# Video-Based Estimation of Lower-Limb Joint Moments During Running Using Constrained Multi-View Keypoint Reconstruction

Gaju Shin, Jungsoo Kim, Yujin Yun

## Abstract

*Estimating lower-limb joint moments during running typically requires laboratory motion capture, ground reaction force measurement, and inverse dynamics. Although smartphone videos offer a scalable alternative, sparse 2D keypoints may not preserve the lower-limb geometry needed for biomechanically meaningful kinetic prediction. This study compares a Hossain-style 2D keypoint-based direct regression baseline with a constrained 3D keypoint-based pipeline using synchronized front and side videos. The proposed pathway lifts 2D keypoints into canonical 3D lower-limb keypoints, applies subject-specific segment-length constraints, and predicts stance-phase hip, knee, and ankle moment waveforms from derived kinematic features. Reference moments were generated from motion-capture-derived kinematics and measured ground reaction forces using OpenSim inverse dynamics. On a held-out test set, the best constrained 3D model achieved a mean per-step Pearson correlation of* $r = 0.948$ *and a mean NRMSE of* $12.5\%$*, outperforming the best sparse 2D baseline (*$r = 0.918$*,* $\mathit{NRMSE} = 14.2\%$*). These results suggest that anatomically constrained pseudo-3D keypoints can provide a useful intermediate representation for supervised video-based joint moment prediction.*

## 1. Introduction

Running is one of the most common forms of human locomotion and physical exercise, but it is also a highly repetitive task in which the lower limbs experience substantial mechanical demand. During each stance phase, the hip, knee, and ankle absorb impact, control body motion, and generate propulsion. These joint-level demands are commonly quantified using joint moments, joint power, and joint work, which describe the net rotational demand and mechanical energy exchange at each joint [4, 5]. Unlike spatiotemporal measures such as cadence, step length, or contact time, joint moments provide a mechanistic description of how mechanical load is distributed across the lower limb. This distinction is important because runners with similar global gait patterns may still exhibit different knee extensor moments, ankle plantarflexor moments, or hip moment strategies.

The standard approach for estimating lower-limb joint moments requires laboratory-based motion analysis. Marker-based motion-capture provides three-dimensional body kinematics, while force plates or an instrumented treadmill measure ground reaction forces (GRF) and center-of-pressure (COP). These signals are then processed through musculoskeletal modeling software such as OpenSim to obtain joint angles through inverse kinematics and joint moments through inverse dynamics [6, 7]. Although this workflow is widely used in biomechanics research, it requires expensive equipment, trained operators, controlled laboratory environments, and substantial pre- and post-processing. In practice, obtaining reliable reference moments also requires marker labeling, trajectory smoothing, force and center-of-pressure correction, filtering, multimodal synchronization, and gait-event detection before inverse dynamics can be performed. These constraints limit the scalability of joint-level running analysis in field, clinical, athletic, and consumer settings.

Smartphone and RGB video-based approaches offer a practical alternative because cameras are inexpensive, portable, and easy to deploy. Recent markerless and smartphone-based systems have shown that video can support musculoskeletal kinematics and dynamics estimation [8–10]. However, estimating joint moments from video remains more difficult than detecting body keypoints. Standard pose-estimation metrics do not necessarily guarantee biomechanical accuracy, and sparse keypoint representations may be insufficient for OpenSim-based kinematic analysis [11]. This issue is particularly important during running, where the stance phase is short, external forces are large, and small errors in joint center location, segment orientation, foot motion, synchronization, or filtering can propagate into inverse dynamics [12–15].

A complementary line of work estimates kinetics directly from video-derived joint dynamics. Hossain et al. proposed a smartphone-video-based framework for estimating knee adduction moment, knee flexion moment, and three-dimensional GRFs from 2D joint center positions augmented with velocity and acceleration features [16]. This study provides an important baseline because it demonstrates that sparse 2D joint dynamics can support direct video-based kinetic estimation without reconstructing a full marker-based

representation [16]. However, direct 2D regression does not explicitly test whether the intermediate pose representation is biomechanically plausible or whether lower-limb segment geometry is consistent with subject-specific anatomy, even though such geometric consistency may affect moment prediction.

In this study, we compare a sparse 2D keypoint-based video-only baseline with a subject-specific 3D keypoint-based prediction pipeline for estimating lower-limb joint moments during treadmill running. The baseline directly maps video-derived 2D joint trajectories to joint moment waveforms, whereas the proposed pathway first reconstructs three-dimensional lower-limb keypoints from synchronized front and side videos and then uses the reconstructed 3D keypoint sequence for moment prediction. To reduce physically implausible lower-limb configurations, we incorporate participant-specific segment-length constraints derived from motion-capture data. The reference hip, knee, and ankle moment waveforms are computed from laboratory motion-capture kinematics and measured GRFs using inverse dynamics.

The main contribution of this study is a controlled comparison between sparse 2D keypoint-based moment regression and biomechanically constrained 3D keypoint-based moment regression for video-based running joint moment estimation. This comparison evaluates whether anatomically constrained 3D keypoint representations provide additional value over sparse 2D joint trajectories for predicting stance-phase lower-limb joint moments. In addition, the study uses participant-specific thigh and shank length constraints to reduce unrealistic hip–knee–ankle configurations before moment prediction. This structure allows the role of 3D reconstruction to be evaluated as an intermediate representation for video-based kinetic prediction, rather than as a direct replacement for laboratory inverse dynamics.

## 2. Prior Works

### 2.1. Laboratory-Based Inverse Dynamics

Lower-limb joint moments are traditionally estimated using marker-based motion capture, external force measurements, and musculoskeletal modeling. In this workflow, a participant-specific model is scaled, inverse kinematics estimates joint angles from marker trajectories, and inverse dynamics computes net joint moments from kinematics and GRF/COP data [6, 7]. This laboratory workflow provides the reference standard for many biomechanics studies, but it requires specialized equipment, calibration, synchronization, and manual preprocessing. These requirements limit the scalability of joint-level running analysis and motivate video-based alternatives.

### 2.2. Video Pose Estimation and Markerless Biomechanics

Human pose estimation provides the visual foundation for markerless biomechanics, but image-space keypoint accuracy does not directly imply biomechanical validity. Biomechanical analysis requires anatomically meaningful joint centers, segment orientations, temporal consistency, and compatibility with musculoskeletal models. OpenCapBench highlights this gap by showing that sparse keypoints may be insufficient for accurate OpenSim-based kinematic analysis [11]. Multi-view and smartphone-based systems such as Pose2Sim, OpenCap, and smartphone-driven musculoskeletal modeling workflows have demonstrated that video can support accessible biomechanical analysis [8–10, 20]. However, validation studies also show that markerless and marker-based systems can differ in lower-limb joint moments and powers during treadmill running, with errors that may depend on running speed [12, 13]. These findings motivate careful evaluation of video-derived representations before using them for joint moment prediction.

### 2.3. Learning-Based Joint Moment Prediction

Learning-based methods estimate joint moments by mapping motion features directly to kinetic outputs. Prior work has shown that kinematic features can contain predictive information about running joint moment waveforms [24]. Hossain et al. recently proposed a smartphone-video-based method for estimating knee moments and three-dimensional GRFs from video-derived 2D joint center dynamics [16]. This work is closely related to our baseline because it demonstrates the feasibility of sparse 2D joint-based kinetic estimation. However, direct 2D regression does not explicitly reconstruct subject-specific 3D lower-limb geometry or enforce anatomically plausible hip–knee–ankle configurations. The remaining question is therefore whether biomechanically constrained 3D keypoint reconstruction provides a more informative intermediate representation for joint moment regression than sparse 2D keypoint trajectories.

## 3. Method

### 3.1. Overview

The proposed framework estimates lower-limb joint moment waveforms during treadmill running from synchronized front and side videos. The overall analysis consists of two video-based prediction pathways evaluated against the same laboratory reference: a sparse 2D keypoint-based direct regression baseline and a constrained 3D keypoint-based prediction pipeline. The sparse 2D baseline is a video-only direct regression model adapted from Hossain et al. [16]; it directly estimates joint moment waveforms from video-derived 2D joint trajectories. The 3D pathway first reconstructs lower-limb keypoints from synchronized front and side videos,

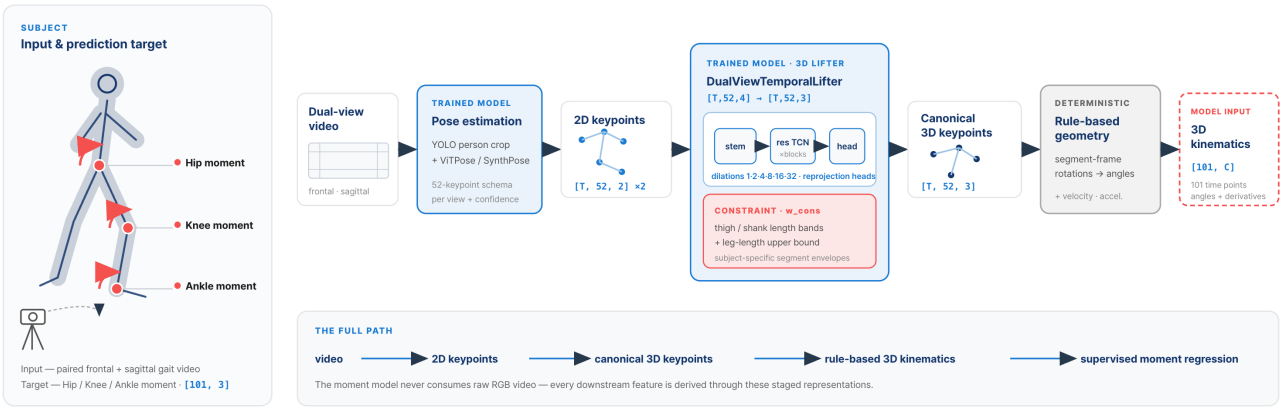



Figure 1. Overview of the proposed pipeline. 2-stage approach with 3D reconstruction by exploiting constrained 3D lifter.

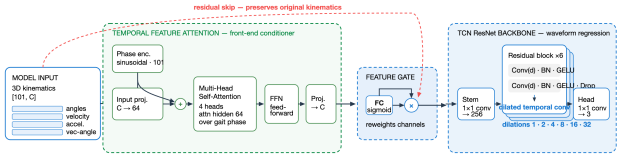



Figure 2. Architecture of joint moment prediction model. Inputs are calculated 3d kinematics and corresponding outputs are joint moment of right ankle, knee, and hip.

applies subject-specific segment-length constraints to reduce physically implausible configurations, and then uses the reconstructed 3D keypoint sequence as the input representation for joint moment prediction.

The primary output variables are stance-phase hip, knee, and ankle joint moment waveforms. The laboratory reference moments are computed from motion-capture-derived kinematics and measured GRF and COP data using OpenSim inverse dynamics [6, 7]. Both video-based pathways are trained and evaluated against these inverse-dynamics-derived reference moment waveforms. This design isolates the effect of the input representation by comparing sparse 2D keypoints with constrained 3D keypoints under the same dataset, target moment definitions, and stance-phase segmentation protocol.

### 3.2. Dataset and Reference Data Preprocessing

The dataset was collected in a treadmill-running laboratory environment using synchronized front and side videos, three-dimensional motion-capture marker trajectories, and ground reaction force and center-of-pressure data from an instrumented treadmill. The front and side cameras captured the runner from complementary viewpoints, while the motion-capture and force systems provided the laboratory reference signals for kinetic analysis.

The dataset included 15 healthy young adult participants, consisting of 10 males and 5 females. Each participant completed a warm-up at a preferred running speed and then performed continuous treadmill running at an individual speed. Running speeds ranged approximately from 1.8 m/s to 3.4 m/s. After data quality inspection, approximately 15 minutes of video and biomechanical data were selected per participant, corresponding to approximately 1,500 detected steps per participant before sample-level quality filtering.

The preprocessing workflow was designed to generate reliable laboratory-reference moment waveforms and to align them with the video-derived keypoint sequences. First, motion-capture marker trajectories were manually labeled and inspected to correct mislabeled or missing markers. Marker trajectories were then smoothed to reduce high-frequency noise while preserving running-related movement patterns. Ground reaction force and center-of-pressure data were also inspected and corrected before filtering, because force and COP artifacts can directly affect inverse-dynamics-derived joint moments. A fourth-order Butterworth low-pass filter was applied to the ground reaction force data.

Heel-strike and toe-off events were identified using a 20 N vertical ground reaction force threshold, and each trial was segmented into stance-phase steps. Synchronization between the video, motion-capture, and force data was performed using a visible reflective marker event at the beginning of data collection. The resulting timing offset was used to align the video frames with the motion-capture and force signals before stance-phase segmentation. This preprocessing stage was necessary because errors in marker labeling, trajectory smoothing, COP correction, filtering, or synchronization can propagate into the reference moment waveforms used to train and evaluate the video-based prediction models.

### 3.3. Reference Moment Generation

After marker labeling, trajectory smoothing, force/COP correction, filtering, synchronization, and stance segmentation, the laboratory reference joint moments were computed from

motion-capture-derived kinematics and measured ground reaction force and center-of-pressure data. OpenSim was used as the biomechanical analysis environment for generating the reference lower-limb joint moments [6, 7]. A generic musculoskeletal model was scaled to each participant to match individual anthropometry, including body size, segment lengths, joint center locations, and marker positions [6, 7]. A Rajagopal-family full-body musculoskeletal model was used as the anatomical model structure for gait-related lower-limb analysis [17].

After model scaling, inverse kinematics estimated the joint angles of the scaled musculoskeletal model by minimizing the discrepancy between experimental marker trajectories and model marker positions. Inverse dynamics then combined the estimated joint kinematics with measured ground reaction force and center-of-pressure data to compute net hip, knee, and ankle joint moments. These inverse-dynamics-derived moment waveforms served as the reference targets for both video-based prediction pathways. All reference waveforms were segmented by stance phase using ground-reaction-force-based gait events and time-normalized from heel-strike to toe-off.

### 3.4. Sparse 2D Keypoint-Based Baseline

We adapted the temporal-spatial modeling architecture of Hossain et al. [16] as a sparse 2D baseline that directly predicts joint moment waveforms from video-derived 2D keypoint trajectories without intermediate 3D reconstruction. The input is a time-normalized keypoint sequence $\mathbf{x} \in \mathbb{R}^{T\times J\times C}$, where $T = 101$, $J = 52$ (ViTPose keypoints), and $C = 4$ for the two-view setting (frontal and sagittal $(x, y)$ coordinates concatenated per joint) or $C = 2$ for single-view ablations. Within the model, finite-difference velocity and acceleration are derived from the position sequence, yielding three modalities (position, velocity, acceleration) each normalized by a dedicated batch-normalization layer.

Each modality is independently encoded by two parallel branches: a two-layer bidirectional LSTM (temporal branch, 128 hidden units per direction in the first layer and 64 in the second) and an attention-weighted graph convolutional network over the $J$ joints (spatial branch). The temporal and spatial features across all three modalities are fused by a single shared sigmoid gate:

$$\boldsymbol{\alpha} = \sigma(W_g\,[\mathbf{f}^p; \mathbf{f}^v; \mathbf{f}^a; \mathbf{g}^p; \mathbf{g}^v; \mathbf{g}^a])\,, \tag{1}$$

$$[\mathbf{h}^p; \mathbf{h}^v; \mathbf{h}^a] = \boldsymbol{\alpha}\odot[\mathbf{f}^p; \mathbf{f}^v; \mathbf{f}^a] + (1-\boldsymbol{\alpha})\odot[\mathbf{g}^p; \mathbf{g}^v; \mathbf{g}^a]\,, \tag{2}$$

where $\mathbf{f}^m$ and $\mathbf{g}^m$ denote the temporal and spatial features for modality $m \in \{p, v, a\}$. The resulting $T \times 384$ representation is refined by a Multi-Fusion Module (MFM) with three parallel paths—multi-head self-attention (MHAF), weighted feature fusion (WFF), and tensor multiplication (TMF)—combined by a late weighted fusion layer and decoded to the three-joint moment waveform by a fully connected layer.

Three adaptations were made relative to the original framework. First, we used 52 ViTPose keypoints instead of 11 OpenPose joints, and a 101-frame stance-phase window instead of the original 50-frame sliding window. Second, keypoints were normalized per step using the median body center and 90th-percentile scale, rather than subject-height normalization. Third, the knowledge-transfer stage requiring synchronized IMU signals was excluded due to data unavailability; the model was instead trained end-to-end on force-plate-derived labels using RMSE loss with Adam ($\mathrm{lr} = 10^{-3}$, batch size 64, 80 epochs). Full architecture equations and implementation details are provided in Appendix B.1.

### 3.5. Multi-View 3D Keypoint Reconstruction and Moment Prediction

The main pipeline reconstructed a canonical 3D pose sequence from synchronized frontal and sagittal videos and then derived biomechanical kinematic features before joint moment prediction. The pipeline consisted of four sequential stages: 2D keypoint preprocessing, weakly supervised dual-view 3D lifting, deterministic segment-kinematic computation, and supervised joint moment regression.

**2D keypoint preprocessing.** For each trial, 52 anatomical keypoints were extracted independently from the frontal and sagittal video streams. Each view was normalized separately to reduce the effects of camera placement and scale differences. Keypoints with confidence values below 0.2 were flagged as invalid, and their coordinates were filled using the median of valid detections. The body center $\boldsymbol{\mu}$ was defined as the median of all valid keypoint coordinates, and the scale $s$ was computed as the 90th percentile of Euclidean distances from $\boldsymbol{\mu}$. The normalized coordinates for each view were computed as $\tilde{\mathbf{u}} = \frac{\mathbf{u}-\boldsymbol{\mu}}{s}$.

The frontal and sagittal sequences were truncated to the minimum of the two available lengths and concatenated along the coordinate dimension: $\mathbf{x} = \left[\tilde{\mathbf{u}}^f; \tilde{\mathbf{u}}^s\right] \in \mathbb{R}^{T\times J\times 4}$.

**Dual-view temporal lifting.** The 3D reconstruction module was implemented as a learned temporal lifting network without camera calibration or geometric triangulation. The model took $\mathbf{x}$ as input and predicted canonical 3D joint coordinates,

$$\hat{\mathbf{X}} \in \mathbb{R}^{T\times J\times 3}. \tag{3}$$

The input was first flattened across the joint and coordinate dimensions and projected by a $1\times1$ temporal convolutional stem. The resulting sequence was processed by $B$ residual temporal convolution blocks with dilations cycling through $\{1, 2, 4, 8, 16, 32\}$, followed by a final $1\times1$ convolution that

predicted 3D coordinates. The output was expressed in a root-relative coordinate system:

$$\hat{\mathbf{X}}_t \leftarrow \hat{\mathbf{X}}_t - \hat{\mathbf{X}}_t^{(0)}, \tag{4}$$

where $\hat{\mathbf{X}}_t^{(0)}$ denotes the root joint at frame $t$.

Because calibrated 3D ground truth was unavailable, the lifter was trained using weak supervision. Two learnable projection heads mapped the predicted 3D coordinates back to each camera plane. The frontal head selected $(X, Y)$, and the sagittal head selected $(Z, Y)$, with both heads initialized from these geometric assumptions. The training objective was defined as

$$\mathcal{L}_{\text{lift}} = \mathcal{L}_{\text{repr}}^{f} + \mathcal{L}_{\text{repr}}^{s} + \lambda_{\text{sm}}\mathcal{L}_{\text{sm}} + \lambda_{\text{bone}}\mathcal{L}_{\text{bone}}, \tag{5}$$

where $\mathcal{L}_{\text{repr}}^{f}$ and $\mathcal{L}_{\text{repr}}^{s}$ are masked L1 reprojection losses computed only on keypoints whose confidence exceeded the threshold in both views. The smoothness term $\mathcal{L}_{\text{sm}}$ penalized the mean absolute second-order temporal differences of the predicted 3D joints to suppress jitter, and the bone-consistency term $\mathcal{L}_{\text{bone}}$ penalized within-window variability in bone lengths. Both regularization weights were set to $\lambda_{\text{sm}} = \lambda_{\text{bone}} = 0.05$.

**Subject-specific segment-length constraints.** In the constrained configuration, denoted as *w_cons*, subject-specific thigh and shank length envelopes derived from motion-capture data were incorporated into the lifting objective. For each segment $s \in \{\text{thigh}, \text{shank}\}$, with predicted length $\ell_s(t)$ and participant-specific plausible range $[\ell_s^{\min}, \ell_s^{\max}]$, a soft band penalty was applied:

$$\mathcal{L}_{\text{seg}} = \sum_{t,s} \left[ \max\big(0, \ell_s^{\min} - \ell_s(t)\big)^2 + \max(0, \ell_s(t) - \ell_s^{\max})^2 \right]. \tag{6}$$

Total leg length, defined from ASIS to ankle, was additionally constrained using a one-sided upper-bound penalty:

$$\max\big(0, \ell_{\text{leg}}(t) - \ell_{\text{leg}}^{\max}\big)^2. \tag{7}$$

These constraints were applied only during lifter training with a weight of 0.1. The downstream moment predictor remained a standard supervised regression model.

**Kinematic feature extraction.** The 3D keypoint sequence was smoothed using a seven-frame moving-average filter. Anatomical segment frames were then constructed for the pelvis, bilateral thighs, shanks, and feet using landmarks including ASIS, PSIS, medial and lateral knee, medial and lateral ankle, toe, metatarsal, and calcaneus points. For each joint $j \in \{\text{hip}, \text{knee}, \text{ankle}\}$ on both sides, the relative rotation from the parent segment frame to the child segment frame was converted to XYZ Euler angles $\boldsymbol{\theta}_j(t) \in \mathbb{R}^3$ in degrees, with phase unwrapping applied before conversion. Angular velocity and angular acceleration were obtained as temporal gradients:

$$\dot{\boldsymbol{\theta}}_j = \nabla_t \boldsymbol{\theta}_j, \qquad \ddot{\boldsymbol{\theta}}_j = \nabla_t \dot{\boldsymbol{\theta}}_j. \tag{8}$$

Three vector-angle features were additionally computed per side from segment direction vectors at the hip, knee, and ankle. The complete feature vector contained 18 Euler-angle channels, 18 angular-velocity channels, 18 angular-acceleration channels, 6 vector-angle channels, and 2 metadata channels for stance side and step duration, yielding 62 channels in total.

**Joint moment prediction.** For each annotated gait step, the kinematic feature sequence was sliced using the recorded start and end frames and resampled to 101 normalized time points by linear interpolation. Both the input features and the target moments

$$\mathbf{x} \in \mathbb{R}^{101\times 62}, \quad \mathbf{y} \in \mathbb{R}^{101\times 3}, \tag{9}$$

were z-scored using training-set statistics.

The moment predictor first applied a temporal attention module. Each of the 101 gait-phase samples was treated as a token, projected into a lower-dimensional space, and augmented with sinusoidal positional encodings. Multi-head self-attention was applied across the temporal dimension, followed by a feedforward block with a residual connection and layer normalization. The attention output was projected back to the original 62-channel dimension and modulated by a learned sigmoid feature gate:

$$\tilde{\mathbf{x}} = \mathbf{x}_{\text{attn}} \odot \left(1 + \sigma(W_{\text{gate}}\mathbf{x}_{\text{attn}})\right), \tag{10}$$

where $\mathbf{x}_{\text{attn}}$ denotes the residual-normalized attention output. The gated sequence was processed by a dilated temporal convolutional residual network with 6 blocks and hidden dimension 256, with dilations cycling through $\{1, 2, 4, 8, 16, 32\}$, followed by a $1\times1$ convolution that output the three-joint moment waveform. The model was trained using MSE loss on normalized targets with Adam, a learning rate of $10^{-3}$, weight decay of $10^{-3}$, and batch size 128 for 100 epochs. The checkpoint with the lowest validation MSE was selected, and predictions were denormalized before reporting.

# 4. Experiment

## 4.1. Baselines and our models

All models were trained to predict stance-phase lower-limb joint moment waveforms from video-derived motion representations. For each gait step, the target was a 101-point time-normalized waveform for the hip, knee, and ankle, i.e. $\mathbb{R}^{101\times3}$. The same subject split was used across all experiments: SUB04 and SUB06–SUB12 for training, SUB13–SUB15 for validation, and SUB01–SUB03 for testing. The

Table 1. Test performance across models on SUB01–SUB03. NRMSE is reported as a percentage.

| | NRMSE (%) ↓ | | | | $r$ ↑ | | | |
|---|---|---|---|---|---|---|---|---|
| Model | Mean | Hip | Knee | Ankle | Mean | Hip | Knee | Ankle |
| Sparse 2D baseline | 14.17 | 17.87 | **11.03** | 13.62 | 0.918 | 0.802 | **0.987** | 0.965 |
| 3D reconstruction + kinematics + MLP | 16.21 | 18.61 | 17.65 | 12.37 | 0.865 | 0.764 | 0.910 | 0.921 |
| 3D reconstruction + kinematics + DLinear | 19.03 | 18.62 | 22.70 | 15.77 | 0.750 | 0.738 | 0.760 | 0.752 |
| 3D reconstruction + kinematics + TCN | 12.94 | 16.43 | 13.99 | 8.41 | 0.931 | 0.836 | 0.981 | 0.975 |
| 3D reconstruction + kinematics + TCN + Attention | **12.49** | **16.77** | 12.53 | **8.18** | **0.948** | **0.883** | 0.982 | **0.978** |

test set contained 917 gait steps. Model checkpoints were selected using the validation loss, and test predictions were generated after reloading the selected checkpoint. Details of implementation are described in Appendix.

**Sparse 2D Baseline** We implemented a sparse 2D keypoint-based baseline by adapting the spatio-temporal GCN and TCN backbone of Hossain et al. [16] to our joint-moment prediction task. Each joint was represented by a four-dimensional feature vector concatenating the $(x, y)$ coordinates from the frontal and sagittal cameras.

**Our 3D reconstruction and kinematics models** Our main models used a staged video-to-kinematics-to-moment pipeline. Frontal and sagittal 2D keypoints were first lifted into canonical 3D keypoints using a dual-view temporal lifter.

We evaluated four model families on this 3D kinematics representation. **The MLP baseline** flattened the full kinematic sequence and predicted the complete moment waveform using fully connected layers. **DLinear** provided a simple time-series linear baseline that projected along the temporal axis and then mapped kinematic features to the three joint-moment channels. The **TCN model** used residual dilated one-dimensional temporal convolution blocks to model local and multi-scale temporal structure over the normalized stance phase. Finally, the **TCN+Attention** model added a temporal feature-attention module before the TCN backbone, allowing the model to reweight kinematic information across the gait cycle before waveform prediction.

**Metrics** We report two complementary metrics evaluated per joint and averaged across hip, knee, and ankle.

**Step-wise Pearson correlation** $r$ measures waveform shape similarity independently for each gait step and averages over all $N$ steps:

$$r = \frac{1}{N}\sum_{s=1}^{N} \frac{\sum_t (\hat{y}_{s,t} - \bar{\hat{y}}_s)(y_{s,t} - \bar{y}_s)}{\sqrt{\sum_t (\hat{y}_{s,t} - \bar{\hat{y}}_s)^2 \sum_t (y_{s,t} - \bar{y}_s)^2}}, \quad (11)$$

where $\hat{y}_{s,t}$ and $y_{s,t}$ denote the predicted and ground-truth moment at normalized time point $t$ of step $s$.

**Normalized RMSE** (NRMSE) expresses prediction error as a percentage of the ground-truth waveform range:

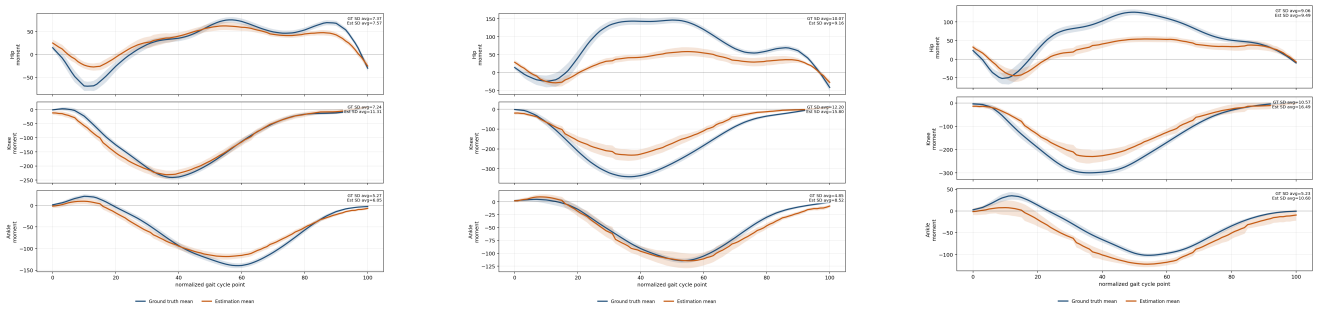
$$\mathrm{NRMSE} = \frac{\sqrt{\frac{1}{NT}\sum_{s,t}(\hat{y}_{s,t} - y_{s,t})^2}}{\max_{s,t} y_{s,t} - \min_{s,t} y_{s,t}} \times 100\%. \quad (12)$$

## 4.2. Experimental Results

**Main Results** Table 1 summarizes the test performance on SUB01–SUB03. Among the 3D reconstruction-based models, temporal modeling was important for joint moment waveform prediction. The MLP baseline achieved a mean $r$ of 0.865 and a mean NRMSE of 16.21%, while DLinear showed lower correlation and higher error, with a mean $r$ of 0.750 and a mean NRMSE of 19.03%. The TCN model substantially improved performance, reaching a mean $r$ of 0.931 and a mean NRMSE of 12.94%. Adding temporal feature attention further improved the best overall performance, with a mean $r$ of 0.948 and a mean NRMSE of 12.49%. This model corresponds to the constrained-lifter input pipeline with Temporal Feature Attention TCN, which was also the best model by mean correlation in the combined model summary.

Compared with the sparse 2D keypoint baseline, the proposed 3D reconstruction + kinematics + TCN+Attention model achieved the strongest overall correlation and the lowest mean error. These results suggest that the constrained 3D kinematic representation provides useful information beyond sparse 2D keypoint dynamics, especially when combined with a temporal attention and TCN regression backbone.

Across joints, knee and ankle moments were predicted most consistently. The best model achieved knee and ankle $r$ values of 0.982 and 0.978, respectively, with NRMSE values of 12.53% and 8.18%. Hip prediction remained more difficult, with a lower $r$ of 0.883 and a higher NRMSE of 16.77%. This joint-wise pattern was also visible in the subject-level waveform plot in Fig. 3: knee and ankle waveforms generally followed the reference mean curves closely, whereas the hip showed larger subject-dependent deviations and wider variability.



(a) Sub.01 (b) Sub.02 (c) Sub.03

Figure 3. Waveform plot of test subjects. Blue lines refer to ground truth wave, while the orange lines correspond to estimated wave.

Table 2. Subject-level comparison of the best models without and with constraint inputs. NRMSE is reported in percentage.

| | | NRMSE (%) ↓ | | | | $r$ ↑ | | | |
|---|---|---|---|---|---|---|---|---|---|
| Setting | Subj. | Mean | Hip | Knee | Ankle | Mean | Hip | Knee | Ankle |
| w/o cons | SUB01 | 8.61 | 10.35 | 7.82 | 7.65 | 0.960 | 0.922 | 0.982 | 0.975 |
| w/o cons | SUB02 | 17.61 | 22.52 | 19.18 | 11.12 | 0.920 | 0.804 | 0.979 | 0.978 |
| w/o cons | SUB03 | 17.31 | 20.96 | 16.10 | 14.88 | 0.892 | 0.724 | 0.982 | 0.969 |
| w/ cons | SUB01 | 8.50 | 10.55 | 7.08 | 7.87 | 0.966 | 0.938 | 0.981 | 0.980 |
| w/ cons | SUB02 | 16.66 | 24.06 | 17.11 | 8.81 | 0.938 | 0.852 | 0.982 | 0.981 |
| w/ cons | SUB03 | 16.40 | 18.66 | 14.62 | 15.91 | 0.928 | 0.831 | 0.986 | 0.967 |

**Effect of leg-length constraints.** We further examined whether subject-specific leg-length constraints improved the downstream moment prediction pipeline. These constraints were applied during the 3D lifter stage rather than directly in the final moment-regression model. For each subject, the running trials were used to estimate thigh and shank length envelopes from the 5th and 95th percentiles, while the full ASIS-to-ankle leg length was used as a one-sided upper bound. We designed the lifter to penalize implausible thigh and shank lengths outside the subject-specific bands, while preventing unrealistically over-extended leg configurations.

Table 2 compares the best model trained without these constraints against the best model trained with constrained-lifter kinematics. It demonstrates that the improvement was not limited to a single test subject. Mean $r$ increased for SUB01, SUB02, and SUB03, and mean NRMSE decreased for all three subjects. The reduction was most visible for SUB02 and SUB03, where the constrained model reduced mean NRMSE from 17.61% to 16.66% and from 17.31% to 16.40%, respectively. Joint-wise, knee and ankle correlations were already high in the unconstrained setting, whereas the constrained model provided clearer gains for the hip correlation. This suggests that leg-length constraints mainly improve the more geometry-sensitive components of the 3D kinematic representation, while preserving the already strong knee and ankle waveform estimates.

# 5. Discussion

## 5.1. Main Finding: Constrained 3D Pose as a Biomechanical Intermediate Representation

The main finding of this study is that anatomically constrained 3D keypoint reconstruction can serve as a more informative intermediate representation for video-based running joint moment prediction than sparse 2D keypoint trajectories alone. The best constrained 3D model achieved a mean per-step Pearson correlation of $r = 0.948$ and a mean NRMSE of $12.5\%$, whereas the best sparse 2D baseline achieved $r = 0.918$ and an NRMSE of $14.2\%$. Although the absolute improvement is moderate, its biomechanical meaning is important: preserving lower-limb geometry before moment regression appears to provide additional information beyond raw image-plane joint motion.

This contribution differs from prior markerless biomechanics pipelines and direct 2D regression approaches. Markerless systems such as Pose2Sim and OpenCap have shown that video can support accessible biomechanical analysis through multi-view reconstruction and musculoskeletal modeling [8, 9, 20], while smartphone-based workflows have extended this idea toward musculoskeletal dynamics [10]. However, these studies primarily focus on building or validating end-to-end markerless biomechanics pipelines. In contrast, our study asks a representation-level question: when predicting inverse-dynamics-derived running joint moments from video, is sparse 2D motion sufficient, or does a constrained 3D keypoint representation provide additional value?

The comparison with the 2D video-only baseline is central to this question [16]. Direct 2D regression demonstrates that video-derived joint dynamics can contain useful kinetic information, but sparse 2D trajectories do not explicitly encode three-dimensional lower-limb geometry, subject-specific segment lengths, or anatomically plausible hip–knee–ankle configurations. Our results suggest that 3D reconstruction is useful not because it replaces laboratory inverse dynamics, but because it organizes video-derived motion into a more biomechanically structured representation before su-

pervised moment prediction.

### 5.2. Biomechanical Interpretation of Joint-Specific Results

The joint-specific results suggest that the value of constrained 3D representation depends on the biomechanical role of each joint. Ankle and knee moments were predicted accurately under both 2D and 3D approaches, likely because their stance-phase waveforms are strongly related to visually prominent sagittal-plane motion, including shank rotation, knee flexion-extension, foot progression, and push-off timing. The 3D pipeline nevertheless reduced ankle NRMSE, suggesting that segment-level kinematic features can improve waveform scaling even when timing is already well captured.

Hip moment prediction was the most difficult task and showed the clearest representational benefit from the 3D pipeline. This is biomechanically plausible because hip moments depend not only on local thigh motion, but also on pelvis orientation, trunk posture, proximal coordination, and whole-body center-of-mass control. These factors are difficult to infer from sparse 2D keypoints alone. The stronger improvement at the hip therefore supports the interpretation that constrained 3D reconstruction is most useful when joint kinetics depend on multi-segment coordination rather than easily visible local joint motion.

### 5.3. Role of Anatomical Constraints and Multi-View Representation

The segment-length constraint results support the importance of anatomical plausibility in video-based kinetic prediction. The best unconstrained 3D model from earlier experimental groups achieved a mean correlation of approximately $r = 0.860$, whereas the constraint-tuned configuration, consistently reached correlations above $r = 0.93$. This indicates that 3D reconstruction alone is not sufficient; the intermediate pose representation must also be biomechanically plausible.

From a biomechanics perspective, thigh and shank lengths should remain approximately constant within a subject. If a 3D lifter produces unrealistic frame-to-frame changes in hip–knee–ankle distance, the resulting segment orientations, Euler angles, and angular velocities become physically inconsistent. These errors can propagate into the downstream moment predictor even when the reconstructed pose appears visually reasonable. The subject-specific soft band constraints reduced such implausible variations and regularized the pose representation before moment regression.

### 5.4. Limitations and Future Directions

Several limitations should be considered. The 3D reconstruction was performed without camera calibration or triangulation, so the derived kinematic features represent a learned intermediate representation rather than metrically calibrated joint kinematics; their quality was assessed only through downstream moment prediction performance, as calibrated 3D annotations corresponding to the video-derived 52-keypoint representation were not available. In addition, the subject-specific leg-length constraints used in this study were derived from motion-capture data. While this design allowed us to test whether anatomically plausible segment geometry improves moment prediction, it also limits direct deployment in fully video-only settings. The sparse 2D baseline also excluded the IMU-based knowledge-transfer stage of the original Hossain et al. framework due to the absence of synchronized IMU data. Finally, the evaluation relied on a fixed three-subject test set, which limits conclusions about population-level generalization.

Future work should examine whether the benefits of constrained 3D reconstruction generalize across subjects, speeds, running styles, and camera configurations using leave-one-subject-out or repeated-split protocols. Datasets that include calibrated 3D keypoint ground truth or motion-capture-derived kinematics would allow direct evaluation of the intermediate representation rather than relying solely on downstream kinetic performance. Additional work is also needed to replace motion-capture-derived segment-length constraints with deployable alternatives, such as video-estimated limb lengths, short static calibration trials, or population-average anthropometric models. Richer biomechanical supervision, such as subject-specific anthropometric priors, ground reaction force, or joint range-of-motion constraints, may further improve the plausibility of the reconstructed pose and the accuracy of the predicted moments.

## References


[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[2] A. Zeng, M. Chen, L. Zhang, and Q. Xu. Are transformers effective for time series forecasting? In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023.

[3] C. Lea, M. D. Flynn, R. Vidal, A. Reiter, and G. D. Hager. Temporal convolutional networks for action segmentation and detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 156–165, 2017.

[4] D. A. Winter. Moments of force and mechanical power in jogging. *Journal of Biomechanics*, 16(1):91–97, 1983. https://doi.org/10.1016/0021-9290(83)90050-7. 1

[5] T. F. Novacheck. The biomechanics of running. *Gait & Posture*, 7(1):77–95, 1998. https://doi.org/10.1016/S0966-6362(97)00038-6. 1

[6] S. L. Delp, F. C. Anderson, A. S. Arnold, P. Loan, A. Habib, C. T. John, E. Guendelman, and D. G. Thelen. OpenSim: Open-source software to create and analyze dynamic simu-

lations of movement. *IEEE Transactions on Biomedical Engineering*, 54(11):1940–1950, 2007. https://doi.org/10.1109/TBME.2007.901024. 1, 2, 3, 4, 10

[7] A. Seth, J. L. Hicks, T. K. Uchida, A. Habib, C. L. Dembia, J. J. Dunne, C. F. Ong, M. S. DeMers, A. Rajagopal, M. Millard, S. R. Hamner, E. M. Arnold, J. R. Yong, S. K. Lakshmikanth, M. A. Sherman, J. P. Ku, and S. L. Delp. OpenSim: Simulating musculoskeletal dynamics and neuromuscular control to study human and animal movement. *PLOS Computational Biology*, 14(7):e1006223, 2018. https://doi.org/10.1371/journal.pcbi.1006223. 1, 2, 3, 4, 10

[8] D. Pagnon, M. Domalain, and L. Reveret. Pose2Sim: An end-to-end workflow for 3D markerless sports kinematics—Part 1: Robustness. *Sensors*, 21(19):6530, 2021. https://doi.org/10.3390/s21196530. 1, 2, 7, 10

[9] S. D. Uhlrich, A. Falisse, Ł. Kidziński, J. Muccini, M. Ko, A. S. Chaudhari, J. L. Hicks, and S. L. Delp. OpenCap: Human movement dynamics from smartphone videos. *PLOS Computational Biology*, 19(10):e1011462, 2023. https://doi.org/10.1371/journal.pcbi.1011462. 7, 10

[10] Y. Peng, W. Wang, L. Wang, H. Zhou, Z. Chen, Q. Zhang, and G. Li. Smartphone videos-driven musculoskeletal multibody dynamics modelling workflow to estimate the lower limb joint contact forces and ground reaction forces. *Medical & Biological Engineering & Computing*, 62:3841–3853, 2024. https://doi.org/10.1007/s11517-024-03171-3. 1, 2, 7, 10

[11] Y. Gozlan, A. Falisse, S. Uhlrich, A. Gatti, M. J. Black, J. Hicks, S. Delp, and A. S. Chaudhari. OpenCapBench: A benchmark to bridge pose estimation and biomechanics. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4056–4065, 2025. 1, 2

[12] H. C. Tang, J.-W. Pan, B. A. Munkasy, K. Duffy, and L. Li. Comparison of lower extremity joint moment and power estimated by markerless and marker-based systems during treadmill running. *Bioengineering*, 9(10):574, 2022. https://doi.org/10.3390/bioengineering9100574. 1, 2, 10

[13] H. C. Tang, B. A. Munkasy, and L. Li. Differences between lower extremity joint running kinetics captured by marker-based and markerless systems were speed dependent. *Journal of Sport and Health Science*, 13(4):569–578, 2024. https://doi.org/10.1016/j.jshs.2024.01.002. 2, 10

[14] P. Mai and S. Willwacher. Effects of low-pass filter combinations on lower extremity joint moments in distance running. *Journal of Biomechanics*, 95:109311, 2019. https://doi.org/10.1016/j.jbiomech.2019.08.005.

[15] E. Kristianslund, T. Krosshaug, and A. J. van den Bogert. Effect of low-pass filtering on joint moments from inverse dynamics: Implications for injury prevention. *Journal of Biomechanics*, 45(4):666–671, 2012. https://doi.org/10.1016/j.jbiomech.2011.12.011. 1

[16] M. S. B. Hossain, H. Choi, Z. Guo, S. Yoo, M.-K. Song, H. Shin, H. Shim, and J. Choi. Knowledge transfer-driven estimation of knee moments and ground reaction forces from smartphone videos via temporal-spatial modeling of augmented joint kinematics. *PLOS ONE*, 20(11):e0335257, 2025. https://doi.org/10.1371/journal.pone.0335257. 1, 2, 4, 6, 7

[17] A. Rajagopal, C. L. Dembia, M. S. DeMers, D. D. Delp, J. L. Hicks, and S. L. Delp. Full-body musculoskeletal model for muscle-driven simulation of human gait. *IEEE Transactions on Biomedical Engineering*, 63(10):2068–2079, 2016. https://doi.org/10.1109/TBME.2016.2586891. 4, 10

[18] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer, 2014. https://doi.org/10.1007/978-3-319-10602-1_48.

[19] Y. Xu, J. Zhang, Q. Zhang, and D. Tao. ViTPose: Simple vision transformer baselines for human pose estimation. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

[20] D. Pagnon, M. Domalain, and L. Reveret. Pose2Sim: An open-source Python package for multiview markerless kinematics. *Journal of Open Source Software*, 7(77):4362, 2022. https://doi.org/10.21105/joss.04362. 2, 7, 10

[21] Z. Zhang. A flexible new technique for camera calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(11):1330–1334, 2000. https://doi.org/10.1109/34.888718. 10

[22] K. Iskakov, E. Burkov, V. Lempitsky, and Y. Malkov. Learnable triangulation of human pose. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7718–7727, 2019. https://doi.org/10.1109/ICCV.2019.00781. 10

[23] K. Song, T. J. Hullfish, R. Scattone Silva, K. G. Silbernagel, and J. R. Baxter. Markerless motion capture estimates of lower extremity kinematics and kinetics are comparable to marker-based across 8 movements. *Journal of Biomechanics*, 157:111751, 2023. https://doi.org/10.1016/j.jbiomech.2023.111751. 10

[24] B. X. W. Liew, D. Rügamer, X. Zhai, Y. Wang, S. Morris, and K. Netto. Comparing shallow, deep, and transfer learning in predicting joint moments in running. *Journal of Biomechanics*, 129:110820, 2021. https://doi.org/10.1016/j.jbiomech.2021.110820. 2

## A. Related Works in Detail

**Laboratory inverse dynamics.** In laboratory biomechanics, reflective markers are attached to anatomical landmarks, three-dimensional marker trajectories are captured using optical motion-capture cameras, and GRFs are measured using force plates or an instrumented treadmill. A musculoskeletal model is scaled to the participant, inverse kinematics estimates joint angles by minimizing marker-position errors, and inverse dynamics computes net joint moments from joint kinematics and external force data [6, 7]. OpenSim is widely used for this workflow because it provides tools for musculoskeletal model scaling, inverse kinematics, inverse dynamics, and movement simulation. The Rajagopal full-body model provides an anatomically grounded model structure for gait-related lower-limb analysis [17].

**Markerless biomechanics systems.** Pose2Sim combines 2D pose estimation, camera calibration, triangulation, and OpenSim-based inverse kinematics into an end-to-end markerless kinematics workflow [8, 20]. Camera calibration estimates the geometric relationship between camera views and the world coordinate system [21], and triangulation reconstructs three-dimensional landmarks from corresponding two-dimensional detections [22]. OpenCap estimates human movement kinematics and dynamics from two or more smartphone videos by combining pose estimation, multi-view reconstruction, biomechanical modeling, and simulation [9]. Peng et al. proposed a dual smartphone video-driven musculoskeletal multibody dynamics workflow for estimating lower-limb mechanics, including joint contact forces and GRFs, during walking and running [10]. Markerless motion capture has been reported to produce lower-extremity kinematic and kinetic estimates comparable to marker-based motion capture across several movements [23], although treadmill-running validation studies show that markerless and marker-based estimates of joint moments and powers can differ and may be speed dependent [12, 13].

## B. Detailed empirical set-up

### B.1. Sparse 2D Keypoint-Based Baseline: Architecture and Adaptation

**Input representation.** For each gait step, 52 anatomical keypoints were extracted from the frontal and sagittal video streams using ViTPose. In the two-view setting, the $(x, y)$ coordinates from both views were concatenated per joint, yielding a four-dimensional feature vector $[x_f, y_f, x_s, y_s]$ per joint and a sequence $\mathbf{x} \in \mathbb{R}^{T\times J\times 4}$ with $T = 101$ and $J = 52$. Single-view ablations used $C = 2$. Step windows were extracted using gait-event timings and resampled to 101 normalized time points by linear interpolation. The original Hossain et al. used 11 OpenPose joints with a 50-frame sliding window; we used 52 ViTPose joints over a full 101-frame stance-phase representation.

**Augmented kinematics.** Within the model, finite-difference velocity and acceleration were computed from the flattened position vector $\mathbf{p}_t \in \mathbb{R}^{JC}$:

$$\mathbf{v}_t = \mathbf{p}_t - \mathbf{p}_{t-1}, \qquad \mathbf{a}_t = \mathbf{v}_{t+1} - \mathbf{v}_{t-1}, \tag{13}$$

with boundary values set to zero. Position $\mathbf{p}$, velocity $\mathbf{v}$, and acceleration $\mathbf{a}$ were treated as three independent modalities, each normalized by a dedicated batch-normalization layer.

**Temporal encoder (Bi-LSTM).** Each modality $m \in \{p, v, a\}$ was passed through a two-layer stacked bidirectional LSTM. The first layer had 128 hidden units per direction; the second layer had 64 hidden units per direction. The final bidirectional output was projected to produce a 128-dimensional temporal feature $\mathbf{f}^m \in \mathbb{R}^{T\times 128}$.

**Spatial encoder (Attention-weighted GCN).** In parallel, a graph convolutional network modeled spatial relationships among the $J$ joints. At each time step, an attention score was first computed per joint by a fully connected layer followed by softmax, then applied as a multiplicative weight before the graph convolution:

$$\mathbf{g}^m = \mathrm{ReLU}(A\,(a^m \odot V^m)\,W + b) \in \mathbb{R}^{T\times 128}, \tag{14}$$

where $A \in \mathbb{R}^{J\times J}$ is the adjacency matrix, $a^m$ is the softmax attention weight, and $V^m$ is the per-modality input reshaped to $(B \cdot T) \times J \times C$. The GCN output was projected and reshaped to match the Bi-LSTM output dimension.

**Gated spatio-temporal fusion.** The temporal and spatial features for all three modalities were concatenated and fused by a single shared sigmoid gate:

$$\boldsymbol{\alpha} = \sigma(W_g\,[\mathbf{f}^p; \mathbf{f}^v; \mathbf{f}^a; \mathbf{g}^p; \mathbf{g}^v; \mathbf{g}^a])\,, \tag{15}$$

$$[\mathbf{h}^p; \mathbf{h}^v; \mathbf{h}^a] = \boldsymbol{\alpha} \odot [\mathbf{f}^p; \mathbf{f}^v; \mathbf{f}^a] + (1 - \boldsymbol{\alpha}) \odot [\mathbf{g}^p; \mathbf{g}^v; \mathbf{g}^a]\,, \tag{16}$$

yielding a fused representation $[\mathbf{h}^p; \mathbf{h}^v; \mathbf{h}^a] \in \mathbb{R}^{T\times 384}$.

**Multi-Fusion Module (MFM).** The fused representation was passed through three parallel fusion paths and then combined by a late weighted fusion (LWF):

$$\mathbf{z}_{\mathrm{mha}} = \mathrm{MHA}([\mathbf{h}^p; \mathbf{h}^v; \mathbf{h}^a])\,, \tag{17}$$

$$\mathbf{z}_{\mathrm{wff}} = \sigma(W_f[\mathbf{h}^p; \mathbf{h}^v; \mathbf{h}^a]) \odot [\mathbf{h}^p; \mathbf{h}^v; \mathbf{h}^a], \tag{18}$$

$$\mathbf{z}_{\mathrm{tmf}} = \mathbf{h}^p \odot \mathbf{h}^v \odot \mathbf{h}^a. \tag{19}$$

The three outputs were concatenated, re-weighted by a sigmoid gate, and summed to produce the MFM output $\mathbf{z}_{\mathrm{mfm}}$.

MHA refers to multi-head self-attention; WFF applies a learned sigmoid importance weight per feature; TMF suppresses modality-specific noise by element-wise multiplication across modalities.

**Output decoder.** The MFM output was decoded by a two-layer MLP to produce the three-joint moment waveform $\mathbf{y} \in \mathbb{R}^{T\times 3}$ (hip, knee, ankle). Target moments were z-scored using training-set statistics and denormalized before evaluation.

**Training.** The model was optimized using RMSE loss on normalized targets with Adam, a learning rate of $10^{-3}$, batch size 64, and 80 epochs. The checkpoint with the lowest validation RMSE was selected.

**Adaptations from Hossain et al.** The following differences were made relative to the original framework.

- **Keypoint detector:** OpenPose (11 joints) $\rightarrow$ ViTPose (52 joints).
- **Input window:** $\Delta T = 50$ frames (sliding) $\rightarrow T = 101$ (one full stance phase, resampled).
- **Keypoint normalization:** Subject height + hip-midpoint origin $\rightarrow$ per-step median center and 90th-percentile scale.
- **Prediction targets:** KFM, KAM, 3D GRF $\rightarrow$ hip, knee, and ankle sagittal moments.
- **Knowledge transfer:** Teacher (IMU) $\rightarrow$ student (video) two-step training $\rightarrow$ excluded (no IMU available); replaced by end-to-end single-step RMSE training.
- **Evaluation:** Leave-one-subject-out (LOSO) $\rightarrow$ fixed split (Train: SUB04, SUB06–12; Val: SUB13–15; Test: SUB01–03).

### B.2. Implementation

All neural moment-prediction models were implemented in Python using PyTorch. The training code loaded the OpenSim-derived ground-truth waveform table and the precomputed 3D kinematics files, matched samples by gait-step identifier, and applied the fixed subject split described above. Input kinematic features were normalized using the training-set mean and standard deviation, and target moment waveforms were also normalized using training-set statistics. Predictions were denormalized before saving test predictions and computing final metrics.

The moment predictors were trained with Adam using MSE loss on the normalized $[101, 3]$ waveform target. Unless otherwise specified, the batch size was 128, the learning rate was $10^{-3}$, and the random seed was fixed to 42. Checkpoints were selected by the lowest validation MSE, and the reported test results were computed after reloading the selected checkpoint. Training outputs included the best and last checkpoints, training history, validation and test predictions, and JSON metric summaries. The experiments were run on a Linux SLURM server environment. GPU jobs were launched on nodes equipped with NVIDIA GeForce RTX 3090 GPUs with 24 GB of memory. The software environment used Linux 6.8, Python 3.13.5, and PyTorch 2.10.0+cu128.

### B.3. Hyperparameter tuning

Hyperparameter tuning was performed in stages. First, we trained the base waveform-regression models on the original 3D kinematics representation without segment-length constraints. This stage included MLP, DLinear, TCN, and an early temporal-feature-attention TCN model. The models used 80 epochs, batch size 128, learning rate $10^{-3}$, weight decay $10^{-4}$, and the same train/validation/test split.

The second tuning stage focused on stronger temporal convolutional models using the original, unconstrained 3D kinematics. We swept TCN hidden dimensions $\{192, 256\}$, residual block counts $\{6, 8\}$, and dropout values $\{0.2, 0.3\}$, with weight decay fixed at $10^{-3}$. All second-trial TCN variants used kernel size 3, BatchNorm, GELU activations, residual temporal blocks, dilations cycling through $\{1, 2, 4, 8, 16, 32\}$, and a $1\times1$ convolutional output head. A one-dimensional U-Net baseline was also trained as an additional waveform-to-waveform comparison.

Finally, we tuned the constrained-lifter setting using kinematic features generated from the segment-length-constrained 3D lifter. The best model by mean correlation was the Temporal Feature Attention TCN with hidden dimension 256, 6 TCN blocks, attention hidden dimension 64, 4 attention heads, dropout 0.2, learning rate $10^{-3}$, and weight decay $10^{-3}$. This model used temporal self-attention and phase encoding as a front-end feature conditioner, followed by the same residual TCN waveform-regression backbone. We selected this configuration for the main comparison because it achieved the highest mean correlation in the combined model summary.

## C. Detailed empricial results

### C.1. Full results

Here, we provide the results for the corresponding comparison with additional models trained without constraint inputs in Table 3. This broader comparison shows that the unconstrained 3D models were not uniformly better than the sparse 2D baseline. For example, the unconstrained MLP and DLinear models achieved mean $r$ values of 0.807 and 0.801, respectively, and the unconstrained temporal-feature-attention TCN reached only 0.815. These results indicate that adding model complexity alone was not sufficient when the upstream 3D kinematics were less constrained.

The strongest unconstrained model in the appendix was the TCN, which achieved the lowest mean NRMSE among the unconstrained 3D variants (12.79%) but a lower mean $r$ of 0.856. In contrast, the constrained TCN+Attention model

achieved a higher mean $r$ of 0.948 while maintaining a competitive mean NRMSE of 12.49%. The appendix also shows that the constrained setting did not uniformly improve every NRMSE entry; in some comparisons, the point-wise magnitude error was similar or slightly worse. However, the constrained models consistently improved the correlation-based metrics, which better reflect waveform shape and temporal trend agreement. Thus, the clearest benefit of the constraint-based pipeline was not simply minimizing point-wise error, but improving recovery of the overall joint-moment waveform pattern across the stance phase. This gap suggests that the final performance gain came from the combination of three factors: a biomechanically more plausible constrained 3D representation, temporal convolutional waveform modeling, and attention-based conditioning of the kinematic sequence. The appendix results therefore support the main conclusion that the intermediate representation quality is as important as the downstream regression architecture for video-based joint moment estimation.

Table 3. Full test performance on SUB01–03, including additional models without constraint inputs. NRMSE is reported in percentage.

| | NRMSE (%) ↓ | | | | $r$ ↑ | | | |
|---|---|---|---|---|---|---|---|---|
| Model | Mean | Hip | Knee | Ankle | Mean | Hip | Knee | Ankle |
| Sparse 2D Baseline | 14.17 | 17.87 | 11.03 | 13.62 | 0.918 | 0.802 | 0.987 | 0.965 |
| 3D reconstruction + kinematics + MLP | 16.21 | 18.61 | 17.65 | 12.37 | 0.865 | 0.764 | 0.910 | 0.921 |
| 3D reconstruction + kinematics + DLinear | 19.03 | 18.62 | 22.70 | 15.77 | 0.750 | 0.738 | 0.760 | 0.752 |
| 3D reconstruction + kinematics + TCN | 12.94 | 16.43 | 13.99 | 8.41 | 0.931 | 0.836 | 0.981 | 0.975 |
| 3D reconstruction + kinematics + TCN + Attention | 12.49 | 16.77 | 12.53 | 8.18 | 0.948 | 0.883 | 0.982 | 0.978 |
| 3D reconstruction + kinematics + MLP (w/o cons) | 14.61 | 17.25 | 15.67 | 10.92 | 0.807 | 0.658 | 0.859 | 0.904 |
| 3D reconstruction + kinematics + DLinear (w/o cons) | 15.54 | 15.78 | 14.22 | 16.63 | 0.801 | 0.693 | 0.893 | 0.818 |
| 3D reconstruction + kinematics + TCN (w/o cons) | 12.79 | 16.46 | 12.49 | 9.44 | 0.856 | 0.720 | 0.911 | 0.938 |
| 3D reconstruction + kinematics + TCN + Attention (w/o cons) | 14.64 | 17.30 | 16.41 | 10.22 | 0.815 | 0.700 | 0.848 | 0.896 |