

# MonoMSK: Monocular 3D Musculoskeletal Dynamics Estimation

Farnoosh Koleini, Hongfei Xue, Ahmed Helmy, Pu Wang

University of North Carolina at Charlotte

{fkoleini, hongfei.xue, ahmed.helmy, pu.wang}@charlotte.edu

## Abstract

*Reconstructing biomechanically realistic 3D human motion—recovering both kinematics (motion) and kinetics (forces)—is a critical challenge. While marker-based systems are lab-bound and slow, popular monocular methods use oversimplified, anatomically-inaccurate models (e.g., SMPL) and ignore physics, fundamentally limiting their biomechanical fidelity. In this work, we introduce MonoMSK, a hybrid framework that bridges data-driven learning and physics-based simulation for biomechanically realistic 3D human motion estimation from monocular video. MonoMSK jointly recovers both kinematics (motions) and kinetics (forces and torques) through an anatomically accurate musculoskeletal model. By integrating transformer-based inverse dynamics with differentiable forward kinematics and dynamics layers governed by ODE-based simulation, MonoMSK establishes a physics-regulated inverse-forward loop that enforces biomechanical causality and physical plausibility. A novel forward-inverse consistency loss further aligns motion reconstruction with the underlying kinetic reasoning. Experiments on BML-MoVi, BEDLAM, and OpenCap show MonoMSK significantly outperforms state-of-the-art methods in kinematic accuracy, while, for the first time, enabling precise monocular kinetics estimations.*

## 1. Introduction

Understanding and reconstructing human motion with biomechanical and physical realism is a long-standing goal in computer vision, biomechanics, and robotics. Biomechanically-accurate 3D human motion estimation aims to recover not only anatomically valid joint configurations but also the underlying physical quantities, such as forces and torques, that drive those motions. Such physically-grounded representations are critical for applications in clinical motion analysis, sports science, rehabilitation, and human–robot interaction [19, 20]. While Biomechanically-accurate motion estimation offers numerous benefits, its widespread adoption is hindered by significant barriers.

The gold standard approach for biomechanics-accurate



Figure 1. MonoMSK is a framework for physically grounded 3D human motion estimation from monocular videos. It couples a transformer-based Inverse Dynamics Transformer (IDT) that infers joint torques and ground reaction forces with a differentiable Forward Dynamics (FD) ODE solver that integrates these forces over time to produce biomechanically consistent motion.

motion estimation combines multi-camera, marker-based motion capture with biomechanical optimization tools, which are expensive, labor-intensive, and time-consuming [5]. In practice, current workflows use multiple cameras to track the markers on the body [22]. The captured marker data are then processed with time-consuming optimization-based simulations such as OpenSim [19]. For example, the state-of-the-art OpenCap system takes approximately 2 minutes for kinematics (e.g., joint rotation angles and velocities) and 35 minutes for kinetics (e.g., joint force and torques) [20].

Learning-based monocular motion estimation frameworks [8, 11, 17] have achieved impressive results in 3D human pose reconstruction by employing deep learning models to recover parametric body representations such as SMPL

[15] from single-camera monocular videos. Despite their strong visual realism and accessibility, these models typically rely on oversimplified skeletal structures with anatomically inaccurate joint positions and bone orientations, fundamentally limiting their biomechanical fidelity. Recent efforts have sought to mitigate these issues by incorporating biomechanically accurate skeletal models [1, 13]. However, they overlook the underlying physical laws that govern the causal relationships between forces (kinetics) and motions (kinematics). Consequently, these approaches are incapable of estimating kinetics and often exhibit reduced accuracy in kinematic reconstruction, as well.

To address these challenges, we introduce MonoMSK, a hybrid framework that integrates learning-based inverse dynamics with a differentiable, anatomically-accurate musculoskeletal (MSK) forward simulation to recover biomechanically-accurate motion dynamics from monocular video. MonoMSK comprises five integrated components: (1) a pretrained human-mesh-recovery model to extract anatomically-grounded virtual markers from the input video; (2) an Inverse Kinematics Transformer (IKT) to predict joint angles from these markers; (3) an Inverse Dynamics Transformer (IDT) to infer kinetic quantities (e.g., joint torques, ground-reaction forces) from the predicted kinematics; (4) a differentiable Forward Kinematics (FK) layer; and (5) a differentiable Forward Dynamics (FD) layer that leverages an ODE solver to simulate physically-consistent trajectories from the inferred kinetics.

This novel integrated kinematics–kinetics design tightly couples data-driven temporal modeling with continuous-time physical simulation. Specifically, our data-driven inverse transformers (IKT, IDT) are trained to infer the kinetic causes (joint torques) from the observed kinematic consequences (joint motion). The differentiable FD and FK layers then act as a physics-based verifier: it uses these inferred kinetics to simulate the resulting motion, ensuring the predicted forces can faithfully reconstruct the original observation. Moreover, by leveraging anatomy-constrained FK and FD layers, this physics-regulated loop embeds domain knowledge during both training and inference, ensuring estimated motion remains physically plausible and anatomically coherent.

Our key contributions are summarized as follows:

- We introduce MonoMSK, a hybrid motion dynamics estimation framework that couples learning-based inverse dynamics with a differentiable, anatomically accurate musculoskeletal forward simulator to recover physically grounded 3D motion dynamics (kinematics and kinetics) from monocular video
- We introduce a biomechanics-informed training scheme, which combines optimal-control biomechanics simulations for ground-truth generations with kinetics and kinematics supervision losses and forward-inverse consistency loss that align the bidirectional translation between kine-

matics (e.g., motion) and kinetics (e.g., force).

- Experiments on BML-MoVi, BEDLAM, and OpenCap show MonoMSK significantly outperforms state-of-the-art methods in kinematic accuracy, while, for the first time, enabling precise monocular kinetics estimations.

## 2. Related Work

### 2.1. Monocular 3D Human Pose Estimation

Monocular 3D human pose estimation has progressed rapidly with the advent of deep learning and parametric body models such as SMPL [15]. Early approaches [14] employ convolutional networks to regress SMPL parameters from single images, while later works extend this to video sequences using temporal transformers [23]. The introduction of transformer-based Human Mesh Recovery (HMR) models, such as HMR2.0 [11], TokenHMR [8], and CameraHMR [17], , have significantly improved human mesh reconstruction accuracy. However, these models are fundamentally limited for the applications that require biomechanical-accurate kinematics estimations. In particular, these models typically rely on oversimplified skeletal structures with anatomically inaccurate joint positions and bone orientations, fundamentally limiting their biomechanical fidelity. To address this limitations, we leverage the 3D human mesh reconstructed from the HMR models to extract virtual tracking markers, which then are leveraged by our MonoMSK model to infer physically grounded kinetics and kinematics estimations based on anatomically-accurate human MSK model. As a result, MonoMSK can serve as the plug-and-play module for any HMR models so that they can produce biomechanically-accurate

### 2.2. Biomechanically-accurate Motion Estimation

The current gold standard for biomechanics analysis combines multi-camera, marker-based motion capture systems with optimization-based musculoskeletal solvers such as OpenSim [19]. These systems rely on retro-reflective markers tracked by synchronized infrared cameras in controlled laboratory environments. The recorded marker trajectories are then processed through inverse kinematics and dynamics pipelines to estimate joint torques, ground reaction forces, and muscle activations. Although these workflows achieve high biomechanical accuracy, they are expensive, labor-intensive, and limited to specialized lab setups. They also require expert supervision, precise calibration, and time-consuming optimal control simulations. For example, OpenCap [20] reports approximately two minutes per trial for kinematic reconstruction and up to thirty-five minutes for kinetic estimation. Such constraints make these methods impractical for large-scale or real-world deployment. Recent work, such as BioPose and D3KE has begun to incorporate biomechanically accurate skeletal models [1, 13]. How-

ever, these methods typically omit explicit physical laws that govern the causal link between forces (kinetics) and motion (kinematics). As a result, they cannot estimate kinetic variables and often yield diminished accuracy in kinematic reconstruction.

In contrast to prior approaches, our proposed *MonoMSK* framework unifies kinematic perception and dynamic reasoning within a single, end-to-end trainable model. By embedding a differentiable physics solver directly into a transformer-based motion prediction architecture, *MonoMSK* simultaneously learns inverse dynamics (force–torque inference) and forward dynamics (motion generation) without separate optimization stages. This integration enables physically stable, temporally coherent, and anatomically consistent motion trajectories—achieving biomechanical accuracy comparable to laboratory systems while maintaining the scalability and flexibility of modern deep learning.

### 3. Proposed Method: MonoMSK

The objective of *MonoMSK* is to estimate biomechanically accurate and physically consistent 3D human motion directly from monocular video. As illustrated in Figure 2, *MonoMSK* builds upon a detailed musculoskeletal (MSK) model and integrates kinematic inference with physics-based dynamic reasoning. The pipeline begins with a pretrained Human Mesh Recovery (HMR) models, which regress pose and shape parameters and generate SMPL meshes whose vertices serve as virtual motion-capture markers (§3.2). These marker trajectories are processed by an Inverse Kinematics Transformer (IKT) (§3.2.1) to estimate the generalized MSK kinematic state  $\mathbf{q} = \{\mathbf{T}, \mathbf{R}, \mathbf{q}^r\}$ , capturing anatomically valid joint rotations. The recovered kinematics are then provided to the Inverse Dynamics Transformer (IDT) (§3.2.2), which predicts internal joint torques  $\boldsymbol{\tau}$  and external ground-reaction forces  $\boldsymbol{\lambda}$  that generate the observed motion. To enforce biomechanical correctness, *MonoMSK* integrates a differentiable Forward Dynamics (FD) layer based on an ODE formulation of the MSK dynamics (§3.2.3). This layer simulates the forward-time evolution of the body under the predicted forces and torques, producing physically coherent kinematic states that are matched back to the IKT estimates through inverse–forward consistency. Together, these components form a unified, end-to-end differentiable framework that tightly couples perception with musculoskeletal physics, enabling accurate, stable, and interpretable human motion reconstruction.

#### 3.1. Human Musculoskeletal Model

The biomechanical human model consists of two core components: the musculoskeletal (MSK) body model and the musculoskeletal (MSK) dynamics model (See Figure 3).

**MSK Body Model.** The MSK body model, e.g., the

widely-adopted OpenSim model [19], represents the body as  $N_s = 24$  rigid bone segments interconnected by anatomically constrained joints. Each joint  $i$  defines motion according to its physiological Degrees of Freedom (DoFs)  $D_i$ , thus determining the human body’s valid kinematic and dynamic configuration space. Each joint  $i$  is parameterized by the anatomy-dependent bone orientation  $q_i^o \in \mathbb{R}^3$ , muscle-induced joint rotation  $q_i^r \in \mathbb{R}^{D_i}$  ( $D_i \leq 3$ ), and bone scaling  $s_i \in \mathbb{R}^3$  [9, 16, 19]. The scaling process tailors a generic anatomical MSK body model to fit subject-specific body geometry, such as bone lengths and muscle attachment points. These parameters ( $\mathbf{q}^o, \mathbf{q}^r, s$ ) collectively defines for the entire skeleton [6]. The 3D motion kinematics can be described compactly in a generalized coordinate system as

$$\mathbf{q} = \{\mathbf{T}, \mathbf{R}, \mathbf{q}^r\}, \quad (1)$$

where  $\mathbf{T} \in \mathbb{R}^3$  and  $\mathbf{R} \in \mathbb{R}^3$  represent global translation and rotation of the root segment, and  $\mathbf{q}^r$  encodes the motion-induced joint rotations of the 24-joint MSK model. Correspondingly,  $\dot{\mathbf{q}}$  and  $\ddot{\mathbf{q}}$  are used to represent the velocity and acceleration of generalized coordinates.

**Muscle-driven Forward Dynamics.** It uses differential-algebraic equations to predict body motion from muscle excitations. Muscles attach to skeleton bones through defined paths to determine fiber and tendon lengths, which, combining with body segment parameters (e.g., mass, center of mass, and inertia), generates the forces and torques applied at joints, according to muscle activation, contraction dynamics, and multibody dynamics [4]. In particular, multi-body dynamics, following the Newton-Euler equations of motion, defines forward dynamics that predict motion from internal joint torques  $\boldsymbol{\tau}$  and external forces  $\boldsymbol{\lambda}$  (e.g., ground reaction forces, GRFs)

$$\mathbf{J}_C^\top \boldsymbol{\lambda} + \boldsymbol{\tau} = \mathbf{M}(\mathbf{q})\ddot{\mathbf{q}} + \mathbf{C}(\mathbf{q}, \dot{\mathbf{q}}) + \mathbf{g}(\mathbf{q}), \quad (2)$$

where  $\mathbf{M}$ ,  $\mathbf{C}$ , and  $\mathbf{g}$  denote the generalized inertia matrix, Coriolis/centrifugal terms, and gravitational generalized torques, respectively.  $\mathbf{J}_C$  is the contact Jacobian matrix that translates the generalized velocities  $\dot{\mathbf{q}}$  to the velocities at the point of contact between foot and ground. Each joint torque  $\tau_i$  is expressed as the net contribution of muscle forces transmitted through moment arms:

$$\tau_i = \sum_{j=1}^{N_m} r_{ij} F_j(a_j, l_j, v_j) + \tau_i^{\text{tm}}, \quad (3)$$

where  $r_{ij}$  is the moment arm of muscle  $j$  about joint  $i$ ,  $N_m$  is the number of muscles,  $a_j$  is the muscle activation,  $l_j$  and  $v_j$  are fiber length and velocity, and  $\tau_i^{\text{tm}}$  denotes an ideal torque motor at major joints (e.g., lumbar, shoulder, elbow) for residual actuation. Muscle forces  $F_j$  follow a Hill-type model [4]:  $F_j(a_j, l_j, v_j) = a_j F_j^{\max} f_l(l_j) f_v(v_j) +$

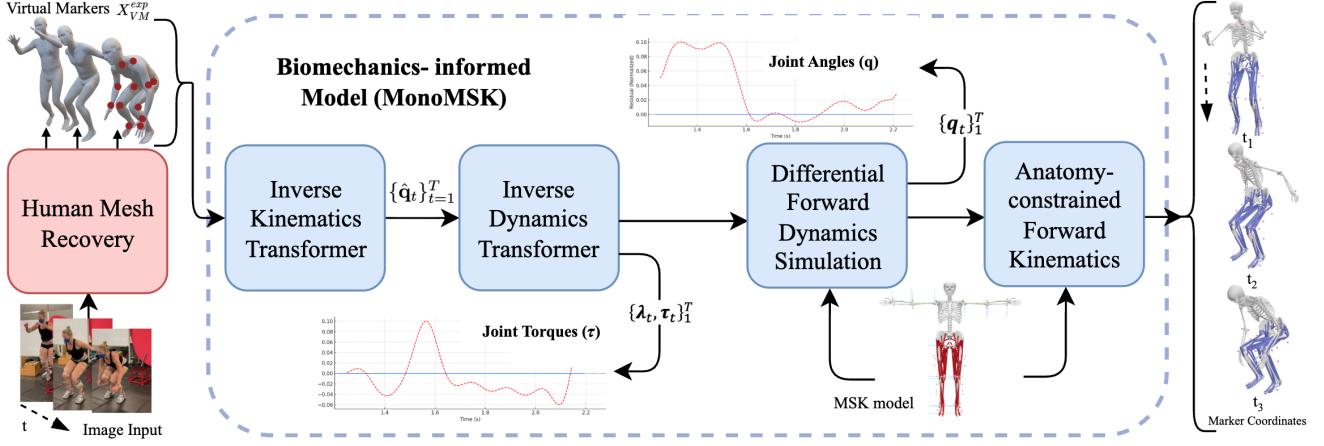


Figure 2. Overview of the MonoMSK pipeline. A monocular video is processed by a pretrained Human Mesh Recovery (HMR) model to obtain 3D meshes and virtual markers. The **Inverse Kinematics Transformer** (IKT) converts these markers into anatomically constrained musculoskeletal joint states  $\mathbf{q}$ . The **Inverse Dynamics Transformer** (IDT) infers the latent dynamic quantities, internal torques  $\boldsymbol{\tau}$  and external ground-reaction forces  $\boldsymbol{\lambda}$ . A differentiable **Forward Dynamics** (FD)-ODE solver integrates these forces through the Euler–Lagrange MSK dynamics to produce physically coherent future motion.

$F_j^{\text{pass}}(l_j)$ , where  $F_j^{\text{max}}$  is maximal isometric force,  $f_l$  and  $f_v$  are the force-length and force–velocity relationships, and  $F_j^{\text{pass}}$  represents passive elastic tension.

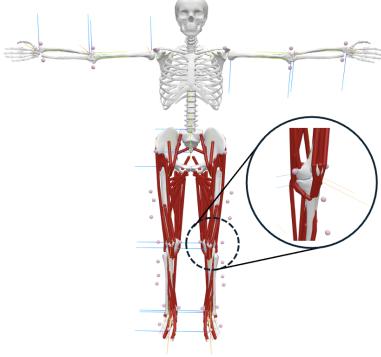


Figure 3. Musculoskeletal (MSK) body model with anatomically precise joint positions, bone orientations, and muscle geometry (red). Pink spheres indicate virtual model markers attached to bone segments for accurate biomechanical tracking. The zoomed region illustrates detailed muscle–joint structure around the knee.

**Ground Reaction Forces.** To simulate realistic foot-ground interactions, ground reaction forces are modeled through six foot-ground contact spheres attached to the foot segments. Each contact sphere  $k$  generates normal and tangential ground reaction forces modeled by the compliant Hunt–Crossley formulation [12]:

$$\mathbf{F}_{n,k} = k_n \delta_k^{1.5} (1 + c_n \dot{\delta}_k) \mathbf{n}, \quad \mathbf{F}_{t,k} = -\mu \|\mathbf{F}_{n,k}\| \frac{\mathbf{v}_{t,k}}{\|\mathbf{v}_{t,k}\| + \epsilon},$$

where  $\delta_k$  and  $\dot{\delta}_k$  are the penetration depth and velocity,  $\mathbf{n}$  the ground normal,  $\mathbf{v}_{t,k}$  the tangential velocity, and  $(k_n, c_n, \mu)$

are stiffness, damping, and friction coefficients. The total ground reaction force is

$$\boldsymbol{\lambda} = \sum_k (\mathbf{F}_{n,k} + \mathbf{F}_{t,k}), \quad (4)$$

This physically grounded contact model captures smooth transitions through heel-strike, mid-stance, and toe-off phases [18, 19].

### 3.2. Biomechanics-informed Motion Estimation Model (MonoMSK)

Built on top of the MSK model, we introduce MonoMSK to estimate kinematics and kinetics from monocular videos. First, we will leverage the existing monocular human mesh recovery (HMR) models, which estimate 3D pose  $\theta$  and shape  $\beta$  parameters to generate a 3D mesh via a parametric model like SMPL. Any subset of the 6,890 vertices on the SMPL mesh can act as virtual motion capture markers. Leveraging these virtual marker tracking data as inputs, MonoMSK will first learn to estimate kinematics  $\mathbf{q} = \{\mathbf{T}, \mathbf{R}, \mathbf{q}^r\}$  via Inverse Kinematics Transformer (IKT). Then, the kinematic estimation  $\mathbf{q}$  serves as the input of the Inverse Dynamics Transformer (IDT) to predict kinetic attributes, including internal joint torques  $\boldsymbol{\tau}$  and external contact forces  $\boldsymbol{\lambda}$ . To incorporate biomechanics priors into the network architecture, we directly inject a physics-based ODE solver into the network, which leverages forward dynamics to transform the estimated joint torques and contact forces  $\{\boldsymbol{\tau}, \boldsymbol{\lambda}\}$  back to joint kinematics  $\mathbf{q}$ . The kinematic estimation  $\mathbf{q}$  will then be translated into joint positions via an anatomical-aware forward kinematic layer. Inverse-forward

consistency training is leveraged to ensure the biomechanically accurate motion estimations.

### 3.2.1. Inverse Kinematics Transformer (IKT)

Given a monocular RGB video sequence  $\{I_t\}_{t=1}^T$ , we employ a pretrained human mesh recovery model to regress per-frame pose and shape parameters  $\{\hat{\theta}_t, \hat{\beta}_t\}_{t=1}^T$ . The pose  $\theta_t \in \mathbb{R}^{23 \times 3}$  encodes axis-angle rotations of the 23 joints, while the shape  $\beta_t \in \mathbb{R}^{10}$  represents low-dimensional body morphology [15]. By combining these pose and shape parameters, the SMPL model produces a detailed 3D mesh  $M(\theta, \beta) \in \mathbb{R}^{3 \times N}$ , where  $N = 6890$  vertices represent the surface of the body. The positions of the body joints  $J \in \mathbb{R}^{3 \times k}$  are then derived as a weighted sum of these vertices, formulated as  $J = MW$ , where  $W \in \mathbb{R}^{N \times 23}$  contains the predefined linear blending weights that map vertices to the joints. With the joint positions  $J$  as inputs, we employ the pretrained global trajectory predictor [3] to estimate per-frame translation and rotation  $\{\hat{T}_t, \hat{R}_t\}_{t=1}^T$ .

Since the SMPL model employs a simplified skeleton rig with inaccurate joint location and bone orientations,  $\theta_t \in \mathbb{R}^{23 \times 3}$  cannot be directly used to estimate anatomical joint rotation  $q^r \in \mathbb{R}^{24 \times 3}$ . Following BioPose [13], we select a set of  $M$  tracking markers  $\mathbf{x} = \{x^i\}_{i=1}^M$  that attached to the bone segments in such a way that each bone segment is associated with at least  $D_i$  markers to ensure the unique solutions of the derived rotation angles at joint  $i$  with  $D_i$  degrees of freedom. The transformer encoder takes a sequence of marker data  $\{\mathbf{x}_t\}_{t=1}^T$  over  $T$  frames as inputs to predict the muscle-induced joint rotation  $\{\hat{q}_t^r\}_{t=1}^T = E_{\text{IKT}}(\{\mathbf{x}_t\}_{t=1}^T)$  which, combining with predicted global translation and rotation, yields the kinematic estimation [23]

$$\{\hat{q}_t\}_{t=1}^T = \{\hat{T}_t, \hat{R}_t, \hat{q}_t^r\}_{t=1}^T$$

### 3.2.2. Inverse Dynamics Transformer (IDT)

Given the kinematic estimations  $\{\hat{q}_t\}_{t=1}^T$  from IKT encoder, the IDT predicts the joint torques and contact forces  $\{\tau_t, \lambda_t\}_{t=1}^T$ . At each time step  $t$ , a motion feature vector  $\mathbf{f}_t \in \mathbb{R}^{F_{\text{dim}}}$  is formed by concatenating the generalized coordinates and their temporal derivatives, i.e.,  $\mathbf{f}_t = [\hat{q}_t, \dot{\hat{q}}_t]$ . Each feature is projected into the transformer embedding space using an MLP encoder:

$$\mathbf{z}_t = \text{MLP}_{\text{enc}}(\mathbf{f}_t) \in \mathbb{R}^{d_{\text{model}}},$$

followed by sinusoidal positional encoding  $\text{PE}(\cdot)$  to maintain temporal order. The encoded sequence  $\mathbf{z}_{1:T}$  is then processed by a transformer encoder:

$$\mathbf{h}_{1:T} = E_{\text{IDT}}(\text{PE}(\mathbf{z}_{1:T})),$$

yielding temporally contextualized motion embeddings  $\mathbf{h}_{1:T}$  that capture biomechanical dependencies and long-range temporal context across the motion sequence. From  $\mathbf{h}_{1:T}$ , a

dedicated MLP head predicts the per-frame kinetics parameters:

$$\{\hat{\lambda}_t, \hat{\tau}_t\}_{t=1}^T = \text{MLP}_{\text{force}}(\mathbf{h}_{1:T}),$$

### 3.2.3. Forward Dynamics (FD) Layer via ODE Solver

Given the predicted kinetics  $\{\hat{\lambda}_t, \hat{\tau}_t\}_{t=1}^T$  from IDT, the FD layer predict continuous-time kinematics via the differential ODE solver, which numerically integrates the human body's motion over time using the Newton-Euler equations of motions defined in Eq. (2), enabling simulation of future body states consistent with physical laws. The kinematic state at time  $t$  is

$$\mathbf{x}_t = [\mathbf{q}_t, \dot{\mathbf{q}}_t]^T,$$

where  $\mathbf{q}_t = \{\mathbf{T}_t, \mathbf{R}_t, \mathbf{q}_t^r\}$  represents generalized states and  $\dot{\mathbf{q}}_t$  their velocities. The second-order dynamics are rewritten as a first-order ODE system  $f_{\text{ODE}}(\mathbf{x}_t)$ :

$$\frac{d}{dt} \begin{bmatrix} \mathbf{q}_t \\ \dot{\mathbf{q}}_t \end{bmatrix} = \begin{bmatrix} \dot{\mathbf{q}}_t \\ \mathbf{M}^{-1}(\mathbf{q}_t)(\mathbf{J}_C^T \boldsymbol{\lambda}_t + \boldsymbol{\tau}_t - \mathbf{C}(\mathbf{q}_t, \dot{\mathbf{q}}_t) - \mathbf{g}(\mathbf{q}_t)) \end{bmatrix}, \quad (5)$$

which defines a differentiable mapping that governs the temporal evolution of the biomechanical system. To compute the predicted trajectory  $\{\hat{\mathbf{q}}_{t+n}\}_{n=1}^N$ , we integrate Eq. (5) forward in time using a differentiable ODE solver. In practice, a fourth-order Runge-Kutta (RK4) or adaptive Dormand-Prince (RK45) solver can be adopted [24], which ensures numerical stability while maintaining differentiability for backpropagation:

$$\mathbf{x}_{t+\Delta t} = \text{ODE}_{\text{solver}}(f_{\text{ODE}}, \mathbf{x}_t, \Delta t; \phi), \quad (6)$$

where  $\phi$  denotes physical parameters (e.g., segmental masses, inertia, joint stiffness, and damping). The solver predicts subsequent generalized states as:

$$\begin{aligned} \mathbf{q}_{t+1}^{fd} &= \dot{\mathbf{q}}_t^{fd} + \ddot{\mathbf{q}}_t^{fd} \Delta t, \\ \mathbf{q}_{t+1}^{fd} &= \mathbf{q}_t^{fd} + \dot{\mathbf{q}}_t^{fd} \Delta t, \end{aligned} \quad (7)$$

This iterative process simulates motion under the learned forces and torques, producing continuous, dynamically coherent trajectories.

### 3.2.4. Anatomical Forward Kinematic (FK) Layer

The FK layer transforms the rest-pose anatomical joint markers to the new positions according to the ODE-simulated kinematic dynamics  $\{\mathbf{q}_i^{fd}\}_{i=1}^T$ . In particular, the global position of joint  $i$  is computed recursively through:

$$\mathbf{J}_i^{fk} = \mathbf{R}_i(\mathbf{J}_i^0 \odot s_i) + \mathbf{J}_{\text{par}(i)}, \quad (8)$$

where  $\mathbf{J}_i^0$  is the local joint position offset,  $\odot$  denotes element-wise bone scaling, and  $\text{par}(i)$  indicates the parent joint. The corresponding rotation is obtained as  $\mathbf{R}_i = \mathbf{R}_{\text{par}(i)} \mathbf{R}(q_i^o) \mathbf{R}(q_i^r)$ . Since we employed a full-body biomechanical MSK model, the FK transformation is inherently constrained by the realistic joint degrees of freedom  $q_i^r \in \mathbb{R}^{D_i}$  ( $D_i \leq 3$ ), bone orientations  $q^o$  and scales  $s_i$ .

### 3.3. Biomechanics-informed Model Training

To enforce biomechanics-accurate estimations, our model is trained with multiple physics-informed losses including supervision losses on kinematic states  $\{\hat{\mathbf{q}}_t\}_{t=1}^T$  and kinetic forces  $\{\hat{\lambda}_t, \hat{\tau}_t\}_{t=1}^T$  along with the consistency loss between inverse and forward motion dynamics.

**Ground-truth Dynamics via Optimal-control Forward Simulation.** To facilitate supervised training, we obtain ground-truth MSK dynamics via optimal control simulations in physics engines like OpenSim-Moco [7]. This has become the gold standard practice for non-invasive measurement of variables such as muscle forces and joint torques, which are difficult to measure directly *in vivo*. In particular, the goal is to find the optimal muscle excitation signals  $e$  that minimize muscular effort and maximize agreement between simulated and observed motion under the constraints of Newton-Euler equations of motion defined in (2), i.e.,

$$\arg \min_e \int_{t_0}^{t_f} \left( w_1 \|\mathbf{e}_t\|_2^2 + w_2 \|\mathbf{q}_t - \tilde{\mathbf{q}}_t\|_2^2 + w_3 \|\dot{\mathbf{q}}_t - \tilde{\dot{\mathbf{q}}}_t\|_2^2 + w_4 \|\ddot{\mathbf{q}}_t - \tilde{\ddot{\mathbf{q}}}_t\|_2^2 \right) dt,$$

where  $w_i$  are weighting coefficients and  $(\tilde{\mathbf{q}}, \tilde{\dot{\mathbf{q}}}, \tilde{\ddot{\mathbf{q}}})$  are ground-truth kinematics derived from inverse kinematics optimization based on the marker tracking data detailed in the Supplementary material. The optimal control simulation can be transcribed into a finite-dimensional nonlinear program using direct collocation, which is then solved using a large-scale nonlinear optimization solver such as IPOPT [21]. The simulation outputs are reference torque  $\tilde{\tau}_t$  and force  $\tilde{\lambda}_t$ .

**Kinetics Losses.** By training the model with supervised losses for the contact force and joint torque, it will learn the physiological dynamics of the MSK system. Based on the reference torque  $\tilde{\tau}_t$  and force  $\tilde{\lambda}_t$ , the training objective is to minimize the mean squared absolute errors

$$\mathcal{L}_{\text{kinetic}} = w_\lambda \mathcal{L}_\lambda + w_\tau \mathcal{L}_\tau, \quad (9)$$

$$\text{where } \mathcal{L}_\lambda = \sum_{t=1}^T \|\hat{\lambda}_t - \tilde{\lambda}_t\|_2^2, \quad \mathcal{L}_\tau = \sum_{t=1}^T \|\hat{\tau}_t - \tilde{\tau}_t\|_2^2$$

$\mathcal{L}_\lambda$  promotes accurate estimation of foot-ground contact, and  $\mathcal{L}_\tau$  enforces the accurate joint torque estimation.

**Inverse-Forward Consistency Losses.** To embed the biomechanics causality into the model, we exploit the property that the translation between kinematics (e.g., motion) and kinetics (e.g., force) should be consistent. In particular, if the model inversely translates from kinematic consequences (e.g., rotation angles and joint positions) to the kinetic causes (e.g., muscle forces and joint torque), it should be able to arrive back at the original kinematic observations from its own kinetics predictions through physics-governed forward

reasoning (e.g., the forward dynamics ODE simulations). In particular, we adopt two consistency losses

$$\mathcal{L}_{\text{con}} = w_q \mathcal{L}_q + w_J \mathcal{L}_J, \quad (10)$$

$$\text{where } \mathcal{L}_q = \sum_{t=1}^T \|\hat{\mathbf{q}}_t^{fd} - \tilde{\mathbf{q}}_t\|_2^2 \quad \mathcal{L}_J = \sum_{t=1}^N \|\hat{\mathbf{J}}_t^{fk} - \tilde{\mathbf{J}}_t\|_2^2.$$

Here,  $\tilde{\mathbf{q}}_t$  and  $\tilde{\mathbf{J}}_t$  denote ground-truth joint rotations and positions, while  $\hat{\mathbf{q}}_t^{fd}$  are joint rotations obtained from the forward dynamics ODE solver and  $\hat{\mathbf{J}}_t^{fk}$  are joint positions derived via anatomical-constrained forward kinematics.

Moreover, the ODE-based forward kinetics module functions as a differentiable physics simulator embedded within the transformer’s decoder loop. During training, the gradients of  $\mathcal{L}_{\text{ODE}}$  are backpropagated through both the neural network parameters and the ODE solver, allowing the model to learn intrinsic dynamic properties—such as stiffness, damping, and contact responses—directly from data. This integration tightly couples data-driven temporal modeling with continuous-time physics, resulting in biomechanically faithful and dynamically smooth human motion trajectories. Furthermore, through the physics-regulated and anatomy-constrained FD and FK layers, domain knowledge is embedded not only during training but also during inference, ensuring consistent physical plausibility and anatomical coherence throughout the generation process.

## 4. Experiments

**Datasets.** During training, we use BML-MoVi [10], which provides rich biomechanical motion capture and video data from multiple actors performing everyday activities. BML-Movi consists of 90 subjects performing 21 different actions, captured using two cameras and a marker-based motion capture system. For evaluation, we utilize the test set of OpenCap and BEDLAM, MonoMSK is *not* trained on OpenCap [20] and BEDLAM [2] datasets to show its cross-dataset generalization performance. OpenCap includes data from ten subjects performing various actions such as walking, squatting, standing up from a chair, drop jumps, and their asymmetric variations. The recordings were made using five RGB cameras alongside a marker-based motion capture system. Additionally, OpenCap offers processed marker data and kinematic annotations for a comprehensive full-body OpenSim skeletal model [19]. BEDLAM dataset comprises synthetic video data featuring a total of 271 subjects, including 109 men and 162 women. It includes monocular RGB videos, covering a diverse range of body shapes, and motions.

**Implementation.** The transformer backbone employs a standard encoder–decoder architecture consisting of 6 layers, 8 attention heads, and an embedding dimension of 1024. The model operates on input sequences of length 16, consistent

Table 1. We evaluate mean per-bony–landmark position error (MPBLPE, mm), joint–angle mean absolute error ( $MAE_{angle}$ , deg), and physics–plausibility metrics acceleration error (ACCL, mm/frame<sup>2</sup>) and velocity error (VEL, mm/frame). Lower values (↓) indicate better performance. Blue percentages represent relative improvements over the best kinematics-only baseline (BioPose). MonoMask uses the MQ-HMR in [13] to extract virtual markers

Methods	BML-MoVi				BEDLAM				OpenCap			
	MPBLPE (↓)	MAE <sub>angle</sub> (↓)	Acc. (↓)	Vel. (↓)	MPBLPE (↓)	MAE <sub>angle</sub> (↓)	Acc. (↓)	Vel. (↓)	MPBLPE (↓)	MAE <sub>angle</sub> (↓)	Acc. (↓)	Vel. (↓)
HMR2.0 [11]	48.32	3.78	8.70	6.28	53.21	3.92	9.58	6.92	50.16	3.81	9.03	6.52
TokenHMR [8]	44.54	3.54	8.02	5.79	48.42	3.69	8.72	6.29	46.17	3.57	8.31	6.00
CameraHMR [17]	39.63	3.28	7.13	5.15	42.17	3.39	7.59	5.48	39.48	3.31	7.11	5.13
D3KE [1]	36.98	3.54	—	—	39.45	6.72	—	—	38.62	5.92	—	—
OpenCap Multi-Camera [20]	—	—	—	—	—	—	—	—	—	4.50	—	—
BioPose [13]	<b>25.76</b>	<b>2.84</b>	<b>6.30</b>	<b>5.18</b>	<b>26.54</b>	<b>3.14</b>	<b>6.84</b>	<b>4.37</b>	<b>26.34</b>	<b>3.19</b>	<b>5.68</b>	<b>4.23</b>
<b>MonoMSK (Ours)</b>	<b>24.36</b> (5.4%↓)	<b>1.93</b> (32.0%↓)	<b>4.38</b> (30.5%↓)	<b>3.17</b> (38.8%↓)	<b>25.62</b> (3.5%↓)	<b>2.57</b> (18.1%↓)	<b>4.61</b> (32.6%↓)	<b>3.33</b> (23.8%↓)	<b>25.28</b> (4.0%↓)	<b>2.84</b> (11.0%↓)	<b>4.55</b> (19.9%↓)	<b>3.29</b> (22.2%↓)

Table 2. Impact of Human Mesh Recovery (HMR) backbones on kinetics estimation of MonoMSK.  $MAE_\lambda$  : ground-reaction forces and  $MAE_\tau$  : joint torques,  $MAE_\tau$ .

HMR Backbone	BML-MoVi		BEDLAM		OpenCap	
	MAE <sub>λ</sub> (↓)	MAE <sub>τ</sub> (↓)	MAE <sub>λ</sub> (↓)	MAE <sub>τ</sub> (↓)	MAE <sub>λ</sub> (↓)	MAE <sub>τ</sub> (↓)
HMR2.0 [11] + MonoMSK	0.0162	0.0584	0.0489	0.0841	0.0398	0.0726
TokenHMR [8] + MonoMSK	0.0150	0.0553	0.0463	0.0802	0.0379	0.0704
CameraHMR [17] + MonoMSK	0.0144	0.0528	0.0441	0.0781	0.0367	0.0689
<b>MQ-HMR [13] + MonoMSK</b>	<b>0.0139</b>	<b>0.0498</b>	<b>0.0422</b>	<b>0.0748</b>	<b>0.0351</b>	<b>0.0675</b>

Table 3. Ablation on the Training Losses for the MonoMSK framework. Lower values (↓) indicate better performance. The first row corresponds to the kinematic-only baseline.

<b>Training Losses</b>	<b>Performance (↓)</b>				$\mathcal{L}_\lambda$	$\mathcal{L}_\tau$	$\mathcal{L}_q$	$\mathcal{L}_J$
	$\mathcal{L}_\lambda$	$\mathcal{L}_\tau$	$\mathcal{L}_q$	$\mathcal{L}_J$				
—	—	—	—	—	—	—	2.84	25.76
✓	—	—	—	—	0.0156	0.0538	2.57	25.62
✓	✓	—	—	—	0.0148	0.0512	2.45	25.28
✓	✓	✓	—	—	0.0142	0.0501	2.17	24.86
✓	✓	✓	✓	✓	<b>0.0139</b>	<b>0.0498</b>	<b>1.93</b>	<b>24.36</b>

with common configurations in prior works. For efficient training, we first use the ground-truth supervision to extract output embeddings for 20 epochs, followed by an additional 5 epochs using predicted embeddings. The Adam optimizer is utilized with a weight decay of  $10^{-4}$ . The initial learning rate is set to  $10^{-5}$  and is decayed by a factor of 0.8 every 5 epochs. Empirically, the hyperparameters are configured as  $\gamma_q = 2 \times 10^3$ ,  $\gamma_J = 1 \times 10^5$ ,  $\gamma_\tau = 5$ ,  $\gamma_\lambda = 1$ ,  $\gamma_v = 100$ , and  $\gamma_z = 200$ .

**Evaluation Metrics.** We evaluate MonoMSK using both kinematic and dynamic measures. Kinematic accuracy is assessed using the MPBLPE (mm), which measures 3D bony–landmark positional error, and  $MAE_{angle}$  (deg), which quantifies joint–angle accuracy. To evaluate physical plausibility, we report the acceleration error (ACCL) and velocity

error (VEL), capturing deviations in joint accelerations and velocities that reflect dynamic consistency. Finally, for direct kinetics estimation, we compute  $MAE_\lambda$  (ground-reaction forces) and  $MAE_\tau$  (joint torques), which assess the correctness of predicted external and internal forces.

## 4.1. Comparison to State-of-the-art Approaches

**Improvements to Kinematics-based Methods.** As shown in Table 1, the proposed MonoMSK framework establishes a new benchmark for biomechanically consistent 3D human motion estimation across the BML-MoVi, BEDLAM, and OpenCap datasets. By combining transformer-based force–torque prediction with physics-aware motion integration through a differentiable ODE solver, our model achieves substantial gains in both kinematic accuracy and physical plausibility compared to existing approaches. On BML-MoVi, MonoMSK reduces the joint-angle error ( $MAE_{angle}$ ) by 32.0% and the bony–landmark position error (MPBLPE) by 5.4% relative to the BioPose baseline, while decreasing the acceleration (ACCL) and velocity (VEL) errors by 30.5% and 38.8%, respectively. Similarly, on BEDLAM, our method achieves an 18.1% improvement in  $MAE_{angle}$ , a 3.5% gain in MPBLPE, and consistent reductions in ACCL (32.6%) and VEL (23.8%), confirming robustness across large-scale synthetic motion data with complex dynamics. For OpenCap, which contains real-world motion capture sequences, MonoMSK continues to outperform BioPose, lowering  $MAE_{angle}$  by 11.0%, MPBLPE by 4.0%, ACCL by 19.9%, and VEL by 22.2%. These consistent improvements across all benchmarks highlight the effectiveness of integrating differentiable physics into transformer architectures, enabling anatomically accurate, and physically interpretable human motion estimation from monocular videos.

## 4.2. Ablation Study

**Ablation on HMR Backbones.** Table 2 presents a quantitative comparison of different Human Mesh Recovery (HMR) backbones used to initialize the kinematic esti-

Table 4. Comparison between Single Frame Out and Multiple Frames Out MonoMSK models across three datasets. Lower values ( $\downarrow$ ) indicate better performance.

Model	BML-MoVi			BEDLAM			OpenCap		
	MAE $_{\lambda}$	MAE $_{\tau}$	MAE $_{angle}$	MAE $_{\lambda}$	MAE $_{\tau}$	MAE $_{angle}$	MAE $_{\lambda}$	MAE $_{\tau}$	MAE $_{angle}$
Multiple Frames Out	0.0156	0.0538	2.45	0.0472	0.0785	2.81	0.0373	0.0682	3.42
Single Frame Out	0.0139	0.0498	<b>1.93</b>	0.0422	0.0748	<b>2.57</b>	0.0351	0.0675	<b>2.84</b>

mates within the MonoMSK framework. The results clearly demonstrate that the accuracy and physical consistency of MonoMSK strongly depend on the quality of the underlying HMR initialization. Across all datasets, performance improves consistently as the backbone transitions from HMR2.0 to TokenHMR, CameraHMR, and finally MQ-HMR. The proposed MQ-HMR-based MonoMSK achieves the lowest physics-based losses, with  $MAE_{\lambda} = 0.0139$  and  $MAE_{\tau} = 0.0498$  on the BML-MoVi dataset, outperforming the next-best CameraHMR variant by a noticeable margin. Similar improvements are observed on BEDLAM ( $MAE_{\lambda} = 0.0422$ ,  $MAE_{\tau} = 0.0748$ ) and OpenCap ( $MAE_{\lambda} = 0.0351$ ,  $MAE_{\tau} = 0.0675$ ), confirming the robustness of our approach across both synthetic and real-world motion capture data. These results highlight that accurate mesh-based kinematic initialization directly enhances the downstream physics-based optimization in MonoMSK, enabling more accurate force–torque prediction.

**Ablation on the Training Losses.** We perform an ablation study to analyze the contribution of each loss component in the MonoMSK training objective, as shown in Table 3. The kinematic-only baseline (without any physics-based supervision) exhibits the highest errors ( $MAE_{angle} = 2.84^\circ$ , MPBLPE = 25.76 mm), confirming that purely kinematic reconstruction fails to capture biomechanical consistency. Introducing the external force loss  $\mathcal{L}_{\lambda}$  significantly reduces both angular and positional errors by enforcing stable contact interactions with the environment. Adding the torque loss  $\mathcal{L}_{\tau}$  further improves internal actuation coherence across the 24-joint biomechanical skeleton, leading to smoother and more physically grounded joint dynamics. Incorporating joint rotation consistency loss  $\mathcal{L}_q$  enhances temporal smoothness and consistency between predicted and physically integrated motion states, while the joint-space consistency regularization  $\mathcal{L}_J$  provides anatomical alignment and prevents kinematic drift. When all four objectives are jointly optimized, MonoMSK achieves the lowest overall errors ( $MAE_{\lambda} = 0.0139$ ,  $MAE_{\tau} = 0.0498$ ,  $MAE_{angle} = 1.93^\circ$ , MPBLPE = 24.36 mm), demonstrating that coupling external (force) and internal (torque) supervision with coordinate- and joint-level consistency is critical for biomechanically accurate and physically coherent motion estimation.

**Single Frame Out vs. Multiple Frames Out.** To examine the effect of temporal prediction strategy within the

transformer-based MonoMSK architecture, we compare two configurations: (1) a *Single Frame Out* model, where the transformer predicts one frame at a time and is jointly optimized with the ODE layer in an end-to-end manner, and (2) a *Multiple Frames Out* variant, which predicts several future frames simultaneously in a two-stage setup without end-to-end physical supervision. As shown in Table 4, the Single Frame Out configuration consistently outperforms its multi-frame counterpart across all datasets. It achieves lower external force and joint-torque losses ( $MAE_{\lambda}$  and  $MAE_{\tau}$ ) and yields notably improved joint-angle accuracy. For example, on the BML-MoVi dataset, Single Frame Out reduces  $MAE_{\lambda}$  from 0.0156 to 0.0139 and  $MAE_{\tau}$  from 0.0538 to 0.0498, while improving  $MAE_{angle}$  from  $2.45^\circ$  to  $1.93^\circ$ . A similar trend is observed for BEDLAM and OpenCap, where consistent reductions in both force–torque and angular errors are obtained. These results indicate that step-wise temporal prediction allows stronger supervision signals from the ODE layer, ensuring smooth physical integration and preserving biomechanical consistency. In contrast, multi-frame forecasting introduces temporal drift due to error accumulation, confirming that autoregressive single-step prediction is more effective for physically grounded human motion modeling.

## 5. Conclusion

We introduce MonoMSK, the first hybrid framework to recover full-body, biomechanically-accurate motion dynamics (kinematics and kinetics) from a monocular video. MonoMSK integrates learning-based inverse dynamics transformers with a differentiable, anatomically-accurate musculoskeletal (MSK) forward simulator. Data-driven inverse transformers infer kinetic causes (torques) from observed kinematic consequences (motion). Differentiable Forward Kinematics and Dynamics (FK and FD) layers act as a physics-based verifier, simulating motion from the inferred kinetics. This physics-regulated loop embeds domain knowledge during training and inference, ensuring physically plausible estimations. We train this architecture with a novel forward-inverse consistency loss, ensuring the simulated motion faithfully reconstructs the original observation. Extensive experiments on BML-MoVi, BEDLAM, and OpenCap datasets demonstrate that MonoMSK not only achieves SOTA kinematic accuracy but also delivers the first precise monocular kinetics estimation.

## References

- [1] Marian Bittner, Wei-Tse Yang, Xucong Zhang, Ajay Seth, Jan van Gemert, and Frans CT Van der Helm. Towards single camera human 3d-kinematics. *Sensors*, 23(1):341, 2022. 2, 7
- [2] Michael J Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. Bedlam: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8726–8737, 2023. 6
- [3] Joao Carreira, Pulkit Agrawal, Katerina Fragkiadaki, and Jitendra Malik. Human pose estimation with iterative error feedback. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4733–4742, 2016. 5
- [4] Matthieu Caruel and Lev Truskinovsky. Physics of muscle contraction. *Reports on Progress in Physics*, 81(3):036602, 2018. 3
- [5] Elena Ceseracciu, Zimi Sawacha, and Claudio Cobelli. Comparison of markerless and marker-based motion capture technologies through simultaneous data collection during gait: proof of concept. *PLoS one*, 9(3):e87640, 2014. 1
- [6] Scott L Delp, Frank C Anderson, Allison S Arnold, Peter Loan, Ayman Habib, Chand T John, Eran Guendelman, and Darryl G Thelen. Opensim: open-source software to create and analyze dynamic simulations of movement. *IEEE transactions on biomedical engineering*, 54(11):1940–1950, 2007. 3
- [7] Christopher L Dembia, Nicholas A Bianco, Antoine Falisse, Jennifer L Hicks, and Scott L Delp. Opensim moco: Musculoskeletal optimal control. *PLOS Computational Biology*, 16(12):e1008493, 2020. 6
- [8] Sai Kumar Dwivedi, Yu Sun, Priyanka Patel, Yao Feng, and Michael J Black. Tokenhm: Advancing human mesh recovery with a tokenized pose representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1323–1333, 2024. 1, 2, 7
- [9] Roy Featherstone. *Rigid body dynamics algorithms*. Springer, 2008. 3
- [10] Saeed Ghorbani, Kimia Mahdaviani, Anne Thaler, Konrad Kording, Douglas James Cook, Gunnar Blohm, and Nikolaus F Troje. Movi: A large multi-purpose human motion and video dataset. *Plos one*, 16(6):e0253157, 2021. 6
- [11] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4d: Reconstructing and tracking humans with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14783–14794, 2023. 1, 2, 7
- [12] Kenneth H Hunt and Frank R Erskine Crossley. Coefficient of restitution interpreted as damping in vibroimpact. 1975. 4
- [13] Farnoosh Koleini, Muhammad Usama Saleem, Pu Wang, Hongfei Xue, Ahmed Helmy, and Abbey Fenwick. Biopose: Biomechanically-accurate 3d pose estimation from monocular videos. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 6330–6339. IEEE, 2025. 2, 5, 7
- [14] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4501–4510, 2019. 2
- [15] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 851–866. 2023. 2, 5
- [16] Richard M Murray, Zexiang Li, and S Shankar Sastry. *A mathematical introduction to robotic manipulation*. CRC press, 2017. 3
- [17] Priyanka Patel and Michael J Black. Camerahmr: Aligning people with perspective. In *2025 International Conference on 3D Vision (3DV)*, pages 1562–1571. IEEE, 2025. 1, 2, 7
- [18] Apoorva Rajagopal, Christopher L Dembia, Matthew S DeMers, Denny D Delp, Jennifer L Hicks, and Scott L Delp. Full-body musculoskeletal model for muscle-driven simulation of human gait. *IEEE transactions on biomedical engineering*, 63(10):2068–2079, 2016. 4
- [19] Ajay Seth, Jennifer L Hicks, Thomas K Uchida, Ayman Habib, Christopher L Dembia, James J Dunne, Carmichael F Ong, Matthew S DeMers, Apoorva Rajagopal, Matthew Millard, et al. Opensim: Simulating musculoskeletal dynamics and neuromuscular control to study human and animal movement. *PLOS computational biology*, 14(7):e1006223, 2018. 1, 2, 3, 4, 6
- [20] Scott D Uhlrich, Antoine Falisse, Łukasz Kidziński, Julie Muccini, Michael Ko, Akshay S Chaudhari, Jennifer L Hicks, and Scott L Delp. Opencap: Human movement dynamics from smartphone videos. *PLOS computational biology*, 19(10):e1011462, 2023. 1, 2, 6, 7
- [21] Andreas Wächter and Lorenz T Biegler. On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. *Mathematical programming*, 106(1):25–57, 2006. 6
- [22] Jacqueline Kory Westlund, Sidney K D’Mello, and Andrew M Olney. Motion tracker: Camera-based monitoring of bodily movements using motion silhouettes. *PLoS one*, 10(6):e0130293, 2015. 1
- [23] Yufei Zhang, Jeffrey O Kephart, Zijun Cui, and Qiang Ji. Physpt: Physics-aware pretrained transformer for estimating human dynamics from monocular videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2305–2317, 2024. 2, 5
- [24] Yufei Zhang, Jeffrey O Kephart, and Qiang Ji. Incorporating physics principles for precise human motion prediction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6164–6174, 2024. 5