OpenCapBench: A Benchmark to Bridge Pose Estimation and Biomechanics

Yoni Gozlan
Stanford University
yonigoz@stanford.edu

Antoine Falisse
Stanford University
afalisse@stanford.edu

Scott Uhlrich
Stanford University
suhlrich@stanford.edu

Anthony Gatti
Stanford University
aagatti@stanford.edu

Michael Black
Max Planck Institute for Intelligent Systems
black@tuebingen.mpg.de

Jennifer Hicks
Stanford University
jenhicks@stanford.edu

Scott Delp
Stanford University
delp@stanford.edu

Akshay Chaudhari
Stanford University
akshaysc@stanford.edu

Abstract

Pose estimation has promised to impact healthcare by enabling more practical methods to quantify nuances of human movement and biomechanics. However, despite the inherent connection between pose estimation and biomechanics, these disciplines have largely remained disparate. For example, most current pose estimation benchmarks use metrics such as Mean Per Joint Position Error, Percentage of Correct Keypoints, or mean Average Precision to assess performance, without quantifying kinematic and physiological correctness - key aspects for biomechanics. To alleviate this challenge, we develop OpenCapBench to offer an easy-to-use unified benchmark to assess common tasks in human pose estimation, evaluated under physiological constraints. OpenCapBench computes consistent kinematic metrics through joints angles provided by an open-source musculoskeletal modeling software (OpenSim). Through OpenCapBench, we demonstrate that current pose estimation models use keypoints that are too sparse for accurate biomechanics analysis. To mitigate this challenge, we introduce SynthPose, a new approach that enables finetuning of pre-trained 2D human pose models to predict an arbitrarily denser set of keypoints for accurate kinematic analysis through the use of synthetic data. Incorporating such finetuning on synthetic data of prior models leads to twofold reduced joint angle errors. Moreover, OpenCapBench allows users to benchmark their own developed models on our clinically relevant cohort. Overall, OpenCapBench bridges the computer vision and biomechanics communities, aiming to drive simultaneous advances in both areas.

1. Introduction

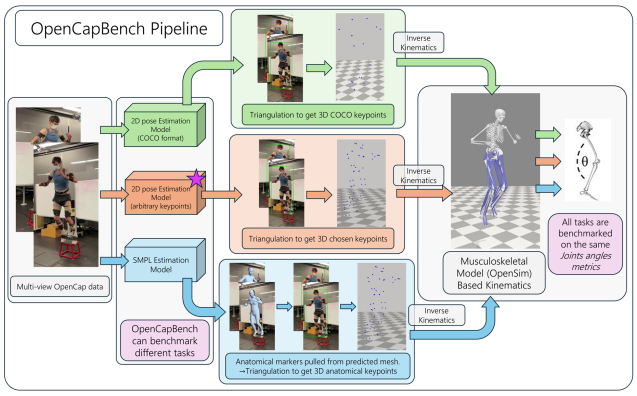

Figure 1. OpenCapBench pipeline. SynthPose, our method to finetune 2D pose estimation models to predict any set of body keypoints (designated by a star here) is detailed in Figure 4.

A major part of kinematic biomechanical analysis is the study of joint angles that are critical for understanding the interplay between body segments for use in applications ranging from diagnostics [2, 10, 36] and intervention strategies [37] to optimizing athletic performance [25]. Traditionally, acquiring high-quality kinematic data for research and clinical studies requires a dedicated gait laboratory with synchronized high-speed cameras, application of multiple optical motion markers, and expert personnel trained in biomechanics. These cumbersome requirements make clinical assessment and large-scale clinical trials cost-prohibitive. Meanwhile, in the domain of computer vision, pose estimation models strive to capture and predict human movement from single or multiple videos. Yet, despite the clear intersections in the objectives of the biomechanics and

pose estimation domains, there remains a disconnect in their methodologies and evaluations. Importantly, biomechanical models for estimating kinematics constrain joints to move in physiologically realistic ways. For example, while biomechanics researchers constrain movement of the knee to only have one degree-of-freedom joint [13], computer vision approaches use physiologically implausible unconstrained three degree-of-freedom motion [33].

Evaluating kinematic metrics is not only important in clinical and sports biomechanics, but can also improve machine learning approaches to pose estimation [11]. Using more physiologic joints may serve as a form of regularization, thus improving estimated poses. Furthermore, kinematic metrics provide a richer, temporally consistent, and functionally relevant evaluation compared to traditional pose estimation metrics like Mean Per Joint Position Error (MPJPE) [17], Percentage of Correct Keypoints (PCK) or mean Average Precision (mAP) [51]. Kinematic metrics such as Root Mean Squared Error (RMSE) of joint angle better encapsulate the complexities and constraints of human motion, and by emphasizing these metrics, models might generalize more effectively across diverse and out-of-distribution poses, view angles, and occlusions [18, 20].

Biomechanical studies [43] show that popular computer vision-based pose estimation models and datasets with sparse keypoint annotations only on joint centers [8, 27] result in large errors in joint angles. These errors are likely owing to the fact that only estimating joint centers leaves identifying specific joint angle contributions from the three anatomical axes unconstrained. Therefore, while typical computer vision metrics focused on keypoints accuracy might be satisfactory, specific joint kinematics can still have large errors, illustrating the need for improved benchmarks and metrics of pose estimation.

Connecting the realms of biomechanical kinematic analysis and pose estimation with computer vision can benefit both fields. Tighter integration can provide real-world benchmarks for computer vision researchers, while translating promising pose estimation models into clinical practice can benefit biomechanics researchers. Against this backdrop, our work aims to bridge the current separation between these disciplines. Our contributions are as follows:

- We introduce **OpenCapBench**, a benchmark to align the fields of biomechanics and pose estimation. OpenCapBench includes a **fully automated pipeline** to streamline the transfer from pose estimation results to the widely-used musculoskeletal modeling and simulation software OpenSim [40]. This integration allows computer vision experts to seamlessly generate kinematic analyses, without requiring expertise in musculoskeletal biomechanical modeling.

- We introduce **SynthPose, a novel method that uses synthetic data to allow efficient finetuning of pretrained pose estimation models to predict a denser set of keypoints** and improve biomechanical analysis.

- Using our new kinematic benchmark, we show that compared to sparse keypoints, our Synthpose method **twofold reduces average joint angle RMSE** and up to fourfold for certain biomechanically-relevant body joints.

The benchmarking pipeline and the different components of SynthPose will be available ¡future github link¿

In uniting the strengths of biomechanics and computer vision, we envision a future where pose estimation models are not just technically proficient, but can also help improve human movement analysis and human health outcomes.

2. Related Work

2D Pose Estimation: Datasets like COCO [27] and MPII [4] include extensive annotated human keypoints for solving monocular pose estimation tasks across a wide array of static images, with annotations for 17 and 16 2D body joint centers, respectively. However, 2D pose estimation datasets do not offer depth information which is crucial for understanding realistic 3D movement patterns.

Video-Based Pose Estimation: Temporal datasets such as JHMDB [19] and PoseTrack [3] have introduced the challenge of maintaining consistency across frames, a step towards measuring continuous motion pertinent to biomechanical research. However, current computer vision benchmarks primarily focus on visual consistency rather than biomechanical accuracy, indicating a gap for new benchmarks, metrics, and datasets that also evaluate temporal kinematic consistency.

3D Pose Estimation: Datasets such as Human3.6M [17] and CMU Panoptic [22] have enabled 3D monocular and multi-view pose estimation methods. However, even translating monocular 3D poses into biomechanically valid models remains a complex task [6]. Current pose estimation metrics primarily focus on joint position accuracy and fail to assess biomechanical factors such as joint angle limits. This leads to the generation of physiologically implausible movement solutions that are inadequate for comprehensive biomechanical analysis. Multi-view 3D pose estimation allows for more precise 3D reconstructions, especially in datasets with occlusions, which often hinder monocular estimations. Yet, there exists no rigorous benchmark specifically designed to evaluate the biomechanical plausibility of these estimations.

Motion Capture Datasets: Conventional motion capture

(MoCap) technology is a specialized and costly resource that requires multiple hours for data collection per subject. Yet, it is indispensable for producing high-fidelity human movement data. MoCap has given rise to datasets such as OpenCap [43], MoVi [15] or PSU-TMM100 [38], which offer detailed continuous kinematic human motion data. While MoCap provides unparalleled accuracy, its reliance on elaborate equipment and controlled environments limits its accessibility and scalability. Despite these limitations, MoCap datasets enable validating and testing pose estimation models, to ensure that the models trained on more generalized data can be benchmarked against the "gold standard" of human movement data, as well as providing the necessary ground truth for benchmarking biomechanically grounded kinematic metrics such as joint angles [41].

Biomechanical Validity and Parametric Models: The development of the SMPL model [29] and its successor SMPL-X [33] has been instrumental in introducing the notion of parametric body shape. These models have facilitated the creation of datasets used for 3D body shape and pose estimation [16, 45] with annotations based on the SMPL framework. Although useful, even such data often lack precision necessary for biomechanically accurate pose estimation. The pursuit of biomechanical validity has thus seen the adoption of models from biomechanical research like the Rajagopal et al. model [34], which is implemented in the OpenSim musculoskeletal modeling platform [40]. The recent introduction of the SKEL model [23], a SMPL-like model with more biomechanically-realistic degrees of freedom represents another step towards biomechanically accurate modeling of human movement.

Synthetic Data for Pose Estimation: Synthetic data has recently found applications in a wide range of fields, including image classification [5], natural language processing [46], healthcare [9] and more. The field of pose estimation is no exception: synthetic datasets using SMPL to model and animate their subjects [7, 32, 44, 47] seek to overcome the lack of labeled multi-view 3D data. Synthetic data enables sampling a wide array of human poses and shapes in diverse environments, and provides rich annotations of different types such as 2D/3D keypoints, SMPL meshes, segmentation masks etc. However, synthetic data still faces the same challenge as real data; they require better evaluation metrics to assess biomechanical outcomes.

Computer Vision for Biomechanical Metrics Estimation: Recent efforts have been channeled into leveraging computer vision techniques to estimate biomechanical metrics directly [24]. The OpenCap [43] platform uses pose estimation outputs for accurate biomechanical analysis in two stages - first by predicting sparse keypoints using pose estimation models followed by lifting from a sparser to denser set of keypoints using MoCap data and recurrent networks. Bittner et al explore the challenge of reconstructing 3D kinematics from monocular video data [6]. BioPose-3D [39] aims to predict 3D biomechanical joint corrections for video-based joint detection methods. These methodologies underscore a growing trend in leveraging computer vision for biomechanical assessment. Our OpenCapBench framework seeks to extend these efforts by providing a holistic, easily accessible evaluation that benchmarks pose estimation models against biomechanically relevant metrics.

3. OpenCapBench Pipeline

OpenCapBench introduces a comprehensive benchmarking pipeline, designed to evaluate the efficacy of pose estimation models in the context of biomechanics.

3.1. Dataset Integration

The foundation of our benchmark is the OpenCap [43] dataset, a biomechanically-focused MoCap and multi-view data collection. This dataset includes exercises commonly used in biomechanical studies, such as squats, sit-to-stand, drop jumps and walking, performed by 10 different subjects. There are 16 movements per person, with durations between 2 to 8 seconds. Each movement is recorded by 5 synchronized and calibrated cameras; we include two cameras in our experiments following results from Uhlrich et al that suggest that two cameras provide comparable accuracy to more cameras [43]. OpenCap includes 3D marker data obtained with an eight-camera MoCap system (Motion Analysis Corp., Santa Rosa, CA, USA) that tracked the positions (100 Hz) of 31 retroreflective markers placed on established anatomical landmarks and 20 tracking markers [43]. The joint angles we use as ground-truth were obtained from the MoCap markers using OpenSim's Inverse Kinematics tool [40] and the Rajagopal model [34]. We also added bounding boxes obtained with high-performing human detection model [28] to the OpenCap videos for subject cropping, in order to provide a fair basis to benchmark all pose estimation models on.

3.2. Benchmarking Pipeline

Our benchmark is characterized by its versatility and modular design, capable of evaluating a wide range of pose estimation tasks. For the purpose of this paper, we focus on evaluation of 2D single-frame pose estimation models and single-frame SMPL pose and shape estimation models. We apply these models on the video sequences from OpenCap Cam1 and Cam3, adopting a common pipeline to go from 2D single frame pose estimation to 3D joints kinematics detailed below. The goal of this common pipeline is to

provide an automated and fair baseline to evaluate the most basic tasks in pose estimation while losing as little information as possible in the process. However, we note that the way OpenCapBench is designed allows adapting to additional tasks, such as 3D multi-view , 3D monocular, 2D/3D temporal pose estimation or even direct kinematics estimation, by selectively bypassing or modifying components in our modular pipeline.

3.2.1 2D keypoints extraction

The first stage of the OpenCapBench pipeline (Figure 1) involves extracting a set of 2D body keypoints for each frame of the two different camera feeds. **This is where users wanting to test their single frame 2D pose or SMPL shape estimation models can easily integrate their models.** 2D keypoint extraction can be performed in one of two ways:

- **Using a 2D pose estimation model predicting any set of 2D body keypoints** (as long as the Inverse Kinematics setup has been defined for this set). The only post-processing done to the outputted predictions is temporal denoising prior to multi-view triangulation using a dual pass low-pass smoothing with a Butterworth filter with a cut-off frequency of 30Hz. We have integrated the MMPose framework [12] 2D pose inference pipeline to OpenCapBench, thus if a model is available on MMPose, it can directly be benchmarked on OpenCapBench without further modification needed. Otherwise, the user can plug-in their model inference function to the pipeline as described in our Github repository.
- **Using a SMPL shape estimation model.** Users can plug-in their SMPL shape estimation model inference function to the pipeline as described in our Github repository. This time, the post-processing consists in projecting a subset of vertices from the predicted SMPL shape onto the image, as illustrated in Figure 3. By default, the projected set of vertices corresponds to a subset of anatomical markers used in MoCap setups [43] and illustrated in Figure 2. This set of marker was manually chosen by experienced biomechanics researchers using SMPL Blender add-on, by selecting the anatomically closest SMPL vertex for each anatomical marker. The anatomical markers are derived specifically to create biomechanically relevant 3D joint coordinate systems per segment, based on the recommendations of the International Society for Biomechanics [48] .The same post-processing step as for 2D pose estimation is applied to the extracted keypoints. Of course, due to the modular nature of OpenCapBench, other marker sets can be used. We include a SKEL tool [23] to either use the default set of markers suggested by SKEL or alternatively, users can define a personalized set and regress the corresponding OpenSim musculoskeletal model markers using the visualization tool.

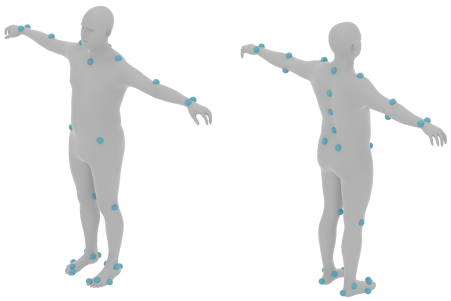

Figure 2. Chosen subset of 35 vertices from SMPL mesh.

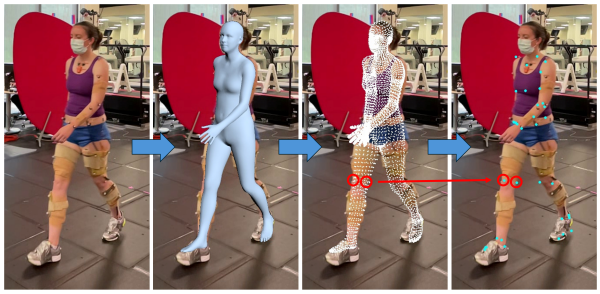

Figure 3. Extracting 2D keypoints from SMPL mesh.

3.2.2 3D Triangulation

Following 2D estimation, the framework employs a deterministic triangulation algorithm to combine multi-view 2D keypoints into 3D keypoints with real-world absolute distances, using viewpoints from two calibrated cameras (Cam1 and Cam3 as defined in OpenCap [43]).

3.2.3 Inverse Kinematics (IK)

The IK step uses the Rajagopal [34] musculoskeletal model to estimate joint kinematics from a sequence of 3D keypoints. This step is performed through OpenSim's python API. Our framework is versatile, providing IK configuration files for common landmark sets (such as COCO, OpenPose, COCO whole body) as well as for the subset of MoCap markers described in Figure 2. Again, we include a SKEL tool [23] which allows the use of a personalized set of SMPL markers and regress the corresponding markers on the OpenSim model.

3.3. Evaluation Metrics

From the outputted joint kinematics, we use a subset of joint angles to be compared with the ground truth joint angles obtained with the MoCap setup of OpenCap [43]. We

use joint angles for *Pelvic Tilt, Pelvic List, Pelvic Rotation, Hip Flexion, Hip Adduction, Hip Rotation, Knee Flexion, Ankle Flexion, Subtalar Inversion/Eversion, Lumbar Extension, Lumbar Bending, and Lumbar Rotation*, which correspond to lower-body kinematics as they are the focus of OpenCap and guaranteed to have accurate ground-truths. Following the literature [14, 35, 43], we use RMSE for the entire waveform for each joint and each trial as metrics.

3.4. Leaderboard

OpenCapBench will feature different leaderboards for individual tasks alongside a global leaderboard encompassing all tasks. This initiative aims to cultivate competition, motivating the computer vision community to develop kinematically accurate pose estimation models, and serve as a resource for biomechanists seeking optimal camera-based kinematics prediction methods. The leaderboard will include all joint angles metrics specified above separately, as well as an average of those to establish a ranking.

4. Arbitrary 2D keypoints estimation using synthetic data

4.1. Motivation

Previous biomechanics studies [31, 43] show the shortcomings of typical pose estimation models which predict sparse sets of keypoints in format such as COCO, OpenPose or MPII. Not only are these models trained and evaluated on manually annotated data, which do not guarantee precise annotations., but the sparse number of keypoints they predict (17-22 keypoints) do not fully characterize the translations and rotations of all body segments. This inadequacy is accentuated between the hips and the shoulders due to the lack of keypoints in this area [43]. Using this limited marker set for inverse kinematics is thus susceptible to result in large angular joint errors.

OpenCap findings [43] show that kinematic metrics can be drastically improved using an LSTM augmenter trained to predict a time series of anatomical markers (corresponding to a subset of MoCap markers) from a time series of sparse keypoints (in COCO-like format). However, this approach is based on MoCap data limited in diversity and captured in controlled environments, and may be prone to overfitting and imprecision since the original image priors are lost in this process. Thus we hypothesize that predicting a denser set of anatomically meaningful keypoints directly from images, akin to MoCap setups, will improve joint angle metrics after inverse kinematics.

To obtain such a model, we can leverage the characteristics of the SMPL model that maintain a fixed mesh topology, meaning that the number of vertices and their connectivity (i.e., the mesh structure) remains unchanged independent of the body shape and pose parameters. This allows for the identification of vertices corresponding to specific anatomical features on the SMPL mesh, such as subsets of MoCap landmarks.

Thus, one potential solution to automatically predict new keypoints involves utilizing an existing model capable of predicting SMPL parameters (or mesh directly [30]) from an image, generating a SMPL mesh from these parameters, extracting a pre-defined subset of SMPL vertices, and projecting them onto the original images, as described in the previous section and in Figure 3. However, this approach presents computational challenges as predicting SMPL parameters and generating meshes are resource-intensive tasks. Moreover, the current state-of-the-art SMPL prediction models do not offer the desired level of accuracy compared to landmark estimation models.

4.2. Leveraging synthetic data to finetune pose estimation models

To address these challenges, we introduce SynthPose, a novel approach for training pose estimation models to predict an arbitrary subset of body keypoints derived from SMPL mesh vertices.

In this work, we focus on predicting bony anatomical keypoints to maximize kinematic accuracy. However, this method can be adapted to specialize models for predicting any sets of body keypoints, such as hands, feet, or head keypoints.

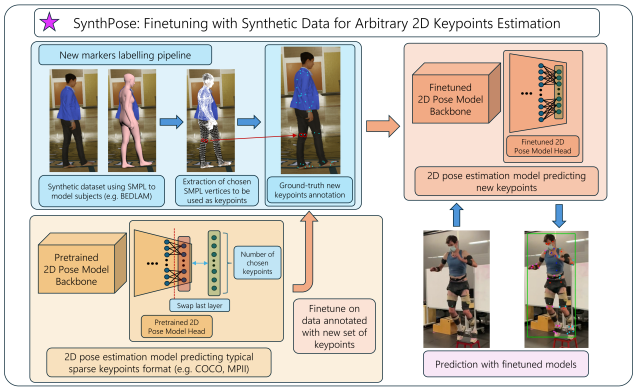

Figure 4. SynthPose: a new method leveraging finetuning of pose estimation models on synthetic data to predict an arbitrary set of 2D keypoints.

Labelling pipeline. As illustrated in Figure 4, a major component of our new method is an automated labelling pipeline, which can create 2D keypoints annotations of a chosen subset of SMPL vertices on any synthetic dataset that uses SMPL to model its subjects. To do so, we generate the SMPL meshes corresponding to the subjects in the synthetic data using the SMPL pose and shape parameters given in the dataset annotations, project them onto the images and extract the specified subset of vertices to be

used as 2D keypoints annotations. The projected keypoints P_{2D} depend on the given camera's intrinsic and extrinsic matrices K and E respectively, and are computed as $P_{2D} = K \times E \times V_{smpl}$ Wwere V_{spml} denotes the subset of vertices of the SMPL model.

Finetuning. The second major component of SynthPose is a finetuning process, where we take a pretrained 2D pose estimation models predicting typical set of keypoints (COCO, MPII etc.), swap the last layer for a layer adapted to the new set of keypoints we want to predict, and finetune this model on our synthetic labelled dataset.

Leveraging synthetic datasets offers several advantages. Besides the benefits of the SMPL model detailed above, these synthetic datasets offer exact annotations, as the mesh of the 3D models depicted in the different images of the datasets are precisely the SMPL meshes. This contrasts with datasets derived from real-world data subsequently annotated with SMPL models, which often exhibit shortcomings in annotation precision, because of the difficulty in precisely manually aligning a SMPL mesh on a 2D image.

By design, SynthPose can leverage learned features from traditional pose estimation training, by using weigths from models trained on datasets like COCO or MPII as initial weights for the new models backbone and prediction head. This transfer learning approach not only significantly reduces training time but also enhances model performance.

In our study, we carefully selected datasets for finetuning our models designed to predict arbitrary keypoints. Each dataset was chosen with specific considerations in mind:

BEDLAM Dataset [7]: This extensive synthetic dataset provides a large-scale foundation for training our models.

Infinity VisionFit [1, 47]: This synthetic dataset, generated using the Infinity VisionFit API (Infinity AI), includes out-of-distribution samples featuring individuals engaged in various exercise routines. This movement diversity is beneficial in enabling our models to generalize effectively across different scenarios.

3DPW Dataset [45]: To address the limitations of our training set which lacks real data and subjects wearing shoes, we incorporate the 3DPW dataset. This dataset contains in-the-wild data of 18 subjects in 60 different scenes. Despite imperfections in annotations, this dataset supplements our training data with valuable real-world examples to help close the sim-to-real gap.

COCO Dataset [27]: In addition to the selected SMPL vertices, our models are designed to also output keypoints in the COCO format. Therefore, we integrate the COCO dataset into our aggregated training dataset. This inclusion also helps mitigate the sim-to-real distribution shift and thus, prevent catastrophic forgetting.

In the next section, we demonstrate that our proposed method significantly outperforms the approach of using state-of-the-art SMPL shape and pose estimation models, followed by projecting a subset of the predicted mesh vertices to predict arbitrary keypoints.

5. Experiments

5.1. New arbitrary 2D keypoints prediction method

In this section, we present experiments which showcase the benefit of using our new method to predict arbitrary 2D keypoints compared to state-of-the-art SMPL estimation models. We will first compare our results on a conventional 2D keypoints estimation benchmark using Percentage of Correct Keypoints (PCK) normalized by bounding boxes at different levels of precision, and then use OpenCapBench for comparison. For this experiment, we choose a subset of 35 SMPL vertices corresponding to a subset of anatomical markers typically used in MoCap setups [34, 43], detailed in Figure 2. We study several state-of-the-art models for each task, namely two SMPL estimation models, CLIFF [26] and VirtualMarkers [30], and two 2D pose estimation models, HRNet-W48+DARK [50] (referred as HRNet-W48 in the tables) and VitPose (Base and Huge) [49]. We choose these 2D pose estimation models to represent diverse SOTA model types (CNN and ViT) and determine potential architecture-based biases. We evaluated two ViT sizes to show that results are coherent with expectations (i.e., Huge outperforms Base).

5.1.1 Results on RICH

	PCK Precision(↑)		
	@0.05	@0.10	@0.20
SMPL (MoCap markers extracted)			
CLIFF [26]	0.640	0.802	0.905
Virtual Markers [30]	0.707	0.844	0.926
SynthPose (predicting MoCap markers)			
HRNet-W48 [50]	0.892	0.958	0.982
ViTPose-B [49]	0.859	0.941	0.971
ViTPose-H [49]	**0.903**	**0.966**	**0.985**

Table 1. Comparison with SOTA SMPL Models on the RICH [16] test set. SynthPose significantly outperforms SMPL estimation based methods. Note that CLIFF, VirtualMarkers and HRNet-W48+DARK share the same HRNet backbone architecture, showing the advantage of Synthpose with similarly-sized models.

We first benchmark different models on the RICH dataset [16]. The RICH dataset captures multi-view out-

Joint angle RMSE (↓)	Pelvis				Hip		Knee	Ankle	Subtalar	Lumbar		
	Tilt	Rotation	List	Flex	Add	Rot				Ext	Bend	Rot
2D pose methods (COCO-wholebody keypoints)												
HRNet-W48	24.8	3.7	4	23.4	5.1	9.9	9.4	8.2	14.2	39.1	5.6	6
RTMW-X [21]	14.4	3.5	4.5	15	5.2	7.9	9.3	7.8	11.4	20.3	5.6	**5.9**
SMPL-based models (MoCap landmarks)												
CLIFF	6.1	4.7	4.7	10.5	6.7	7.4	11.9	11.8	13.2	6.5	6.3	13.3
Virtual Markers	6.0	4.2	4.9	8.7	5.5	8.1	9.5	11.9	10.9	8.5	6.3	9.7
***Our work:** SynthPose (MoCap landmarks)*												
HRNet-W48	5.9	2.9	3.5	8.9	**4.9**	7.3	8.9	8.7	9.5	7.5	4.7	8.3
ViTPose-B	5.2	3.3	3.6	**8.3**	5.3	8	8.6	8.7	9.3	6.6	5.1	10
ViTPose-H	**5.1**	**2.8**	**3.4**	**8.3**	**4.9**	**7.3**	**8.3**	**7.6**	**9.1**	**6.2**	**4.6**	8.7
OpenCap method (using LSTM augmenter trained on 108 hours of MoCap data) [43]												
HRNet-W48	7.4	2.4	3.5	6.8	3.2	4.7	4.3	6.2	6.7	9.2	3.4	5

Table 2. Cross-Comparison of Results. SynthPose significantly improves over computing kinematics from only COCO keypoints, and outperforms SMPL estimation based 2D MoCap landmarks predictions on all joint angles metrics. We emphasize in green results improving over the method used in OpenCap [43], which uses an LSTM augmenter trained on 108 hours of MoCap data.

door and indoor video sequences of diverse subjects performing different physical activities. Although not reaching the same level of precision as synthetic data annotations, the SMPL annotations provided that we use to extract ground-truth 2D keypoints (as in Figure 3) represent the current best achievable quality in datasets containing real-world data. These high-quality data are the result of the fitting of SMPL meshes to 3D human bodies captured by markerless motion capture and 3D body scans. RICH also includes high-resolution 3D scene scans, which allow for accurate vertex-level contact labels on the body. Therefore, we have deemed this dataset the ideal choice for evaluating our novel arbitrary keypoint prediction technique.

We outperform SOTA SMPL mesh prediction-based models by 26% in PCK at 0.05, showing the clear advantage of our method for this particular task.

5.1.2 Results on OpenCapBench

In addition to traditional computer vision benchmarking, we benchmark our method on OpenCapBench using RMSE of kinematic joint angles obtained from OpenSim for each individual trial. We also add 2D pose estimation models which predict COCO-wholebody subset of keypoints (from which we do not use the hands or the face landmarks) to the comparison.
The results are summarized in Table 2.

The results illustrate the advantage of our method. Indeed, we observe that keypoint-based methods obtain better joint angles prediction overall compared to models predicting SMPL meshes. We also show that predicting a MoCap subset of landmarks over COCO keypoints enables clear improvement on all metrics except Lumbar Rotation, with 3-5x reduced RMSEs for Pelvis Tilt, Hip Flexion, and Lumbar Extension metrics.

We underline the fact that we are using the same set of keypoints (described in Figure 2) to compute inverse kinematics in OpenSim for both the SMPL-based method and our newly introduced method in these specific experiments. However, tools such as the one introduced with SKEL [23] allow users to specify any set of SMPL vertices they want to use for inverse kinematics, and visually regress the corresponding markers on the OpenSim model. The set of anatomical keypoints we use is much sparser than the one used by default in SKEL to perform inverse kinematics, which may provide an unfair edge to our method. We simply illustrate here that, with the same set of predicted SMPL vertices, our method performs better than the SMPL-based method. However, we encourage the community to test the denser keypoints suggested by SKEL to determine their effect on performance in the SMPL leaderboard of OpenCap-Bench.

Our proposed synthetic data approach challenges the method used in OpenCap on several metrics, namely Pelvis Tilt, Pelvis Rotation, and Lumbar Extension. Importantly, we only low pass filter raw pose prediction results for our inverse kinematics input. In contrast, OpenCap performs post-processing on the predicted keypoints based on the predictions' confidence and use an LSTM keypoints augmenter model which was trained on 108 hours of motion capture data [43]. This augmenter converts the predicted 3D keypoints from COCO or OpenPose format to MoCap keypoints, leveraging temporal priors in marker prediction.

5.2. Benefits of OpenCapBench

Here, we aim to illustrate how OpenCapBench can offer insights on models that current benchmarks and metrics cannot. To show this, we propose an ablation study on the aggregated dataset on which we finetune the arbitrary keypoints prediction models, by removing one of the dataset for each finetuning run.

We perform this ablation study on both RICH dataset and OpenCapBench. The results are summarized in Tables 3 and 4.

Joint angles RMSE (↓)	Pelvis			Hip			Knee	Ankle	Subtalar	Lumbar		
	Tilt	Rotation	List	Flex	Add	Rot				Ext	Bend	Rot
Full agg. dataset												
HRNet-W48	5.9	2.9	3.5	8.9	4.9	7.3	8.9	8.7	9.5	7.5	4.7	8.3
ViTPose-B	5.2	3.3	3.6	8.3	5.3	8	8.6	8.7	9.3	6.6	5.1	10
ViTPose-H	5.1	2.8	3.4	8.3	4.9	7.3	8.3	7.6	9.1	6.2	4.6	8.7
Without COCO dataset												
HRNet-W48	-0.1	+0.1	+0.1	+0.2	+0.1	+0.2	+1.1	+0.3	+0.3	-0.1	+0.1	-0.1
ViTPose-B	+0.4	0.0	+0.2	+1.2	+0.2	-0.1	+0.6	+0.7	+0.4	+0.3	+0.1	0.0
ViTPose-H	-0.1	0.0	0.0	+0.1	0.0	-0.4	+0.3	0.0	0.0	-0.1	0.0	-0.2
Without BEDLAM Dataset												
HRNet-W48	-0.4	+0.1	+0.3	+0.4	+0.1	+0.3	+0.4	0.0	-0.1	+0.6	+0.2	+0.3
ViTPose-B	0.0	0.0	+0.2	+0.6	+0.1	+0.4	+0.4	+0.8	+0.4	+0.8	+0.1	-0.5
ViTPose-H	-0.5	+0.1	+0.1	-0.1	+0.0	-0.3	+0.1	+0.3	0.0	+0.1	+0.1	+0.1
Without Infinity Data												
HRNet-W48	+0.4	+0.1	+0.1	+0.7	+0.3	+0.8	+0.6	+1.0	+0.6	+1.1	+0.0	+0.2
ViTPose-B	+2.9	+0.2	0.0	+2.8	+0.4	+0.8	-0.1	+0.4	+0.5	+2.3	+0.1	-0.1
ViTPose-H	+2.2	+0.2	-0.1	+2.2	+0.3	+0.5	+0.0	+0.6	+0.3	+2.3	0.0	+0.2
Without 3DPW Dataset												
HRNet-W48	-0.6	-0.1	+0.2	+0.1	+0.0	+0.6	+0.2	-0.1	+1.1	-1.1	+0.1	-0.5
ViTPose-B	0.0	0.0	+0.2	+0.4	+0.1	+0.1	+0.5	-0.2	+0.5	+0.7	+0.1	-0.2
ViTPose-H	+0.3	0.0	+0.2	+1.0	+0.0	+0.2	+0.8	-0.2	+1.0	+0.6	+0.0	-0.7

Table 3. Ablation study on OpenCapBench using SynthPose. Significant ($\geq$ 1.0) decrease/increase in performance over baseline are highlighted in red/green. The study shows the importance of Infinity data when it comes to prediciting accurate kinematics.

	PCK Precision(↑)		
	@0.05	@0.1	@0.2
Full agg. dataset			
HRNet-W48	0.89	0.96	0.98
ViTPose-B	0.86	0.94	0.97
ViTPose-H	**0.90**	**0.97**	**0.99**
Without COCO Dataset			
HRNet-W48	0.87 (-0.02)	0.94 (-0.02)	0.96 (-0.02)
ViTPose-B	0.83 (-0.03)	0.91 (-0.03)	0.95 (-0.02)
ViTPose-H	0.90 (0.00)	0.96 (-0.01)	0.98 (-0.01)
Without BEDLAM Dataset			
HRNet-W48	0.87 (-0.02)	0.95 (-0.01)	0.97 (-0.01)
ViTPose-B	0.85 (-0.01)	0.93 (-0.01)	0.97 (0.00)
ViTPose-H	0.89 (-0.01)	0.96 (-0.01)	0.98 (-0.01)
Without Infinity Data			
HRNet-W48	0.89 (0.00)	0.96 (0.00)	0.98 (0.00)
ViTPose-B+DARK	0.86 (0.00)	0.94 (0.00)	0.97 (0.00)
ViTPose-H+DARK	0.90 (0.00)	0.97 (0.00)	0.99 (0.00)
Without 3DPW Dataset			
HRNet-W48	0.90 (+0.01)	0.97 (+0.01)	0.98 (0.00)
ViTPose-B	0.88 (+0.2)	0.95 (+0.01)	0.97 (0.00)
ViTPose-H	0.91 (+0.01)	0.97 (0.00)	0.99 (0.00)

Table 4. Ablation Study results on RICH [16] test set using SynthPose. Decrease/increase in performance over baseline are indicated in red/green. The study indicates slight negative impact when removing COCO and BEDLAM, and slight positive impact when removing 3DPW from the training set.

In comparing the effects of various datasets on model performance, OpenCapBench offers detailed insights that are not as evident in the RICH ablation study.

The RICH ablation study, using PCK metric at different precision levels, shows no or very slight decrease in performance with the exclusion of each dataset, except without 3DPW, which appears to increase the models' performance.

OpenCapBench, on the other hand, provides a more detailed perspective, particularly highlighting the importance of the Infinity dataset for enhancing predictions on specific anatomical features such as Pelvis Tilt, Hip Flexion, and Lumbar Extension, potentially due to its focus on exercise-related data. It also reveals that while 3DPW may negatively impact some metrics, it is crucial for improving the Subtalar metric, which we hypothesize is due to the fact that 3DPW addresses the lack of subjects wearing shoes in the other datasets of the aggregated training set. This demonstrates OpenCapBench's ability to offer nuanced insights that traditional pose estimation benchmarks cannot provide, into how different datasets uniquely contribute to model performance on biomechanical relevant metrics.

6. Discussions

OpenCapBench represents a step towards integrating kinematics and pose estimation, while introducing SynthPose, a method for estimating arbitrary keypoints which benefits both fields. This approach yields detailed insights into the performance of pose estimation models and the importance of diverse and comprehensive training data in refining these models.

Despite the benefits of OpenCapBench, the dataset diversity within OpenCapBench currently lacks breadth in terms of subject variety, environmental settings, and the range of activities covered, which will be a focus of future work. Integrating additional datasets which use MoCap as ground truth such as MoYo [42] or PSU-TMM100 [38] may extend the benchmark's applicability and relevance across broader kinematic studies.

At present, OpenCapBench primarily focuses on lower body kinematics. Adding upper body kinematics and including upper limb assessments could help characterize more holistic view of human motion.

Finally, the open-source aspect and the versatility of OpenCapBench presents an opportunity for the community to engage with it through other pose estimation tasks such as 3D keypoint estimation and temporal predictions, or through experimenting with different subsets of keypoints and new setups for inverse kinematics.

References

[1] Infinity AI VisionFit API. https://infinity.ai/visionfit. 6

[2] K N An. Kinematic analysis of human movement. *Annals of Biomedical Engineering*, 12(6):585–597, Nov. 1984. 1

[3] Mykhaylo Andriluka, Umar Iqbal, Eldar Insafutdinov, Leonid Pishchulin, Anton Milan, Juergen Gall, and Bernt Schiele. Posetrack: A benchmark for human pose estimation and tracking, 2018. 2

[4] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014. 2

[5] Shekoofeh Azizi, Simon Kornblith, Chitwan Saharia, Mohammad Norouzi, and David J. Fleet. Synthetic data from diffusion models improves imagenet classification, 2023. 3

[6] Marian Bittner, Wei-Tse Yang, Xucong Zhang, Ajay Seth, Jan van Gemert, and Frans C. T. van der Helm. Towards single camera human 3d-kinematics. *Sensors*, 23(1), 2023. 2, 3

[7] Michael J. Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. BEDLAM: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 8726–8737, June 2023. 3, 6

[8] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields, 2019. 2

[9] Pierre Chambon, Christian Bluethgen, Jean-Benoit Delbrouck, Rogier Van der Sluijs, Małgorzata Połacin, Juan Manuel Zambrano Chaves, Tanishq Mathew Abraham, Shivanshu Purohit, Curtis P. Langlotz, and Akshay Chaudhari. Roentgen: Vision-language foundation model for chest x-ray generation, 2022. 3

[10] Laurence Chèze. *Some Clinical Applications*, chapter 5, pages 73–101. John Wiley & Sons, Ltd, 2014. 1

[11] Ross Clark, Yong-Hao Pua, Cristino Carneiro Oliveira, Kelly Bower, Shamala Thilarajah, Rebekah McGaw, Ksaniel Hasanki, and Benjamin Mentiplay. Reliability and concurrent validity of the microsoft xbox one kinect for assessment of standing balance and postural control. *Gait & Posture*, 42, 04 2015. 2

[12] MMPose Contributors. Openmmlab pose estimation toolbox and benchmark. https://github.com/open-mmlab/mmpose, 2020. 4

[13] S L Delp, J P Loan, M G Hoy, F E Zajac, E L Topp, and J M Rosen. An interactive graphics-based model of the lower extremity to study orthopaedic surgical procedures. *IEEE Trans Biomed Eng*, 37(8):757–767, Aug. 1990. 2

[14] Mohsen Gholami, Christopher Napier, and Carlo Menon. Estimating lower extremity running gait kinematics with a single accelerometer: A deep learning approach. *Sensors*, 20(10), 2020. 5

[15] 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, June 2021. 3

[16] Chun-Hao P. Huang, Hongwei Yi, Markus Höschle, Matvey Safroshkin, Tsvetelina Alexiadis, Senya Polikovsky, Daniel Scharstein, and Michael J. Black. Capturing and inferring dense full-body human-scene contact. In *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 13274–13285, June 2022. 3, 6, 8

[17] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, 2014. 2

[18] Joseph H. R. Isaac, M. Manivannan, and Balaraman Ravindran. Corrective filter based on kinematics of human hand for pose estimation. *Frontiers in Virtual Reality*, 2, 2021. 2

[19] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black. Towards understanding action recognition. In *International Conf. on Computer Vision (ICCV)*, pages 3192–3199, Dec. 2013. 2

[20] Zhangjian Ji, Zilong Wang, Ming Zhang, Yapeng Chen, and Yuhua Qian. 2d human pose estimation with explicit anatomical keypoints structure constraints, 2022. 2

[21] Tao Jiang, Xinchen Xie, and Yining Li. Rtmw: Real-time multi-person 2d and 3d whole-body pose estimation, 2024. 7

[22] Hanbyul Joo, Tomas Simon, Xulong Li, Hao Liu, Lei Tan, Lin Gui, Sean Banerjee, Timothy Scott Godisart, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social interaction capture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. 2

[23] Marilyn Keller, Keenon Werling, Soyong Shin, Scott Delp, Sergi Pujades, Liu C. Karen, and Michael J. Black. From skin to skeleton: Towards biomechanically accurate 3d digital humans. In *ACM ToG, Proc. SIGGRAPH Asia*, volume 42, Dec. 2023. 3, 4, 7

[24] Lukasz Kidzinski, Bryan Yang, Jennifer Hicks, Apoorva Rajagopal, Scott Delp, and Michael Schwartz. Deep neural networks enable quantitative movement analysis using single-camera videos. *Nature Communications*, 11:4054, 08 2020. 3

[25] Tron Krosshaug, Atsuo Nakamae, Barry Boden, Lars Engebretsen, Gerald Smith, James Slauterbeck, Timothy Hewett, and Roald Bahr. Mechanisms of anterior cruciate ligament injury in basketball: Video analysis of 39 cases. *The American journal of sports medicine*, 35:359–67, 04 2007. 1

[26] Zhihao Li, Jianzhuang Liu, Zhensong Zhang, Songcen Xu, and Youliang Yan. Cliff: Carrying location information in full frames into human pose and shape estimation, 2022. 6

[27] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. 2, 6

[28] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[29] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, Oct. 2015. 3

[30] Xiaoxuan Ma, Jiajun Su, Chunyu Wang, Wentao Zhu, and Yizhou Wang. 3d human mesh estimation from virtual markers, 2023. 5, 6

[31] Laurie Needham, Murray Evans, Darren P. Cosker, Logan Wade, Polly M. McGuigan, James L. Bilzon, and Steffi L. Colyer. The accuracy of several pose estimation methods for 3d joint centre localisation. *Scientific Reports*, 11(1):20673, Oct 2021. 5

[32] Priyanka Patel, Chun-Hao P. Huang, Joachim Tesch, David T. Hoffmann, Shashank Tripathi, and Michael J. Black. AGORA: Avatars in geography optimized for regression analysis. In *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2021. 3

[33] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, 2019. 2, 3

[34] 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. 3, 4, 6

[35] Eric Rapp, Soyong Shin, Wolf Thomsen, Reed Ferber, and Eni Halilaj. Estimation of kinematics from inertial measurement units using a combined deep learning and optimization framework. *Journal of Biomechanics*, 116:110229, 01 2021. 5

[36] Christoph Reinschmidt, Anton J. van den Bogert, Benno M. Nigg, Arne Lundberg, and Norman Murphy. Effect of skin movement on the analysis of skeletal knee joint motion during running. *Journal of biomechanics*, 30 7:729–32, 1997. 1

[37] Douglas Robertson, Graham Caldwell, Joseph Hamill, Gary Kamen, and Saunders Whittlesey. *Research Methods in Biomechanics: Second edition (eBook)*. 11 2013. 1

[38] Jesse Scott. *Dynamic Stability Monitoring of Complex Human Motion Sequences via Precision Computer Vision*. PhD thesis, 2022 2022. Copyright - Database copyright ProQuest LLC; ProQuest does not claim copyright in the individual underlying works; Last updated - 2023-11-10. 3, 8

[39] Jesse Scott, Bharadwaj Ravichandran, Christopher Funk, Robert T Collins, and Yanxi Liu. From image to stability: Learning dynamics from human pose. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIII 16*, pages 536–554. Springer, 2020. 3

[40] 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. Opensim: Simulating musculoskeletal dynamics and neuromuscular control to study human and animal movement. *PLOS Computational Biology*, 14(7):1–20, 07 2018. 2, 3

[41] Tian Tan, Anthony Gatti, Bingfei Fan, Kevin Shea, Seth Sherman, Scott Uhlrich, Jennifer Hicks, Scott Delp, Peter Shull, and Akshay Chaudhari. Towards out-of-lab anterior cruciate ligament injury prevention and rehabilitation assessment: A review of portable sensing approaches, 10 2022. 3

[42] Shashank Tripathi, Lea Müller, Chun-Hao P. Huang, Taheri Omid, Michael J. Black, and Dimitrios Tzionas. 3D human pose estimation via intuitive physics. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4713–4725, 2023. 8

[43] Scott D. Uhlrich, Antoine Falisse, Łukasz Kidziński, Julie Muccini, Michael Ko, Akshay S. Chaudhari, Jennifer L. Hicks, and Scott L. Delp. Opencap: 3d human movement dynamics from smartphone videos. *bioRxiv*, 2022. 2, 3, 4, 5, 6, 7

[44] Gül Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *CVPR*, 2017. 3

[45] Timo von Marcard, Roberto Henschel, Michael Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *European Conference on Computer Vision (ECCV)*, sep 2018. 3, 6

[46] Jerry Wei, Da Huang, Yifeng Lu, Denny Zhou, and Quoc V. Le. Simple synthetic data reduces sycophancy in large language models, 2023. 3

[47] Andrew Weitz, Lina Colucci, Sidney Primas, and Brinnae Bent. Infiniteform: A synthetic, minimal bias dataset for fitness applications, 2021. 3, 6

[48] Ge Wu, Sorin Siegler, Paul Allard, Chris Kirtley, Alberto Leardini, Dieter Rosenbaum, Mike Whittle, Darryl D D'Lima, Luca Cristofolini, Hartmut Witte, Oskar Schmid, and Ian Stokes. Isb recommendation on definitions of joint coordinate system of various joints for the reporting of human joint motion—part i: ankle, hip, and spine. *Journal of Biomechanics*, 35(4):543–548, 2002. 4

[49] 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*, 2022. 6

[50] Feng Zhang, Xiatian Zhu, Hanbin Dai, Mao Ye, and Ce Zhu. Distribution-aware coordinate representation for human pose estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 6

[51] Ce Zheng, Wenhan Wu, Chen Chen, Taojiannan Yang, Sijie Zhu, Ju Shen, Nasser Kehtarnavaz, and Mubarak Shah. Deep learning-based human pose estimation: A survey, 2023. 2