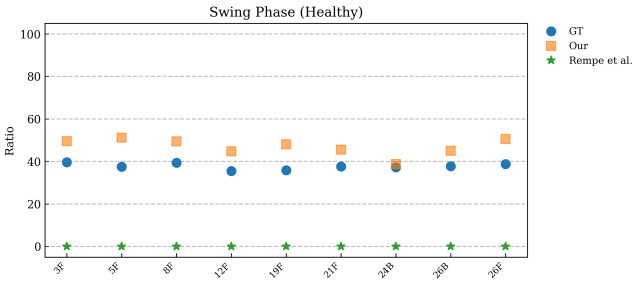

Figure 15: *Quantitative comparison of the swing phase ratio based on the right foot during a single gait cycle in healthy adults.*

As shown in Figures 14 and 15, the stance phase ratios and swing phase ratios estimated by our framework (represented by orange squares) closely match the ground-truth values (represented by blue circles), demonstrating the accuracy of our method. In contrast, the method by Rempe et al. [RGH*20] showed an estimation of 100% for stance phase ratios and 0% for swing phase ratios (represented by green stars). This issue arises due to the imprecise detection of heel contact points. The model frequently misclassified frames where the heel joint was in contact as non-contact. Many frames are also classified as undefined. As a result, the method fails to accurately estimate the swing phase, which is reflected in the 0% estimation for swing phase ratios.

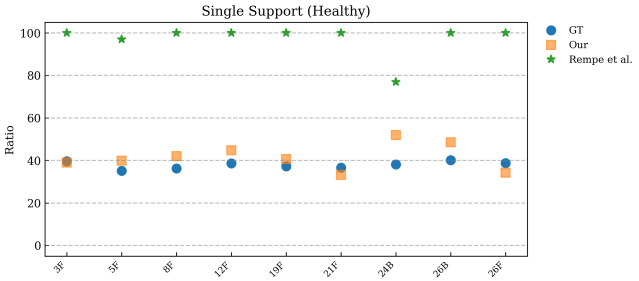

Figure 16: *Quantitative comparison of the single support ratio based on the right foot during a single gait cycle in healthy adults.*

The Figure 16 presents single support ratios per video, showing that our model yields results consistently closer to the ground truth.

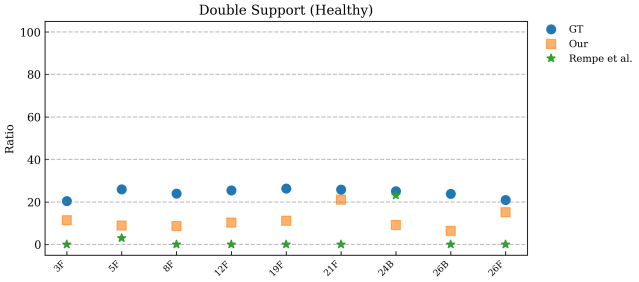

Figure 17: *Quantitative comparison of the double support ratio based on the right foot during a single gait cycle in healthy adults.*

The Figure 17 presents double support ratios per video. The ground truth values were distributed between 20% and 25%, whereas our framework tended to underestimate the ratio, with values falling in the range of 10% to 20%. Among the four gait parameters, the estimation error for double support ratios was the largest when compared to the ground truth values, with a maximum error of 12.9%, as shown in Table 3. This underestimation may be attributed to the brief duration of the double support phase within the gait cycle. The double-support phase occurs in a very brief period within the gait cycle, corresponding to only 2-3 frames in the video recorded at 29.97 fps, which is roughly 0.067 to 0.1 seconds. If contact is misclassified during these 2-3 frames, even a small error in detecting the foot-ground contact can have a significant impact on the estimation of the double-support ratio. This is why the estimation error for the double support ratio is particularly large.

Appendix C: Details of Reward Terms Used in Motion Reconstruction

The reward function in Equation 6 is composed of a weighted sum of seven individual reward terms. Each reward term is designed to minimize the error with respect to the reference motion, and they can be broadly categorized into three groups: motion tracking, physical plausibility, and video style preservation.

$$\begin{aligned} R = w_q r_q + w_v r_v + w_e r_e + w_{fc} r_{fc} + \\ w_{up} r_{up} + w_{root\_ori} r_{root\_ori} + w_{torque} r_{torque}, \end{aligned} \tag{6}$$

where the specific values of the reward weights used in the experiments are as follows: $w_{\text{q}} = 0.3$, $w_{\text{v}} = 0.05$, $w_{\text{e}} = 0.05$, $w_{\text{fc}} = 0.3$, $w_{\text{torque}} = 0.05$, $w_{\text{root\_ori}} = 0.15$, and $w_{\text{up}} = 0.1$.

1. Motion Tracking These reward terms r_q, r_v, and r_e encourage the character to closely track the reference motion by matching joint angles, joint velocities, and end-effector positions, respectively.

$$\begin{aligned} r_q &= \exp\left(-\alpha_q \left\|\hat{\mathbf{q}} - \mathbf{q}\right\|^2\right), \\ r_v &= \exp\left(-\alpha_v \left\|\hat{\mathbf{v}} - \mathbf{v}\right\|^2\right), \\ r_e &= \exp\left(-\alpha_e \left\|\hat{\mathbf{p}}_e - \mathbf{p}_e\right\|^2\right). \end{aligned} \tag{7}$$

- **Pose reward** ($\mathbf{r}_q$) encourages the character's joint angles ($\mathbf{q}$) to match the target angles ($\hat{\mathbf{q}}$) from the reference motion. It serves as the most fundamental tracking term and primarily determines the overall shape of the motion.
- **Velocity reward** ($\mathbf{r}_v$) encourages the character's joint angular velocities ($\mathbf{v}$) to follow the target joint velocities ($\hat{\mathbf{v}}$) from the reference motion, helping to replicate the timing of the movement.
- **End-effector reward** ($\mathbf{r}_e$) minimizes the discrepancy between the pelvis-relative positions of the character's end-effectors (hands and feet), $\mathbf{p}_e$, and their target positions in the reference motion, $\hat{\mathbf{p}}_e$. It helps directly correct positional errors that often occur at the distal parts of long limbs.

2. Physical plausibility These reward terms r_{fc}, r_{root_ori}, and r_{torque} ensure that the simulated character's motion adheres to physical laws and appears stable.

$$\begin{aligned} r_{fc} &= \exp\left(-\alpha_{\text{fc}} \sum_{l \in \{\text{LT,RT,LH,RH}\}} \left\|\text{xor}(\hat{c}_l, c_l)\right\|^2\right), \\ r_{root\_ori} &= \exp(-\alpha_{root\_ori} \cdot (1 - \mathbf{z}_{root} \cdot \hat{\mathbf{z}}_{world})^2), \\ r_{torque} &= \exp\left(-\alpha_{torque} \left|\tau\right|^2\right). \end{aligned} \tag{8}$$

- **Stability reward** ($\mathbf{r}_{\text{root\_ori}}$) promotes alignment between the character's local z-axis ($\mathbf{z}_{root}$) and the world's z-axis ($\hat{\mathbf{z}}_{world}$), providing continuous balance during dynamic motions.
- **Torque minimization reward** ($\mathbf{r}_{torque}$) minimizes excessive joint torques (τ), encouraging energy-efficient and smooth motions. It prevents the agent from applying unrealistic forces.

3. Video style preservation This reward term r_{up} helps preserve the unique motion style present in the original video, which is difficult to capture using only 3D joint data.

$$r_{\text{up}} = \exp\left(-\alpha_{\text{up}} \left\|\hat{\theta} - \theta\right\|^2\right) \tag{9}$$

- **Upper Body Posture Reward** ($\mathbf{r}_{up}$) encourages the simulated character's upper body tilt to match that computed from the 2D pose in the original video. Specifically, in simulation, the angle, θ, between the pelvis-head vector and the vertical axis of the world coordinate system is calculated, while in the 2D video, the angle, $\hat{\theta}$, between the pelvis-neck vector and the vertical image axis is computed. Minimizing the difference between these two angles guides the simulation to learn postures such as leaning forward or backward consistent with the original video.

References

[ALX*19] Abdolhosseini F., Ling H. Y., Xie Z., Peng X. B., Van de Panne M.: On learning symmetric locomotion. In *Proceedings of the 12th ACM SIGGRAPH Conference on Motion, Interaction and Games* (2019), pp. 1–10. 3

[ÅOÅ*21] Åberg A. C., Olsson F., Åhman H. B., Tarassova O., Arndt A., Giedraitis V., Berglund L., Halvorsen K.: Extraction of gait parameters from marker-free video recordings of timed up-and-go tests: Validity, inter-and intra-rater reliability. *Gait & Posture 90* (2021), 489–495. 3

[ASY*19] Akiba T., Sano S., Yanase T., Ohta T., Koyama M.: Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2019). 6

[BBS*08] Bamberg S. J. M., Benbasat A. Y., Scarborough D. M., Krebs D. E., Paradiso J. A.: Gait analysis using a shoe-integrated wireless sensor system. *IEEE transactions on information technology in biomedicine 12*, 4 (2008), 413–423. 3

[CHS*19] Cao Z., Hidalgo G., Simon T., Wei S.-E., Sheikh Y.: Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE transactions on pattern analysis and machine intelligence 43*, 1 (2019), 172–186. 2, 3, 4, 5, 6, 7

[CPKC24] Cimorelli A., Patel A., Karakostas T., Cotton R. J.: Validation of portable in-clinic video-based gait analysis for prosthesis users. *Scientific reports 14*, 1 (2024), 3840. 3

[CSH*24] CHEN R., SHI M., HUANG S., TAN P., KOMURA T., CHEN X.: Taming diffusion probabilistic models for character control. In *ACM SIGGRAPH 2024 Conference Papers* (2024), pp. 1–10. 3

[EKA*17] ELTOUKHY M., KUENZE C., ANDERSEN M. S., OH J., SIGNORILE J.: Prediction of ground reaction forces for parkinson's disease patients using a kinect-driven musculoskeletal gait analysis model. *Medical engineering & physics 50* (2017), 75–82. 3

[ENW*24] EVANS M., NEEDHAM L., WADE L., PARSONS M., COLYER S., MCGUIGAN P., BILZON J., COSKER D.: Synchronised video, motion capture and force plate dataset for validating markerless human movement analysis. *Scientific Data 11*, 1 (Nov. 2024). Publisher Copyright: © The Author(s) 2024. doi:10.1038/s41597-024-04077-3. 2, 3, 4, 5, 13

[HGZ*23] HII C. S. T., GAN K. B., ZAINAL N., MOHAMED IBRAHIM N., AZMIN S., MAT DESA S. H., VAN DE WARRENBURG B., YOU H. W.: Automated gait analysis based on a marker-free pose estimation model. *Sensors 23*, 14 (2023), 6489. 3

[HST*23] HAN X., SENDERLING B., TO S., KUMAR D., WHITING E., SAITO J.: Groundlink: A dataset unifying human body movement and ground reaction dynamics. In *SIGGRAPH Asia 2023 Conference Papers* (2023), pp. 1–10. 1, 3

[IPOS13] IONESCU C., PAPAVA D., OLARU V., SMINCHISESCU C.: Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence 36*, 7 (2013), 1325–1339. 2

[JJLK14] JUNG Y., JUNG M., LEE K., KOO S.: Ground reaction force estimation using an insole-type pressure mat and joint kinematics during walking. *Journal of biomechanics 47*, 11 (2014), 2693–2699. 3

[JP20] JEONG H., PARK S.: Estimation of the ground reaction forces from a single video camera based on the spring-like center of mass dynamics of human walking. *Journal of Biomechanics 113* (2020), 110074. 3

[JŽD19] JURCA A., ŽABKAR J., DŽEROSKI S.: Analysis of 1.2 million foot scans from north america, europe and asia. *Scientific reports 9*, 1 (2019), 19155. 8

[KEY*24] KIM C., EOM H., YOO J. E., CHOI S., NOH J.: Interactive locomotion style control for a human character based on gait cycle features. In *Computer Graphics Forum* (2024), vol. 43, Wiley Online Library, p. e14988. 3

[KG03] KOVAR L., GLEICHER M.: Flexible automatic motion blending with registration curves. In *Symposium on Computer Animation* (2003), vol. 2, San Diego, CA, USA. 2

[KLVDP20] KWON T., LEE Y., VAN DE PANNE M.: Fast and flexible multilegged locomotion using learned centroidal dynamics. *ACM Transactions on Graphics (TOG) 39*, 4 (2020), 46–1. 3

[KPLK15] KIM J., PARK H., LEE J., KWON T.: Human motion control with physically plausible foot contact models. *The Visual Computer 31*, 6 (2015), 883–891. 2

[KSG02a] KOVAR L., SCHREINER J., GLEICHER M.: Footskate cleanup for motion capture editing. SCA '02, Association for Computing Machinery, p. 97–104. URL: https://doi.org/10.1145/545261.545277, doi:10.1145/545261.545277. 2

[KSG02b] KOVAR L., SCHREINER J., GLEICHER M.: Footskate cleanup for motion capture editing. In *Proceedings of the 2002 ACM SIGGRAPH/Eurographics symposium on Computer animation* (2002), pp. 97–104. 3

[KSW*25] KOLEINI F., SALEEM M. U., WANG P., XUE H., HELMY A., FENWICK A.: Biopose: Biomechanically-accurate 3d pose estimation from monocular videos. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)* (2025), IEEE, pp. 6330–6339. 3

[Lab03] LAB C. M. U. G.: Cmu graphics lab motion capture database. http://mocap.cs.cmu.edu/, 2003. Accessed: 2025-08-26. 2

[LCR*02] LEE J., CHAI J., REITSMA P. S. A., HODGINS J. K., POLLARD N. S.: Interactive control of avatars animated with human motion data. *ACM Trans. Graph. 21*, 3 (July 2002), 491–500. URL: https://doi.org/10.1145/566654.566607, doi:10.1145/566654.566607. 2

[LDZ*24] LI H., DAI J., ZENG R., BAI J., CHEN Z., PAN J.: Foot-constrained spatial-temporal transformer for keyframe-based complex motion synthesis. *Computer Animation and Virtual Worlds 35*, 1 (2024), e2217. 3

[LJFN21] LEAL-JUNIOR A., FRIZERA-NETO A.: *Optical fiber sensors for the next generation of rehabilitation robotics*. Academic Press, 2021. 13

[LS99] LEE J., SHIN S. Y.: A hierarchical approach to interactive motion editing for human-like figures. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques* (1999), pp. 39–48. 2

[LSC*19] LI Z., SEDLAR J., CARPENTIER J., LAPTEV I., MANSARD N., SIVIC J.: Estimating 3d motion and forces of person-object interactions from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019), pp. 8640–8649. 2

[LXC*21] LI J., XU C., CHEN Z., BIAN S., YANG L., LU C.: Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2021), pp. 3383–3393. 3, 8

[LZD*22] LIANG S., ZHANG Y., DIAO Y., LI G., ZHAO G.: The reliability and validity of gait analysis system using 3d markerless pose estimation algorithms. *Frontiers in Bioengineering and Biotechnology 10* (2022), 857975. 3

[MHCH22] MOUROT L., HOYET L., CLERC F. L., HELLIER P.: Underpressure: Deep learning for foot contact detection, ground reaction force estimation and footskate cleanup. In *Computer Graphics Forum* (2022), vol. 41, Wiley Online Library, pp. 195–206. 1, 2, 3, 4, 5, 13

[MKD*20] MUNDT M., KOEPPE A., DAVID S., BAMER F., POTTHAST W., MARKERT B.: Prediction of ground reaction force and joint moments based on optical motion capture data during gait. *Medical Engineering & Physics 86* (2020), 29–34. 3

[MRC*17] MEHTA D., RHODIN H., CASAS D., FUA P., SOTNYCHENKO O., XU W., THEOBALT C.: Monocular 3d human pose estimation in the wild using improved cnn supervision. In *3D Vision (3DV), 2017 Fifth International Conference on* (2017), IEEE. URL: http://gvv.mpi-inf.mpg.de/3dhp_dataset, doi:10.1109/3dv.2017.00064. 2

[NBU99] NAWOCZENSKI D. A., BAUMHAUER J. F., UMBERGER B. R.: Relationship between clinical measurements and motion of the first metatarsophalangeal joint during gait. *JBJS 81*, 3 (1999), 370–6. 8

[NCCV21] NANDY A., CHAKRABORTY S., CHAKRABORTY J., VENTURE G.: *Modern methods for affordable clinical gait analysis: theories and applications in healthcare systems*. Academic Press, 2021. 13

[RFCA10] ROUHANI H., FAVRE J., CREVOISIER X., AMINIAN K.: Ambulatory assessment of 3d ground reaction force using plantar pressure distribution. *Gait & posture 32*, 3 (2010), 311–316. 3

[RGH*20] REMPE D., GUIBAS L. J., HERTZMANN A., RUSSELL B., VILLEGAS R., YANG J.: Contact and human dynamics from monocular video. In *Proceedings of the European Conference on Computer Vision (ECCV)* (2020). 2, 3, 7, 9, 10, 11, 12, 14

[SAA*20] SHI M., ABERMAN K., ARISTIDOU A., KOMURA T., LISCHINSKI D., COHEN-OR D., CHEN B.: Motionet: 3d human motion reconstruction from monocular video with skeleton consistency. *Acm transactions on graphics (tog) 40*, 1 (2020), 1–15. 3

[SGX*21] SHIMADA S., GOLYANIK V., XU W., PÉREZ P., THEOBALT C.: Neural monocular 3d human motion capture with physical awareness. *ACM Transactions on Graphics (ToG) 40*, 4 (2021), 1–15. 3

[SJL*24] SONG W., JIN X., LI S., CHEN C., HAO A., HOU X., LI N., QIN H.: Arbitrary motion style transfer with multi-condition motion latent diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2024), pp. 821–830. 3

[TPXL24] TRUONG T. E., PISENO M., XIE Z., LIU K.: Pdp: Physics-based character animation via diffusion policy. In *SIGGRAPH Asia 2024 Conference Papers* (2024), pp. 1–10. 3

[VSP*17] VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER Ł., POLOSUKHIN I.: Attention is all you need. *Advances in neural information processing systems 30* (2017). 2, 7

[WL19] WON J., LEE J.: Learning body shape variation in physics-based characters. *ACM Transactions on Graphics (TOG) 38*, 6 (2019), 1–12. 3

[XLKVDP20] XIE Z., LING H. Y., KIM N. H., VAN DE PANNE M.: Allsteps: curriculum-driven learning of stepping stone skills. In *Computer Graphics Forum* (2020), vol. 39, Wiley Online Library, pp. 213–224. 3

[YPL21] YU R., PARK H., LEE J.: Human dynamics from monocular video with dynamic camera movements. *ACM Trans. Graph. 40*, 6 (2021). 3, 8

[YTL18] YU W., TURK G., LIU C. K.: Learning symmetric and low-energy locomotion. *ACM Transactions on Graphics (TOG) 37*, 4 (2018), 1–12. 3

[ZLHA24] ZHANG Z., LIU R., HANOCKA R., ABERMAN K.: Tedi: Temporally-entangled diffusion for long-term motion synthesis. In *ACM SIGGRAPH 2024 Conference Papers* (2024), pp. 1–11. 3

[ZYC*20a] ZOU Y., YANG J., CEYLAN D., ZHANG J., PERAZZI F., HUANG J.-B.: Reducing footskate in human motion reconstruction with ground contact constraints. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)* (March 2020). 2, 3

[ZYC*20b] ZOU Y., YANG J., CEYLAN D., ZHANG J., PERAZZI F., HUANG J.-B.: Reducing footskate in human motion reconstruction with ground contact constraints. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (2020), pp. 459–468. 2, 3