# 1. Appendix

## 1.1. Overview

The appendix is organized into the following sections:

- Section 1.2: Implementation Details
- Section 1.3: Datasets
- Section 1.4: Evaluation Metrics
- Section 1.5: Data Augmentation
- Section 1.6: Camera Model
- Section 1.7: Impact of Backbones
- Section 1.8: Impact of Deformable Cross-Attention Layers
- Section 1.9: Effect of losses on NeurIK model
- Section 1.10: Multi-frame out vs Single frame out
- Section 1.11: Impact of Number of Frames in NeuralIK
- Section 1.12: Qualitative Results

## 1.2. Implementation Details

**MQ-HMR**. Our MQ-HMR model is implemented in PyTorch. To achieve this, we utilized multi-resolution feature maps at 4×, 8×, and 16× scales, ensuring the model captures both local detail and global structure. The total loss function in MQ-HMR is defined as $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{SMPL}} + \mathcal{L}_{\text{3D}} + \mathcal{L}_{\text{2D}}$. This combines 3D loss ($\mathcal{L}_{\text{3D}}$), 2D loss ($\mathcal{L}_{\text{2D}}$), and SMPL parameter loss ($\mathcal{L}_{\text{SMPL}}$) to optimize the shape ($\beta$) and pose ($\theta$) parameters in the SMPL space. The loss weights for this stage are carefully tuned to balance each objective, where the pose component is weighted at $\lambda_\theta = 1 \times 10^{-3}$ and the shape component at $\lambda_\beta = 5 \times 10^{-4}$. For the 3D loss, $\lambda_{\text{3D}} = 5 \times 10^{-2}$, and for the 2D loss, $\lambda_{\text{2D}} = 1 \times 10^{-2}$. The architecture incorporates 96 Pose Token Queries ($Q$) and 4 deformable cross-attention layers as default, which enables the model to attend to relevant spatial information across different scales. The model was trained for 100K iterations using the Adam optimizer with a batch size of 48 and a learning rate of $1 \times 10^{-5}$.

**NeurIK**. Our NeurIK module is implemented in PyTorch and processes virtual markers $\mathbf{X}_{\text{VM}}^{\text{exp}} \in \mathbb{R}^{142\times3}$ extracted from the 3D mesh produced by MQ-HMR. The architecture includes a Spatial Convolution Encoder that utilizes 1D convolutional layers for spatial feature extraction and a Temporal Transformer Encoder that employs multi-head self-attention to model temporal dependencies across multiple frames. The total loss function is defined as $\mathcal{L}_{\text{neurIK}} = \lambda_j\mathcal{L}_{\text{j}} + \lambda_m\mathcal{L}_{\text{m}} + \lambda_s\mathcal{L}_{\text{s}} + \lambda_q\mathcal{L}_{\text{q}}$, with the weights set as $\lambda_j = 1.0$, $\lambda_m = 2.0$, $\lambda_s = 0.1$, and $\lambda_q = 0.06$. The model was trained for 25 epochs using the Adam optimizer with an initial learning rate of 0.001, decaying to $5 \times 10^{-6}$, and a batch size of 128. Data augmentation techniques such as scaling, rotation, translation, and noise injection were applied to increase the model's robustness to occlusions and real-world variations.

To generate the input for our network, the SMPL mesh for each video frame was recovered using a test-time optimized HMR 2.0 model. The vertex indices of virtual markers on the SMPL mesh were used to calculate marker locations, which served as critical inputs for the subsequent spatio-temporal modeling. During training, we employed data augmentation methods such as scaling, rotation, translation, and noise to mimic occlusions, following the approach used in [1]. Our model was trained using the following hyperparameters and loss functions. We used an Adam optimizer with a weight decay of 0.001 and a batch size of 128. The learning rate decayed exponentially from an initial rate of 0.001 to a final rate of $5 \times 10^{-6}$ over 25 epochs. For the spatio-temporal model, we set the hyperparameters experimentally, adjusting key parameters as needed throughout the training process.

## 1.3. Datasets

We utilized videos from BMLmovi [4], OpenCap [7] and BEDLAM [2]datasets. We trained our NeurIK model on BMLmovi data and tested on OpenCap and BEDLAM datasets.

**BML-MoVi:** BMLMovi consists of 90 subjects performing 21 different actions, captured using two cameras and a marker-based motion capture system. As BMLMovi lacks kinematic annotations, we processed the available marker data through the OpenSim Scale and OpenSim IK tools to accurately determine joint angles and body segment dimensions for ground truth measurements. [4].

**OpenCap:** OpenCap includes data from ten subjects performing various actions such as walking, squatting, standing up from a chair, drop jumps, and their asymmetric variations. The recordings were made using five RGB cameras alongside a marker-based motion capture system. Additionally, OpenCap offers processed marker data and kinematic annotations for a comprehensive full-body OpenSim skeletal model.

**BEDLAM:** BEDLAM dataset comprises synthetic video data featuring a total of 271 subjects, including 109 men and 162 women. It includes monocular RGB videos paired with ground-truth 3D bodies in SMPL-X format, covering a diverse range of body shapes, skin tones, and motions. The dataset features realistic animations with detailed hair and clothing simulated using physics, enhancing the realism of the data. To obtain kinematic annotations, we leverage the vertices of SMPL-X mesh to generate virtual markers and processed the available marker data through the OpenSim Scale and OpenSim IK tools to determine joint angles and body segment dimensions as the ground-truth data. [2].

### 1.4. Evaluation Metrics

For NeurIK, we employ metrics that focus on biomechanical accuracy. These include Mean Per Bony Landmarks Position Error (MPBLPE), which measures the accuracy of predicted bony landmarks against ground truth positions, and Mean Absolute Error for body scale ($\text{MAE}_{body}$), which evaluates the correctness of predicted body segment dimensions by comparing their longest axes in millimeters. Additionally, we report Mean Absolute Error for joint angles ($\text{MAE}_{angle}$), which assesses the precision of joint angle predictions in degrees, critical for biomechanical simulations such as joint force and muscle force analysis.

**MPBLPE:** Mean Per Bony Landmarks Position Error, measures the accuracy of predicted bony landmark positions by comparing them to ground truth data. This metric is inspired by the mean per joint position error (MPJPE) which is often used in 3D pose estimation. MPBLPE involves aligning both the predicted and actual positions at a common root point and then calculating the average Euclidean distance between corresponding landmarks, i.e.,

$$\text{MPBLPE}(\text{pred}, \text{target}) = \frac{1}{N} \sum_{i=1}^{N} \sqrt{\sum_{j=1}^{3} (\text{pred}_{ij} - \text{target}_{ij})^2}$$

where $i$ indexes the joints, and $j$ indexes the spatial dimensions (X, Y, Z). For each joint $i$, $\text{pred}_{ij}$ and $\text{target}_{ij}$ represent the predicted and target positions in dimension $j$, respectively. The entire expression is divided by $N$, the number of bony landmarks, to calculate the average Euclidean distance per landmark, which normalizes the loss to account for differences in the number of joints among different datasets or models.

**$\text{MAE}_{body}$:** The axis corresponding to the longest dimension of each body segment is selected, and its scale is converted into millimeters to calculate the Mean Absolute Error ($\text{MAE}_{body}$). Specifically, the x-axis is used for the skull, toes, and calcaneus; the y-axis is chosen for the spine, lower limbs, and upper limbs; the z-axis is applied for the jaw, scapula, and clavicle; and for the pelvis, all three axes are considered [1].

**$\text{MAE}_{angle}$:** $\text{MAE}_{angle}$ represents the mean absolute error of the joint angle, measured in degrees [1, 7].

In this study, $\text{MAE}_{angle}$ is prioritized over MPBLPE because joint angles, rather than marker positions, will be utilized in subsequent applications of musculoskeletal models, such as simulating joint reaction forces and analyzing muscle forces. Consequently, ensuring the accuracy of joint angles is crucial for producing precise force simulations, making it more significant than the accuracy of marker positions.

### 1.5. Data Augmentation

To enhance MQ-HMR's robustness and generalization, we applied extensive data augmentation during training. This included random scaling, rotation, horizontal flips, and color jittering on both images and poses. These augmentations help the model better handle real-world challenges like occlusions and incomplete body information. As a result, the data augmentation process significantly contributes to MQ-HMR's improved performance in human mesh reconstruction by making the model more adaptable and resilient to diverse and unpredictable inputs.

### 1.6. Camera Model

In our approach, MQ-HMR utilizes a weak perspective camera model with a fixed focal length 5000 and an intrinsic matrix $K \in \mathbb{R}^{3\times3}$. To simplify the computation, the rotation matrix $R$ is set to the identity matrix $I_3$, allowing us to focus solely on the translation vector $T \in \mathbb{R}^3$. The 3D joints $J_{3\text{D}}$ are then projected onto 2D coordinates $J_{2\text{D}}$ using the equation $J_{2\text{D}} = \Pi(K(J_{3\text{D}} + T))$, where $\Pi$ denotes the perspective projection based on camera intrinsics $K$. This simplification reduces the number of parameters involved, improving the computational efficiency of our human mesh recovery process.

### 1.7. Impact of Backbones

The comparison of backbones in MQ-HMR highlights that the ViT-H backbone consistently delivers superior performance in terms of MPJPE on both the 3DPW and EMDB datasets (Table 1). Specifically, ViT-H achieves the lowest MPJPE, with 69.0 mm on 3DPW and 92.5 mm on EMDB, indicating its strength in accurate pose estimation. While HRNet-w48 shows a slight advantage in MVE on 3DPW, with 79.5 mm compared to ViT-H's 79.8 mm, this minor improvement in mesh vertex error does not offset its higher MPJPE. On the more complex EMDB dataset, ViT-H maintains a clear edge in both MPJPE and MVE, further solidifying its effectiveness in capturing detailed pose structures. Thus, ViT-H is the more robust backbone for 3D human pose estimation, especially when precision across both metrics is essential.

Table 1. Impact of Backbones on MQ-HMR

| | 3DPW | | EMDB | |
|---|---|---|---|---|
| | MPJPE (↓) | MVE (↓) | MPJPE (↓) | MVE (↓) |
| ViT-H 96 | **69.0** | 79.8 | **92.5** | **98.9** |
| HRNet-w48 | 70.3 | **79.5** | 93.5 | 100.5 |

### 1.8. Impact of Deformable Cross-Attention Layers

The impact of deformable cross-attention layers in MQ-HMR reveals that the optimal number of layers for accu-

rate 3D pose estimation is 4 (Table 2). With 4 layers, the model achieves the lowest MPJPE on both the 3DPW (69.0 mm) and EMDB (92.5 mm) datasets, capturing pose information effectively across different complexities. Increasing the number of layers to 6 offers a slight improvement in mesh vertex error (MVE), particularly reducing it to 78.9 mm on 3DPW, but at the cost of a higher MPJPE. However, further increasing the number of layers to 8 results in diminished performance, as seen with the 72.4 mm MPJPE for 3DPW, suggesting that too many layers may introduce redundancy and inefficiencies. Additionally, the higher number of cross-attention layers increases the computational burden without significant accuracy gains. Therefore, 4 deformable cross attention layers provide the best balance between computational efficiency and performance, avoiding the computationally heavy overhead of higher-layer configurations.

Table 2. Impact of # of Deformable Cross Attention Layers in MQ-HMR.

| | 3DPW | | EMDB | |
|---|---|---|---|---|
| # of Deformable Cross Attention Layers | MPJPE | MVE | MPJPE | MVE |
| 2 | 75.2 | 87.1 | 98.1 | 105.6 |
| 4 | **69.0** | 79.8 | **92.5** | **98.9** |
| 6 | 70.4 | **78.9** | 93.1 | 103.1 |
| 8 | 72.4 | 84.1 | 93.5 | 102.5 |

## 1.9. Effect of Losses on NeurIK Model

We assess the effect of various loss terms as presented in Table 3, by progressively introducing each loss during the model training process. Initially, the model is trained using only $L_{\text{m}}$ (marker loss), followed by the sequential addition of $L_{\text{j}}$ (joint loss), $L_{\text{q}}$ (angle loss), and finally $L_{\text{s}}$ (body scale loss). The results show that training with only $L_{\text{m}}$ achieves the lowest MPBLPE (21.57), indicating good alignment of bony landmarks. However, this configuration yields relatively high errors in body scale ($\text{MAE}_{\text{body}}$ = 8.46) and joint angles ($\text{MAE}_{\text{angle}}$ = 7.58), highlighting its limitations in capturing accurate body dimensions and angles. Introducing $L_{\text{j}}$ into the training improves both the body and angle predictions, reducing $\text{MAE}_{\text{body}}$ to 7.18 and $\text{MAE}_{\text{angle}}$ to 6.43, although MPBLPE slightly increases to 22.26. Further, the inclusion of $L_{\text{q}}$ significantly improves joint angle accuracy, achieving the lowest $\text{MAE}_{\text{angle}}$ (2.34), though this comes at the cost of a slight increase in body scale error ($\text{MAE}_{\text{body}}$ = 6.21). Finally, adding $L_{\text{s}}$ results in the best body scale accuracy, with $\text{MAE}_{\text{body}}$ reduced to 3.97, though the joint angle error slightly increases to $\text{MAE}_{\text{angle}}$ = 2.84. These findings demonstrate that while each loss term optimizes different aspects of the model's performance, utilizing all four losses together achieves a balanced improvement across both body scales and joint angles.

Table 3. Effect of losses on NeurIK model

| | $L_{\text{m}}$ | $L_{\text{j}}$ | $L_{\text{q}}$ | $L_{\text{s}}$ |
|---|---|---|---|---|
| $L_{\text{m}}$ | ✓ | ✓ | ✓ | ✓ |
| $L_{\text{j}}$ | | ✓ | ✓ | ✓ |
| $L_{\text{q}}$ | | | ✓ | ✓ |
| $L_{\text{s}}$ | | | | ✓ |
| MPBLPE | **21.57** | 22.26 | 24.08 | 25.76 |
| $\text{MAE}_{\text{body}}$ | 8.46 | 7.18 | 6.21 | **3.97** |
| $\text{MAE}_{\text{angle}}$ | 7.58 | 6.43 | **2.34** | 2.84 |

## 1.10. Multi-frame out vs Single frame out

Table 4 compares the performance of NeurIK using two different temporal models: multiple frame out and single last frame out, across three datasets: BML-MoVi, BEDLAM, and OpenCap. In the multi-frame temporal model, instead of predicting just the last frame in a sequence, the model predicts all 64 frames. The results show that the single last frame out model consistently outperforms the multiple frame out model across all metrics and datasets. For instance, in BML-MoVi, the single frame out model achieves lower $\text{MAE}_{body}$ (3.97 vs. 4.01) and $\text{MAE}_{angle}$ (2.84 vs. 2.95). Similar improvements are seen in BEDLAM and OpenCap, suggesting that the single frame out model offers better accuracy, making it the preferred choice for NeurIK. The single frame out model likely outperforms the multiple frame out model for several reasons. First, by focusing solely on predicting the final frame, the model can concentrate its capacity on optimizing that specific output, resulting in more precise predictions. In contrast, the multiple frame out model must predict the entire sequence, which can introduce cumulative error across frames. Additionally, predicting all frames may create temporal inconsistencies, as the model tries to maintain coherence throughout the sequence. The single frame out model avoids these challenges, offering a simplified learning objective that reduces complexity and leads to better performance, especially in terms of MAE. This makes it a more efficient choice for tasks like NeurIK, where accuracy in the final frame is critical.

## 1.11. Impact of Number of Frames in NeuralIK

Table 5 illustrates the impact of varying the number of frames used in the temporal model on the performance of NeurIK across three datasets: BML-MoVi, BEDLAM, and OpenCap. The results show that the number of frames significantly affects the accuracy of the model, with optimal performance generally observed at 64 frames across all datasets. For the BML-MoVi dataset, increasing the

Table 4. Impact of different temporal model on NeurIK

| Temporal model | BML-MoVi | | | BEDLAM | | | OpenCap | | |
|---|---|---|---|---|---|---|---|---|---|
| | $MAE_{body}$ | MPBLPE | $MAE_{angle}$ | $MAE_{body}$ | MPBLPE | $MAE_{angle}$ | $MAE_{body}$ | MPBLPE | $MAE_{angle}$ |
| Multiple frame out | 4.01 | 25.84 | 2.95 | 4.41 | 26.61 | 3.32 | 5.12 | 26.72 | 3.44 |
| Single frame out | **3.97** | **25.76** | **2.84** | **4.28** | **26.54** | **3.14** | **4.87** | **26.34** | **3.19** |

Table 5. Impact of number of frames in temporal model on NeurIK

| # of Frames | BML-MoVi | | | BEDLAM | | | OpenCap | | |
|---|---|---|---|---|---|---|---|---|---|
| | $MAE_{body}$ | MPBLPE | $MAE_{angle}$ | $MAE_{body}$ | MPBLPE | $MAE_{angle}$ | $MAE_{body}$ | MPBLPE | $MAE_{angle}$ |
| 16 | 5.08 | 26.83 | 4.12 | 5.76 | 27.62 | 4.35 | 5.98 | 27.82 | 4.56 |
| 32 | 4.52 | 26.28 | 3.62 | 5.23 | 27.19 | 3.82 | 5.53 | 27.15 | 3.97 |
| 64 | **3.97** | **25.76** | **2.84** | **4.28** | **26.54** | **3.14** | **4.87** | **26.34** | **3.19** |
| 128 | 4.67 | 26.41 | 3.43 | 5.15 | 27.12 | 3.94 | 5.52 | 27.04 | 4.13 |

number of frames from 16 to 64 leads to a consistent reduction in both $MAE_{body}$ and $MAE_{angle}$, with the lowest errors achieved at 64 frames (3.97 for $MAE_{body}$ and 2.84 for $MAE_{angle}$). Similarly, the MPBLPE decreases as the number of frames increases, reaching 25.76 at 64 frames. However, performance begins to degrade at 128 frames, where both $MAE_{body}$ and $MAE_{angle}$ increase. A similar trend is observed in the BEDLAM and OpenCap datasets. For BEDLAM, the $MAE_{body}$ reduces from 5.76 at 16 frames to 4.28 at 64 frames, with a corresponding improvement in $MAE_{angle}$ (from 4.35 to 3.14). In the OpenCap dataset, the optimal performance is also achieved at 64 frames, with $MAE_{body}$ reaching 4.87 and $MAE_{angle}$ improving to 3.19. This ablation study shows that 64 frames strike the right balance between capturing enough temporal information and maintaining computational efficiency. While performance improves when increasing frames from 16 to 64, using 128 frames offers no further gains and even slightly degrades performance. The likely reason is that 64 frames provide sufficient motion dynamics and biomechanical patterns for accurate pose estimation, while fewer frames lack context, and more frames introduce redundant information. Thus, 64 frames offer the optimal amount of temporal data without adding unnecessary complexity or noise.

### 1.12. Qualitative Results

We present qualitative results of MQ-HMR in Figures 1 and 2, showcasing the model's capability in handling extreme poses and partial occlusions. These results demonstrate the effectiveness of MQ-HMR, where the 3D reconstructions align well with the input images and maintain accuracy when viewed from different perspectives. A key factor behind this success is MQ-HMR's multi-query deformable attention mechanism, which efficiently manages uncertainty during the 2D-to-3D mapping process. MQ-HMR is able to overcome challenges that typically affect other state-of-the-art models. This approach ensures that MQ-HMR produces accurate and consistent 3D reconstructions, even in complex or ambiguous scenarios where traditional methods often struggle. Also, We show the qualitative results of BioPose in Figure 3, highlighting how our model is able to predict different poses, very close to the ground truth. The figures show multiple actions like squatting and drop jumping.

## References


[1] Marian Bittner, Wei-Tse Yang, Xucong Zhang, Ajay Seth, Jan van Gemert, and Frans CT van der Helm. Towards single camera human 3d-kinematics. *Sensors*, 23(1):341, 2022. 1, 2

[2] Michael J Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. Bedlam: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8726–8737, 2023. 1

[3] Sai Kumar Dwivedi, Yu Sun, Priyanka Patel, Yao Feng, and Michael J Black. Tokenhmr: Advancing human mesh recovery with a tokenized pose representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1323–1333, 2024. 5

[4] Saeed Ghorbani, Kimia Mahdaviani, Anne Thaler, Konrad Kording, Douglas James Cook, Gunnar Blohm, and Nikolaus F Troje. Movi: A large multi-purpose human motion and video dataset. *Plos one*, 16(6):e0253157, 2021. 1

[5] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4d: Reconstructing and tracking humans with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14783–14794, 2023. 5

[6] Sam Johnson and Mark Everingham. Learning effective human pose estimation from inaccurate annotation. In *CVPR 2011*, pages 1465–1472. IEEE, 2011. 6

[7] Scott D Uhlrich, Antoine Falisse, Łukasz Kidziński, Julie Muccini, Michael Ko, Akshay S Chaudhari, Jennifer L Hicks, and Scott L Delp. Opencap: Human movement dynamics from smartphone videos. *PLoS computational biology*, 19(10):e1011462, 2023. 1, 2

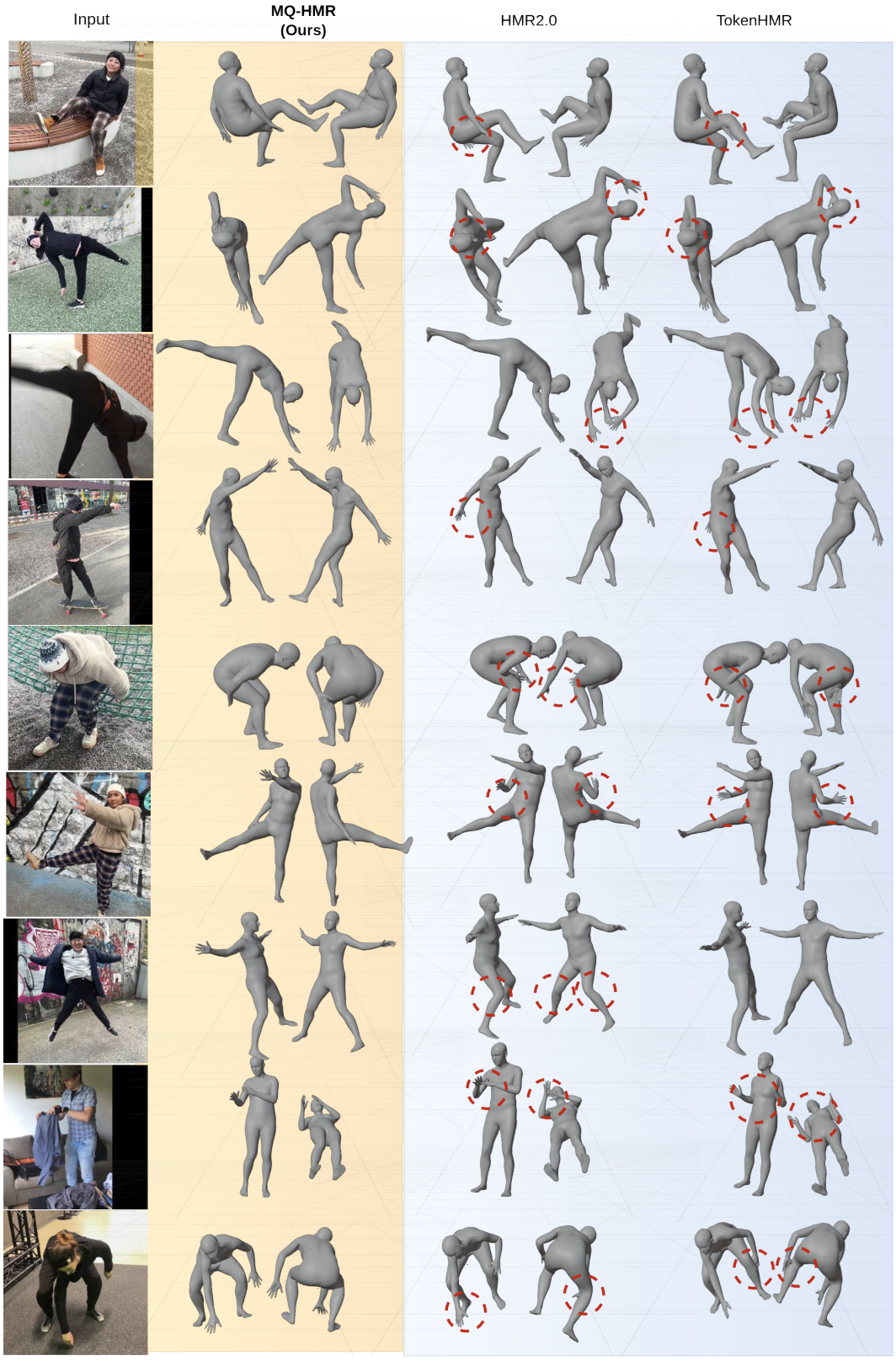


Figure 1. Comparison of state-of-the-art methods, HMR2.0 [5] and TokenHMR [3], which use vision transformers for 3D human mesh recovery from a single image. Red circles highlight errors in these methods when dealing with complex or ambiguous poses. In contrast, our MQ-HMR method addresses these challenges by incorporating a multi-query deformable transformer, leveraging multi-scale feature maps and a deformable attention mechanism to deliver more accurate and anatomically consistent pose estimations, even in difficult scenarios.


Figure 2. Qualitative results of our approach on challenging poses from the LSP [6] dataset.

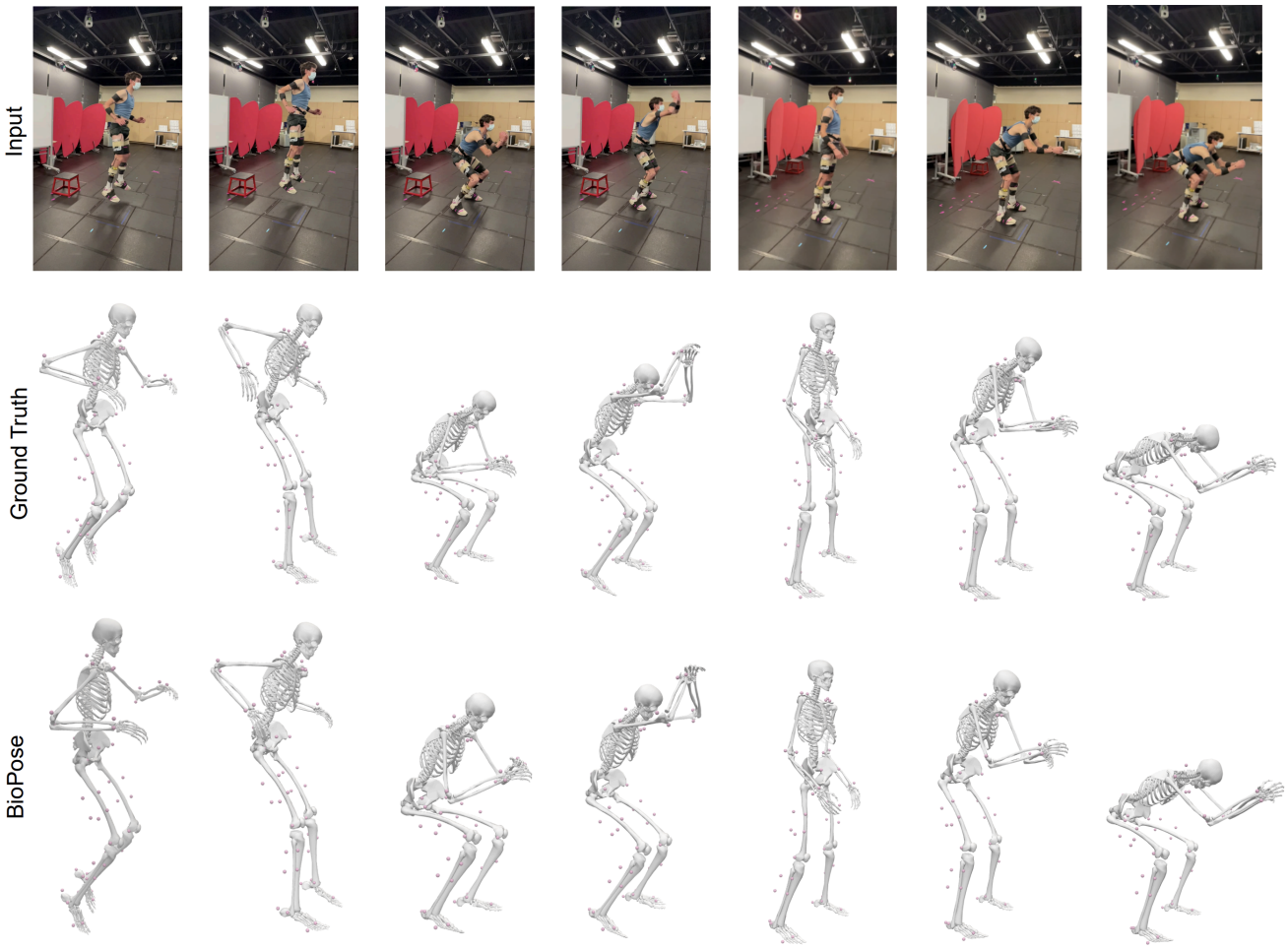


Figure 3. Qualitative results of our proposed method BioPose and comparison with ground truth. These pictures include multiple actions such as squatting and drop jumps.

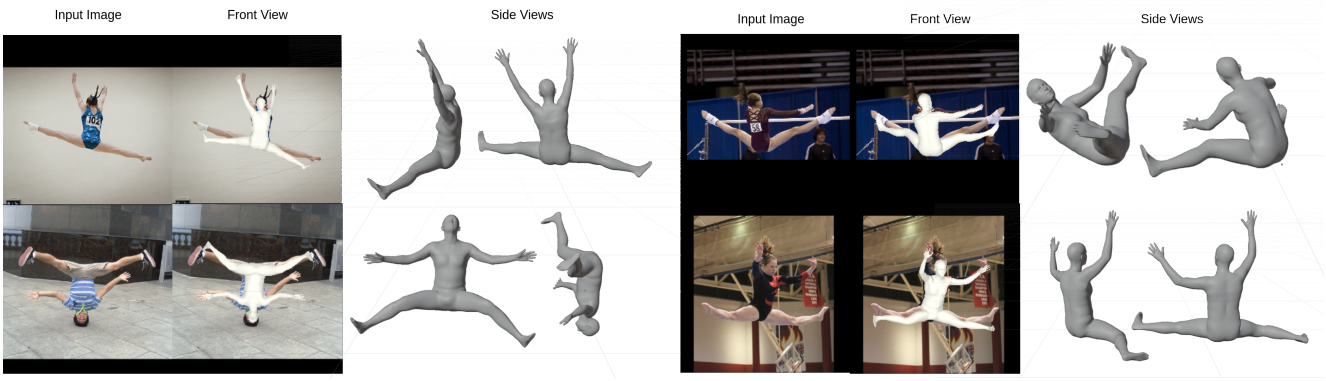


Figure 4. Failure Cases of MQ-HMR in 3D Human Reconstruction: MQ-HMR frequently struggles with handling unusual body movements and the complex layering of body parts in three-dimensional space. These difficulties often lead to inaccurate 3D pose estimations and invalid results. The main reason for these limitations is the model's dependence on the SMPL parametric model, which fails to adequately represent the complexity of extreme or rare human pose.