BIOMECHANICALLY GROUNDED 3D HUMAN MESH RECONSTRUCTION VIA SPATIO-TEMPORAL ADAPTERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Recovering 3D human pose and shape from monocular input is central to computer vision, yet for anatomy-aware applications such as clinical motion analysis and sports biomechanics, geometric accuracy alone is insufficient: reconstructions must also be biomechanically valid. Most methods build on SMPL, whose joints are unconstrained ball-and-sockets and often produce physically impossible rotations. The SKEL model instead provides a physiologically grounded skeleton that is valid by construction, but learning to predict SKEL has remained limited by the scarcity of SKEL annotations. We introduce BioHMR, a SKEL-based framework that is simultaneously scalable, temporally coherent, and efficient. BioHMR reconstructs in an explicit skeletal latent space: a compact set of learnable joint tokens, one per SKEL joint, gathers evidence from image patches, reasons over the SKEL kinematic tree, and tracks each joint through time within a Vision Transformer, giving temporally coherent reconstruction without dense spatio-temporal attention. To overcome the annotation scarcity, a lightweight SMPL-to-SKEL mapper, trained by mesh consistency, generates pseudo-SKEL supervision from abundant SMPL annotations roughly $300\times$ faster than prior conversion, making large-scale SKEL training practical. The same joint-token attention drives an optional pose-aware patch pruning for a tunable compute-accuracy trade-off at inference. Across COCO, LSP-Extended, 3DPW, and MOYO, BioHMR attains state-of-the-art accuracy, efficiency, and temporal coherence among SKEL-based methods, with the lowest MPJPE on both 3DPW and MOYO and 2D accuracy on par with the strongest SMPL-based baselines. Moreover, its reconstructions stay biomechanically valid by construction, cutting per-joint violation rates on MOYO from 44–56% for SMPL-based HMR2.0 to 0–4%.

1 INTRODUCTION

Reconstructing the 3D body pose and shape of a person from monocular images or video is one of the central problems in computer vision, with numerous applications in healthcare (Armstrong et al., 2024; Liu et al., 2024a), sport analytics (Liu et al., 2024b; Zhang et al., 2023; Guo et al., 2024), animation (Saito et al., 2021; Guo et al., 2025), AR/VR (Song et al., 2023; Zhao et al., 2024), and human-computer interaction (DeVrio et al., 2023; Son Nguyen et al., 2025). Great progress has been made in human mesh recovery thanks to parametric body models such as SMPL (Loper et al., 2015) and SMPL-X (Pavlakos et al., 2019) and end-to-end regression architectures that directly learn shape and pose parameters from pixels (Kanazawa et al., 2018; Kolotouros et al., 2019; Goel et al., 2023). However, this progress is measured by geometric accuracy, and even highly accurate reconstructions are frequently biomechanically implausible.

This limitation stems in part from the widespread use of SMPL as the underlying representation. Although SMPL offers a convenient low-dimensional parameterization, its kinematic tree is not biomechanically accurate: each joint is modeled as an unconstrained 3-DoF ball-and-socket articulation, ignoring physiological rotation limits and true skeletal structure. As a result, even the best models produce nonphysical twisting, bending, or hyperextension. These artifacts may remain invisible to vertex- or keypoint-based metrics but are critical in anatomy-aware settings such as biomechanics and clinical motion analysis.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

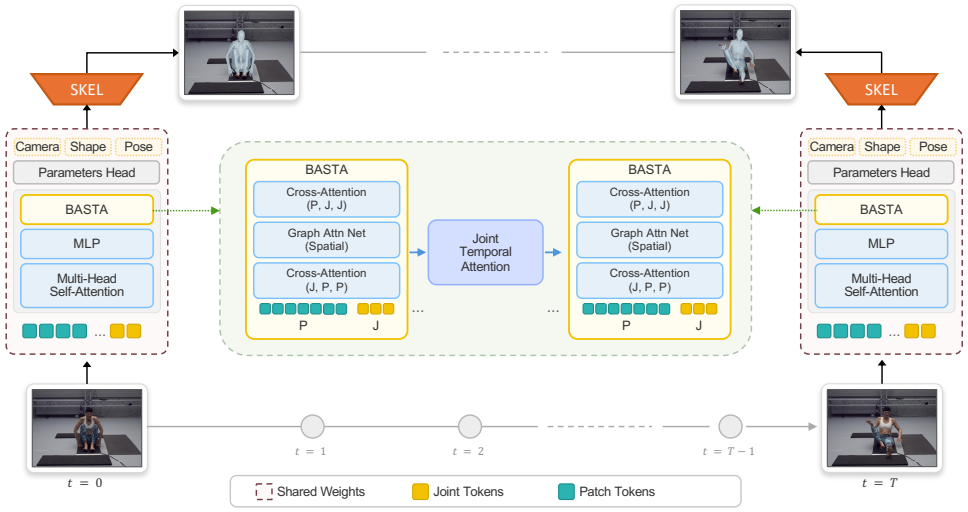


Figure 1: **BioHMR architecture.** For each frame $t \in \{0, \ldots, T\}$, the backbone couples ViT patch tokens with learnable joint tokens via BASTA (biomechanics-aware spatio-temporal adapter): joint tokens attend to patch tokens, are refined by graph attention over the SKEL kinematic tree, are fused across time by joint temporal attention, and write back to patch tokens. The resulting pose-aware representation is decoded into per-frame SKEL pose, shape, and camera parameters, from which the SKEL layer produces physiologically constrained 3D joints and meshes.

The recently proposed SKEL (Keller et al., 2023) model addresses this gap by integrating a biomechanically accurate skeleton beneath the SMPL surface, providing physiologically meaningful joint degrees of freedom and explicit rotation limits. Because SKEL constrains every joint to physiologically valid degrees of freedom, any model that predicts SKEL parameters rules out the gross, anatomically impossible articulations that pervade SMPL-based reconstruction *by construction*; the open challenge is making SKEL-based reconstruction accurate, temporally coherent, and trainable at scale. Nevertheless, SKEL-based learning has been held back by the scarcity of SKEL-annotated data. HSMR (Xia et al., 2025) is the most closely related prior work, exploring SKEL-based reconstruction from single images by training on pseudo-SKEL parameters obtained from SMPL via direct optimization; this conversion is computationally expensive ($\sim$180 s per frame) and the method operates on individual frames, limiting both scalability and temporal consistency.

In this work, we introduce BioHMR, a biomechanically informed transformer that predicts SKEL parameters from both images and videos. Our starting point is the observation that the binding constraint on SKEL-based learning is *supervision*, not architecture: the optimization-based SMPL-to-SKEL conversion of prior work (Keller et al., 2023; Xia et al., 2025) ($\sim$180 s/frame) would take thousands of GPU-days to label a corpus the size of standard human-mesh training sets, which is a primary reason SKEL-based learning has remained small-scale and single-image. We therefore learn a lightweight SMPL-to-SKEL mapping network that converts widely available SMPL annotations into pseudo-SKEL supervision at 0.01 s/frame (0.58 s with offline refinement), making large-scale SKEL supervision practical for the first time.

To turn this supervision into accurate, temporally coherent predictions, BioHMR reasons about the body in an explicit skeletal latent space through the Biomechanics-Aware Spatio-Temporal Adapter (BASTA): a compact set of learnable joint tokens, one per SKEL joint, that mediates between image features and SKEL parameters. Within each transformer block, these tokens aggregate evidence from image patches via cross-attention, reason over the SKEL kinematic tree with graph attention, and track each joint through time, then inject this structured skeletal and temporal context back into the backbone, giving physiologically structured, temporally coherent reconstruction without the cost of dense spatio-temporal attention. Finally, joint-token attention enables an optional pose-aware patch-selection mechanism for a tunable compute-accuracy trade-off at inference. Together, scalable supervision, joint-based temporal modeling, and pose-aware pruning make BioHMR the first SKEL framework that is simultaneously scalable, temporally coherent, and efficient.

BioHMR makes the following contributions.

- **Scalable SKEL supervision.** A learned SMPL-to-SKEL mapper, trained from mesh consistency, that generates pseudo-SKEL labels roughly $300\times$ cheaper than optimization-based conversion (up to four orders of magnitude for the raw mapper) and makes large-scale SKEL training feasible.
- **Temporal SKEL reconstruction.** BASTA modules that add kinematic-tree graph reasoning and per-joint temporal attention to a ViT backbone, yielding the first temporally coherent, video-based SKEL reconstruction without dense spatio-temporal attention.
- **Efficient inference.** An optional pose-aware patch-selection mechanism that prunes uninformative patches via joint-token attention, giving a tunable compute-accuracy trade-off with no extra parameters or retraining.
- Experiments on 2D and 3D benchmarks showing the best accuracy and biomechanical validity among SKEL-based methods, on par with the strongest SMPL-based baselines on 2D pose, at a fraction of the supervision and inference cost.

2 Related Work

2.1 Parametric Human Body Models

Parametric models have played a central role in 3D human reconstruction. Early work such as SCAPE (Anguelov et al., 2005) introduced data-driven surface modeling, while SMPL (Loper et al., 2015) and its extensions (*e.g.*, SMPL-H (Romero et al., 2017), SMPL-X (Pavlakos et al., 2019), and GHUM (Xu et al., 2020)) provided low-dimensional, differentiable representations that enabled direct regression from images. Despite their effectiveness, these models adopt simplified kinematic structures with ball-and-socket joints and unconstrained rotation axes, and therefore often produce physiologically implausible joint configurations. The SKEL (Keller et al., 2023) model instead embeds a biomechanically accurate skeleton with physiologically valid degrees of freedom and explicit joint limits, making predicted poses valid by construction; its main obstacle has been the scarcity of SKEL annotations.

2.2 SMPL-Based Human Mesh Recovery

Early methods recovered SMPL parameters by optimization over silhouettes or keypoints (Bogo et al., 2016), whereas modern methods regress them end-to-end (Kanazawa et al., 2018; Kolotouros et al., 2019). HMR (Kanazawa et al., 2018) predicts SMPL directly from an image with a 2D reprojection loss and adversarial priors; SPIN (Kolotouros et al., 2019) and CLIFF (Li et al., 2022) improve accuracy through iterative fitting and stronger training data, and transformer backbones such as HMR2.0 (Goel et al., 2023) further raise mesh accuracy. To exploit temporal cues, video-based methods aggregate information across frames using recurrent units (Kocabas et al., 2020; Luo et al., 2020), convolutional temporal models (Kanazawa et al., 2019), temporal encoders (Choi et al., 2021), non-local attention (Wei et al., 2022), or multi-scale fusion (Shen et al., 2023) to reduce jitter and improve stability. A parallel line of work lifts detected 2D keypoints to 3D pose with temporal convolutions (Pavllo et al., 2019) or spatio-temporal transformers (Zheng et al., 2021; Zhu et al., 2023). These methods are effective but share two limitations relevant to us: they regress SMPL joints or pose and thus inherit physiologically unconstrained kinematics, and their temporal variants typically rely on full spatio-temporal attention, keypoint-lifting pipelines, or sequence-level models that are computationally expensive. BioHMR is not another generic spatio-temporal transformer: its joint tokens are structured by the SKEL kinematic graph rather than a flat token sequence, they exchange information *bidirectionally* with image patches inside the backbone rather than lifting precomputed keypoints, and they are trained with biomechanical SKEL-parameter objectives rather than purely geometric ones.

2.3 Biomechanically Accurate Reconstruction

A recent line of work targets physiological validity directly by predicting biomechanical body models. HSMR (Xia et al., 2025) reconstructs SKEL from single images, but generates its

pseudo-labels through SKEL's expensive per-frame optimization and performs no temporal modeling. BioPose (Koleini et al., 2025) targets biomechanical accuracy from video but builds on a *different* biomechanical skeleton model rather than SKEL, through a heavy multi-stage inverse-kinematics pipeline. Thus every existing SKEL-based reconstructor is single-image and optimization-supervised, and no biomechanical method offers scalable supervision, efficient inference, and temporal modeling together. BioHMR is the first to provide all three: it learns SKEL supervision at scale from abundant SMPL annotations, reasons in an explicit skeletal latent space for temporally coherent video reconstruction, and prunes uninformative tokens for efficient inference.

3 METHOD

This section describes BioHMR's biomechanical representation, its architectural components, the scalable supervision strategy for obtaining SKEL training data, and the training objectives.

3.1 OVERVIEW

BioHMR estimates physiologically accurate 3D human meshes from monocular video or single images by predicting the parameters of the SKEL model. As shown in Fig. 1, reconstruction is carried out in an explicit skeletal latent space of learnable joint tokens, one per SKEL joint, which inside each Vision Transformer block couple with the patch tokens through structured cross-attention, graph reasoning over the SKEL kinematic tree, and per-joint temporal attention (our *Biomechanics-Aware Spatio-Temporal Adapter*, BASTA), injecting skeletal structure and temporal consistency without full spatio-temporal attention. A transformer-based regression head maps the final features to SKEL pose, shape, and camera parameters.

3.2 BODY MODELS: SMPL AND SKEL

Both SMPL (Loper et al., 2015) and SKEL (Keller et al., 2023) are differentiable parametric body models: given pose and shape parameters, each applies blend shapes, forward kinematics over a kinematic tree, and linear blend skinning to produce a mesh, so gradients flow from mesh-level losses back to the parameters for end-to-end training. We denote their differentiable forward functions $\mathcal{S}_{\text{SMPL}}$ and $\mathcal{S}_{\text{SKEL}}$.

SKEL (Keller et al., 2023) combines the widely used SMPL surface (with the same $N = 6890$ vertices) with a physiologically grounded skeletal structure; since it preserves the SMPL mesh topology, a regressor can be applied directly to obtain the 3D positions of the 24 joints. The key distinction lies in the pose representation: SMPL models each joint as a 3-DoF ball joint regardless of the underlying anatomy, whereas SKEL parameterizes each joint by its true biomechanical behavior, assigning only the physiologically valid degrees of freedom with realistic rotation limits. This yields a more compact pose vector (46 parameters vs. 72 in SMPL) and ensures that joint rotations comply with human kinematic constraints. A practical challenge, however, is the limited availability of datasets annotated directly with SKEL parameters; we address this in Section 3.5.

3.3 BIOHMR ARCHITECTURE

BioHMR represents the body in a low-dimensional joint-token space and reasons over it within a Vision Transformer (ViT) backbone. Given an input image or video clip, the network outputs SKEL pose and shape parameters, which are then decoded into a biomechanically valid 3D mesh.

3.3.1 BACKBONE ENCODER.

We adopt a ViT-style encoder following (Goel et al., 2023; Xia et al., 2025): each of the T input frames is resized to 256×192 and embedded into $H \times W$ patch tokens (16×16 patches), which are concatenated across frames into a length-$T \times H \times W$ sequence and processed by standard self-attention blocks.

3.3.2 BIOMECHANICS-AWARE SPATIO-TEMPORAL ADAPTERS (BASTA).

A plain ViT decoder regresses SKEL parameters from a flat set of patch features and has no notion of the body's articulated structure: it cannot express that an error at one joint should be resolved using its anatomical neighbors, nor that a joint should move smoothly across frames. BASTA injects exactly these two inductive biases, and does so in a compact joint space so they remain cheap even on video. Each transformer block is augmented with a lightweight BASTA module organized around a small set of learnable *joint tokens*, one per SKEL joint, which act as an explicit skeletal bottleneck between the image features and the SKEL parameters. While learnable joint tokens for single-image 2D pose were introduced by UniTransPose (Kinfu & Vidal, 2026), BASTA differs in three ways central to biomechanically accurate, temporal SKEL recovery: it (i) routes information *bidirectionally* between patches and joints, (ii) reasons over the *SKEL kinematic graph*, and (iii) propagates each joint *through time*. Each module has four steps:

1. **Down-projection.** Patch tokens are linearly projected from the ViT dimension C to adapter dimension C_a via a learned matrix $\mathbf{W}_{\text{down}} \in \mathbb{R}^{C \times C_a}$.
2. **Patch-to-joint aggregation.** J learnable *joint tokens* queries the projected patch tokens via cross-attention to produce joint-specific features of shape $T \times J \times C_a$.
3. **Spatio-temporal skeletal reasoning.** The joint tokens are updated by graph-attention layers over the SKEL kinematic graph, followed by a temporal transformer that models per-joint dynamics across the T frames.
4. **Joint-to-patch redistribution.** Patch tokens query the refined joint tokens via cross-attention, absorbing the structured skeletal information; the result is projected back to dimension C via $\mathbf{W}_{\text{up}} \in \mathbb{R}^{C_a \times C}$ and added to the ViT token stream through a gated residual connection.

This design allows BASTA to inject biomechanically structured information into every selected transformer block while the gated residual connection lets it integrate gradually for stable training.

Joint tokens and skeletal graph. The J joint tokens are a single shared set of learnable vectors: every BASTA module initializes its tokens from this set and refines them independently, so all layers operate on a common skeletal representation rather than separate per-layer token banks. The graph used in the spatio-temporal reasoning step is a fixed undirected adjacency derived from the SKEL kinematic tree, over which we apply multi-head graph attention (Brody et al., 2022) so that information propagates along anatomically connected parts (*e.g.*, pelvis-hip-knee-ankle chains). Because messages flow only along anatomical connections, a locally ambiguous joint, such as an occluded or motion-blurred knee, is constrained by its parent and child joints along the kinematic chain, a dependency that patch-level self-attention cannot express and that directly targets the implausible articulations motivating SKEL.

Temporal modeling. For video ($T > 1$), the spatio-temporal reasoning step applies a per-joint temporal transformer that treats each joint as a length-T sequence. Attending over the $J = 24$ joint tokens rather than the full $T \times H \times W$ patch grid reduces the cost of temporal modeling by orders of magnitude, making end-to-end temporal SKEL reconstruction tractable where dense spatio-temporal attention would be prohibitive. For single-frame input ($T = 1$), the temporal branch is an identity and BASTA reduces to a purely spatial skeletal module.

3.3.3 REGRESSION HEAD AND OUTPUTS.

BioHMR predicts SKEL parameters through two complementary decoding mechanisms.

Primary decoder. The primary decoder applies a transformer decoder with a single learnable query token to the patch-token features produced by the backbone; the resulting query embedding is mapped via linear layers to SKEL pose ($\mathbf{q}$), shape ($\boldsymbol{\beta}$), and weak-perspective camera parameters. Poses are estimated in the continuous 6D rotation representation (144D for 24 joints) and then converted to Euler-angle representation. The SKEL layer subsequently applies joint-specific degree-of-freedom constraints, selecting only the 46 physiologically valid axes; the remaining axes are fixed by the SKEL model's kinematic structure.

Auxiliary joint-token decoder. In parallel, an auxiliary decoder attention-pools the BASTA joint tokens with a learned query and regresses pose and shape through an MLP, supervised against

the pseudo-ground-truth targets. This branch is used *only during training*: its supervision ensures that joint tokens encode joint-specific information, strengthening the alignment between the learned skeletal latent space and the SKEL parameterization.

3.4 Pose-Aware Patch Selection

Beyond enriching backbone representations, the joint tokens provide a natural signal for identifying the image patches most relevant to the pose. Inspired by Kinfu & Vidal (2023), we exploit the cross-attention scores between joint and patch tokens at inference time to estimate patch importance: patches with low attention responses are pruned, discarding uninformative regions while preserving pose-relevant context and enabling a tunable compute-accuracy trade-off.

Concretely, each patch receives an importance score combining the maximum and the mean cross-attention it receives from the joint tokens, favoring patches critical for at least one joint while accounting for global relevance. Starting from a predefined backbone depth, the lowest-scoring patches (subject to a minimum budget) are *frozen*, stored without further updates by subsequent blocks, and reinserted at their original locations at the end of the backbone, so the decoder still operates on a dense feature map while depth-wise computation concentrates on pose-relevant patches. The full scoring and masking procedure is given in Appendix C.3. We emphasize that patch selection is an *optional* inference-time mechanism enabled by the joint-token design. Section 4.5 analyzes the trade-off in detail.

3.5 SKEL Supervision via SMPL-to-SKEL Mapping Network

While SKEL provides biomechanically accurate pose representation, datasets annotated directly with SKEL parameters are extremely limited. The SKEL framework provides an optimization-based conversion (Keller et al., 2023), employed by HSMR (Xia et al., 2025), that directly optimizes SKEL parameters to match SMPL meshes using gradient-based methods (*e.g.*, L-BFGS). However, this process requires approximately 180 s ($\sim$3 min) per frame, making the conversion of standard human-mesh training corpora (millions of frames) computationally prohibitive; this bottleneck has confined learning-based SKEL recovery to small datasets and single images. To make large-scale training feasible, we introduce a lightweight mapping network that converts SMPL parameters into SKEL counterparts at negligible cost ($\sim$0.01 s per frame). Together with the offline geometric refinement described below, the full conversion pipeline produces the pseudo-SKEL labels used for training roughly $300\times$ faster than the original optimization-based procedure ($\sim$0.58 s vs. $\sim$180 s per frame); the raw mapper alone is up to four orders of magnitude faster.

Training Objective. The mapper $\mathcal{M}$, a compact MLP whose architecture is detailed in Appendix B.4, is trained to enforce geometric consistency between the SMPL mesh and the SKEL mesh produced from the predicted parameters, generated with the differentiable forward functions of the two models. The objective combines per-vertex and per-joint alignment terms with a Laplacian surface regularizer that suppresses the conversion failure modes observed in Keller et al. (2023), and selective joint masking down-weights regions where the two models differ systematically (*e.g.*, upper limbs) while retaining dense vertex supervision; the full formulation is given in Appendix B.4. Critically, this objective is *self-supervised by the SMPL source* and uses no SKEL ground truth: the converter never sees an optimization-derived SKEL label, so building it does not reintroduce the expensive optimization it is meant to replace. It can therefore be trained once on a small set of SMPL meshes and then applied to convert arbitrarily large SMPL-annotated corpora. We evaluate the accuracy and speed of our learned mapper against the optimization-based baseline in Section 4.4.

Pseudo-Label Refinement. While the learned mapping network provides fast and reasonably accurate pseudo-SKEL labels, the conversion is imperfect; we apply two complementary refinement stages (detailed in Appendix B.4). *Offline*, a lightweight L-BFGS optimization over the SKEL parameters improves mesh alignment with the SMPL source under the same mesh-consistency objective; warm-started from the mapper's near-optimal output, it converges in a handful of iterations rather than the hundreds needed from scratch, and the combined mapper + refinement pipeline runs in $\sim$0.58 s per frame, $300\times$ faster than direct optimization, with substantially better alignment (Section 4.4, Table 4). *Online*, during BioHMR training we periodically apply SKELify (Xia et al., 2025), a SMPLify-style (Bogo et al., 2016) 2D-reprojection optimization under SKEL's physiological joint-limit constraints, to a small random subset of samples, warm-started from the pseudo-labels

Table 1: **Quantitative comparison on 2D and 3D benchmarks.** PCK@0.05/@0.1 on COCO and LSP-Extended (higher ↑ is better); MPJPE and PA-MPJPE (mm) on 3DPW and MOYO, with MPVPE and PA-MPVPE (mm) on MOYO (lower ↓ is better). Baseline numbers are taken from the respective papers (TokenHMR COCO PCK as reported by CameraHMR (Patel & Black, 2024)); "–" marks results not reported.

Method	COCO		LSP-Extended		3DPW		MOYO			
	@0.05↑	@0.1↑	@0.05↑	@0.1↑	MPJPE↓	PA-MPJPE↓	MPJPE↓	PA-MPJPE↓	MPVPE↓	PA-MPVPE↓
PARE (Kocabas et al., 2021)	0.72	0.91	0.27	0.60	82.0	50.9	165.6	117.1	174.5	121.9
CLIFF (Li et al., 2022)	0.64	0.88	0.32	0.66	69.0	43.0	154.6	109.3	–	–
HybrIK (Li et al., 2021)	0.61	0.80	0.37	0.69	80.0	48.8	140.1	93.2	143.6	94.4
PLIKS (Shetty et al., 2023)	0.62	0.90	0.26	0.66	–	–	132.6	91.8	136.7	94.8
HMR2.0b (Goel et al., 2023)	**0.86**	**0.96**	**0.53**	**0.82**	81.3	54.3	123.3	90.4	142.2	103.4
TokenHMR (Dwivedi et al., 2024)	0.80	0.95	–	–	71.0	44.3	–	–	–	–
HSMR (Xia et al., 2025)	0.85	**0.96**	0.51	0.81	81.5	54.8	104.5	79.6	120.1	90.7
BioHMR (ours)	0.84	0.95	0.52	**0.82**	**65.3**	**39.9**	**90.3**	**57.8**	**98.2**	**64.5**

and with half the iterations used by HSMR, a reduction made practical by the mapper's high-quality initialization.

3.6 Training Objectives

Our training objective follows the formulation introduced in HSMR (Xia et al., 2025), extended with temporal consistency and auxiliary joint-token supervision. We supervise SKEL pose, shape, and camera parameters, as well as their induced 2D and 3D joints, using ground-truth labels when available and pseudo-labels otherwise.

Parameter and Keypoint Losses. Let $\mathbf{q} \in \mathbb{R}^{46}$ and $\boldsymbol{\beta} \in \mathbb{R}^{10}$ denote the SKEL pose and shape parameters predicted by the primary decoder. We first convert $\mathbf{q}$ to rotation matrices $\mathbf{q}_{mat}$, and apply parameter losses $\mathcal{L}_q = \|\mathbf{q}_{mat} - \mathbf{q}^*_{mat}\|_F^2$ and $\mathcal{L}_\beta = \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2^2$, where $\mathbf{q}^*_{mat}, \boldsymbol{\beta}^*$ are the pseudo-ground-truth labels. We further supervise the predicted 3D joints $\mathbf{X}$ and their 2D projections $\pi(\mathbf{X})$ with $\mathcal{L}_{\text{kp3D}} = \|\mathbf{X} - \mathbf{X}^*\|_1$ and $\mathcal{L}_{\text{kp2D}} = \|\pi(\mathbf{X}) - \mathbf{x}^*\|_1$.

Temporal Consistency. For videos, we impose smoothness across adjacent frames via $\mathcal{L}^q_{\text{temp}} = \sum_t \|\mathbf{q}^t_{mat} - \mathbf{q}^{t-1}_{mat}\|_F^2$ and $\mathcal{L}^\beta_{\text{temp}} = \sum_t \|\boldsymbol{\beta}_t - \boldsymbol{\beta}_{t-1}\|_2^2$. These losses discourage abrupt parameter fluctuations and improve temporal stability; since body shape is constant within a clip, $\mathcal{L}^\beta_{\text{temp}}$ acts as a noise-suppressing regularizer.

Auxiliary Joint-Token Supervision. The auxiliary joint-token decoder produces predictions $(\mathbf{q}^{\text{aux}}, \boldsymbol{\beta}^{\text{aux}})$, supervised with the same parameter losses: $\mathcal{L}_{\text{aux}} = \|\mathbf{q}^{\text{aux}}_{mat} - \mathbf{q}^*_{mat}\|_F^2 + \|\boldsymbol{\beta}^{\text{aux}} - \boldsymbol{\beta}^*\|_2^2$. Although unused at inference, this loss regularizes the joint tokens, encouraging them to encode joint-specific biomechanical information aligned with the SKEL skeleton.

Full Objective. The final training loss is:

$$\mathcal{L} = \lambda_q \mathcal{L}_q + \lambda_\beta \mathcal{L}_\beta + \lambda_{\text{3D}} \mathcal{L}_{\text{kp3D}} + \lambda_{\text{2D}} \mathcal{L}_{\text{kp2D}} + \lambda_{\text{temp}} \left(\mathcal{L}^q_{\text{temp}} + \mathcal{L}^\beta_{\text{temp}} \right) + \lambda_{\text{aux}} \mathcal{L}_{\text{aux}}. \quad (1)$$

4 Experiments and Results

4.1 Experimental Setup

Training Datasets. We train BioHMR on Human3.6M (Ionescu et al., 2014), MPI-INF-3DHP (Mehta et al., 2017), COCO (Lin et al., 2014), MPII (Andriluka et al., 2014), AI Challenger (Wu et al., 2017), AVA (Gu et al., 2018), and InstaVariety (Kanazawa et al., 2019). Unlike several prior works, neither 3DPW (von Marcard et al., 2018) nor the synthetic BEDLAM (Black et al., 2023) dataset is used for training: all 3DPW results are zero-shot, and our training corpus matches that of HSMR (Xia et al., 2025) for a fair comparison. These datasets provide diverse coverage of indoor, outdoor, and in-the-wild human poses. We convert the SMPL parameters into SKEL pseudo-labels using our lightweight SMPL-to-SKEL mapping network (Sec. 3.5).

Evaluation Datasets. We evaluate on COCO and LSP-Extended for 2D keypoint estimation from images, and on 3DPW and MOYO for 3D pose estimation from videos.

Table 2: **Joint-limit violations on MOYO.** Percentage of frames whose joint rotations exceed physiological limits by $> 10°$ and $> 20°$. SMPL-based methods violate frequently, whereas SKEL's hinge knees and elbows rule out these gross violations by construction, keeping HSMR and BioHMR near-zero.

	$> 10°$				$> 20°$			
Method	L.elb.	R.elb.	L.knee	R.knee	L.elb.	R.elb.	L.knee	R.knee
PARE	36.4	42.4	20.0	23.2	14.6	15.4	3.2	3.8
CLIFF	34.2	33.0	28.3	31.0	13.0	12.4	4.8	4.5
HybrIK	58.7	60.9	52.9	48.6	29.4	34.6	30.7	27.0
PLIKS	41.6	44.7	47.4	43.8	17.9	22.7	18.2	17.6
HMR2.0b	47.6	44.3	45.7	56.4	19.8	19.6	6.4	11.6
HSMR	**0.0**	**0.0**	3.9	4.5	**0.0**	**0.0**	0.2	0.5
BioHMR	**0.0**	**0.0**	**2.1**	**4.2**	**0.0**	**0.0**	**0.0**	**0.2**

Table 3: **Broad-body validity on MOYO.** Frames with violations $> 0°$/$> 5°$ for HSMR and BioHMR. BioHMR is the most biomechanically valid on every joint.

	HSMR		BioHMR	
Joint	$> 0°$	$> 5°$	$> 0°$	$> 5°$
L. knee	16.0	9.1	**15.0**	**8.4**
R. knee	14.8	8.6	**14.5**	**8.5**
L. hip	6.0	4.1	**2.0**	**0.8**
R. hip	7.8	5.6	**5.1**	**3.5**
Elbows	1.6	**0.1**	**0.4**	**0.1**

Table 4: **SMPL-to-SKEL conversion.** MPVPE (mm) and mean joint violation rate (Viol., %, $>10°$) per dataset; our mapper + refine is $\sim$300$\times$ faster and improves both.

	Direct Optim. (Keller et al., 2023)		MLP+Refine (ours)	
Dataset	MPVPE↓	Viol.↓	MPVPE↓	Viol.↓
COCO	27.10	0.40	**13.04**	**0.20**
MPII	32.27	0.80	**15.99**	**0.75**
3DHP	37.42	0.93	**18.11**	**0.65**
Time	180 s/frame		**0.58 s/frame**	

Figure 2: **Conversion quality.** MLP mapper (center) misaligns wrist/knee/foot (red); MLP with offline refinement (right) corrects them.

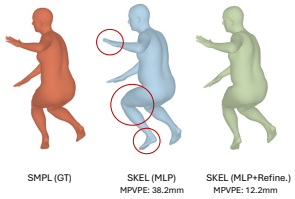


Table 5: **Temporal modeling (3DPW).** BioHMR improves MPJPE over per-frame and video baselines, with competitive ACCEL.

Method	MPJPE↓	ACCEL↓
HSMR	81.5	25.2
VIBE	82.9	23.4
VIMO	72.7	9.8
BioHMR	**65.3**	10.2

Metrics. We use the following standard metrics. *2D pose:* Percentage of Correct Keypoints (PCK) at thresholds 0.05 and 0.1. *3D joints:* Mean Per-Joint Position Error (MPJPE, mm) and its Procrustes-aligned variant (PA-MPJPE). *Surface:* Mean Per-Vertex Position Error (MPVPE, mm) and PA-MPVPE. *Biomechanical validity:* percentage of frames with joint rotations exceeding physiological limits, reported at thresholds from the strict $> 0°$ up to $> 30°$ (Xia et al., 2025). *Temporal:* per-joint acceleration error (ACCEL, mm s^{-2}).

4.2 Comparison with State of the Art

Table 1 reports quantitative results on the 2D (COCO, LSP-Extended) and 3D (3DPW, MOYO) benchmarks; BioHMR uses per-joint temporal attention when video is available and is otherwise run per frame. On 2D pose estimation, BioHMR is within 0.02 PCK@0.05 of the strongest baseline (HMR2.0b) on COCO and within 0.01 on LSP-Extended, tying it at LSP-Extended PCK@0.1, showing that pseudo-SKEL supervision does not compromise 2D accuracy.

On 3D pose estimation, BioHMR is the strongest 3D method, attaining the lowest MPJPE and PA-MPJPE on both 3DPW (65.3/39.9) and MOYO (90.3/57.8), as well as the best MOYO MPVPE and PA-MPVPE, all without ever training on 3DPW. It also surpasses the recent SMPL-based regressor TokenHMR (Dwivedi et al., 2024) on 3DPW despite predicting the more constrained SKEL parameterization. Together with its superior biomechanical validity (Table 3) and far lower supervision cost, BioHMR delivers the best overall accuracy, validity, and efficiency among SKEL-based methods. On video, BioHMR also attains the best 3DPW MPJPE among temporal baselines with competitive acceleration error (Table 5).

4.3 Physiological Validity

We distinguish two levels of biomechanical validity. At the *structural* level, SKEL models the knee and elbow as constrained hinges rather than ball-and-socket joints, so the gross out-of-plane twisting and abduction that SMPL permits are impossible by construction. Table 2 confirms this at

the coarse $> 10^\circ$ and $> 20^\circ$ thresholds: SMPL-based methods violate frequently (*e.g.*, HMR2.0b: 45.7% left-knee violations $> 10^\circ$), whereas the SKEL-based HSMR and BioHMR are near-zero. Staying within SKEL's finer anatomical *angle limits*, however, is not automatic: a regressor can still hyperextend a joint past its physiological range, showing that adopting SKEL alone does not guarantee fine-grained validity. Table 3 reports per-joint violations (hips included) at the strict $> 0^\circ$ and $> 5^\circ$ thresholds for the SKEL-based methods; BioHMR is the most biomechanically valid on every joint, pairing SKEL's structural guarantees with the lowest residual within-limit violations, even under $300\times$ cheaper supervision. Full per-threshold results are in the appendix (Table 6).

4.4 SMPL-to-SKEL Mapping Evaluation

Table 4 compares the accuracy and speed of our learned SMPL-to-SKEL mapper against the optimization-based conversion from SKEL (Keller et al., 2023) used by HSMR (Xia et al., 2025). The full offline pipeline (mapper + geometric refinement) is approximately $300\times$ faster than direct optimization (0.58 s vs. 180 s per frame; the raw mapper alone achieves 18,000$\times$ by running in $\sim$0.01 s) while achieving substantially better alignment, roughly halving MPVPE on all three datasets (*e.g.*, 27.10 to 13.04 mm on COCO). Crucially, Table 4 also reports the *biomechanical validity* of the generated labels: our pipeline matches or improves the direct-optimization baseline on every dataset, so the $300\times$ speed-up does not sacrifice label validity. Figure 2 shows a representative example, where offline refinement corrects the raw mapper's misalignments; additional comparisons are in the appendix.

4.5 Inference Efficiency via Pose-Aware Patch Selection

Pose-aware patch selection gives a tunable compute-accuracy trade-off at inference with no retraining. With aggressive pruning, GFLOPs drop from 134.7 to 98.7 (27%) and inference improves from 100.05 to 76.45 ms/batch on an A100 GPU. The accuracy cost is modest at practical thresholds (COCO PCK@0.10 $0.96 \rightarrow 0.94$); the larger drop appears only under the stricter PCK@0.05. Full accuracy-efficiency curves are in the appendix (Fig. 4).

Additional results. Implementation details, full component ablations (Table 7), full per-threshold validity results, and qualitative results (comparisons with HMR2.0 and HSMR, multi-view renderings, joint-token attention visualizations, and failure cases) are provided in the appendix; a video demo highlighting the temporal coherence of our reconstructions is included in the supplementary material.

5 Conclusion

We present **BioHMR**, a biomechanically informed transformer for 3D human mesh reconstruction from monocular images and video, driven by three design choices. First, through our BASTA mechanism, BioHMR reasons in an explicit skeletal latent space where learnable joint tokens perform cross-attention, kinematic-graph reasoning, and per-joint temporal attention, enabling temporal coherence without dense spatio-temporal attention. Second, a lightweight SMPL-to-SKEL mapping network accelerates pseudo-label generation by roughly $300\times$, allowing scalable training on existing SMPL-annotated datasets. Third, pose-aware patch selection prunes uninformative patches via joint-token attention for a tunable compute-accuracy trade-off at inference. Experiments on COCO, LSP-Extended, 3DPW, and MOYO demonstrate state-of-the-art accuracy, efficiency, and temporal coherence among SKEL-based methods, with the lowest MPJPE on both 3DPW and MOYO and 2D accuracy on par with the strongest SMPL-based baselines, while the SKEL parameterization rules out gross anatomical violations by construction even under scalable learned supervision. Remaining failure modes (inverted poses, extreme contortion, and heavy self-occlusion) are analyzed in the appendix (Figure 9); errors in these cases stem from inherent monocular ambiguity, while the reconstructions remain biomechanically valid. Scalable pseudo-label generation with physiologically grounded models such as SKEL offers a promising path toward biomechanically faithful human modeling at scale.

AI USE STATEMENT

In this work, we used generative AI tools to assist with editing and polishing the manuscript text and with routine implementation tasks in the experimental pipeline. We have not used generative AI tools for research ideation, experimental design, or the analysis and interpretation of results. We have reviewed all AI-assisted work: AI-assisted text was checked for factual accuracy, and AI-assisted code was verified and tested for correctness by the authors. The authors take responsibility for the final content of this work.

REFERENCES

Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2D human pose estimation: New benchmark and state of the art analysis. In *CVPR*, pp. 3686–3693, 2014.

Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. SCAPE: Shape completion and animation of people. In *ACM SIGGRAPH*, pp. 408–416, 2005.

Kai Armstrong, Lei Zhang, Yan-mao Wen, Alexander P. Willmott, Paul Lee, and Xujiong Ye. A marker-less human motion analysis system for motion-based biomarker identification and quantification in knee disorders. *Frontiers in Digital Health*, 6, 2024.

Michael J. Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. BEDLAM: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8726–8737, 2023.

Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J. Black. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *ECCV*, pp. 561–578, 2016.

Shaked Brody, Uri Alon, and Eran Yahav. How attentive are graph attention networks? In *ICLR*, 2022.

Hongsuk Choi, Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. Beyond static features for temporally consistent 3D human pose and shape from a video. In *CVPR*, pp. 1964–1973, 2021.

Nathan DeVrio, Vimal Mollyn, and Chris Harrison. SmartPoser: Arm pose estimation with a smartphone and smartwatch using UWB and IMU data. In *Proc. ACM Symp. User Interface Software and Technology (UIST)*, pp. 1–11, 2023.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

Sai Kumar Dwivedi, Yu Sun, Priyanka Patel, Yao Feng, and Michael J. Black. TokenHMR: Advancing human mesh recovery with a tokenized pose representation. In *CVPR*, pp. 1323–1333, 2024.

Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4D: Reconstructing and tracking humans with transformers. In *ICCV*, pp. 14783–14794, 2023.

Chunhui Gu, Chen Sun, David A. Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, et al. AVA: A video dataset of spatio-temporally localized atomic visual actions. In *CVPR*, pp. 6047–6056, 2018.

Wenxuan Guo, Zhiyu Pan, Ziheng Xi, Alapati Tuerxun, Jianjiang Feng, and Jie Zhou. Sports analysis and VR viewing system based on player tracking and pose estimation with multimodal and multiview sensors. *arXiv preprint arXiv:2405.01112*, 2024.

Zhiyang Guo, Jinxu Xiang, Kai Ma, Wengang Zhou, Houqiang Li, and Ran Zhang. Make-it-Animatable: An efficient framework for authoring animation-ready 3D characters. In *CVPR*, pp. 10783–10792, 2025.

Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. *IEEE TPAMI*, 36 (7):1325–1339, 2014.

Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *CVPR*, pp. 7122–7131, 2018.

Angjoo Kanazawa, Jason Y. Zhang, Panna Felsen, and Jitendra Malik. Learning 3D human dynamics from video. In *CVPR*, pp. 5614–5623, 2019.

Marilyn Keller, Keenon Werling, Soyong Shin, Scott L. Delp, Sergi Pujades, C. Karen Liu, and Michael J. Black. From skin to skeleton: Towards biomechanically accurate 3D digital humans. *ACM TOG*, 42(6):1–12, 2023.

Kaleab A. Kinfu and René Vidal. Transformers with joint tokens and local-global attention for efficient human pose estimation. *IEEE TPAMI*, 2026.

Kaleab Alamayehu Kinfu and René Vidal. Efficient vision transformer for human pose estimation via patch selection. In *BMVC*, 2023.

Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. VIBE: Video inference for human body pose and shape estimation. In *CVPR*, pp. 5253–5263, 2020.

Muhammed Kocabas, Chun-Hao P. Huang, Otmar Hilliges, and Michael J. Black. PARE: Part attention regressor for 3D human body estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11127–11137, 2021.

Farnoosh Koleini, Muhammad Usama Saleem, Pu Wang, Hongfei Xue, Ahmed Helmy, and Abbey Fenwick. BioPose: Biomechanically-accurate 3D pose estimation from monocular videos. In *IEEE/CVF Winter Conf. on Applications of Computer Vision (WACV)*, 2025.

Nikos Kolotouros, Georgios Pavlakos, Michael J. Black, and Kostas Daniilidis. Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In *ICCV*, pp. 2252–2261, 2019.

Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, and Cewu Lu. HybrIK: A hybrid analytical-neural inverse kinematics solution for 3D human pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3383–3393, 2021.

Zhihao Li, Jianzhuang Liu, Zhensong Zhang, Songcen Xu, and Youliang Yan. CLIFF: Carrying location information in full frames into human pose and shape estimation. In *ECCV*, pp. 590–606, 2022.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, pp. 740–755, 2014.

Benjamin Liu, Gilles Soenens, Joshua Villarreal, Jeffrey Jopling, Isabelle Van Herzeele, Anita Rau, and Serena Yeung-Levy. A human mesh-centered approach to action recognition in the operating room. *Artificial Intelligence Surgery*, 4(2):92–108, 2024a. doi: 10.20517/ais.2024.19.

Jingyuan Liu, Nazmus Saquib, Zhutian Chen, Rubaiat Habib Kazi, Li-Yi Wei, Hongbo Fu, and Chiew-Lan Tai. PoseCoach: A customizable analysis and visualization system for video-based running coaching. *IEEE TVCG*, 30:3180–3195, 2024b.

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM TOG*, 34(6):248:1–248:16, 2015.

Zhengyi Luo, S. Alireza Golestaneh, and Kris M. Kitani. 3D human motion estimation via motion compression and refinement. In *ACCV*, 2020.

Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5442–5451, 2019.

Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3D human pose estimation in the wild using improved CNN supervision. In *Int. Conf. 3D Vision (3DV)*, pp. 506–516, 2017.

Priyanka Patel and Michael J. Black. CameraHMR: Aligning people with perspective. *arXiv preprint arXiv:2411.08128*, 2024.

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *CVPR*, pp. 10967–10977, 2019.

Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3D human pose estimation in video with temporal convolutions and semi-supervised training. In *CVPR*, pp. 7753–7762, 2019.

Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM TOG*, 36(6):222:1–222:17, 2017.

Shunsuke Saito, Jinlong Yang, Qianli Ma, and Michael J. Black. SCANimate: Weakly supervised learning of skinned clothed avatar networks. In *CVPR*, pp. 2885–2896, 2021.

Xiaolong Shen, Zongxin Yang, Xiaohan Wang, Jianxin Ma, Chang Zhou, and Yi Yang. Global-to-local modeling for video-based 3D human pose and shape estimation. In *CVPR*, pp. 8887–8896, 2023.

Karthik Shetty, Annette Birkhold, Srikrishna Jaganathan, Norbert Strobel, Markus Kowarschik, Andreas Maier, and Bernhard Egger. PLIKS: A pseudo-linear inverse kinematic solver for 3D human body estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 574–584, 2023.

Hong Son Nguyen, DaEun Cheong, Andrew Chalmers, Myoung Gon Kim, Taehyun Rhee, and JungHyun Han. Interaction with virtual objects using human pose and shape estimation. *Computer Animation and Virtual Worlds*, 36(3), 2025.

Hail Song, Boram Yoon, Woojin Cho, and Woontack Woo. RC-SMPL: Real-time cumulative SMPL-based avatar body generation. *IEEE Int. Symp. Mixed and Augmented Reality (ISMAR)*, pp. 89–98, 2023.

Timo von Marcard, Roberto Henschel, Michael J. Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3D human pose in the wild using IMUs and a moving camera. In *ECCV*, 2018.

Wen-Li Wei, Jen-Chun Lin, Tyng-Luh Liu, and Hong-Yuan Mark Liao. Capturing humans in motion: Temporal-attentive 3D human pose and shape estimation from monocular video. In *CVPR*, pp. 13211–13220, 2022.

Jiahong Wu, He Zheng, Bo Zhao, Yixin Li, Baoming Yan, Rui Liang, Wenjia Wang, Shipei Zhou, Guosen Lin, Yanwei Fu, et al. AI Challenger: A large-scale dataset for going deeper in image understanding. *arXiv preprint arXiv:1711.06475*, 2017.

Yan Xia, Xiaowei Zhou, Etienne Vouga, Qixing Huang, and Georgios Pavlakos. Reconstructing humans with a biomechanically accurate skeleton. In *CVPR*, pp. 5355–5365, 2025.

Hongyi Xu, Eduard Gabriel Bazavan, Andrei Zanfir, William T. Freeman, Rahul Sukthankar, and Cristian Sminchisescu. GHUM & GHUML: Generative 3D human shape and articulated pose models. In *CVPR*, pp. 6184–6193, 2020.

Haotian Zhang, Ye Yuan, Viktor Makoviychuk, Yunrong Guo, Sanja Fidler, Xue Bin Peng, and Kayvon Fatahalian. Learning physically simulated tennis skills from broadcast videos. *ACM TOG*, 42:1–14, 2023.

Amy Zhao, Chengcheng Tang, Lezi Wang, Yijing Li, Mihika Dave, Lingling Tao, Christopher D. Twigg, and Robert Y. Wang. Egobody3M: Egocentric body tracking on a VR headset using a diverse dataset. In *ECCV*, pp. 375–392, 2024.

Ce Zheng, Sijie Zhu, Matias Mendieta, Taojiannan Yang, Chen Chen, and Zhengming Ding. 3D human pose estimation with spatial and temporal transformers. In *ICCV*, pp. 11656–11665, 2021.

Wentao Zhu, Xiaoxuan Ma, Zhaoyang Liu, Libin Liu, Wayne Wu, and Yizhou Wang. MotionBERT: A unified perspective on learning human motion representations. In *ICCV*, pp. 15085–15099, 2023.

A Dataset Preparation and Preprocessing

A.1 Training Datasets

BioHMR is trained on Human3.6M (Ionescu et al., 2014), MPI-INF-3DHP (Mehta et al., 2017), COCO (Lin et al., 2014), MPII (Andriluka et al., 2014), AI Challenger (Wu et al., 2017), AVA (Gu et al., 2018), and InstaVariety (Kanazawa et al., 2019), spanning controlled indoor mocap, in-the-wild images, and video. Neither 3DPW (von Marcard et al., 2018) nor the synthetic BEDLAM (Black et al., 2023) dataset is used for training; 3DPW is held out entirely and used only for zero-shot evaluation. We follow the training splits of HMR2.0 (Goel et al., 2023) and HSMR (Xia et al., 2025) for the shared datasets to ensure comparability.

A.2 SMPL-to-SKEL Conversion Pipeline

We rely on the pseudo-ground-truth SMPL (Loper et al., 2015) labels from HMR2.0. Since SKEL annotations are not available for any of these datasets, we convert all SMPL parameters into SKEL pseudo-labels using our learned SMPL-to-SKEL mapping network (see Sec. 3.5), which is roughly $300\times$ faster than the optimization-based conversion proposed in SKEL (Keller et al., 2023) and used in HSMR (Xia et al., 2025). While the mapper occasionally produces suboptimal results, we apply filtering procedures following HMR2.0 and HSMR to remove extreme failure cases.

A.3 Data Preprocessing

Following standard practice in human mesh recovery, we operate on person-centric crops. For each detected person, we use the bounding box to extract a tight crop around the body, which is then resized to the input resolution of the ViT backbone (256×192). The same preprocessing is applied for both image-based and video-based experiments. All additional low-level preprocessing and data augmentation follow the public HSMR (Xia et al., 2025) implementation.

For video datasets, we extract sliding windows of 16 frames with 50% overlap. Joint-token temporal attention and temporal losses are applied only when consecutive frames are available for a given sequence.

B Implementation Details

B.1 Architecture

BioHMR consists of a ViT backbone augmented with learnable joint tokens and BASTA adapters, along with two SKEL parameter decoders.

Backbone. The backbone is a ViT-H (Dosovitskiy et al., 2021) operating on 256×192 person-centric crops with patch size of 16×16. After patch embedding, each frame has a sequence of $L = H_p \times W_p$ tokens in a $C = 1280$-dimensional space. We use 32 transformer blocks with 16 attention heads and an MLP ratio of 4. Positional embeddings are learned and added to all tokens.

Learnable Joint Tokens. BioHMR introduces a set of learnable joint tokens $\mathbf{J} \in \mathbb{R}^{J \times C_a}$, where $J = 24$ corresponds to the SMPL/SKEL joints and $C_a = 384$ is the adapter channel dimension. This single shared set is broadcast to every frame, and each BASTA module initializes its tokens

from it and refines them independently, so all blocks operate on a common skeletal representation rather than separate per-layer token banks. Their purpose is to maintain an explicit skeletal latent space aligned with the SKEL kinematic structure.

B.2 BASTA ADAPTER

Each transformer block is augmented with a BASTA module (after the MLP) that couples the patch tokens with the joint tokens. Its computation consists of four steps:

Patch-to-Joint Aggregation. Let $\mathbf{X} \in \mathbb{R}^{(BT)\times L\times C}$ denote the patch tokens for a batch of B sequences of length T, and let $\mathbf{X}_a = \mathbf{X}\mathbf{W}_{\text{down}} \in \mathbb{R}^{(BT)\times L\times C_a}$ be their linear projection into adapter space. Joint tokens $\mathbf{J}$ are expanded to $\mathbf{J}_{\text{bt}} \in \mathbb{R}^{(BT)\times J\times C_a}$ and query the patch tokens via multi-head cross-attention:

$$\tilde{\mathbf{J}} = \text{Attn}_{\text{patch}\to\text{joint}}(\mathbf{Q} = \mathbf{J}_{\text{bt}}, \mathbf{K} = \mathbf{X}_a, \mathbf{V} = \mathbf{X}_a), \tag{2}$$

resulting in refined joint features $\tilde{\mathbf{J}} \in \mathbb{R}^{(BT)\times J\times C_a}$.

Spatial Graph Reasoning. We apply $L_s = 3$ spatial graph-attention layers via GATv2 (Brody et al., 2022) to the joint tokens per frame using the SKEL kinematic graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, with $|\mathcal{V}| = 24$ joints and edges $\mathcal{E}$ from the kinematic tree:

$$\tilde{\mathbf{J}}_{b,t,:,:} = \text{GATv2}\left(\tilde{\mathbf{J}}_{b,t,:,:}, \mathcal{E}\right), \forall b \in [1, B], t \in [1, T]. \tag{3}$$

Each joint token receives message-passing updates from its neighbors in the skeletal tree.

Temporal Joint Attention. For video ($T > 1$), after spatial refinement, temporal dynamics across frames are modeled independently for each joint. The joint tokens form a sequence $\tilde{\mathbf{J}}_{:,:,j,:} \in \mathbb{R}^{B\times T\times C_a}$ and are processed by a temporal transformer with $L_t = 3$ layers. For images ($T = 1$), this step is an identity.

Joint-to-Patch Propagation. The refined joint tokens are used to update patch tokens so that biomechanically structured and temporally coherent information propagates back to the ViT stream:

$$\tilde{\mathbf{X}}_a = \text{Attn}\big(\mathbf{Q} = \mathbf{X}_a, \mathbf{K} = \tilde{\mathbf{J}}, \mathbf{V} = \tilde{\mathbf{J}}\big), \tag{4}$$

followed by up-projection to $\mathbb{R}^C$ and a gated residual connection.

B.3 SKEL DECODING HEADS

Main Decoder. We apply a transformer decoder (6 layers, 8 heads) with a single learnable query token, following HMR2.0 (Goel et al., 2023) and HSMR (Xia et al., 2025). The output vector is mapped to:

- 6D joint rotations for 24 joints (converted to the Euler-angle format required by SKEL's 46-DoF parameterization),
- 10 shape coefficients,
- 3 weak-perspective camera parameters.

Auxiliary Joint-Token Decoder. A second decoder operates on the final joint tokens. A learned query performs attention pooling over J joints, followed by an MLP predicting pose and shape parameters directly against the pseudo-ground-truth targets. This head is used only during training.

B.4 SMPL-TO-SKEL MAPPING NETWORK

The mapping function $\mathcal{M} : \mathbb{R}^{157} \to \mathbb{R}^{147}$ is a compact MLP with residual connections and layer normalization. The input consists of SMPL pose in 6D rotation representation ($144D$), shape parameters $\boldsymbol{\beta} \in \mathbb{R}^{10}$, and global translation $\mathbf{t} \in \mathbb{R}^3$. The network predicts the corresponding SKEL pose in 6D form ($144D$) together with translation ($3D$). The mapper is trained on SMPL sequences from AMASS (Mahmood et al., 2019) without requiring optimization-derived SKEL ground truth (see Sec. 3.5). It is frozen during BioHMR training.

Training objective. Given SMPL parameters $(\boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{t})$, we jointly predict the corresponding SKEL pose and a refined translation, $(\hat{\mathbf{q}}, \hat{\mathbf{t}}) = \mathcal{M}(\boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{t})$, and generate meshes using the differentiable forward functions of the two models:

$$\mathbf{M}_{\text{SMPL}} = \mathcal{S}_{\text{SMPL}}(\boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{t}), \qquad \hat{\mathbf{M}}_{\text{SKEL}} = \mathcal{S}_{\text{SKEL}}(\hat{\mathbf{q}}, \boldsymbol{\beta}, \hat{\mathbf{t}}). \tag{5}$$

We supervise the mapping with a combination of vertex and joint alignments along with a Laplacian smoothness regularizer:

$$\mathcal{L}_{\text{map}} = \lambda_v \, \mathcal{L}_{\text{verts}} + \lambda_j \, \mathcal{L}_{\text{joints}} + \lambda_s \, \mathcal{L}_{\text{lap}}, \tag{6}$$

where the vertex loss measures per-vertex alignment between the two meshes, $\mathcal{L}_{\text{verts}} = \|\hat{\mathbf{M}}_{\text{SKEL}} - \mathbf{M}_{\text{SMPL}}\|_2^2$, and the joint loss the corresponding joint alignment, $\mathcal{L}_{\text{joints}} = \|\hat{\mathbf{J}}_{\text{SKEL}} - \mathbf{J}_{\text{SMPL}}\|_2^2$, with $\hat{\mathbf{J}}_{\text{SKEL}}$ and $\mathbf{J}_{\text{SMPL}}$ the joints obtained via the respective body-model regressors. To avoid the conversion failure modes observed in Keller et al. (2023), the Laplacian surface regularizer penalizes deformation relative to the reference SMPL mesh,

$$\mathcal{L}_{\text{lap}} = \|\mathbf{L}\,(\hat{\mathbf{M}}_{\text{SKEL}} - \mathbf{M}_{\text{SMPL}})\|_F^2, \tag{7}$$

where $\mathbf{L}$ is the normalized mesh Laplacian; this suppresses high-frequency surface artifacts while preserving the anatomical geometry captured by the reference mesh. Since SMPL and SKEL differ systematically in certain anatomical regions (*e.g.*, upper limbs), we apply selective joint masking to down-weight inconsistent areas while retaining dense vertex supervision.

Offline geometric refinement. After the mapper, we run a lightweight L-BFGS optimization over the SKEL parameters to improve geometric mesh alignment with the SMPL source, using the same mesh-consistency objective as the mapper training. Because the mapper provides a near-optimal initialization, this refinement converges in a handful of iterations rather than the hundreds needed from scratch. It is applied to the full training set during pre-processing and does not use 2D image evidence; the combined mapper + refinement pipeline runs in approximately 0.58 s per frame.

Online SKELify refinement. SKELify (Xia et al., 2025) adapts SMPLify (Bogo et al., 2016) to SKEL parameters, minimizing a 2D reprojection objective with respect to detected keypoints while enforcing SKEL's physiological joint-limit constraints as penalty terms. During BioHMR training, it is applied periodically to a small random subset of samples, using the pre-generated pseudo-labels as warm-start initialization; compared to HSMR, we halve the number of SKELify iterations, reducing online overhead while retaining its supervision benefit.

Figure 3 illustrates the resulting conversion quality: offline geometric refinement corrects the residual wrist, knee, and foot misalignments of the raw mapper, achieving tighter alignment than direct optimization at $300\times$ lower cost.

B.5 Training

BioHMR is trained following protocols closely aligned with HMR2.0 (Goel et al., 2023) and HSMR (Xia et al., 2025). We use the AdamW optimizer with a base learning rate of 1×10^{-5}, weight decay of 1×10^{-4}, and default β values $(0.9,\ 0.999)$. Training uses mixed precision (FP16).

Training is performed on 7 NVIDIA A100 GPUs. Each GPU processes a batch size of 128 image crops (or 8 sequences of 16 frames for video). This results in an effective batch size of 896 images or 56 video windows per iteration. The sequence length is fixed at 16 frames with a stride of 8.

We train for a total of 160k iterations, with SKELify refinement applied every 230 iterations to improve supervision quality. SKELify performs L-BFGS optimization over camera, orientation, pose, and shape parameters for up to 10 inner-loop steps.

Loss weights: $\lambda_{3\text{D}} = 0.05$ for 3D keypoints, $\lambda_{2\text{D}} = 0.01$ for 2D keypoints, $\lambda_{\text{orient}} = 0.002$ for global orientation, $\lambda_q = 0.001$ for body pose, $\lambda_\beta = 0.0005$ for shape parameters, $\lambda_{\text{temp}} = 0.001$ for temporal consistency, and $\lambda_{\text{aux}} = 0.001$ for auxiliary joint-token supervision.

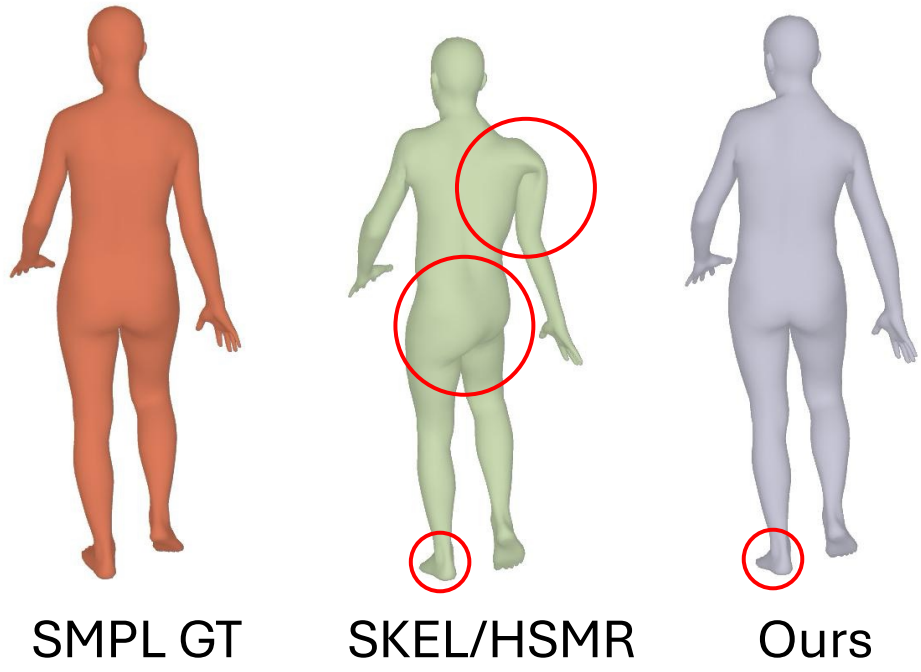


Figure 3: **SMPL-to-SKEL conversion quality.** Mesh alignment for the SMPL source (left), raw mapper output (center), and mapper + offline refinement (right). Our pipeline achieves tighter alignment than direct optimization at 300× lower cost.

Table 6: **Frequency of unnatural rotations on MOYO** for all methods at three violation thresholds (%).

	violation $> 10°$ ↓				violation $> 20°$ ↓				violation $> 30°$ ↓			
Methods	left elb.	right elb.	left knee	right knee	left elb.	right elb.	left knee	right knee	left elb.	right elb.	left knee	right knee
PARE	36.4%	42.4%	20.0%	23.2%	14.6%	15.4%	3.2%	3.8%	5.5%	4.8%	0.3%	0.4%
CLIFF	34.2%	33.0%	28.3%	31.0%	13.0%	12.4%	4.8%	4.5%	5.2%	5.2%	0.5%	0.3%
HybrIK	58.7%	60.9%	52.9%	48.6%	29.4%	34.6%	30.7%	27.0%	16.4%	21.0%	20.0%	17.5%
PLIKS	41.6%	44.7%	47.4%	43.8%	17.9%	22.7%	18.2%	17.6%	8.3%	11.4%	8.5%	8.5%
HMR2.0b	47.6%	44.3%	45.7%	56.4%	19.8%	19.6%	6.4%	11.6%	8.5%	8.8%	1.0%	1.6%
HSMR	**0.0%**	**0.0%**	3.9%	4.5%	**0.0%**	**0.0%**	0.2%	0.5%	**0.0%**	**0.0%**	**0.0%**	**0.0%**
BioHMR	**0.0%**	**0.0%**	**2.1%**	**4.2%**	**0.0%**	**0.0%**	**0.0%**	**0.2%**	**0.0%**	**0.0%**	**0.0%**	**0.0%**

C Extended Quantitative Results

C.1 Full Biomechanical Validity

Table 6 reports the per-joint violation rates on MOYO at the $> 10°$, $> 20°$, and $> 30°$ thresholds for all methods; the main text (Table 2) shows the $> 10°$ and $> 20°$ thresholds, and Table 3 the strict $> 0°/> 5°$ comparison.

C.2 Ablation Studies

Table 7 reports the contribution of each BioHMR component on COCO (2D) and 3DPW (3D); the same trends hold on the other benchmarks. Removing the BASTA modules (eliminating both graph attention over the kinematic tree and joint-token cross-attention) reduces COCO PCK@0.05 from 0.84 to 0.82 and increases 3DPW MPJPE by 5.2 mm, confirming that structured skeletal reasoning contributes meaningfully beyond the ViT's standard self-attention. Disabling per-joint temporal attention increases 3DPW MPJPE by 5.6 mm and substantially increases acceleration error on MOYO (ACCEL 7.86 → 10.83 and V-ACCEL 9.19 → 12.61 mm/s^2, roughly a 38% increase in both), confirming that temporal attention is the primary driver of smooth, physically consistent predic-

Table 7: **Component ablation.** COCO reports PCK@0.05↑; 3DPW reports MPJPE↓ (mm). Each row removes one component from the full model.

Variant	COCO	3DPW
w/o BASTA adapters	0.82	70.5
w/o temporal attention	0.84	70.9
w/o auxiliary decoder	0.83	66.4
w/o SKELify refinement	0.81	68.2
BioHMR (full)	**0.84**	**65.3**

tions. Removing the training-only auxiliary joint-token decoder modestly degrades accuracy across all benchmarks, most noticeably on 3D pose metrics. Finally, removing online SKELify refinement degrades 3DPW MPJPE from 65.3 to 68.2 mm: the learned mapper's high-quality initializations make training without refinement feasible, while SKELify's periodic L-BFGS steps (applied to a random subset every 230 iterations) correct residual pose errors the mapper cannot resolve from SMPL parameters alone.

C.3 Inference Efficiency: Patch Selection Trade-off

Scoring and masking procedure. Let $\mathcal{A}_t \in \mathbb{R}^{J\times P}$ denote the joint-to-patch cross-attention matrix at frame t, where J is the number of joints and P the number of patch tokens; each entry $\mathcal{A}_t^{(j,p)}$ measures the relevance of patch p to joint j. We compute a single importance score per patch,

$$s_p = \alpha_{\max} \max_j \mathcal{A}_t^{(j,p)} \; + \; \alpha_{\text{mean}} \frac{1}{J} \sum_j \mathcal{A}_t^{(j,p)}, \tag{8}$$

where $\alpha_{\max} = 0.7$ and $\alpha_{\text{mean}} = 0.3$ are fixed hyperparameters; this favors patches critical for at least one joint while accounting for global relevance to the body. Starting from a predefined depth, we maintain a binary mask $m_p^{(\ell)} \in \{0, 1\}$ indicating whether patch p remains active in layer ℓ: at each such layer we rank patches by s_p and keep the top-K_ℓ tokens per image, subject to a minimum budget $K_{\min}$. Inactive patches ($m_p^{(\ell)} = 0$) have their feature vectors stored and frozen, receiving no further updates from subsequent blocks, while active patches receive full attention and MLP updates; at the end of the backbone, frozen features are reinserted at their original spatial locations so the decoder operates on a dense feature map.

Trade-off. Figure 4 characterizes the compute-accuracy trade-off of pose-aware patch selection, with pruning applied from transformer block 18 onward. Under moderate pruning, FLOPs drop to 108 GFLOPs (20% reduction) with COCO PCK@0.10 decreasing only from 0.96 to 0.95. Under aggressive pruning (27% reduction, 98.7 GFLOPs), COCO PCK@0.10 drops to 0.94. The strict PCK@0.05 metric is more affected (0.85→0.72), as expected for a fine-grained threshold.

D Qualitative Analyses

D.1 Additional Qualitative Results

Figure 5 presents additional BioHMR reconstructions on diverse in-the-wild images, rendered from frontal, side, and top-down views to illustrate the quality of 3D geometry recovery.

D.2 Comparative Visualization

Figure 6 compares HMR2.0 (Goel et al., 2023), HSMR (Xia et al., 2025), and BioHMR across challenging in-the-wild scenes. BioHMR consistently produces more physiologically valid articulation, particularly under extreme bending and high-dynamic motion.

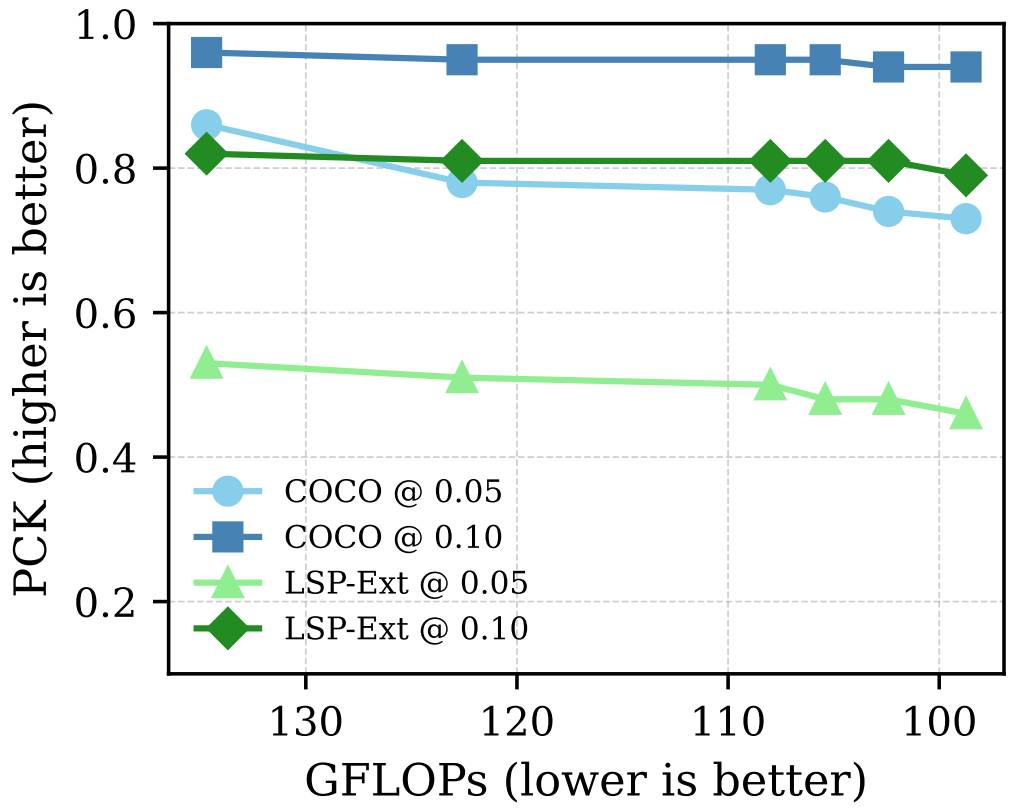


Figure 4: **Compute-accuracy trade-off of pose-aware patch pruning.** FLOPs vs. PCK under progressive pruning. Up to 27% FLOPs reduction incurs only modest accuracy degradation.

D.3 Joint-Token Attention Maps

Figure 7 visualizes the cross-attention maps from the patch-to-joint aggregation step. For each example, we show the input frame, the predicted skeleton, and the aggregated attention heatmap overlaid on the image. Attention maps are computed by averaging attention weights from all joint tokens at the final BASTA block. The visualizations confirm that joint tokens consistently attend to body-relevant image regions and ignore irrelevant background.

D.4 Robustness to Viewpoint and Occlusion

Figure 8 renders predictions from multiple viewpoints to assess the self-consistency of the predicted 3D geometry. BioHMR maintains coherent body geometry across views, demonstrating reliable 3D reasoning even under large viewpoint rotations.

D.5 Failure Case Analysis

Although BioHMR significantly reduces biomechanically invalid rotations, several challenging scenarios produce imperfect reconstructions (Figure 9).

Extreme articulation and inversion. In highly inverted or fast-moving poses, foreshortening and limited visual cues make it difficult to recover precise axial twist. BioHMR maintains valid joint limits but may misestimate small rotations around the spine or shoulders.

Heavy self-occlusion. When a limb is fully hidden behind the body, the model sometimes infers incorrect depth ordering or slight misplacement.



Figure 5: **Additional qualitative results.** For each input image (left), we show the reconstructed SKEL mesh and skeleton from frontal, side, and top viewpoints.

Loose clothing and weak boundaries. Loose-fitting clothing or low-texture backgrounds can obscure joint boundaries, leading to mild inaccuracies in limb orientation even when the global pose is correct.

These remaining errors stem from inherent monocular ambiguities rather than violations of anatomical constraints; BioHMR preserves valid joint angles and kinematic structure even in these cases.

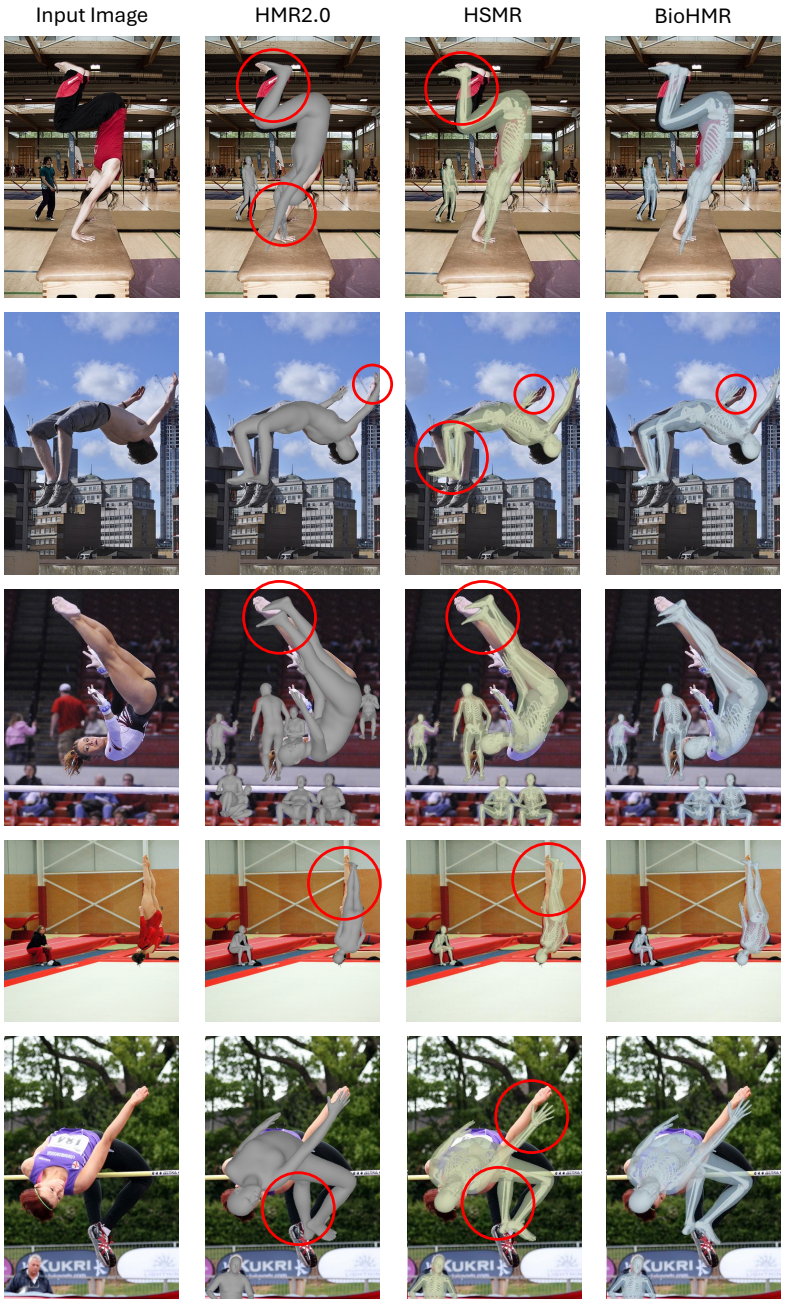


Figure 6: **Qualitative comparison.** Side-by-side comparisons of HMR2.0, HSMR, and BioHMR on challenging in-the-wild scenes. Red circles highlight errors in prior methods. BioHMR produces more physiologically valid articulation, particularly for extreme poses.

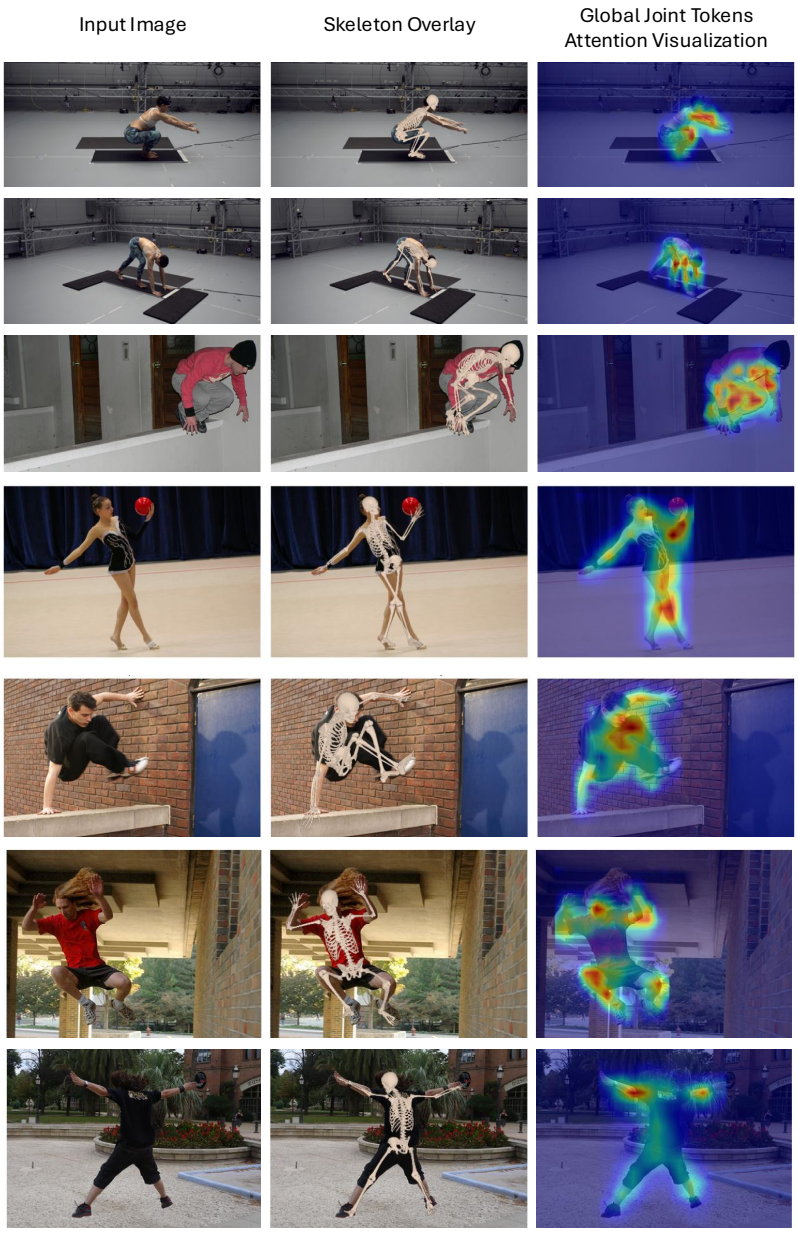


Figure 7: **Joint-token attention visualization.** For each input image, we show (left) the input, (middle) the predicted SKEL skeleton, and (right) the aggregated patch-to-joint attention map. Joint tokens attend to limb endpoints, hinge regions, and areas of meaningful articulation.

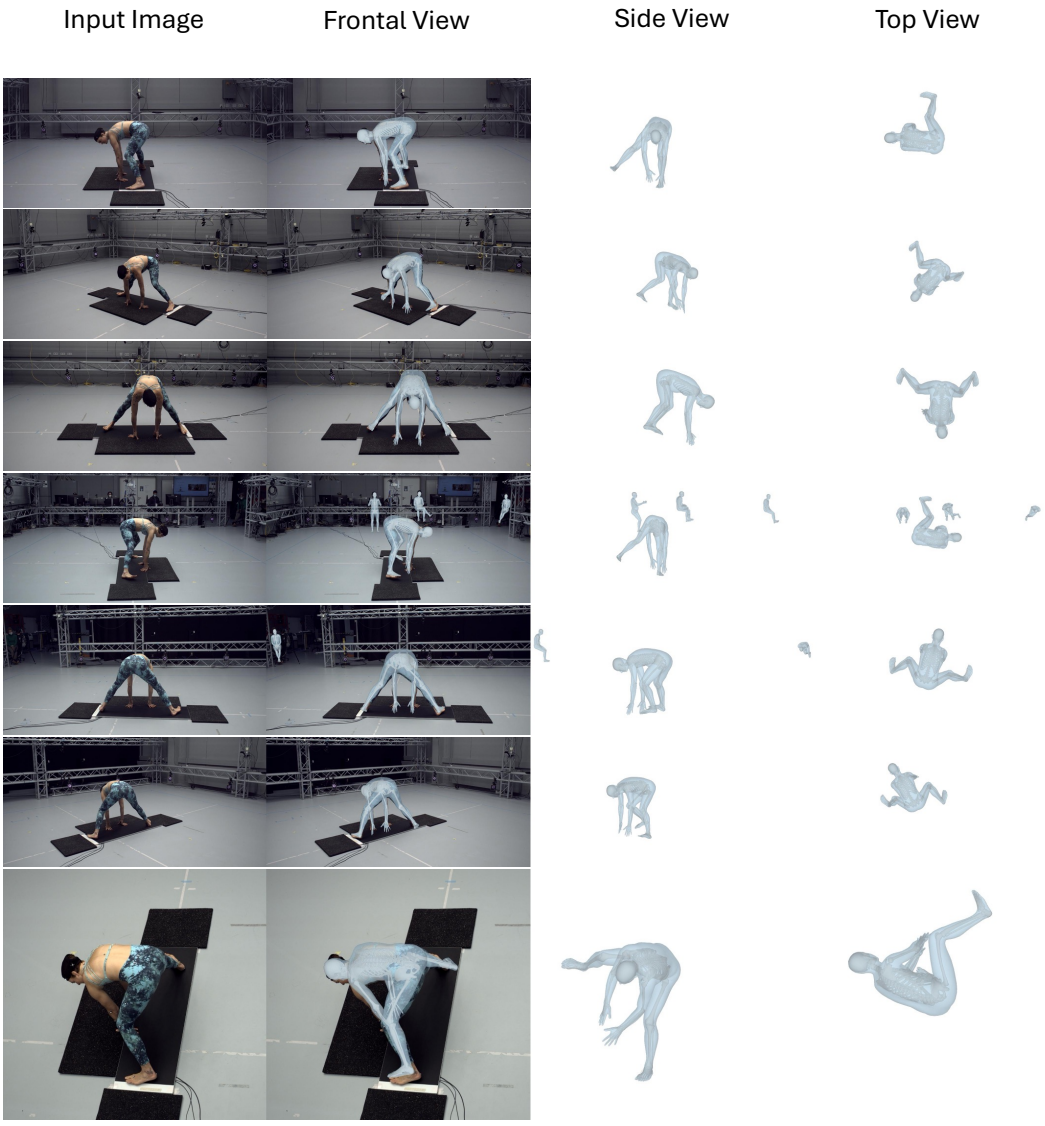


Figure 8: **Multi-view visualization.** For each input image (left), we show the reconstructed SKEL mesh from frontal, side, and top viewpoints. Predicted geometry remains stable under large viewpoint rotations, demonstrating strong 3D consistency.



Figure 9: **Failure cases.** While reconstructions remain biomechanically valid, errors arise in rare scenarios involving inversion, extreme contortion, and heavy self-occlusion.