OpenCap Monocular: 3D Human Kinematics and Musculoskeletal Dynamics from a Single Smartphone Video

Selim Gilon Emily Y. Miller Scott D. Uhlrich
Department of Mechanical Engineering, University of Utah, Salt Lake City, UT 84112
selim.gilon@utah.edu emily.miller@utah.edu scott.uhlrich@utah.edu

Abstract

Scalable measurement of human movement (kinematics) and musculoskeletal forces (kinetics), for example, estimating quadriceps force during sit-to-stand—could transform the prediction, treatment, and monitoring of mobility-related conditions. Yet these analyses still rely on costly, time-intensive laboratory workflows, limiting clinical translation. We present OpenCap Monocular, a pipeline that estimates 3D skeletal kinematics and musculoskeletal kinetics from a single static smartphone video. The method refines 3D global pose outputs from a monocular pose estimation model (WHAM) through a physics-inspired pose optimization that enforces reprojection consistency and foot-floor contact constraints. Keypoints are extracted from the mesh using the refined global pose, and skeletal kinematics of a 33-degree-of-freedom musculoskeletal model are obtained using inverse kinematics. Kinetics (i.e., ground forces and joint moments) are estimated via physics-based simulation and machine learning, without force plates. Validated against marker-based motion capture and force plates for walking, squatting, and sit-to-stand, OpenCap Monocular achieves 4.8° mean absolute error (MAE) for rotational kinematics and 3.4 cm for translations, corresponding to 48% and 69% lower error than a direct computer vision baseline. It estimates walking ground reaction forces with 9.7% bodyweight MAE, comparable to a two-camera system. We demonstrate that the algorithm estimates important kinetic outcomes with clinically meaningful accuracy in applications related to frailty (knee extension, hip extension, and ankle plantarflexion moments during sit-to-stand transitions) and knee osteoarthritis (knee adduction moment during walking). OpenCap Monocular is deployed via a smartphone app, a web app, and secure cloud computing (https://opencap.ai), enabling free, accessible single-smartphone biomechanical assessments. Our code is available at github.com/utahmobl/opencap-monocular.

1. Introduction

Quantitative analysis of human movement provides critical information for rehabilitation, sports science, and the treatment of musculoskeletal and neuromuscular disorders. Biomechanical measures of kinematics (joint angles, velocities) and kinetics (joint moments, ground reaction forces, muscle forces) can predict injury risk, track functional recovery, and evaluate interventions [9, 14, 16, 22]. The gold standard for measuring these quantities is laboratory-based motion capture with reflective markers and force plates. This approach requires expensive equipment (often exceeding $150,000), dedicated laboratory space, and specialized expertise. As a result, its use in clinical and large-scale settings remains limited [19, 21]. Most biomechanics studies are conducted in marker-based motion capture labs with a median of only 12–21 participants [5, 13].

Video-based approaches, leveraging advancements in computer vision and deep learning for human pose estimation, offer a path toward scalable biomechanical analysis [2, 3, 25]. For example, we developed OpenCap [21], a cloud-based platform that estimates musculoskeletal kinematics and kinetics using two or more smartphone videos. It is used by 14,000 researchers worldwide who collect 1,000 motion trials per day. However, even a two-camera setup requires tripod-mounted devices, calibration, and a laptop, creating barriers for routine clinical or at-home use.

3D pose estimation from monocular video removes these logistical barriers, but introduces fundamental challenges. Without multiple views, absolute depth is ambiguous and scale is unknown. Monocular pose estimation models such as WHAM [18] estimate the global pose of SMPL [7] body models, but they can suffer from translational drift and physically implausible foot-floor interactions. These artifacts prevent direct use for downstream biomechanical analysis and compromise kinetic estimates. Accurate joint moments and muscle forces require physically realistic, biomechanically constrained kinematics.

Here, we present OpenCap Monocular, a pipeline that bridges monocular pose estimation and musculoskeletal

biomechanics (Fig. 1). Our core contribution is a physics-inspired pose refinement optimization that corrects WHAM estimates for physical plausibility by minimizing reprojection error and enforcing foot-floor contact constraints. Refined kinematics drive an OpenSim [17] musculoskeletal model via inverse kinematics, and kinetics are estimated via physics-based simulation [4, 21] and machine learning [20]. The complete pipeline is deployed as a free cloud application accessible at opencap.ai, requiring under one minute to set up and under two minutes to process a 10-second video.

2. Methods

The OpenCap Monocular pipeline converts a single static smartphone video into 3D joint kinematics and musculoskeletal dynamics through five steps shown in Fig. 1: initial 3D pose estimation, physics-inspired pose refinement, virtual marker extraction, inverse kinematics, and dynamics estimation.

For validation, we used the public OpenCap dataset [21]: 10 healthy adults (5 females; age 26±4 years; mass 74±8 kg). Participants performed level walking, five bodyweight squats, and five sit-to-stand transitions, plus modified tasks to increase biomechanical variability. These included squats with one-foot offloading, sit-to-stand with increased trunk flexion/angular velocity (as seen in older adults with quadriceps weakness [10, 23]), and trunk-sway gait modifications. We compared OpenCap Monocular against (1) marker-based motion capture and force plates (gold standard), (2) a computer vision + inverse kinematics (CV+IK) baseline, and (3) the two-camera OpenCap system [21]. The CV+IK baseline used the same downstream marker extraction, OpenSim inverse kinematics, and dynamics estimation as OpenCap Monocular, but operated directly on the initial ViTPose+WHAM pose estimate without the physics-inspired pose refinement stage. Kinematic accuracy was quantified as mean absolute error (MAE) across 18 rotational degrees-of-freedom and 3 pelvic translational degrees-of-freedom. All recordings used the 45° anterolateral camera view, which minimized occlusions and can observe motion in both the sagittal and frontal planes.

2.1. Initial 3D Pose Estimation

We use ViTPose [26] to estimate 2D keypoint locations and confidence scores. Then, we use WHAM [18] to create an initial estimate of the global 3D human pose as a sequence of SMPL body-model parameters [7]: body shape $\boldsymbol{\beta}_0$, pose $\boldsymbol{\theta}_0$, global translation $\boldsymbol{\tau}_0$, and global orientation $\boldsymbol{\Gamma}_0$. WHAM also provides camera extrinsic parameters $\boldsymbol{\xi}$ and ground contact probabilities for the heel and toe.

While WHAM provides a strong initial estimate, its outputs can suffer from translational drift over time and physically implausible foot-floor interactions (*e.g.*, foot sliding, floor penetration), which motivate our refinement step.

2.2. Physics-Inspired Pose Refinement

To enforce physical plausibility, we perform a two-stage optimization. We assume a static camera with known intrinsic parameters (retrieved from our iOS device database for all iPhone/iPad models released since 2018) and a known participant height (entered at recording time in the web application). These assumptions, enabled by the structured OpenCap acquisition workflow [21], simplify the depth estimation problem. Activity-specific optimization weights are selected automatically by classifying the activity in the video using VideoLLaMA3 [27].

Stage 1: shape and camera calibration. We refine body shape $\boldsymbol{\beta}$ and camera extrinsics $\boldsymbol{\xi}$, holding the global pose fixed:

$$\mathcal{J}_1 = w_r\, L_{\text{repr}} + w_h\, L_{\text{height}} + w_\beta\, L_\beta\,, \tag{1}$$

where L_{repr} is the confidence-weighted sum of squared 2D reprojection errors across all keypoints, L_{height} penalizes the squared deviation of the estimated body height from the known participant height, and L_β regularizes shape parameters toward the WHAM estimate.

Stage 2: global pose refinement. With body shape fixed, we jointly refine the global pose ($\boldsymbol{\theta}$, $\boldsymbol{\tau}$, $\boldsymbol{\Gamma}$) and camera extrinsics:

$$\begin{aligned}\mathcal{J}_2 = {} & w_r\, L_{\text{repr}} + w_c\, L_{\text{cam}} + w_v\, L_{\text{fvel}} \\ & + w_s\, L_{\text{fslide}} + w_f\, L_{\text{flat}} + w_{sm}\, L_{\text{smooth}}\,,\end{aligned} \tag{2}$$

where L_{cam} penalizes changes in camera extrinsics from Stage 1; L_{fvel} penalizes non-zero heel and toe velocities during ground contact; L_{fslide} penalizes drift in heel/toe positions across continuous contact bouts; L_{flat} enforces a consistent vertical ground height across all contact events; and L_{smooth} penalizes large joint velocities to encourage temporal smoothness.

2.3. Marker Extraction

From the optimized SMPL pose, we extract 38 virtual surface markers as specific mesh vertices characterizing the forearm, upper arm, torso, pelvis, thigh, shank, and foot segments. We then fit a 33-degree-of-freedom (DOF) musculoskeletal model [6, 15, 17] to the extracted markers using the OpenSim Scale tool.

2.4. Inverse kinematics

We then run inverse kinematics using OpenSim [17]. Unlike the SMPL model, whose joints are rotationally unconstrained, the musculoskeletal model enforces biomechanically realistic joint movement. The output is a trajectory of biomechanically plausible joint angles and pelvis translations.

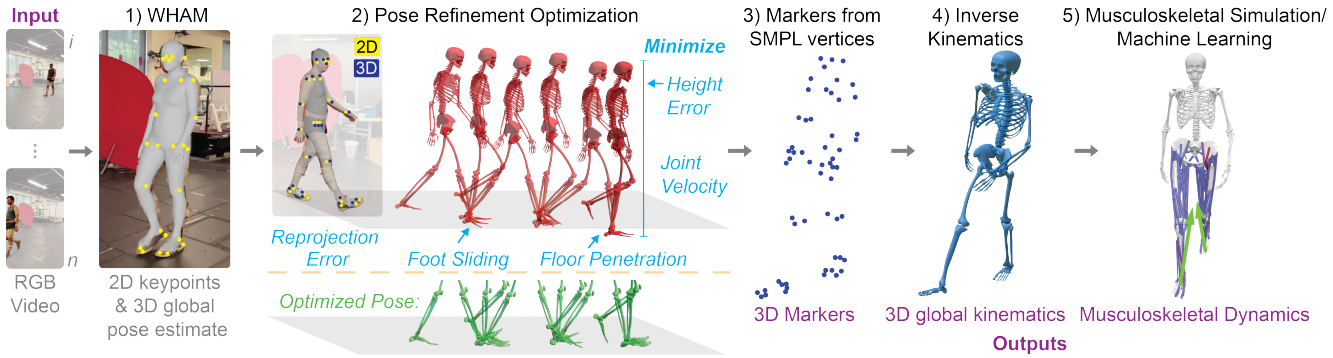

Figure 1. **OpenCap Monocular Pipeline.** (1) ViTPose [26] estimates 2D keypoints; WHAM [18] estimates an initial global 3D pose as SMPL parameters [7]. (2) A two-stage physics-inspired optimization refines the pose (red → green skeleton) by minimizing terms including reprojection error and foot-floor contact constraints, reducing translational drift and foot sliding. (3) Virtual surface markers are extracted from the refined SMPL mesh. (4) OpenSim Scale tool is used to scale a 33-degree-of-freedom musculoskeletal model [15, 17] to the markers. These are tracked using OpenSim Inverse Kinematics yielding biomechanically constrained joint kinematics. (5) Kinetics (ground reaction forces, joint moments, muscle forces) are estimated via physics-based simulation [4, 21] and machine learning [8, 20], without force plates.

2.5. Musculoskeletal Dynamics

For sit-to-stand, we estimate ground reaction forces (GRF), joint moments, and muscle forces via a muscle-driven direct collocation simulation [4, 21] that tracks the monocular kinematics. For walking, we used a hybrid machine learning–simulation approach. We predict GRFs using the GaitDynamics [20] model, a Transformer model that predicts ground forces from kinematics. We predict center of pressure from kinematics and predicted ground forces using another small machine learning model [8]. Finally, to obtain dynamically consistent joint moments, we track predicted ground forces and centers of pressure, along with monocular kinematics, in a torque-driven physics simulation [4, 8].

3. Results

3.1. Kinematic Accuracy

Across all activities, OpenCap Monocular achieved an MAE of 4.8° for rotational kinematics and 3.4 cm for pelvic translations, compared to marker-based motion capture (Fig. 2). These errors are 48% ($p = 0.036$) and 69% ($p < 0.001$) lower than the CV+IK baseline, respectively. Rotational accuracy was within 1° of the two-camera OpenCap system, and translational accuracy within 2 cm.

OpenCap Monocular also reduced translational drift. After five consecutive sit-to-stand repetitions, the CV+IK pelvis drifted 56.9 cm from the ground-truth position on average, whereas OpenCap Monocular's drift was 4.9 cm.

3.2. Kinetic Accuracy

Using a hybrid kinetics pipeline [8], physics-based simulation combined with the GaitDynamics model [20], OpenCap Monocular estimated vertical ground reaction forces during walking with an MAE of 9.7% BW, where BW denotes body weight. This represents a 29% improvement over the CV+IK baseline (13.6% BW; $p = 0.002$). OpenCap Monocular's vertical GRF errors were slightly lower than two-camera OpenCap (12.2% BW), likely due to improved center-of-mass kinematics resulting from the pose refinement step.

3.3. Practical Efficiency

In the deployed cloud application, the full workflow required under one minute of setup and under two minutes to process a 10-second video. These runtimes suggest that OpenCap Monocular can support practical biomechanics analysis outside the laboratory, in addition to achieving clinically meaningful accuracy.

3.4. Clinical Use Cases

Joint Moments during Chair Rise. The knee extension moment, averaged over the chair-rise phase and used as a proxy for quadriceps force, was estimated with an MAE of 5.8 Nm ($r^2 = 0.64$), below the 11 Nm threshold that differentiates older adults with and without early signs of frailty [16] (Table 1). OpenCap Monocular also detected the expected redistribution of lower-extremity joint moments with an exaggerated trunk-lean strategy. Knee extension moment decreased ($p = 0.015$), while hip extension moment and ankle plantarflexion moments increased ($p = 0.003$, and $p = 0.044$, respectively). All changes were consistent in direction with lab-based inverse dynamics (marker-based motion capture and force plates, $p < 0.05$).

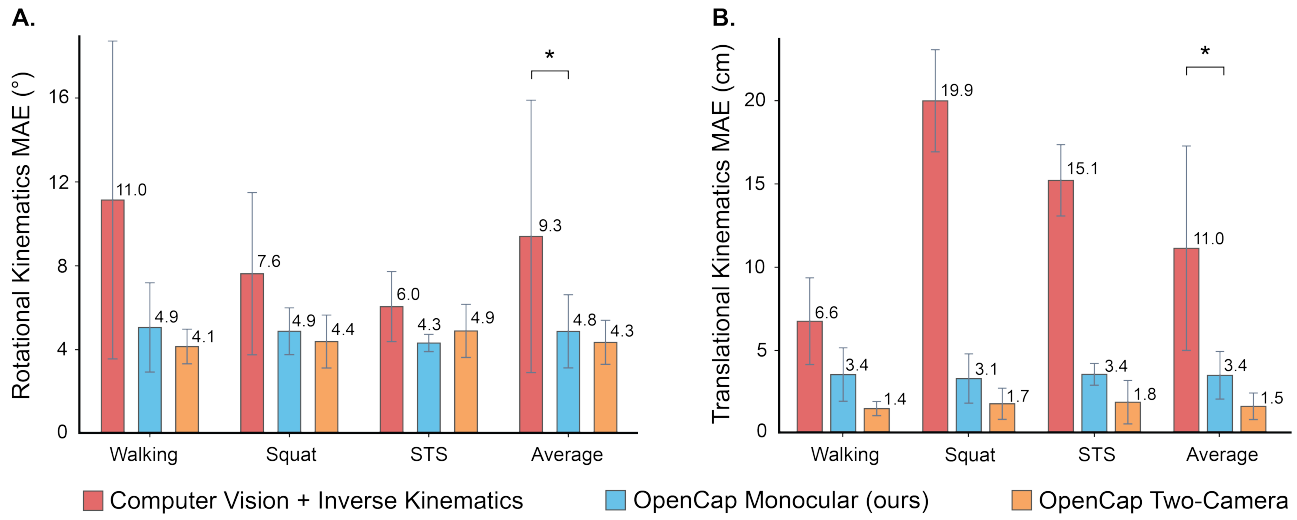

Figure 2. **Kinematic accuracy** (MAE, mean ± SD) across activities compared to marker-based motion capture. OpenCap Monocular reduces rotational error by 48% and translational error by 69% relative to the CV+IK baseline ($^*p < 0.05$), and is within 1° of two-camera OpenCap. STS = sit-to-stand.

Table 1. **Kinetics and downstream clinical tasks.** Two-camera OpenCap (2-Cam) used the same kinetics pipeline as OpenCap Monocular.

Metric	Ours	2-Cam [21]
Vert. GRF MAE (% BW, walking)	**9.7**	12.2
Knee ext. moment MAE (Nm, STS)	**5.8**	—
Clinical threshold (Nm)	< 11	—
1st-peak KAM MAE (% BW·ht)	**0.36**	0.41
Clinical threshold	< 0.5	< 0.5

Knee Loading during Walking. The first-peak knee adduction moment (KAM) during walking is a measure of medial compartment loading associated with osteoarthritis progression [1, 9, 11]. OpenCap Monocular estimated it with an MAE of 0.36% BW·ht, where BW·ht denotes body weight times height, below the 0.5% BW·ht clinically meaningful threshold [9, 11, 12] and comparable to the two-camera OpenCap system (0.41% BW·ht).

4. Discussion

In this study, we developed and validated OpenCap Monocular, a cloud-deployed pipeline that estimates 3D musculoskeletal kinematics and kinetics from a single smartphone video. By combining monocular pose estimation, pose refinement, and musculoskeletal simulation, the method produces physically plausible biomechanics with accuracy sufficient for clinically relevant tasks. OpenCap Monocular bridges computer-vision pose outputs and biomechanics workflows, enabling rapid, accessible movement analysis at scale.

A central finding is the critical role of pose refinement for both kinematic and kinetic accuracy. Without it, the CV+IK baseline produced large translational errors and substantial drift, rendering the kinematics unsuitable for physics-based simulation. Our optimization reduced rotational and translational error by 48% and 69%, respectively. This improvement directly enabled accurate kinetics. OpenCap Monocular achieves accuracy at clinically meaningful levels for two relevant downstream tasks: detecting reduced quadriceps forces when standing with a strategy commonly used in individuals with frailty [16, 24], and estimating knee loading during walking, a key predictor of osteoarthritis progression and an interventional target [9, 22]. Estimating *kinetic* quantities with clinically meaningful accuracy shows that OpenCap Monocular has the potential to support biomedical research and clinical decision making. This biomechanical validation is an advance beyond the way that computer-vision models are typically benchmarked (*e.g.*, mean per-joint position error). The practical runtime of under two minutes for a 10-second video in the deployed cloud application further supports its use in scalable, real-world workflows.

Several limitations warrant future work. First, the validation cohort consisted of healthy young adults; the algorithm's performance in clinical populations with pathological gait remains to be evaluated. Next, WHAM's ground-contact probabilities degrade during activities with significant flight phases (*e.g.*, running, jumping), limiting applicability to those activities currently. Future work should also

investigate the sensitivity of performance to camera position, since all recordings here used a 45° anterolateral view.

Because the pipeline is modular, the pose estimation module can be replaced by newer models as the field advances, without modifying downstream biomechanical components.

5. Conclusion

We present OpenCap Monocular, a pipeline for quantifying 3D human musculoskeletal kinematics and kinetics from a single smartphone video. By combining monocular pose estimation (WHAM [18]) with a pose refinement optimization and a biomechanically constrained musculoskeletal model ([17]), the pipeline produces physically realistic kinematics and clinically meaningful kinetics. It outperforms direct CV model outputs for all reported accuracy metrics and matches the performance of a two-camera setup. Deployed freely at opencap.ai, OpenCap Monocular lowers the barrier to quantitative movement analysis, enabling population-scale biomechanical assessment from a single smartphone.

Acknowledgements

This work was funded by grants from the Myotonic Dystrophy Foundation, the Wu Tsai Human Performance Alliance Agility Project Program, and the NIH Restore Center Pilot Project Program.

References

[1] Shreyasee Amin, Niyom Luepongsak, Chris A. McGibbon, Michael P. LaValley, David E. Krebs, and David T. Felson. Knee adduction moment and development of chronic knee pain in elders. *Arthritis Care and Research*, 51(3):371–376, 2004. 4

[2] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. OpenPose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43:172–186, 2019. 1

[3] Yann Desmarais, Denis Mottet, Pierre Slangen, and Philippe Montesinos. A review of 3d human pose estimation algorithms for markerless motion capture. *Computer Vision and Image Understanding*, 212:103275, 2021. 1

[4] Antoine Falisse, Gil Serrancolí, Christopher Dembia, Joris Gillis, Ilse Jonkers, and Friedl De Groote. Rapid predictive simulations with complex musculoskeletal models suggest that diverse healthy and pathological human gaits can emerge from similar control strategies. *Journal of The Royal Society Interface*, 16, 2019. 2, 3

[5] Duane V. Knudson. Authorship and sampling practice in selected biomechanics and sports science journals. *Perceptual and Motor Skills*, 112(3):838–844, 2011. 1

[6] Adrian K. M. Lai, Allison S. Arnold, and James M. Wakeling. Why are antagonist muscles co-activated in my simulation? A musculoskeletal model for analysing human locomotor tasks. *Annals of Biomedical Engineering*, 45(12): 2762–2774, 2017. 2

[7] 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*. Association for Computing Machinery, New York, NY, USA, 1 edition, 2023. 1, 2, 3

[8] Emily Y. Miller, Tian Tan, Antoine Falisse, and Scott D. Uhlrich. Integrating machine learning with musculoskeletal simulation improves OpenCap video-based dynamics estimation. *bioRxiv*, 2025. 3

[9] T. Miyazaki, M. Wada, H. Kawahara, M. Sato, H. Baba, and S. Shimada. Dynamic load at baseline can predict radiographic disease progression in medial compartment knee osteoarthritis. *Annals of the Rheumatic Diseases*, 61(7):617–622, 2002. 1, 4

[10] Julie Moreland, Julie Richardson, Charlie Goldsmith, and Catherine Clase. Muscle weakness and falls in older adults: A systematic review and meta-analysis. *Journal of the American Geriatrics Society*, 52:1121–1129, 2004. 2

[11] Annegret Mündermann, Chris Dyrby, Debra Hurwitz, Leena Sharma, and Thomas Andriacchi. Potential strategies to reduce medial compartment loading in patients with knee osteoarthritis of varying severity: Reduced walking speed. *Arthritis and Rheumatism*, 50:1172–1178, 2004. 4

[12] Annegret Mündermann, Chris O. Dyrby, and Thomas P. Andriacchi. Secondary gait changes in patients with medial compartment knee osteoarthritis: increased load at the ankle, knee, and hip during walking. *Arthritis and Rheumatism*, 52 (9):2835–2844, 2005. 4

[13] Anderson Souza Oliveira and Cristina Ioana Pirscoveanu. Implications of sample size and acquired number of steps to investigate running biomechanics. *Scientific Reports*, 11 (1):3083, 2021. 1

[14] Mark Paterno, Laura Schmitt, Kevin Ford, Mitchell Rauh, Gregory Myer, Bin Huang, and Timothy Hewett. Biomechanical measures during landing and postural stability predict second anterior cruciate ligament injury after anterior cruciate ligament reconstruction and return to sport. *The American Journal of Sports Medicine*, 38:1968–1978, 2010. 1

[15] 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. 2, 3

[16] T. Seko, H. Akasaka, M. Koyama, N. Himuro, S. Saitoh, S. Ogawa, S. Miura, M. Mori, and H. Ohnishi. The contributions of knee extension strength and hand grip strength to factors relevant to physical frailty: The tanno-sobetsu study. *Geriatrics*, 9(1):9, 2024. 1, 3, 4

[17] 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, Samuel R. Hamner, Edith M. Arnold, Jennifer R. Yong, Shrinidhi K. Lakshmikanth, Michael A. Sherman, Joy P. Ku, and Scott L. Delp. Open-

Sim: Simulating musculoskeletal dynamics and neuromuscular control to study human and animal movement. *PLOS Computational Biology*, 14(7):1–20, 2018. 2, 3, 5

[18] Soyong Shin, Juyong Kim, Eni Halilaj, and Michael Black. WHAM: Reconstructing world-grounded humans with accurate 3d motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2070–2080, 2024. 1, 2, 3, 5

[19] Julie Stebbins, Marian Harrington, and Caroline Stewart. Clinical gait analysis 1973–2023: Evaluating progress to guide the future. *Journal of Biomechanics*, 160:111827, 2023. 1

[20] Tian Tan, Tom Van Wouwe, Keenon F. Werling, C. Karen Liu, Scott L. Delp, Jennifer L. Hicks, and Akshay S. Chaudhari. GaitDynamics: A generative foundation model for analyzing human walking and running. *Nature Biomedical Engineering*, 2026. 2, 3

[21] 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):1–26, 2023. 1, 2, 3, 4

[22] Scott D Uhlrich, Valentina Mazzoli, Amy Silder, Andrea K Finlay, Feliks Kogan, Garry E Gold, Scott L Delp, Gary S Beaupre, and Julie A Kolesar. Personalised gait retraining for medial compartment knee osteoarthritis: a randomised controlled trial. *The Lancet Rheumatology*, 7(10):e708–e718, 2025. 1, 4

[23] Marion Van der Heijden, Kenneth Meijer, Paul Willems, and Hans Savelberg. Muscles limiting the sit-to-stand movement: An experimental simulation of muscle weakness. *Gait & Posture*, 30:110–114, 2009. 2

[24] Eline Van der Kruk, Anne Silverman, Peter Reilly, and Anthony Bull. Compensation due to age-related decline in sit-to-stand and sit-to-walk. *Journal of Biomechanics*, 122: 110411, 2021. 4

[25] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao. Deep high-resolution representation learning for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 1

[26] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. ViTPose: Simple vision transformer baselines for human pose estimation. In *Advances in Neural Information Processing Systems (NeurIPS)*, Red Hook, NY, USA, 2022. Curran Associates Inc. 2, 3

[27] Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, Peng Jin, Wenqi Zhang, Fan Wang, Li Bing, and Deli Zhao. VideoLLaMA 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025. 2