

Biomechanical 3D Body: Self-Supervised Distillation of Biomechanical Pose from a 3D Body Foundation Model

R. James Cotton^{1,2}, J.D. Peiffer^{1,3}, Lucinda Williamson⁴, John Leske^{1,3},
and Georgios Pavlakos⁵

¹ Shirley Ryan AbilityLab, Chicago, IL, USA

² Department of Physical Medicine and Rehabilitation, Northwestern University,
Chicago, IL, USA

³ Department of Biomedical Engineering, Northwestern University, Evanston, IL,
USA

⁴ Department of Biomedical Engineering, University of Illinois Chicago, Chicago, IL,
USA

⁵ Department of Computer Science, University of Texas at Austin, Austin, TX, USA

Abstract. State-of-the-art monocular body recovery methods predict mesh vertices and angles on the corresponding kinematic tree, but their outputs lack biomechanically defined joint angles that downstream applications like clinical and biomechanical analyses require. We extend an existing foundation model, SAM-3D-Body, with an additional biomechanical prediction head that, from a single RGB image, regresses the joint angles and scales of a biomechanical model. Training this model presents a challenge, as there are limited datasets of paired images and biomechanical fits. To overcome this, we supervise biomechanical outputs with in-loop optimized targets from a Levenberg-Marquardt solver performing inverse kinematics fits against markers from the mesh predictions. This allows distilling the biomechanical head from the mesh head, even from unlabeled images. To make this work with GPU-optimized biomechanical models in MuJoCo, the entire model was implemented in JAX using Equinox. We trained this distilled output head on the publicly released SAM-3D-Body dataset. We then validated this model on biomechanical fits to two publicly available marker-based datasets, MoVi and BioCV, as well as movements from a clinical cohort captured with multiview markerless motion capture. The resulting model outperforms existing models for direct regression of biomechanics from images while only slightly underperforming the state-of-the-art monocular biomechanics method that performs more costly inference-time optimization of entire trajectories.

Keywords: biomechanics, pose estimation, foundation models, distillation, inverse kinematics, model-fitting in the loop, SPIN, MuJoCo

1 Introduction

Movement directly reflects neurological and musculoskeletal health, yet it is rarely measured in routine clinical care. Multiview markerless motion capture (MMMC) has matured to the point where, on well-observed movements, it agrees closely with marker-based optical systems for gait and upper-limb analysis [4, 11, 16, 34]. This progress rides on the rapid evolution of monocular human pose estimation, which now culminates in large foundation models that accurately estimate 2D and 3D keypoints [19, 20, 31]. These keypoint trajectories can then be used to recover biomechanically defined joint angles following the ISB standards for the lower and upper limb [37, 38], either by inverse kinematics (IK) on triangulated virtual keypoint trajectories [33] or by directly optimizing biomechanical fits to minimize reprojection error [4, 8].

In parallel, there has been substantial progress in estimating high-quality 3D mesh representations of humans from monocular images [15, 29, 40]. SAM-3D-Body is a particularly powerful recent state-of-the-art method that predicts the parameters of the Momentum Human Rig (MHR) [10], a model that decouples a posable skeleton from a shape-dependent surface. However, the kinematic trees of MHR and other widely used mesh models such as SMPL do not directly respect the anatomical joint definitions or rotational degrees of freedom required for clinical and biomechanical analyses [25]. This mismatch has motivated recent work on models such as SKEL, which incorporates skeletal bones into a surface mesh representation and begins to bridge computer-vision body models with biomechanical analysis [17, 18].

Several recent monocular systems have built on these advances to estimate biomechanical quantities from video. Human Skeleton and Mesh Recovery (HSMR) [39], for example, uses the SKEL model but addresses the lack of paired image/SKEL labels by converting SMPL pseudo-ground-truth from existing pose datasets into SKEL parameters. It then refines these labels during training with SKELify, a SMPLify-style iterative optimizer that fits SKEL parameters to 2D keypoints. HSMR is optimized for keypoint position accuracy, which is an important computer-vision metric but does not reliably predict joint-angle accuracy [35]. Recently, OpenCap released a monocular version of its software [13] that computes IK fits to SMPL estimates from WHAM [29] with additional physics-constrained refinement. Finally, our recent Portable Biomechanics Laboratory (PBL) [28], uses 2D and 3D keypoint locations from an existing foundation model [31] and optimizes a biomechanical fit to these data by minimizing reprojection error and a 3D keypoint loss while enforcing a single skeleton scaling across multiple trials. Although PBL reports state-of-the-art accuracy and was validated against both marker-based and MMMC in a large, clinically representative population, its optimization-based fitting process remains relatively slow.

This motivated us to investigate whether we could adapt a state-of-the-art vision model for estimating mesh representations from video, SAM-3D-Body, to output biomechanical estimates in a feed-forward manner while achieving accuracy comparable to optimization-based approaches. Two obstacles stand between a mesh foundation model and feed-forward biomechanics. First, there is no large

paired dataset of images annotated with biomechanical ground truth, so a regressor cannot simply be trained against joint-angle labels. We recently showed that accurate biomechanical estimates can be obtained from SAM-3D-Body MHR estimates using IK fits to individual MHR poses [6], similarly to SKELify. Second, popular biomechanical simulators such as OpenSim [7] and Nimblephysics/AddBiomechanics [36] are not built to integrate into model training. MuJoCo [32], through its JAX accelerated MJX backend, makes forward kinematics (FK) differentiable and highly performant, but using it would require linking the Torch-based SAM-3D-Body model to a JAX-based biomechanical model.

Our approach addresses both obstacles at once. We port SAM-3D-Body and MHR to JAX/Equinox [3, 21] so that the frozen foundation model and the differentiable MJX FK live in a single differentiable framework. Onto the *frozen* backbone and decoder we graft a small trainable biomechanical head that reads the existing decoder tokens and regresses the generalized coordinates (qpos), per-group segment scales, and anatomical marker offsets of a MuJoCo biomechanical model. To address the absence of paired labels, we draw on model-in-the-loop optimization (SPIN) [23], which during training optimizes a model’s outputs to better align with a more accurate teacher, mirroring the refinement HSMR obtains from 2D keypoints with SKELify. Concretely, a loss function quantifies MHR-to-biomechanics alignment by enforcing pose similarity between site-marker positions (computed via differentiable MuJoCo FK) and MHR-derived markers, requiring no ground-truth biomechanical labels. An inline Levenberg–Marquardt IK solver optimizes this objective in the training loop and provides pseudo-labels for qpos and scale. We show the resulting model, which we term Biomechanical 3D Body or B3D for short, is competitive with SOTA monocular biomechanical methods that utilize extensive optimization.

This paper makes three contributions:

1. **Dual-head foundation model.** A small trainable biomechanical head on a *frozen* SAM-3D-Body backbone adds clinically interpretable, MuJoCo-compatible biomechanical outputs with minimal parameter and compute cost.
2. **Label-free distillation.** The head is trained entirely from the frozen MHR head through an expectation-maximization (EM)-fitted marker mapping and differentiable MuJoCo FK, requiring no paired image–biomechanics dataset. While this is trained on a dataset with ground truth MHR pseudo-labels, these labels are not used in training, only the SAM-3D-Body predictions.
3. **Model-optimization-in-the-loop biomechanical training.** An inline Levenberg–Marquardt IK solver supplies cold-to-warm pseudo-labels and, by learning to seed it, the head drives the IK past what cold starts reach.

2 Related Work

Surface-mesh body recovery. Many monocular 3D body methods recover a surface mesh, with SMPL [25] parameterizing the body surface by pose and

shape. SMPLify [2] demonstrated fitting SMPL meshes to images through optimization against 2D keypoints constrained by a plausible human pose prior. Human mesh recovery (HMR) then showed a model could be trained to directly regress those parameters from an image [15]. SPIN [23] combined these methods to run an SMPLify-style optimizer *inside* the training loop to produce refined pseudo-labels that supervise the regressor, improving performance without full 3D ground truth. Recently, SAM-3D-Body [40] trained a state-of-the-art foundation model for mesh recovery on a massive labeled dataset using a DINOv3 backbone [30] and an MHR output head [10], where MHR inherits the skeleton/shape decoupling of ATLAS [27]. All these emit a deformable surface rather than the biomechanical joint angles and segment lengths this work targets.

Markerless biomechanical estimators. OSSO [18] learned from DXA scans to infer an internal skeleton from the body surface, and SKEL [17] then re-rigged the SMPL body with a biomechanically accurate skeleton — optimized inside registered SMPL meshes — yielding a skinned skeleton model whose joints follow biomechanical conventions. HSMR [39] leverages SKEL by first generating a pseudo-labeled training set with SKELify (a SMPLify-style optimizer that fits SKEL parameters to trustworthy 2D keypoints). Following the SPIN-style approach described above, it periodically refines these labels during training to learn a direct regressor from images to SKEL parameters, yielding both a full body mesh and biomechanically constrained joint angles. Our method adopts the same in-the-loop refinement, but supervises against a 3D biomechanical teacher through differentiable simulation rather than 2D keypoints. OpenCap-Monocular [13] first recovers a per-frame SMPL body mesh from video with WHAM [29], then extracts virtual markers from that mesh to scale and fit an OpenSim [7] musculoskeletal model to the recovered keypoints, which amounts to an IK step on the WHAM output. Similarly, BioPose [22] regresses anatomically constrained joint angles with a neural IK network over mesh-vertex virtual markers, but its code and weights were not released, so we do not compare against it. The Portable Biomechanics Laboratory (PBL) [28] shows state-of-the-art performance for recovering biomechanical kinematics from monocular video by leveraging a vision foundation model for 2D and 3D keypoints [31] and then optimizes kinematic trajectories end-to-end to minimize the 3D and reprojection errors; it is the strongest baseline we compare against here, and its trajectory optimizer also produces the reference fits for two of our evaluation cohorts. In further prior work we showed that accurate biomechanical joint angles, including individual fingers, can be recovered from MHR meshes by minimizing a pose-similarity distance with a Levenberg–Marquardt IK optimizer [6]; that solver is the one we run inside the training loop here.

3 Methods

Overview. We graft a small trainable *biomechanical head* onto a frozen SAM-3D-Body backbone [40] and train it to regress the joint angles and body scale parameters for a MuJoCo-MJX biomechanical model [32] by distilling from the

frozen MHR head [10]. To overcome the lack of paired image-to-biomechanics labels, the biomechanical head’s outputs are pushed through differentiable MuJoCo forward FK and matched to MHR-derived targets. A SPIN-style [23] optimization in the training loop is performed using a Levenberg–Marquardt [24, 26] IK solver to supply pseudo-labels. The entire model is implemented in JAX/Equinox [3, 21] so gradients flow from image features through to the biomechanical model running in MJX. Figure 1 summarizes the architecture.

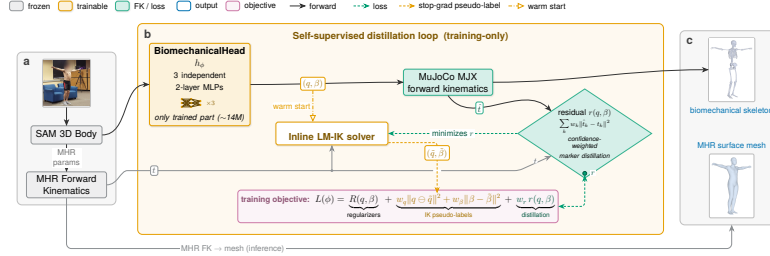


Fig. 1: Method overview. A person crop drives the frozen SAM-3D-Body backbone to a pose token, read by the frozen MHR head and trainable biomechanical head. MJX FK turns the predicted parameters into site positions, matched to MHR-derived targets via a confidence-weighted loss. An inline SPIN-style IK solver supplies warm-started pseudo-labels. At inference a single feed-forward pass emits both the MHR mesh and the biomechanical skeleton.

3.1 Biomechanical model

For the biomechanical model, we use a full-body MuJoCo humanoid with a flexible lumbar spine, ball-and-socket hips and shoulders, and fully articulated hands (99 generalized coordinates). It carries 102 anatomical marker sites (87 body, 10 fingertip, 5 face), rigged to the scaled model via per-marker offsets that capture individualized surface geometry. We disabled knee internal/external rotation, varus/valgus, and metatarsophalangeal flexion. We exclude finger pose from validation.

FK is implemented by MJX, a differentiable MuJoCo simulator running on the GPU. Let q be the generalized coordinates and, following the SMPL-style $\mathcal{M}(\theta, \beta)$ convention [25], $\beta \in \mathbb{R}^{14+102 \times 3}$ the model’s shape parameters — the 14 per-segment scale parameters concatenated with the 102 per-marker offsets. We write the model scaling together with its MJX FK as a single differentiable operator $\mathcal{B}$ that returns the world-frame joint centers J and the site-marker positions x :

$$(J, x) = \mathcal{B}(q, \beta). \quad (1)$$

Although we write $\mathcal{B}(q, \beta)$ by analogy with $\mathcal{M}(\theta, \beta)$, the parameters mean different things. In a SMPL model the pose θ is a set of per-joint rotations of a tree fit to registered scans and the shape β is a coefficient vector in a learned basis, so joint locations and segment lengths are implied by the surface. Here q is the generalized coordinates of a biomechanical model, and thus are inherently clinically defined and interpretable. Correspondingly β is anatomical rather than statistical: per-segment scale factors and per-marker offsets, so segment lengths are explicit and individually scaled.

3.2 MHR Model

MHR is a parametric body model [10] whose posable kinematic skeleton is decoupled from a shape-dependent surface, unlike SMPL. While the scale parameters are more directly interpretable than SMPL β parameters, the kinematic tree and joint angles are not biomechanically defined. Writing its pose and shape parameters as θ_M and β_M , MHR provides posed mesh vertices V , regressed keypoints K , and kinematic joint centers G :

$$(V, K, G) = \mathcal{M}(\theta_M, \beta_M). \quad (2)$$

3.3 MHR-to-biomechanical marker mapping

The MHR and MuJoCo biomechanical models do not share marker sites, joint definitions, or generalized coordinates, so we first construct a correspondence between their observable 3D locations. We use the BML-MoVi marker set as the target marker layout because it provides a dense set of 87 anatomical body markers that we have shown to support accurate joint-angle estimation compared with sparser marker sets [4, 5, 12, 28].

For each biomechanical marker, we assign a corresponding MHR output, similar to our recent work [6]. Some assignments are direct: for example, kinematic joint centers such as the hip or knee are closely matched to the corresponding MHR joint centers. Surface markers are less obvious, and for these markers we estimate correspondences between MoVi marker locations and SAM-3D-Body mesh vertices by comparing their image-plane projections across many video sequences, weighting candidate matches by reprojection agreement. This produces a fixed mapping in which each biomechanical marker corresponds to one MHR output: either a kinematic joint center, a mesh-regressed keypoint, or a mesh vertex.

We then refine the mapping with an offline EM procedure on the SAM-3D-Body dataset [40]. In the E-step, the current MHR targets are fit with IK over the biomechanical model $\mathcal{B}$, solving for joint angles q while sharing body scale and marker offsets across frames from the same subject. In the M-step, the pooled residuals are used to update the population-level marker template and a confidence weight w_k for each marker. Markers with small, consistent residuals are assigned higher confidence weights, whereas markers with noisier correspondences are down-weighted.

The final correspondence is represented by two gather operators that select matched points from each model in the same marker order. The biomechanical prediction, $\hat{t}(q, \beta) = \Pi_B(J, x)$, selects the corresponding marker locations from the biomechanical joint centers and site positions returned by Eq. 1. The MHR target, $t = \Pi_M(V, K, G)$, selects the matched vertices, keypoints, or joint centers from the MHR outputs in Eq. 2. The IK solver then minimizes the confidence-weighted distillation residual

$$r(q, \beta) = \sum_{k=1}^N w_k \|\hat{t}_k(q, \beta) - t_k\|^2, \quad (3)$$

where N is the number of matched markers and w_k are the marker-specific confidence weights learned by the EM procedure.

3.4 Dual-head grafting

The biomechanical head reads the frozen decoder’s 1024-d pose token concatenated with the frozen MHR head’s output (body pose and camera) projected to token space (2048-d input), through three parallel two-layer MLPs (hidden width 2048, ReLU), one per output: generalized coordinates (6D continuous rotations [41] for the root and ball joints, mapped to unit quaternions by Gram–Schmidt), per-segment body scales, and per-marker offsets. This adds ~ 13.6 M trainable parameters while keeping perception and the MHR head frozen (full dimensions in Appendix A.12).

3.5 Training objective and inline Levenberg–Marquardt IK

During training we obtain pseudo-labels from a Levenberg–Marquardt IK solve run every step, mirroring SPIN’s inner optimizer [23]. The Levenberg–Marquardt IK minimizes r Eq. 3 over (q, β) in staged fashion (root, then full pose, then scale and offsets), following our prior work [6]. It is seeded from the neutral pose and, as the head improves, warm-started from its biomechanical predictions; the solution $(\tilde{q}, \tilde{\beta})$ is stop-gradient before use as a target.

The biomechanical head is trained to minimize a single self-supervised objective built on the distillation residual r of Eq. 3. Let $(q, \beta) = h_\phi(z)$ be the head’s prediction from the pose token z , the training objective is

$$\mathcal{L}(\phi) = \underbrace{w_r r(q, \beta)}_{\text{distillation}} + \underbrace{w_q \|q \ominus \tilde{q}\|^2 + w_\beta \|\beta - \tilde{\beta}\|^2}_{\text{IK pseudo-labels}} + \underbrace{\mathcal{R}(q, \beta)}_{\text{regularizers}}. \quad (4)$$

The distillation term backpropagates r through the differentiable FK of $\mathcal{B}$ into the head, requiring no biomechanical ground truth; the pseudo-label terms pull the same prediction toward the inline-IK optimum. Here $\ominus$ is the Euclidean difference on translational/hinge coordinates and a quaternion distance on the orientation coordinates (root and ball joints). $\mathcal{R}(q, \beta)$ collects the anatomical regularizers (joint-limit and ball-cone penalties, scale and quaternion normalization priors; see Appendix A.8).

3.6 Datasets, augmentation, and metrics

We train on the public SAM-3D-Body dataset [1, 40] — **1,486,720 distinct images (3,722,776 person-crop training samples)** drawn from five of its constituent image sources — with mild augmentation (color and bounding-box jitter; no flips or rotations); see Appendix A.7 for the per-source breakdown. We evaluate on **entirely separate biomechanical datasets** to test generalization. First, BML-MoVi (85 participants) provides marker-based tracking with synchronized monocular videos; we represent the reference kinematic trajectory as a learned implicit function following [28] and score monocular inference against it. Second, because public datasets under-represent movement impairments — a key motivation here — we evaluate on synchronized MMMC data with monocular smartphone videos, 1693 trials from 109 participants across four clinical populations (controls, lower-limb prosthesis users, participants with neurological injury, and pediatric inpatients). Third, we add the BioCV dataset [9] — 308 trials from 14 participants performing walking, running, hopping, and both self-selected and maximal effort counter-movement jumps — a second marker-based reference, with monocular inference scored against a MuJoCo IK reconstruction (the optical-mocap gold standard). 308 trials remain after quality gating of the 362 recordings (static poses, non-converged reference fits, and un-trackable clips removed; details in Appendix A.10.1). The feed-forward regressors (ours, HSMR) run per-frame; the optimization methods (PBL, OpenCap) fit the full video. OpenCap has additional filtering based on keypoint detection thresholds that causes it to drop some trials entirely (see Appendix A.10).

For joint angles on each trial we measure the median joint-angle error (MJAE), the median over frames of the absolute signed error in degrees. We report the cohort statistic as the median of per-trial MJAE and with normalized inter-quartile range (nIQR). For joint centers we report the Procrustes-aligned mean per-joint position error (PA-MPJPE), the mean Euclidean residual after a per-frame similarity alignment (scale, rotation, translation). PA-MPJPE is computed over the 19 common lower- and upper-limb joint centers shared by all methods (torso, head, and the dense finger joints excluded; full list in Appendix A.10).

3.7 Inference and IK refinement

At inference our method is a stateless per-frame forward pass on the evaluation crops to (q, β) , with no temporal window (unlike PBL and OpenCap). Because distillation does not fully reach the warm-IK quality seen in training, we optionally apply the same warm Levenberg–Marquardt-**IK** refinement at inference and report results with and without it.

3.8 Comparisons

We compare against the Portable Biomechanics Laboratory (PBL) [28], HSMR [39], and OpenCap Monocular [13] — the existing monocular-video biomechanics methods with released code and weights. Each method and reference uses a

different biomechanical skeleton definition, which we reconcile using ISB conventions [37, 38]. A single fixed per-(method, joint) offset, pooled across the whole cohort, then removes the residual constant cross-skeleton convention difference without absorbing genuine per-subject pose error. Full per-joint correspondence detail is in Appendix A.13.4.

4 Results

We compare our monocular method against three baselines (HSMR, OpenCap-Monocular, and the optimization-based PBL) on three datasets. To render metrics equivalent across the slightly different skeleton formats and avoid penalizing the non-MuJoCo models for rigging differences, we report bias-corrected metrics alongside raw joint-angle metrics, subtracting a single per-joint median offset (pooled over all trials against the reference) from each method before scoring.

4.1 Qualitative comparison

Figure 2 shows qualitative results from each method on frames from the BML-MoVi and MMMC datasets. While all methods perform fairly well most of the time, our method shows slightly better alignment than HSMR in places (e.g., feet); OpenCap-Monocular provides no overlay code, so it cannot be shown. Kinematic waveforms are shown in Figure 3 on an example walking trial, with our method and PBL track the reference more tightly than HSMR and OpenCap.



Fig. 2: Qualitative overlays on BML-MoVi (top two rows) and MMMC clinical data (bottom two rows). OpenCap-Monocular does not provide overlay code.

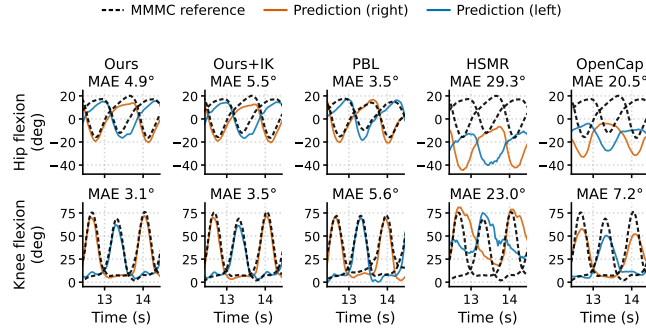


Fig. 3: Per-method gait waveforms on one clinical walking trial from a participant with no movement impairment. Top: hip flexion; bottom: knee flexion. Predictions are shown **raw (without bias correction)** so each method’s native joint-angle offset is visible; the per-panel MAE is therefore uncorrected and computed over this two-second window alone, and is not comparable to the cohort metrics in Table 2.

4.2 Accuracy against marker-based references (BML-MoVi and BioCV)

Both BML-MoVi and BioCV provide synchronized marker-based motion capture and monocular video, though each uses a different reference-fitting approach: BML-MoVi’s reference kinematics are reconstructed from the markers as a learned implicit function [28], while BioCV’s reference is a marker-based MuJoCo IK reconstruction. Monocular inference is scored against each. The clinical cohort (next section) is referenced against MMMC.

Our method tracks the dominant lower-limb angles tightly and recovers joint centers to within several centimetres on both references (Table 1). It is competitive with the optimization method PBL — which reports current SOTA performance for monocular biomechanics — and outperforms both HSMR and OpenCap-Monocular across all metrics. Figure 4a breaks the BML-MoVi results out by individual joint, where these trends hold. On BML-MoVi, enabling the optional inline IK at inference had minimal impact.

4.3 Accuracy on the clinical MMMC cohort

We evaluated on a MMMC reference spanning four clinical populations (controls, lower-limb prosthesis users, participants with neurological injury, and pediatric inpatients). Table 2 reports the same metrics as on the marker-based references and shows the same trend: our method is highly competitive against the other baselines except PBL, which outperforms it. Again, IK during inference had minimal impact. Figure 4 breaks down the average joint-angle error by cohort (including the two marker-based references for comparison), where our advantage over HSMR and OpenCap-Monocular becomes more pronounced in several groups. OpenCap produced no scored fit for 400 of 1693 inputs — its front-end

Table 1: Cross-method comparison against the two marker-based references (BML-MoVi optical mocap; BioCV marker-based MuJoCo IK). Rows are median (nIQR) bias-corrected joint-angle error (degrees) pooled per limb, and PA-MPJPE (mm) over the common joint centers. The joint-angle rows score every method at the same instants — those at which the subsampled feed-forward methods actually ran (Appendix A.10.2); OpenCap is the one column this leaves thin on BML-MoVi, since its reconstructions span only part of each trial and reach a median of 22 of the 149 instants. Best per row in **bold**; lower is better. † marks the optimization-with-kinematic-fitting methods (PBL, OpenCap), vs the feed-forward regressors; PBL’s optimization runs orders of magnitude slower (see Inference speed below). Per-joint detail in Appendix A.1.

Metric	Ours	Ours+IK	PBL†	HSMR	OpenCap†
<i>BML-MoVi (optical mocap, 85 subjects)</i>					
Lower-limb	3.25 (1.38)	3.15 (1.43)	2.73 (1.29)	4.76 (3.81)	3.84 (2.86)
Upper-limb	6.44 (5.57)	6.31 (5.52)	7.25 (6.41)	12.05 (9.54)	10.31 (5.97)
PA-MPJPE	37.1	36.0	33.9	49.8	47.2
<i>BioCV (marker-based MuJoCo IK, 14 subjects, 308 trials)</i>					
Lower-limb	4.04 (2.15)	3.96 (2.29)	4.10 (2.43)	6.11 (4.12)	5.39 (3.88)
Upper-limb	6.25 (6.38)	6.36 (6.05)	6.69 (6.91)	12.05 (12.57)	10.35 (9.98)
PA-MPJPE	62.2	62.7	52.5	67.4	64.9

input-quality gate rejects the posterior and self-occluded views common here (details in Appendix A.10) — excluding them from its analysis and possibly biasing its reported accuracy favorably.

Table 2: Cross-method comparison on the clinical gait cohort vs the multi-camera markerless reference, pooled across four clinical populations. Columns and metrics as in Table 1. Best per row in **bold**; lower is better. † marks the optimization-with-kinematic-fitting methods (PBL, OpenCap), vs the feed-forward regressors; PBL’s optimization runs orders of magnitude slower (see Inference speed below). Per-population detail in Figure 4b; per-joint detail in Appendix A.1.

Metric	Ours	Ours+IK	PBL†	HSMR	OpenCap†
Lower-limb	3.66 (1.61)	3.59 (1.66)	3.24 (1.96)	6.36 (4.35)	7.16 (5.77)
Upper-limb	9.93 (7.35)	9.97 (7.39)	7.43 (6.81)	13.68 (11.60)	18.88 (20.79)
PA-MPJPE	41.9	42.0	27.6	75.8	82.4
Rejected	0	0	0	0	400

4.4 Training dynamics and fine tuning the MHR head

Two features of training are visible in Figure 5. The inline IK marker-fit residual drops sharply at the cold-to-warm transition, when the solver switches from a neutral seed to a warm start from the head’s own prediction (panel a; schedule detail in Appendix A.8). As training proceeds the head’s own marker error falls toward this IK floor but does not reach it (panel b), which motivated the optional

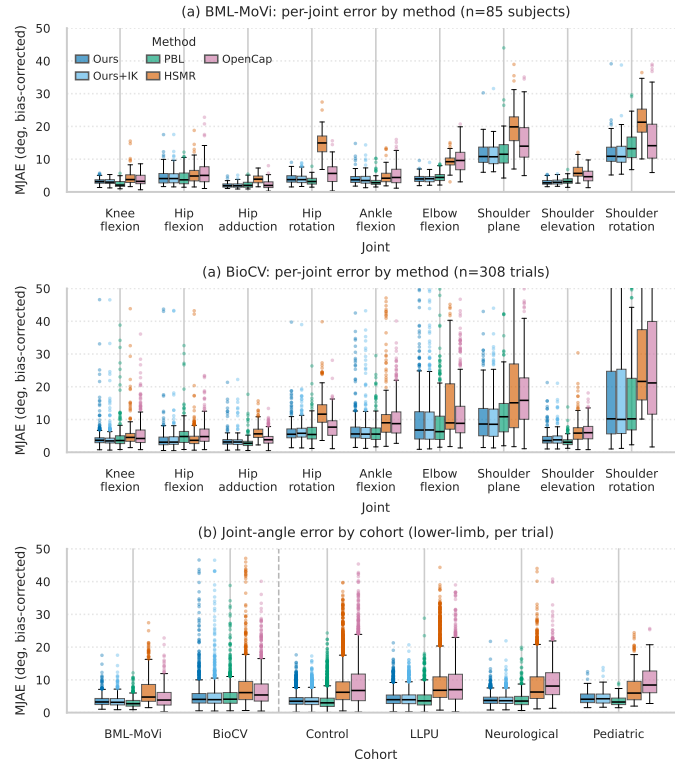


Fig. 4: Bias-corrected joint-angle error (MJAE, deg) across methods. **(a)** Per-subject error by joint on the external BML-MoVi cohort. **(b)** Per-trial lower-limb error by cohort — BML-MoVi (leftmost, left of the dashed guide), the four clinical gait populations, and the BioCV cohort — vs their respective references. Our method stays flat from MoVi through the impairment cohorts while HSMR inflates; PBL is tightest on MoVi and the clinical populations. OpenCap appears where its front-end produced a fit.

IK mode at inference. Separately, fine-tuning the MHR head improved these in-distribution metrics while the recovered mesh degraded on off-axis views, so we kept the MHR head frozen throughout (see Appendix A.4). The full model trained in 134 GPU-hours (5.6 days, 35 epochs) on a single GPU; disabling the per-step inline IK (the ablation below) cuts this to 43 GPU-hours at the same step budget, so the inner solve accounts for roughly two thirds of training cost.

4.5 Inference speed

Our method is feed-forward: a single pass through the frozen backbone and grafted head, with an optional lightweight inline Levenberg–Marquardt-IK refinement. Measured warm over the BML-MoVi and clinical videos, it runs at

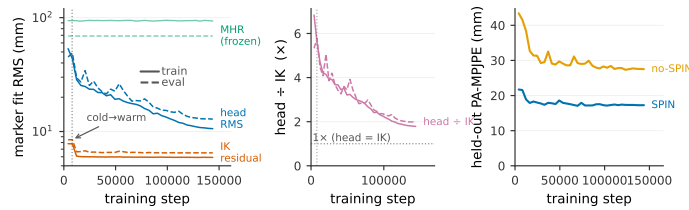


Fig. 5: Training dynamics (train solid, eval dashed) and the SPIN ablation. **(a)** The IK marker-fit residual drops at the cold-to-warm transition. **(b)** The head-to-IK error ratio falls toward $1\times$ as the feed-forward head approaches the in-the-loop optimizer. **(c)** Ablation: held-out biomechanical joint-center PA-MPJPE, full model (SPIN) vs the offset-regularized SPIN-disabled baseline.

roughly 15–20 frames/s in batched mode, dominated by the SAM-3D-Body backbone forward; the biomechanical head adds negligible cost. PBL, in contrast, is a whole-clip optimization and runs at roughly 0.03–0.14 frames/s, two to three orders of magnitude slower.

4.6 Ablation: model-optimization-in-the-loop (SPIN)

To isolate the contribution of the inline IK, we retrain the model with the SPIN pseudo-labels disabled ($w_q = w_\beta = 0$ and the inner solve skipped), so the head is trained on the distillation residual r Eq. 3 alone, under the same schedule and step budget. The pseudo-label helps regularize the per-marker offsets, so this baseline additionally regularizes them to prevent divergence. SPIN improves both the marker fit ($21.1 \rightarrow 12.9$ mm) and the biomechanical joint-center PA-MPJPE ($27.5 \rightarrow 17.3$ mm, Figure 5c).

A complementary ablation removes the biomechanical head rather than the pseudo-labels, running the same solver cold on the frozen MHR output at inference that is used during training. Median joint errors are similar, but the cold solve’s error tail is much heavier. Furthermore, the feed-forward head reaches this accuracy without the cost of performing inverse kinematics at inference — 41 ms/frame against 180 ms/frame for the cold solve, benchmarked at a full batch (Appendix A.5).

5 Discussion

We show that it is possible to graft a biomechanical head onto the SAM-3D-Body model and train a model to produce high quality biomechanical estimates from monocular video through self-distillation. We tested our new model with extensive validation on external, real-world datasets that provided either marker-based or MMC reference values. We found it outperforms an existing human skeleton regression model and is competitive with a state-of-the-art monocular biomechanics method that requires more expensive inference-time optimization.

Importantly for our application domain, we compared methods by measuring the joint angle errors as our primary metric in addition to the MPJPE more common in the human pose estimation literature. This revealed substantial limitations from HSMR, which tends to keep many joints overly flexed, and OpenCap. All methods showed worse accuracy in the arms than the legs, with the shoulder plane and rotation being worst estimated. Accuracy is not uniform across the upper limb, however: elbow flexion and shoulder elevation are recovered to within a few degrees, while the shoulder plane and axial rotation are several times worse. Whether that is sufficient depends on the clinical question — it may support gross upper-limb screening while falling short of what analyses of humeral plane or axial rotation require. Determining exactly why and when the shoulder and arm are less accurate is left to future work. Viewpoint is one candidate, since our earlier monocular work found frontal-plane angles were best recovered from frontal views and sagittal-plane angles from sagittal ones [28]; there, however, the effect on lower-limb angles was small, so viewpoint alone seems unlikely to account for errors of this size. We also tested generalization to clinical populations with movement impairments, often recorded in the frontal plane: several comparison methods showed substantially higher errors in these cases, while our method remained robust and competitive with the optimization-based PBL.

Central to our method is the in-loop Levenberg–Marquardt IK: this SPIN-style optimization [23] let us distill the biomechanical head using none of the SAM-3D-Body labels, suggesting it could refine the model on additional external video datasets where MHR regression remains robust, though this remains to be tested. An additional benefit of directly regressing the biomechanics, beyond faster inference, is that we saw fewer outliers than when cold-starting inverse kinematics on each frame, which avoids some of the difficulties of optimizing inverse kinematics for high-degree-of-freedom models.

The accuracy of our IK fits are bounded by the loss that aligns the biomechanical model to MHR Eq. 3, which we optimized via a large EM over MHR labels from the SAM-3D-Body dataset; it remains an opportunity for improvement, as some bones slightly extend beyond the MHR mesh. Enabling more degrees of freedom, refining the scaling, or adopting a more complete shoulder with scapulothoracic and glenohumeral motion [11, 14] could help, ideally guided by high-quality bone-movement annotations such as biplanar fluoroscopy.

We also compared across different biomechanical models and simulators — SKEL [17] in HSMR [39] (which has no biomechanical simulator) and OpenSim [7] in OpenCap [33] — and our model has more spine degrees of freedom than the reference model [28] used to fit MoVi and MMMC. We converted all models’ conventions to ISB standards [37, 38], but since our reference values are all MuJoCo fits, HSMR and OpenCap may be disadvantaged; the per-joint bias correction mitigates this, and the joint-angle errors track the PA-MPJPE errors, further supporting our conclusions.

While our experiment fine-tuning the MHR head alongside the biomechanical head did not show a benefit (Appendix A.4), optimizing this more carefully and

training more of the network is a potential future direction. Furthermore, our earlier IK optimization also enabled accurate monocular finger tracking against MMMC [6], so extending our head to the MHR hand regression is a natural next step.

While faster than our optimization-based PBL [28] the present method did not exceed its accuracy despite a SOTA vision backbone. PBL remains ahead on Procrustes-aligned joint-center error against all three references, and on the lower limb of both BML-MoVi and the clinical cohort; we are ahead on the upper limb of BML-MoVi and on both limbs of BioCV, which is enough to carry the BML-MoVi joint-angle average. Against the marker-based references the margins in either direction are small enough that the two are best read as comparable; PBL’s advantage is clearest on the clinical cohort, and there most of all in the upper limb.

This method also remains a single-image method, ignoring the inter-frame motion, texture, and appearance cues that can improve biomechanical estimates from video; point-tracking, dynamic scene understanding, and powerful video backbones are all promising directions. Because each frame is estimated independently, nothing in the model enforces temporal continuity. PBL gains accuracy in part through bundle adjustment over reprojection errors with the skeleton scale constrained across frames, whereas our method here is computed independently on each frame. However, our regressed (q, β) and the SAM-3D-Body mesh estimates could also seed a similar bundle-adjustment refinement.

6 Conclusion

Biomechanical 3D Body (B3D) demonstrates that self-distillation using model in the loop optimization can train a biomechanical output head on a SAM 3D Body backbone that provides performance competitive with state-of-the-art optimization-based methods while massively accelerating inference speed.

Acknowledgements

Research reported in this publication was supported by the Eunice Kennedy Shriver National Institute Of Child Health & Human Development of the National Institutes of Health under Award Number R01HD114776. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

- [1] AI at Meta: SAM 3D Body Dataset. HuggingFace Datasets (2025), <https://huggingface.co/datasets/facebook/sam-3d-body-dataset>, the public SAM-3D-Body training dataset, released as

- facebook/sam-3d-body-dataset. Annotations are built on eight public image sources (3DPW, AI Challenger, COCO, EgoExo4D, EgoHumans, Harmony4D, MPII, SA-1B). Accessed June 2026.
- [2] Bogo, F., Kanazawa, A., Lassner, C., Gehler, P., Romero, J., Black, M.J.: Keep It SMPL: Automatic Estimation of 3D Human Pose and Shape from a Single Image. In: ECCV 2016: Computer Vision – ECCV 2016, pp. 561–578. Springer, Cham (2016)
 - [3] Bradbury, J., Frostig, R., Hawkins, P., Johnson, M.J., Leary, C., Maclaurin, D., Necula, G., Paszke, A., VanderPlas, J., Wanderman-Milne, S., Zhang, Q.: JAX: Composable transformations of Python+NumPy programs (2018), <https://github.com/google/jax>, accessed 2026-08-15
 - [4] Cotton, R.J.: Differentiable Biomechanics Unlocks Opportunities for Markerless Motion Capture. In: 2025 International Conference On Rehabilitation Robotics (ICORR). pp. 44–51 (2025), <https://ieeexplore.ieee.org/document/11063174>
 - [5] Cotton, R.J., DeLillo, A., Cimorelli, A., Shah, K., Peiffer, J.D., Anarwala, S., Abdou, K., Karakostas, T.: Optimizing Trajectories and Inverse Kinematics for Biomechanical Analysis of Markerless Motion Capture Data. In: 2023 International Conference on Rehabilitation Robotics (ICORR). pp. 1–6. IEEE (2023)
 - [6] Cotton, R.J., Firouzabadi, P., Murray, W.M.: Monocular biomechanical tracking of fingers with inverse kinematics to foundation models (2026), <https://arxiv.org/abs/2605.09258>, arXiv:2605.09258
 - [7] Delp, S.L., Anderson, F.C., Arnold, A.S., Loan, P., Habib, A., John, C.T., Guendelman, E., Thelen, D.G.: OpenSim: Open-source software to create and analyze dynamic simulations of movement. IEEE Transactions on Biomedical Engineering 54(11), 1940–1950 (2007)
 - [8] Donahue, S., Djuraskovic, I., Shah, K., Sinz, F., Chafetz, R., Cotton, R.J.: Calibrated uncertainty for trustworthy clinical gait analysis using probabilistic multiview markerless motion capture. IEEE Transactions on Biomedical Engineering (2026), eMBS Special Issue; online ahead of print; PMID:42096387; arXiv:2601.22412
 - [9] Evans, M., Cazzola, D., et al.: A synchronised marker-based and markerless dataset of three-dimensional human movement. Scientific Data 11, 1300 (2024), pMC11604968; data at <https://researchdata.bath.ac.uk/1258>
 - [10] Ferguson, A., Osman, A.A.A., Bescos, B., Stoll, C., Twigg, C., Lassner, C., Otte, D., Vignola, E., Prada, F., Bogo, F., Santesteban, I., Romero, J., Zarate, J., Lee, J., Park, J., Yang, J., Doublestein, J., Venkateshan, K., Kitani, K., Kavan, L., Farra, M.D., Hu, M., Cioffi, M., Fabris, M., Ranieri, M., Modarres, M., Kadlec, P., Khiradkar, R., Abdrashitov, R., Prevost, R., Rajbhandari, R., Mallet, R., Pearsall, R., Kao, S., Kumar, S., Parrish, S., Yu, S.I., Saito, S., Shiratori, T., Wang, T.L., Tung, T., Xu, Y., Dong, Y., Chen, Y., Xu, Y., Ye, Y., Jiang, Z.: MHR: Momentum Human Rig (2025), <http://arxiv.org/abs/2511.15586>, arXiv:2511.15586
 - [11] Firouzabadi, P., Murray, W., Sobinov, A.R., Peiffer, J., Shah, K., Miller, L.E., Cotton, R.J.: Biomechanical arm and hand tracking with multiview

- markerless motion capture. IEEE International Conference on Biomedical Robotics and Biomechatronics (BioRob) (2024), reference for multiview validation methodology
- [12] Ghorbani, S., Mahdavian, K., Thaler, A., Kording, K., Cook, D.J., Blohm, G., Troje, N.F.: MoVi: A large multi-purpose human motion and video dataset. PLOS ONE 16(6), e0253157 (2021)
 - [13] Gilon, S., Miller, E.Y., Uhlich, S.D.: OpenCap Monocular: 3D human kinematics and musculoskeletal dynamics from a single smartphone video (2026), <https://arxiv.org/abs/2603.24733>, arXiv:2603.24733
 - [14] Holzbaaur, K.R.S., Murray, W.M., Delp, S.L.: A model of the upper extremity for simulating musculoskeletal surgery and analyzing neuromuscular control. Annals of Biomedical Engineering 33(6), 829–840 (2005)
 - [15] Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-End Recovery of Human Shape and Pose. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7122–7131 (2018)
 - [16] Kanko, R.M., Laende, E.K., Davis, E.M., Selbie, W.S., Deluzio, K.J.: Concurrent assessment of gait kinematics using marker-based and markerless motion capture. Journal of Biomechanics 127, 110665 (2021)
 - [17] Keller, M., Werling, K., Shin, S., Delp, S., Pujades, S., Liu, C.K., Black, M.J.: From skin to skeleton: Towards biomechanically accurate 3D digital humans. ACM Transactions on Graphics (SIGGRAPH Asia) 42(6) (2023)
 - [18] Keller, M., Zuffi, S., Black, M.J., Pujades, S.: OSSO: Obtaining skeletal shape from outside. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022), arXiv:2204.10129
 - [19] Khirodkar, R., Bagautdinov, T., Martinez, J., Zhaoen, S., James, A., Selednik, P., Anderson, S., Saito, S.: Sapiens: Foundation for human vision models. In: Computer Vision – ECCV 2024. Lecture Notes in Computer Science, vol. 15062. Springer (2024), arXiv:2408.12569
 - [20] Khirodkar, R., Wen, H., Martinez, J., Dong, Y., Zhaoen, S., Saito, S.: Sapiens2. In: International Conference on Learning Representations (ICLR) (2026), arXiv:2604.21681
 - [21] Kidger, P., Garcia, C.: Equinox: neural networks in JAX via callable PyTrees and filtered transformations. Differentiable Programming workshop at Neural Information Processing Systems 2021 (2021), <https://arxiv.org/abs/2111.00254>
 - [22] Koleini, F., Saleem, M.U., Wang, P., Xue, H., Helmy, A., Fenwick, A.: BioPose: Biomechanically-accurate 3D Pose Estimation from Monocular Videos. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 6330–6339 (2025), arXiv:2501.07800
 - [23] Kolotouros, N., Pavlakos, G., Black, M.J., Daniilidis, K.: Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2019), <http://arxiv.org/abs/1909.12828>, arXiv:1909.12828; code at <https://github.com/nkolot/SPIN>
 - [24] Levenberg, K.: A method for the solution of certain non-linear problems in least squares. Quarterly of Applied Mathematics 2(2), 164–168 (1944)

- [25] Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: A skinned multi-person linear model. In: *ACM Transactions on Graphics (Proc. SIGGRAPH Asia)* (2015)
- [26] Marquardt, D.W.: An algorithm for least-squares estimation of nonlinear parameters. *Journal of the Society for Industrial and Applied Mathematics* 11(2), 431–441 (1963)
- [27] Park, J., Romero, J., Saito, S., Prada, F., Shiratori, T., Xu, Y., Bogo, F., Yu, S.I., Kitani, K., Khirodkar, R.: ATLAS: Decoupling skeletal and shape parameters for expressive parametric human modeling. In: *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*. pp. 6508–6518 (2025), [arXiv:2508.15767](https://arxiv.org/abs/2508.15767)
- [28] Peiffer, J.D., Shah, K., Djuraskovic, I., Anarwala, S., Abdou, K., Patel, R., Jayabalan, P., Pennicooke, B., Cotton, R.J.: Portable biomechanics laboratory enables clinically accessible movement analysis from a handheld smartphone. *npj Digital Medicine* (2026)
- [29] Shin, S., Kim, J., Halilaj, E., Black, M.J.: WHAM: Reconstructing world-grounded humans with accurate 3d motion. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2024), [arXiv:2312.07531](https://arxiv.org/abs/2312.07531)
- [30] Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., Massa, F., Haziza, D., Wehrstedt, L., Wang, J., Darcet, T., Moutakanni, T., Sentana, L., Roberts, C., Vedaldi, A., Tolan, J., Brandt, J., Couprie, C., Mairal, J., Jégou, H., Labatut, P., Bojanowski, P.: DINOv3 (2025), [http://arxiv.org/abs/2508.10104](https://arxiv.org/abs/2508.10104), [arXiv:2508.10104](https://arxiv.org/abs/2508.10104)
- [31] Sárándi, I., Hermans, A., Leibe, B.: Learning 3d human pose estimation from dozens of datasets using a geometry-aware autoencoder to bridge between skeleton formats. In: *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. pp. 2956–2966 (2023), [http://arxiv.org/abs/2212.14474](https://arxiv.org/abs/2212.14474)
- [32] Todorov, E., Erez, T., Tassa, Y.: MuJoCo: A physics engine for model-based control. In: *IEEE/RSJ International Conference on Intelligent Robots and Systems* (2012)
- [33] Uhlich, S.D., Falisse, A., Kidziński, Ł., Muccini, J., Ko, M., Chaudhari, A.S., Hicks, J.L., Delp, S.L.: OpenCap: Human movement dynamics from smartphone videos. *PLOS Computational Biology* 19(10), e1011462 (2023)
- [34] Unger, T., Moslehian, A.S., Peiffer, J., Ullrich, J., Gassert, R., Lamercy, O., Cotton, R.J., Easthope, C.A.: Differentiable biomechanics for markerless motion capture in upper limb stroke rehabilitation: a comparison with optical motion capture. *IEEE Transactions on Medical Robotics and Bionics* (2025)
- [35] Wang, J., Kupperman, N., Baek, S.: Does lower MPJPE mean better biomechanics? evaluating joint angle fidelity of state-of-the-art 3D pose estimation models. In: *Computer Vision for Biomechanics Workshop (CVBW), IEEE/CVF Conference on Computer Vision and Pattern Recognition*

- (CVPR) (2026), <https://cvbw2026.github.io/assets/papers/12.pdf>, cVPR 2026 Workshop paper; not archived with a DOI. Accessed 2026-08-15
- [36] Werling, K., Bianco, N.A., Raitor, M., Stingel, J., Hicks, J.L., Collins, S.H., Delp, S.L., Liu, C.K.: AddBiomechanics: Automating model scaling, inverse kinematics, and inverse dynamics from human motion data through sequential optimization. *PLOS ONE* 18(11), e0295152 (2023)
 - [37] Wu, G., van der Helm, F.C., Veeger, H.D., Makhsous, M., Van Roy, P., Anglin, C., Nagels, J., Karduna, A.R., McQuade, K., Wang, X., Werner, F.W., Buchholz, B.: ISB recommendation on definitions of joint coordinate systems of various joints for the reporting of human joint motion—part II: shoulder, elbow, wrist and hand. *Journal of Biomechanics* 38(5), 981–992 (2005)
 - [38] Wu, G., Siegler, S., Allard, P., Kirtley, C., Leardini, A., Rosenbaum, D., Whittle, M., D’Lima, D.D., Cristofolini, L., Witte, H., Schmid, O., Stokes, I.: 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)
 - [39] Xia, Y., Zhou, X., Vouga, E., Huang, Q., Pavlakos, G.: Reconstructing Humans with a Biomechanically Accurate Skeleton. In: 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5355–5365 (2025), arXiv:2503.21751
 - [40] Yang, X., Kukreja, D., Pinkus, D., Sagar, A., Fan, T., Park, J., Cao, J., Liu, J., Ugrinovic, N., Feiszli, M., Malik, J., Dollar, P., Kitani, K.: SAM 3D Body: Robust Full-Body Human Mesh Recovery (2026), <http://arxiv.org/abs/2602.15989>, arXiv:2602.15989
 - [41] Zhou, Y., Barnes, C., Lu, J., Yang, J., Li, H.: On the continuity of rotation representations in neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019), arXiv:1812.07035

Supplementary Material — Biomechanical 3D Body: Self-Supervised Distillation of Biomechanical Pose from a 3D Body Foundation Model

R. James Cotton^{1,2}, J.D. Peiffer^{1,3}, Lucinda Williamson⁴, John Leske^{1,3},
and Georgios Pavlakos⁵

¹ Shirley Ryan AbilityLab, Chicago, IL, USA

² Department of Physical Medicine and Rehabilitation, Northwestern University,
Chicago, IL, USA

³ Department of Biomedical Engineering, Northwestern University, Evanston, IL,
USA

⁴ Department of Biomedical Engineering, University of Illinois Chicago, Chicago, IL,
USA

⁵ Department of Computer Science, University of Texas at Austin, Austin, TX, USA

1 Supplementary Materials

Detailed results and qualitative comparisons

This appendix presents the main-text comparison in fuller form. It gives the complete per-joint metric tables behind the summary numbers and additional qualitative overlays, then turns to three analyses the main text only summarises: the near-constant knee-flexion bias that characterises HSMR, the deceptively strong in-distribution metrics that led us to keep the MHR head frozen, and bias-corrected gait waveforms that separate tracking shape from constant convention offset.

A.1 Full per-joint metric tables

Per-joint breakdowns of the main-paper summary tables are given in Table [S1](#) (held-out BML-MoVi) and Table [S2](#) (clinical MMMC). Lower is better throughout. Joint-angle cells report MJAE in degrees, with nIQR in parentheses; both position columns, PA-MPJPE and Rigid MPJPE, are in mm.

The two position metrics measure different joint sets. PA-MPJPE is pooled over the 19 joint centers common to all methods (as in the main text), whereas Rigid MPJPE spans each method’s full exposed joint set. They are therefore not strictly comparable across methods, and for a method whose extra joints track well its PA-MPJPE can slightly exceed its Rigid MPJPE.

For joint angles, **Avg MJAE** is bias-corrected and pooled over the DOFs shared by all methods, which differs by dataset: on BML-MoVi every method reports the elbow, so it joins the hip, knee, ankle and shoulder DOFs, whereas

on the clinical cohort PBL exposes no elbow, so the pooled set there is the hip, knee, ankle and shoulder DOFs alone. The uncorrected row gives each method’s raw per-group mean. Two symbols flag special cases: † marks optimization-with-kinematic-fitting methods (PBL, OpenCap), and ‡ marks cells reported without bias correction (no per-frame series available), which should not be compared against bias-corrected values.

Cohort sizes are 85 subjects (BML-MoVi) and 1693 trials (clinical MMMC); OpenCap is scored only on the 1293 trials its WHAM front-end accepted (see `Inputs rejected`).

Table S1: Per-joint cross-method comparison on held-out BML-MoVi. Joint-angle cells are MJAE (deg) with nIQR in parentheses; PA-MPJPE / Rigid MPJPE in mm. Best per row in bold; lower is better.

Metric	Ours	Ours+IK	PBL†	HSMR	OpenCap†
<i>Lower-limb joints</i>					
Knee flexion	3.15 (0.77)	2.99 (0.87)	2.09 (1.10)	3.77 (1.73)	3.22 (1.90)
Hip flexion	4.08 (2.14)	4.08 (2.06)	3.69 (2.42)	4.86 (2.37)	5.05 (3.39)
Hip adduction	1.89 (0.73)	1.88 (0.75)	1.99 (1.12)	3.89 (1.53)	1.98 (1.30)
Hip rotation	3.75 (1.45)	3.75 (1.45)	3.18 (1.30)	14.95 (3.57)	5.64 (3.34)
Ankle flexion	3.70 (1.35)	3.48 (1.54)	2.82 (0.90)	4.18 (1.91)	4.41 (3.04)
<i>Upper-limb joints</i>					
Elbow flexion	3.97 (1.12)	4.03 (1.15)	4.42 (1.36)	9.23 (1.56)	9.60 (3.94)
Shoulder plane	10.80 (3.45)	10.73 (3.76)	11.50 (4.16)	19.86 (5.39)	13.97 (6.67)
Shoulder elevation	2.77 (0.84)	2.93 (0.88)	3.14 (1.01)	5.67 (1.99)	4.65 (2.17)
Shoulder rotation	10.88 (3.19)	10.79 (3.61)	13.22 (4.66)	21.29 (5.21)	14.12 (7.68)
<i>Aggregate (pooled angle / MPJPE)</i>					
Avg MJAE (°)	5.00	4.96	5.12	9.75	6.96
Avg MJAE, uncorrected (°)	7.65	7.58	7.68	14.21	8.29
PA-MPJPE	37.1	36.0	33.9	49.8	47.2
Rigid MPJPE	43.7	43.5	54.8	59.8	56.2
Instant	149	149	150	149	22
scored/trial (angles)					
Frames	9003	9003	18631	8987	1285
scored/trial (MPJPE)					

A.2 Qualitative skeleton overlays

Figure S1 and Figure S2 show the predicted biomechanical skeleton of each method on the clinical MMMC cohort and on held-out BML-MoVi respectively.

A.3 HSMR knee-flexion (crouch) bias

Across the clinical MMMC cohort, HSMR exhibits a large, near-constant knee-flexion bias: it predicts a knee that almost never extends, holding the limb in

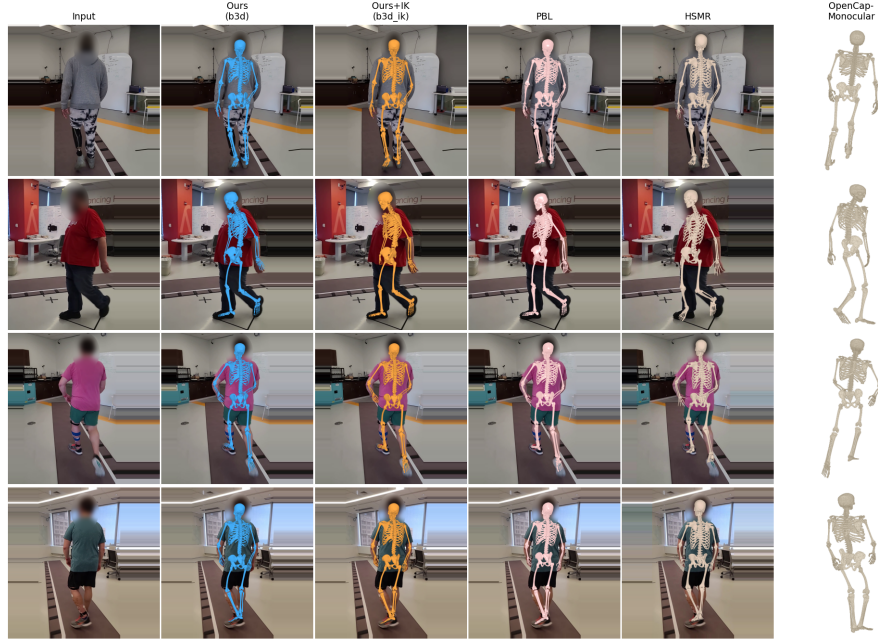


Fig. S1: Cross-method skeleton overlays on representative clinical-gait frames (MMMC cohort). Columns: Input · Ours · Ours+IK · PBL · HSMR · OpenCap. Ours and PBL register the lower limb tightly to the MMC reference, while HSMR's SKEL skeleton carries a visible knee/trunk offset.

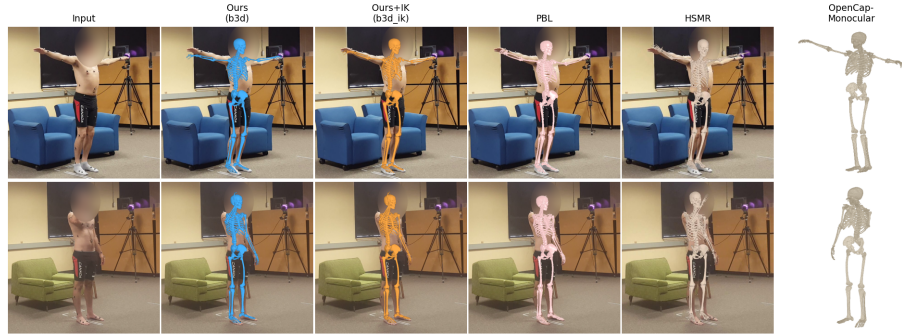


Fig. S2: Cross-method skeleton overlays on held-out BML-MoVi frames. Columns as above. On these well-observed lab frames Ours, Ours+IK, and PBL register the limb segments closely to the optical-marker fit; HSMR's residual knee/trunk offset persists even here, where the view is frontal and unoccluded.

Table S2: Per-joint cross-method comparison on the clinical gait cohort against the MMMC reference. Joint-angle cells are MJAE (deg) with nIQR in parentheses; PA-MPJPE / Rigid MPJPE in mm. Best per row in bold; lower is better.

Metric	Ours	Ours+IK	PBL†	HSMR	OpenCap†
<i>Lower-limb joints</i>					
Knee flexion	4.33 (1.72)	4.33 (1.73)	3.99 (1.60)	17.70 (5.12)	10.93 (6.89)
Hip flexion	3.65 (1.63)	3.60 (1.65)	3.50 (1.84)	5.94 (3.45)	9.58 (6.26)
Hip adduction	2.30 (0.80)	2.09 (0.80)	1.66 (0.70)	3.52 (1.34)	3.80 (2.11)
Hip rotation	3.58 (1.02)	3.53 (0.98)	2.59 (0.92)	6.93 (3.02)	5.90 (3.20)
Ankle flexion	4.59 (1.63)	4.58 (1.67)	5.13 (1.95)	6.32 (1.73)	9.02 (5.66)
<i>Upper-limb joints</i>					
Elbow flexion	3.87 (1.84)	3.81 (1.90)	—	6.03 (3.81)	6.22 (3.73)
Shoulder plane	13.21 (5.56)	13.29 (5.52)	10.08 (6.09)	19.36 (8.23)	27.48 (15.26)
Shoulder ele- vation	3.62 (1.92)	3.62 (1.87)	2.00 (1.10)	4.26 (2.45)	4.55 (2.58)
Shoulder rota- tion	12.60 (5.19)	12.61 (5.08)	9.65 (5.25)	17.63 (8.25)	29.31 (15.80)
<i>Aggregate (pooled angle / MPJPE)</i>					
Avg MJAE (°)	5.98	5.96	4.82	10.21	12.57
Avg MJAE, uncorrected (°)	10.58	10.56	4.98	18.66	16.75
PA-MPJPE	41.9	42.0	27.6	75.8	82.4
Rigid MPJPE	50.4	50.5	27.2	82.6	90.3
Instantants scored/trial (angles)	240	240	240	239	161
Frames scored/trial (MPJPE)	239	239	240	239	161
Inputs rejected	0	0	0	0	400

a permanently flexed, forward-leaning *crouch* even on able-bodied control gait where the reference knee reaches near-full extension at every stance phase (Figure S3).

Quantitatively, the error decomposes into two parts. The dominant part is a large constant offset: a single per-(method, joint) offset pooled across all trials is **+30.4° at the knee**, confined to the sagittal-flexion degrees of freedom (hip flexion +26.8°, but hip adduction +0.8° and ankle +5.1°). Subtracting this one number removes most of the error — cohort knee MJAE drops from 31.4° (raw) to 17.7° (bias-corrected), and hip flexion from 27.4° to 5.9°. What remains, the residual $\approx 18^\circ$, is a genuine frame-to-frame tracking error that no constant can fix: the per-frame correlation with the reference is poor (median Pearson $r \approx 0.32$, versus ≈ 0.94 for our method). This is why the constant-offset curve in Figure S3 aligns the *mean* of the HSMR waveform with the reference but not its *shape*.

This does not contradict HSMR’s reported results. HSMR’s knee benchmark is a joint-limit *violation frequency* (how often the predicted angle exceeds the SKEL skeleton’s anatomical bounds), not a knee-angle accuracy measure: a constant $\sim 30^\circ$ flexion offset sits well inside the knee’s $[0^\circ, 135^\circ]$ limit and therefore never registers as a violation. Its accuracy metrics are 3D joint/vertex positions evaluated in SMPL space — insensitive to a knee-convention rotation — and are

reported only on able-bodied laboratory data, with the model trained against fits to a pseudo-ground-truth body model rather than independent biomechanical motion capture. The bias is likewise invisible from the camera itself: from the camera view the mesh registration looks plausible, and only the off-axis (sagittal) renders make the mechanism legible — the flexed knees and forward trunk lean evident at every gait phase. A systematic knee-flexion-*angle* bias of this kind is thus structurally outside everything those benchmarks measure, which is precisely why an independent multi-camera biomechanical reference surfaces it.

A.4 Why the MHR head is kept frozen: deceptive in-distribution metrics

Figure S4 shows training dynamics from an experiment that fine-tuned the MHR head during distillation. Every in-distribution curve improves monotonically, yet off-axis viewpoints show the recovered mesh collapses. Figure S5 shows that the fine-tuned mesh looks plausible frontally, but rotated off-axis it folds and twists. The in-distribution metrics give no warning, so we keep the MHR head frozen.

A.5 What the biomechanical head adds over IK alone

The head is trained to reproduce what the inline solver computes, so a natural question is whether it earns its place: the same Levenberg–Marquardt IK could simply be run on the frozen MHR output at inference, with no learned head at all. We test this directly by scoring a third variant on BML-MoVi that runs the identical solver seeded from the neutral pose rather than from the head’s prediction, leaving the head bypassed entirely (Table S3).

On typical frames the two are close: over the full 85-subject cohort the cold-started solver lands within about a tenth of a degree of the head on the lower limb (3.60° against 3.70° , and 3.48° for the head with its warm refine) and within a millimetre on joint centers. The difference is in the tail. Cold-started IK reaches the 99th-percentile lower-limb error at 26° against the head’s 15° , and on the median subject produces gross frame failures (absolute error above 30°) on 0.8% of frames where the head produces none at all. This is the behaviour that motivates seeding the solve at all: a jointlimited human skeleton observed through a single marker set admits multiple configurations that explain the markers almost equally well, and from a neutral seed the solver settles into the wrong one often enough to matter. The head supplies a consistent, image-conditioned starting point that keeps the solve in the right basin — the same effect visible at training time, where the inline solver’s marker residual drops sharply once it switches from a neutral seed to a warm start from the head’s own prediction (main-paper training-dynamics figure).

The cost is the second half of the argument. Recovering pose from a neutral seed takes substantially more solver work than refining a warm start, and that cost is paid on every frame at inference. Benchmarked on a single GPU with the compile amortized and a full batch of 32 person-crops, the feed-forward pass costs 41 ms/frame while adding the cold solve takes end-to-end inference to 180

HSMR knee-flexion bias: a permanently flexed (crouched) knee on able-bodied control gait

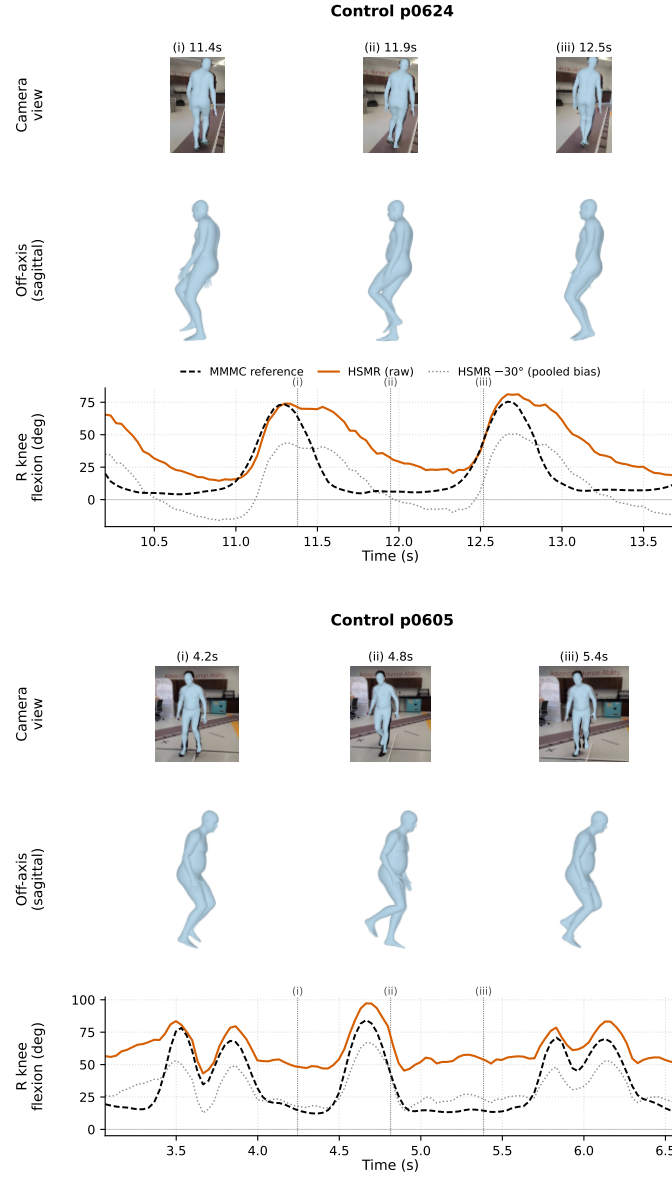


Fig. S3: HSMR knee-flexion bias on two able-bodied control gait trials (stacked). For each trial: the HSMR skeleton mesh shown image-registered on the camera frame (top row) and from an off-axis sagittal camera (middle row) at three sampled gait phases (i, ii, iii), over the right-knee-flexion waveform (bottom). Black dashed = the MMC reference; orange = raw HSMR; faint dotted = HSMR after subtracting the single pooled $+30.4^\circ$ knee bias. Vertical lines mark the time of each sampled mesh. The off-axis profiles reveal the permanently flexed (crouched) knee that produces the bias; the constant offset aligns the waveform mean but not its per-frame shape.

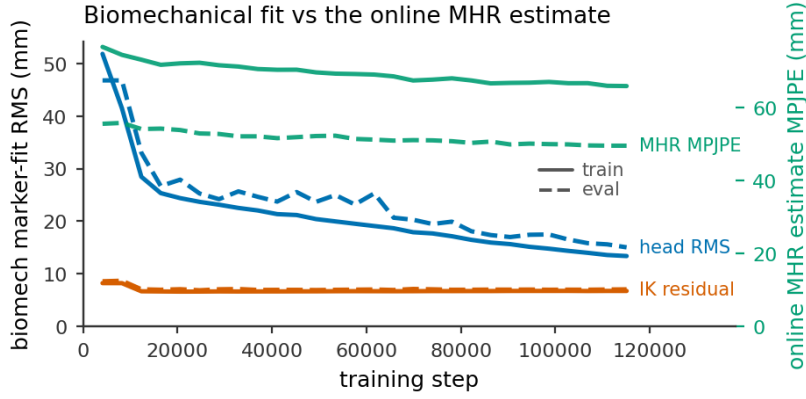


Fig. S4: Training dynamics of the discarded MHR-head fine-tuning run (train solid, eval dashed). Online MHR keypoint MPJPE (green), biomechanical marker RMS (blue), and inline-IK residual (orange) all improve in-distribution throughout training, while the mesh degrades catastrophically on off-axis views.

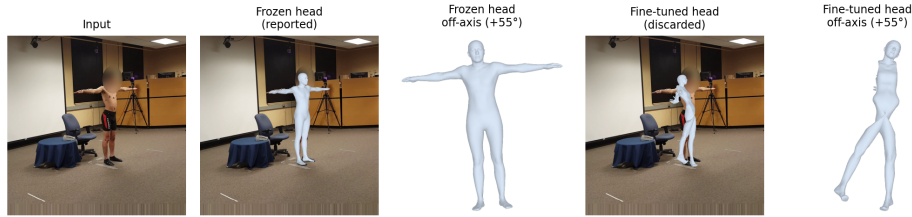


Fig. S5: Frozen vs fine-tuned MHR mesh on the worst-case (maximum frozen-vs-fine-tuned vertex-divergence) held-out BML-MoVi frame. Columns: input · frozen (frontal) · frozen (+55°) · fine-tuned (frontal) · fine-tuned (+55°). Both look plausible frontally; off-axis the fine-tuned mesh collapses while the frozen mesh stays anatomical. This frame illustrates the failure mode at its most severe; the typical off-axis drift of the fine-tuned head is smaller but still substantial.

ms/frame, a factor of 4.4. That factor is batch-dependent and we quote the conservative end of it: the Levenberg–Marquardt solve is dominated by a fixed per-solve cost rather than by its arithmetic, so a solve holding one frame costs almost as much as one holding 32, and the same comparison at batch 1 inflates to roughly $21\times$ — a statement about batching rather than about the head. The head reaches comparable accuracy at the lower cost either way, so distilling the solver into it buys robustness and speed together rather than trading one for the other.

Table S3: Feed-forward head vs cold-started IK on the frozen MHR output, BML-MoVi. All three variants are scored on the same subjects, at the same instants, and with the same pooled per-(method, joint) bias correction as the main tables. The joint-angle rows are computed at the frames where the network actually ran; the PA-MPJPE rows are computed over each variant’s full resampled series, identically for all three. Lower is better throughout.

Metric n=85)	(BML-MoVi, Ours	Ours+IK	MHR + cold IK
Lower-limb (deg)	MJAE 3.70 (1.35)	3.48 (1.45)	3.60 (1.41)
Upper-limb (deg)	MJAE 7.39 (2.16)	7.38 (2.38)	7.90 (2.46)
PA-MPJPE, (mm)	median 37.2	36.4	37.9
PA-MPJPE, (mm)	mean 38.1	37.5	39.2
PA-MPJPE, p90 (mm)	45.1	44.9	47.5
Lower-limb error p95 (deg)	9.25	9.13	11.78
Lower-limb error p99 (deg)	15.11	15.18	26.09
Frames >15 deg (%)	1.07	1.08	3.04
Frames >30 deg (%)	0.00	0.00	0.80
Inference (s/frame)	2.73	2.73	5.06
Relative cost, in harness	1.0x	1.0x	1.9x

The per-frame times are wall-clock inside the evaluation harness, which decodes video and pads every chunk to a fixed batch, so they are not a clean measure of the solver alone: at this dataset’s frame sampling each padded chunk carries about one scored frame, so the cold solve runs essentially unbatched. They bound the cost as evaluated here (2.3 s/frame added over the head) rather than the cost a batched deployment would pay. See the standalone benchmark for a component-level split of forward pass versus IK.

A.6 Per-method gait waveforms across trials

Figure S6 shows hip and knee kinematic waveforms for four clinical walking trials. Unlike the raw main-paper overlay, these supplementary waveforms are bias-corrected (each method’s single pooled per-joint offset removed), so they isolate waveform *shape* rather than constant convention offset. All five methods track in phase across cohorts; the prosthetic trial shows HSMR’s characteristic breakdown on prosthetic gait — its knee waveform decorrelates from the reference (across all prosthetic walking trials HSMR’s per-frame knee correlation is median $r \approx 0.32$ vs ≈ 0.94 for our method on the same reference), so a generic SMPL/SKEL body prior does not capture prosthetic-limb knee kinematics — while our method and the optimization-based PBL prior stay close to the reference.

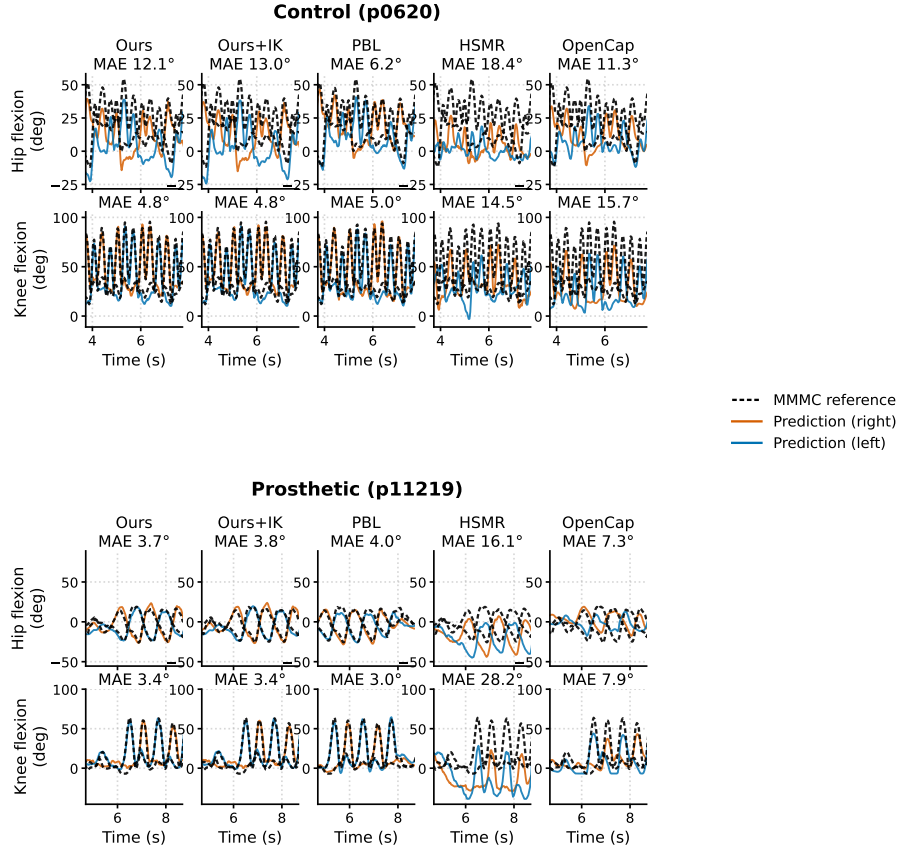


Fig. S6: Per-method gait waveforms on two clinical walking trials (one control and one lower-limb prosthesis user; hip above knee per trial; columns Ours / Ours+IK / PBL / HSMR / OpenCap), 4 s windows. Black dashed: sign-corrected reference; red and blue: predicted right and left; predictions bias-corrected as in the result tables. Panel titles give the per-leg-averaged error. HSMR’s knee decorrelates on the prosthetic trial while Ours and PBL stay close to the reference.

Methods, metrics, and training detail

This appendix records how the model was trained and evaluated. It describes how the distillation dataset was assembled, the training schedule and loss weights, the metric conventions applied throughout, and a check confirming that the JAX/Equinox port stays faithful to the reference implementation.

A.7 Training dataset detail

We distill from the public SAM-3D-Body dataset [11], released on HuggingFace as `facebook/sam-3d-body-dataset` [1]. The dataset does not provide its own

images; it supplies SAM-3D-Body / MHR annotations layered on top of eight publicly available image sources. Our training index uses **five** of these — SA-1B, Harmony4D, COCO, 3DPW, and MPII; the remaining three (AI Challenger, EgoExo4D, EgoHumans) are not included (Table S4 gives the per-source image and person-crop counts). We use it purely as a source of paired images and MHR outputs for distillation, applying only mild colour and bounding-box jitter (no flips or rotations) so the geometric correspondence between image and target is preserved. No separate dataset DOI is published; we cite the HuggingFace release. We deliberately do not treat the SAM-3D-Body validation losses as our headline metric, because they conflate biomechanical accuracy with the quality of the SAM-3D-Body / MHR outputs and the IK approximation; all reported accuracy is measured on the independent biomechanical datasets described in the main text.

Table S4: Per-source composition of the SAM-3D-Body training data. The dataset layers MHR annotations on eight public image sources; our training index draws from five. Counts are from the training index: “person-crops” are individual bounding-box training samples, “images” are distinct source images. The held-out evaluation split (195,696 crops / 78,249 images) is excluded.

Image source	Distinct images	Person-crops	Used
SA-1B	1,206,635	3,185,516	yes
Harmony4D	235,831	471,662	yes
COCO	22,589	37,632	yes
3DPW	16,508	21,232	yes
MPII	5,157	6,734	yes
AI Challenger	—	—	no
EgoExo4D	—	—	no
EgoHumans	—	—	no
Total (used)	1,486,720	3,722,776	

A.8 Training schedule, loss weights, and $\mathcal{R}(q, \beta)$

The reported model was trained for 35 epochs; the values below are its exact configured hyperparameters.

Optimization. 35 epochs $\times$ 4,111 steps/epoch = 143,885 total optimizer steps; batch size 16; learning rate 3×10^{-4} with linear warmup over 50 steps and cosine decay to a floor ratio of 10^{-3} ; seed 42. The MHR head is frozen throughout. Training ran for ≈ 134 hours (≈ 5.6 days) of wall-clock time on a single NVIDIA L40S GPU (the inline Levenberg–Marquardt–IK inner solve, run every step, dominates the per-step cost).

Loss-term weights. The training objective (the loss defined in the main text) decomposes into the distillation residual r , the inline-IK pseudo-label terms, and the regularizer $\mathcal{R}(q, \beta)$:

Mapping these to the main-text objective (Eq. 4 of the main paper): the distillation weight w_r is **w_marker** = 1.0; the coordinate weight w_q is **w_joint**

Table S5: Loss-term weights for the reported run, grouped by their role in the main-text training objective. Config names are the training flags.

Group	Term	Flag	Weight
Distillation r	marker residual	<code>w_marker</code>	1.0
IK pseudo-labels	joint angles q	<code>w_joint</code>	2.5
	root pose	<code>w_root</code>	2.5
	segment scale β	<code>w_scale</code>	1.0
	marker offsets	<code>w_offset_sup</code>	10.0
Regularizer $\mathcal{R}$	joint-limit + ball-cone	<code>w_joint_limits</code>	100.0
	6D column-norm prior	<code>w_quat_reg</code>	0.1
	scale prior	<code>w_scale_reg</code>	0.05
	offset prior (inactive)	<code>w_offset_reg</code>	0.0
	FK scale-clip range	<code>fk_scale_clip</code>	[0.05, 5.0]

= 2.5 (with a separate root sub-term `w_root` = 2.5); and the shape weight w_β decomposes into `w_scale` = 1.0 for the per-segment scales and `w_offset_sup` = 10.0 for the marker offsets (recall β bundles both).

Terms of $\mathcal{R}(q, \beta)$. The regularizer collects: (1) a **joint-limit penalty** on the hinge/slide DOFs (mean squared violation of the anatomical lower/upper bounds); (2) a **ball-cone penalty** on the ball joints (hip, shoulder) penalizing rotation beyond the allowed cone half-angle; (3) a **6D column-norm prior** pulling the two predicted 6D-rotation columns of each joint toward unit norm (the rotation representation is 6D, `use_6d`); (4) a **scale prior** pulling the per-segment scales toward 1; and (5) **FK scale clipping**, a hard clamp of every per-body scale multiplier inside FK (range in Table S5; gradient zeroed on saturation), which suppresses a self-distillation scale runaway. An **offset-magnitude prior** (`w_offset_reg`) is available but disabled (weight 0) in the reported run; offset shrinkage is instead supplied through the IK pseudo-label path (`w_offset_sup`).

Inline Levenberg–Marquardt–IK staging and the cold-to-warm transition. The inline solver runs three Levenberg–Marquardt stages each step — (i) root, (ii) pose + scale, (iii) full pose + marker offsets — with damping 2.0 (a fourth hand-refinement stage is disabled, as the hand pose is not refined or evaluated here). For the first `cold_warm_step` = 8,000 optimizer steps the solver is **cold-started** from the neutral pose with $[133, 133, 134] = 400$ Levenberg–Marquardt iterations across the three stages; once training passes 8,000 steps — at the epoch 2 $\rightarrow$ 3 boundary, given 4,111 steps/epoch — it switches to a **warm start** seeded from the head’s own biomechanical prediction with $[100, 100, 100] = 300$ iterations. The warm start escapes the poor local minima a neutral seed settles into on the highly coupled kinematic tree (training-dynamics figure, main text).

A.9 Scale groups and the scaling matrix

Rather than scaling each segment independently, we group anatomically related segments and predict one scalar per group. The **scaling matrix** maps these

group scalars to the per-body scale vector used by FK, enforcing left/right symmetry and keeping the scale parameterisation low-dimensional.

There are 14 groups: trunk (pelvis + head), thigh/shin/foot per leg, upper-arm/forearm/hand per arm, and one overall body scale. The hands have their own dedicated groups so hand-marker scale mismatches are absorbed independently from the arm. Group scalars are regularized toward 1 and hard-clamped inside FK (see Table S5).

A.10 Metric conventions: bias correction and the 19-joint set

Bias-corrected angle metrics subtract one fixed per-(method, joint-group) offset before differencing — a single value per method, pooled over the whole cohort rather than fit per participant, so it removes only each skeleton’s constant convention or frame offset and cannot absorb genuine per-subject pose error. The same correction is applied identically to every method, so no method is advantaged; uncorrected rows report each method’s raw per-group mean for reference.

Each joint center used in PA-MPJPE is the world-frame position of a named body’s local reference frame origin after FK. For example, the femur body’s frame origin serves as the hip joint center, the tibia body’s as the knee joint center, and so on, following the anatomical placement built into each model’s kinematic tree during its original construction. PA-MPJPE pools the 19 such centers common to every method: the pelvis, and bilaterally the femur, tibia, talus, calcaneus and toes in the lower limb and the humerus, ulna, radius and hand in the upper limb. The torso and head are excluded because not all baselines expose them, and the fingers are excluded because finger pose is not fitted here (the inline solver’s hand-refinement stage is disabled, so the finger coordinates are observed only by the ten fingertip markers and their residuals are dominated by the body markers), so a finger joint-center error would not measure anything the method controls. The per-frame similarity transform (scale, rotation, translation) is fit using each method’s full exposed joint set, not just the 19 joint centers, and the resulting residual is then averaged over the 19 common centers. Restricting the fit itself to the 19 centers changes the score by less than 0.2 mm, so this is a reproducibility choice rather than one that affects the reported numbers (the baselines do not cache per-frame joint positions beyond their native output, so fitting on the full set is what can actually be reproduced for every method).

The feed-forward methods (ours, HSMR) run per-frame and are scored directly on the selected reference frames. OpenCap and PBL are not per-frame methods. OpenCap runs WHAM (a temporal 3D-pose model) followed by OpenSim scaling and IK over a contiguous person-cropped clip spanning the subject’s full in-view presence (clamped to 600 source frames by GPU memory), producing a single full-video trajectory that is then linearly resampled onto the reference frame grid for scoring; reference frames outside that trajectory’s support are excluded with no extrapolation. WHAM’s native keypoint gate (≥ 21 of 25 OpenPose keypoints at confidence ≥ 0.8) is the final arbiter of which frames survive, so OpenCap’s actually-compared frame set is a subset of the shared grid, and a fraction of trials are dropped entirely where it does not produce a usable

trajectory. On the clinical cohort this gate rejected 400 of 1693 inputs — the posterior and self-occluded views common in our data often fail the keypoint-visibility check — consistent with OpenCap-Monocular being validated for a 45° anterolateral view [3]; excluding these may bias its reported accuracy favorably. PBL is a precomputed whole-clip monocular optimization fit, re-scored here by sampling its stored trajectory at the reference frames.

A.10.1 BioCV evaluation cohort: synchronisation and quality gating

BioCV [2] is a University of Bath dataset in which a single acquisition trigger hardware-synchronises nine 200 Hz cameras with a Qualisys optical marker system, so video frame n and marker frame n are the same instant with no software time-alignment — the strongest synchronisation guarantee of our three references. We run monocular inference from a single anterolateral camera (cam01) with a tight per-frame person-detection crop and score against a marker-based MuJoCo implicit kinematic reconstruction across walking (129), running (81), self-selected and maximal effort counter-movement jumps (45 + 44), and hopping (9).

From the 362 recorded trials we exclude static poses (not a movement task), 43 trials whose reference kinematic fits failed to converge — identified by the reference fit’s own median marker residual exceeding 50 mm, which cleanly separates the converged references (< 20 mm) from the diverged ones (> 250 mm) across an empty gap — and 3 trials for which the subject was never tracked within the analysis window (in-window bounding-box coverage below 25%), leaving the monocular front-end with no valid frames. Every exclusion is applied identically to all methods so it cannot bias the comparison, leaving 308 scored trials. The reference-health gate thresholds a median, so a trial whose reference degrades only briefly survives it: four of the retained trials carry a 90th-percentile marker residual between 50 and 66 mm against a typical retained trial’s 13 mm. Excluding all four shifts every method’s joint-angle rows by at most 0.09° and its PA-MPJPE by 0.2–0.7 mm, in the same direction for every column, so we report the full 308.

The PBL column on BioCV is the lab’s single-camera whole-clip optimisation fit of the same cam01 video (2D keypoints plus a monocular 3D lift; no marker input), so it is a monocular baseline on the same footing as the other columns. It covers 267 of the 308 trials: two participants were never fit at cam01, and their trials therefore carry no PBL value. We deliberately do not report the dataset’s marker-based optimisation fit in this column — it is fit to the very Qualisys markers the reference is built from, so scoring it against that reference measures agreement between two fits of the same input rather than monocular accuracy.

A.10.2 Sampling density, per-trial caps, and the matched scoring in-stants The methods do not observe a trial equally densely, and the joint-angle metrics are computed so that this does not become an accuracy difference.

The feed-forward methods (ours, HSMR) pass every scored frame through a network, so for inference compute we subsample the high-rate datasets rather

than running every frame. On BML-MoVi we take every 15th video frame — 2 Hz against a 120 Hz marker reference — and cap each trial at 150 such frames, which on these 75–85 s trials is the binding limit; on BioCV, every 20th frame of the 200 Hz capture, i.e. 10 Hz, a median of 35 frames per trial with the cap never reached. On the clinical cohort, whose videos are 30 Hz, no subsampling is applied and the median trial contributes 240 frames. Both feed-forward methods use the same sampling on the same trials.

The two optimization methods run at the full frame rate by design and are not subsampled: PBL fits an entire clip in one optimization, and OpenCap’s WHAM front-end runs per video frame. Their per-trial limits are therefore set by different constraints than ours — OpenCap processes a contiguous in-view window clamped to 600 source frames by GPU memory — so the caps themselves are *not* matched across methods, and we do not claim they are.

What is matched is where the comparison is evaluated. Every method is scored at the same instants: those at which the subsampled feed-forward methods actually ran, recovered from their stored per-frame outputs. Each method contributes its own value at those instants, and a dense method loses nothing by it, since it has a real estimate there too. Without this step roughly 95% of the frames the feed-forward error was computed over would be linear interpolations between predictions up to half a second apart, while the optimization methods were scored almost entirely on real estimates — so the columns would answer two different questions, one about the reconstructed trajectory and one about the estimate at an instant.

This equalizes sampling density, not temporal coverage. A method is matched only where its own reconstruction exists, so one that covers part of a trial is scored on that part: OpenCap’s BML-MoVi trajectories span about 10.7 s of a 75 s trial and reach a median of 22 of the 149 instants. The realized count per method is reported in the per-joint tables of Appendix A.1.

The joint-centre metrics (PA-MPJPE, rigid MPJPE) are the exception and keep the older basis, scored over each method’s full resampled series. Not every baseline retains per-frame 3D joint positions locally, so those rows cannot all be re-cut to the matched instants; they are computed identically for every method, and the tables report their frame count on a separate row so the two bases are not confused.

A.11 Port faithfulness — JAX/Equinox vs the PyTorch teacher

Because our biomechanical head distills from a frozen SAM-3D-Body trunk that we re-implemented in JAX/Equinox (for a single MJX-differentiable graph), we first establish that the port itself is not a source of error. We run the Meta **PyTorch SAM-3D-Body** model and our **JAX/Equinox port** on the SAME samples drawn from the SAM-3D-Body *dataset* (a seeded permutation over the full eval index, spanning SA-1B / Harmony4D / COCO / 3DPW; n=250), and score BOTH with the identical training metric (MPJPE / PA-MPJPE / PCK over the 70 MHR joints, pelvis-rooted, full-skeleton Procrustes for PA). The ground truth is the dataset pseudo-GT (the Meta model’s stored fits).

The raw torch-converted trunk — before any distillation — is scored against the PyTorch teacher on the same 250 crops (Table S6). The port costs only **+0.08 mm MPJPE** / **+0.13 mm PA-MPJPE** (direct PT↔JAX disagreement 3.42 mm over all 70 joints, ≈ 2.1 mm after Procrustes) — two orders of magnitude below the biomechanical errors in the main-text comparison. The reimplementation is faithful and contributes negligibly to reconstruction error.

Table S6: PyTorch vs JAX/Equinox port parity for the SAM-3D-Body trunk (n=250 crops, pre-distillation). MPJPE/PA-MPJPE are pelvis-rooted over the 70 MHR joints; PCK is 2D normalised by bounding-box diagonal. Δ = JAX - PyTorch; lower magnitude is better.

Metric	PyTorch	JAX/Equinox port	Δ (port cost)
MPJPE (mm)	64.32	64.40	+0.08
PA-MPJPE (mm)	43.55	43.68	+0.13
PCK @ 0.01	0.3645	0.3625	-0.0020
PCK @ 0.025	0.6982	0.6971	-0.0011
PCK @ 0.05	0.8201	0.8199	-0.0002
PCK @ 0.075	0.8574	0.8578	+0.0004
PCK @ 0.1	0.8799	0.8804	+0.0005
PCK (avg)	0.7240	0.7235	-0.0005

A.12 Biomechanical head architecture

The trainable biomechanical head is a lightweight (~ 13.6 M-parameter) module that reads a single 1024-d pose token from the frozen SAM-3D-Body decoder. With MHR conditioning enabled, the frozen MHR head’s output (body pose and camera, 136-d) is mapped to token space by a learned linear layer and concatenated with the pose token (2048-d input), so the head’s task reduces to remapping the MHR prediction into MuJoCo coordinates. Three parallel two-layer MLPs (hidden width 2048, ReLU, no residual or normalization) produce the three outputs. The generalized coordinates are emitted in a 109-d network space that uses 6D continuous rotations [12] for the root free joint and the four hip/shoulder ball joints; a Gram–Schmidt step maps each 6D block to a unit quaternion, giving the 99-d physical MuJoCo `qpos`, while hinge joints pass through as scalar angles. Scale clipping and the offset prior act downstream in FK (§A.8), not inside the head.

A.13 Kinematic model definitions and joint-angle comparability

To establish a valid comparison between each biomechanical framework, the kinematic trees and joint coordinate systems of each benchmark method are evaluated using the original model source files (OpenSim .osim, SKEL/BSM, MuJoCo MJCF) alongside ISB standards designed to ensure joint motions are clinically interpretable. Each method utilizes different parameterizations due to

Table S7: Biomechanical-head outputs. The three regressors are independent two-layer MLPs off the shared 2048-d input; the only extra parameters are the MHR-conditioning projection.

Regressor	Output	Dim
Generalized coordinates	6D rotations (root + 4 ball joints) + hinge scalars	109 (99 physical)
Per-segment scale	zero-centered scale deltas	14
Marker offsets	per-marker XYZ, metres	306 (102×3)

their distinct kinematic trees, rotation axes, Euler sequences, sign conventions, and default neutral poses. Consequently, two methods can agree on the underlying physical motion yet report numerically different angles. This appendix translates the underlying mathematical formulations of each framework to provide per-model and per-joint standardization, ensuring that main-text comparisons are defensible and that residual discrepancies are correctly attributed to structural convention rather than genuine tracking error.

A.13.1 Conventions and notation

- **ISB joint coordinate systems (JCS).** The hip JCS [9] is a Cardan sequence with the *first* axis fixed to the pelvis (Z : flexion/extension), a *second* floating axis orthogonal to the others (X : adduction/abduction), and a *third* axis fixed to the femur (Y : internal/external rotation) — effectively **ZXY** about the body-fixed axes. The spine JCS is similarly defined using a **ZXY** sequence. The shoulder [8] is **YXY**. The knee is based on the Grood–Suntay definition which was adopted into the ISB standards as a **ZXY** sequence with a medial/lateral flexion axis. The ankle is the oblique talocrural axis based on the Rajagopal definitions.
- **Euler order matters.** Only the *first* rotation in a Cardan triple is order-independent. For example, hip flexion is robust but add/abduction and int/external rotation require the **same** decomposition order to compare across models.
- **Angles are frame-invariant; positions are not.** Discrepancies between the underlying global reference frames such as MuJoCo’s Z -up versus OpenSim’s Y -up convention fundamentally shift absolute spatial locations. This global mismatch affects joint-center MPJPE, which we corrected by a rigid Procrustes alignment. Relative rotations are invariant to the world coordinate system so joint-angle MJAE remains unaffected once local frame signs and Euler orders are reconciled.
- **Different default model poses.** A non-zero neutral-pose default or a rotated segment frame produces a constant offset that MJAE (no bias removal) penalizes in full but the per-frame Pearson r and bias-corrected MAE expose as a pure offset — the dominant signature of OpenCap’s hip-flexion error (§A.13.4).

- **Left-side coordinate mirroring.** Each framework negates or mirrors left-limb coordinates to preserve positive flexion bilaterally, so a raw joint value is only meaningful with the side (and, for SKEL, the horizontal-flip state).

A.13.2 The three biomechanical models OpenSim-OpenCap. The runtime model used by OpenCap-Monocular is OpenSim’s `LaiUhlrich2022.osim` (lineage: Rajagopal 2016, Lai 2017, Uhlrich 2022 [5, 6, 7]). Its global reference follows the standard OpenSim body-segment convention: +X anterior/posterior, +Y superior/inferior, +Z medial/lateral (subject’s right positive). The model is comprised of 22 bodies connected by 33 degrees of freedom, with angles reported as OpenSim generalized coordinates (rad). The patellofemoral joint is removed during scaling. **Default-pose offset:** some coordinates carry non-zero neutral defaults (`arm_add` = -1.43 rad, `pelvis_ty` = 0.93 m), producing a constant offset relative to a true zero pose. **HSMR.** HSMR regresses pose parameters for SKEL, a two-layer model built on **BSM** (`bsm.osim`), a genuine OpenSim tree combining a lower limb from Rajagopal 2016 with a shoulder from Seth 2016 [4, 10], with `skel_model.py`. Because BSM is itself OpenSim-derived, it shares the axis convention described above. The model comprises 24 bodies connected by 46 degrees of freedom. **Structural difference:** unlike a conventional 3-hinge Euler spine, SKEL represents the spine as a single constant-curvature joint with an added translation component, and its differentiable re-implementation drops BSM’s Walker knee coupling. **MuJoCo.** Mujoco models use OpenSim derived bone meshes and were used for the trainable biomechanical head and the three validation datasets: the BML-MoVi, BioCV, and MMC-PBL dataset. MuJoCo’s global reference frame is +X anterior/posterior, +Y medial/lateral (subject’s left positive), +Z superior/inferior. A fixed $R_x(+90^\circ)$ rotation at the pelvis root re-aligns every segment’s local axes back to the same convention used above (+X anterior, +Y superior, +Z mediolateral), so segment-level axis meanings match despite the different world orientation. The B3D model has 65 bodies and 99 degrees of freedom where the hips and shoulders are defined as three separate hinges in the XML but composed into MuJoCo ball (quaternion) joints at load time, then decoded back to ISB hinge angles (ZXY for hips, YXY for shoulders) for comparison; the knee remains a direct hinge. Knee varus/valgus, internal/external rotation, and metatarsophalangeal flexion/extension are disabled. The mujoco model used to compute kinematics from the validation datasets contains 22 bodies connected by 41 degrees of freedom. Hip and shoulders are defined as hinge joints and do not use the quaternion implementation.

A.13.3 Cross-model joint comparability Table S8 summarizes, joint by joint, how each model’s axis, Euler order, and sign compare to the reference, and whether the resulting angle is directly comparable, comparable after a routine adjustment such as a sign flip, or not comparable at all. Rather than hard-coding these signs, we resolve them from each model’s own live axes: a joint’s sign is +1 when its axis points the same way as the reference’s (`dot(axis_ours, axis_ref) > 0`) and -1 otherwise. In the table, “(ball)” marks a joint that was

initialized as a MuJoCo ball (quaternion) joint and was decoded into the listed Cardan order for comparison.

Table S8: Cross-model joint parameterization and comparability. Each model cell gives the axis / Euler order / sign.

Joint	OpenCap	HSMR	Ours	Verdict
Hip flexion	ZXY, Z, +	ZXY, Z, +	ZXY, + (ball)	Directly comparable
Hip adduction	ZXY, X, +	ZXY, X, +	ZXY, + (ball)	Comparable
Hip rotation	ZXY, Y, +	ZXY, Y, +	ZXY, + (ball)	Comparable
Knee flexion	Walker, + (L neg)	hinge, +	hinge, -1 vs ref	Comparable
Ankle	oblique hinge	oblique hinge	oblique hinge	Directly comparable
Lumbar / spine	ZXY hinges	const-curv transl	+ lumbar stack	Not comparable (type)
Shoulder	ZXY	XYZ ball	YXY (ball)	Comparable
Elbow	oblique pin	oblique pin	hinge	Directly Comparable (MoVi only)

Rule of thumb. In practice, an angle compares cleanly across models when three things line up: the joint is either single-DOF or we’re looking at the first axis of its Cardan sequence (this covers hip, knee, and ankle flexion), the Euler order matches (all three models use ISB’s ZXY for the hip), and the sign has been reconciled, which in this work refers to the knee.

A.13.4 Implications for the reported metrics

- **Hip-flexion offset (OpenCap, empirical).** OpenCap’s hip-flexion MJAE is dominated by a per-subject *offset* (high Pearson r , low bias-corrected MAE; one PBL control trial at 85° with $r = 0.98$): OpenSim’s pelvis/femur neutral frame and non-zero defaults produce a constant offset relative to the MuJoCo reference that MJAE (no bias removal) penalizes in full.
- **Joint-center MPJPE vs joint-angle MJAE.** Because the knee’s coupled translation differs across models, a method can match flexion angles yet differ in tibial joint-center position — why we report both PA-MPJPE and MJAE, and why the intra-skeleton PBL PA-MPJPE is a lower bound (it shares the reference’s exact coupled knee).

References

- [1] AI at Meta: SAM 3D Body Dataset. HuggingFace Datasets (2025), <https://huggingface.co/datasets/facebook/sam-3d-body-dataset>, the public SAM-3D-Body training dataset, released as [facebook/sam-3d-body-dataset](https://huggingface.co/datasets/facebook/sam-3d-body-dataset). Annotations are built on eight public image sources (3DPW, AI Challenger, COCO, EgoExo4D, EgoHumans, Harmony4D, MPII, SA-1B). Accessed June 2026.

- [2] Evans, M., Cazzola, D., et al.: A synchronised marker-based and markerless dataset of three-dimensional human movement. *Scientific Data* 11, 1300 (2024), pMC11604968; data at <https://researchdata.bath.ac.uk/1258>
- [3] Gilon, S., Miller, E.Y., Uhlich, S.D.: OpenCap Monocular: 3D human kinematics and musculoskeletal dynamics from a single smartphone video (2026), <https://arxiv.org/abs/2603.24733>, arXiv:2603.24733
- [4] Keller, M., Werling, K., Shin, S., Delp, S., Pujades, S., Liu, C.K., Black, M.J.: From skin to skeleton: Towards biomechanically accurate 3D digital humans. *ACM Transactions on Graphics (SIGGRAPH Asia)* 42(6) (2023)
- [5] Lai, A.K.M., Arnold, A.S., Wakeling, J.M.: 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)
- [6] Rajagopal, A., Dembia, C.L., DeMers, M.S., Delp, D.D., Hicks, J.L., Delp, S.L.: Full-body musculoskeletal model for muscle-driven simulation of human gait. *IEEE Transactions on Biomedical Engineering* 63(10), 2068–2079 (2016)
- [7] Uhlich, S.D., Falisse, A., Kidziński, Ł., Muccini, J., Ko, M., Chaudhari, A.S., Hicks, J.L., Delp, S.L.: OpenCap: Human movement dynamics from smartphone videos. *PLOS Computational Biology* 19(10), e1011462 (2023)
- [8] Wu, G., van der Helm, F.C., Veeger, H.D., Makhsoos, M., Van Roy, P., Anglin, C., Nagels, J., Karduna, A.R., McQuade, K., Wang, X., Werner, F.W., Buchholz, B.: ISB recommendation on definitions of joint coordinate systems of various joints for the reporting of human joint motion—part II: shoulder, elbow, wrist and hand. *Journal of Biomechanics* 38(5), 981–992 (2005)
- [9] Wu, G., Siegler, S., Allard, P., Kirtley, C., Leardini, A., Rosenbaum, D., Whittle, M., D’Lima, D.D., Cristofolini, L., Witte, H., Schmid, O., Stokes, I.: 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)
- [10] Xia, Y., Zhou, X., Vouga, E., Huang, Q., Pavlakos, G.: Reconstructing Humans with a Biomechanically Accurate Skeleton. In: 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5355–5365 (2025), arXiv:2503.21751
- [11] Yang, X., Kukreja, D., Pinkus, D., Sagar, A., Fan, T., Park, J., Cao, J., Liu, J., Ugrinovic, N., Feiszli, M., Malik, J., Dollar, P., Kitani, K.: SAM 3D Body: Robust Full-Body Human Mesh Recovery (2026), <https://arxiv.org/abs/2602.15989>, arXiv:2602.15989
- [12] Zhou, Y., Barnes, C., Lu, J., Yang, J., Li, H.: On the continuity of rotation representations in neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019), arXiv:1812.07035