# SPUD: SKELETAL PHYSICS-BASED HUMAN POSE ESTIMATION USING DIFFUSION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

We propose SPUD as a diffusion-based framework for physics-based human pose estimation (PHPE), regularized by a novel kinetics prior and a novel SKEL-based pose prior. We also present a fencing motion-capture video dataset, synchronized with force plates, to allow for quantitative evaluation of kinetics estimation for athletic human movements. PHPE methods aim to improve realism by enforcing the laws of dynamics. However, inverse dynamics optimization without force annotations is challenging, as it is an underconstrained problem. To address this, we turn to diffusion models and their expressiveness in learning complex distributions, as recent studies have demonstrated impressive results for solving inverse problems via variational diffusion sampling. We build on these frameworks for inverse kinematics, and extend them to inverse dynamics by training a novel kinetics prior on simulation data. Furthermore, for improved compatibility with downstream biomechanics analysis, SPUD is the first PHPE method to infer human dynamics for the SKEL human model. Lastly, we conduct extensive experiments on various inverse kinematics and inverse dynamics tasks to demonstrate SPUD's versatility, and on datasets with disparate human movements to demonstrate improved adaptability. This includes a force-annotated fencing dataset we collected, which can facilitate more robust evaluations for future PHPE research.

## 1 INTRODUCTION

While "*Citius, Altius, Fortius*" is the Olympic motto, or "*Faster, Higher, Stronger*" in English, most of us likely won't prioritize these three objectives for many daily activities, such as brushing our teeth or reading a book. In other words, different activities can be accompanied with very different types of movements and goals. In the context of monocular human pose estimation (HPE), although regression-based methods dominate, they can be constrained by their training data, and performance can drop when tested on new domains (Jiang et al., 2024). From an application standpoint, this means that the user may need to collect new training data from said domain for retraining or finetuning, which can be extremely time consuming due to the processing of motion-capture (mocap) data. To this end, following the success of Lu et al. (2025) in solving inverse kinematics tasks via variational diffusion sampling (Mardani et al., 2024) regularized by an unconditional pose prior, we extend their test-time optimization (TTO) framework to inverse dynamics for monocular physics-based HPE (PHPE). We call our method SPUD (Skeletal PHPE Using Diffusion).

PHPE (Zhu et al., 2025; Zhang et al., 2024; Gartner et al., 2022b; Rempe et al., 2020; Shimada et al., 2020) aims to improve realism by incorporating the underlying dynamics equations (7) that govern human motion. As such, in addition to kinematics, PHPE involves estimating kinetics, which becomes an under-constrained problem when multiple external forces are applied to the human simultaneously at different locations. In parallel to how pose priors (Lu et al., 2025; Pavlakos et al., 2019; Kanazawa et al., 2018; Bogo et al., 2016) regularize the ill-posed depth estimation problem for purely kinematic methods, we handle this with a novel diffusion-based kinetics prior, DForce, which we pretrained on the MinT dataset (Schneider et al., 2024).

Furthermore, the human models used for dynamics modeling in PHPE are often incompatible with common biomechanics standards (Wu et al., 2005; 2002). Existing methods either employ a human model constructed from shape primitives, or directly optimize dynamics for the SMPL human model (Loper et al., 2015). Instead, we infer dynamics based on the kinematic tree of the more anatomically



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.

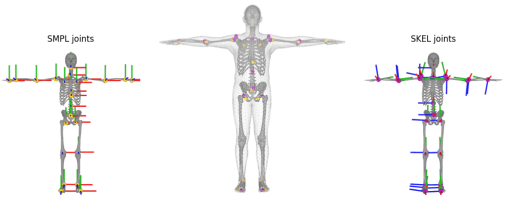


Figure 1: SMPL's joint centers (yellow) overlayed with SKEL's (purple). As with common biomechanics standards, SKEL's joint centers and local coordinate systems are based on bone landmarks and orientations (*right*); whereas SMPL's local coordinates are based on aligning its "T-pose" with the world's coordinate system (*left*).

accurate SKEL model (Keller et al., 2023), which was recently introduced to bridge the gap between vision-based HPE and biomechanics. SKEL shares SMPL's surface mesh, but with a biomechanical skeleton based on human anatomy (Fig. 1). And since existing pose priors are SMPL-based, we also pretrained a novel SKEL-based pose prior, DBone, on the training set in (Xia et al., 2025). To estimate anthropometrics, we follow the approach in (Zhu et al., 2025) that adjusts predictions based on values from biomechanics studies, and further expand this biomechanics database to better fit our dynamics modeling framework.

Lastly, since large-scale human motion video datasets do not include force labels, the predicted kinetics can be difficult to evaluate. As such, we collected a marker-based mocap dataset, synchronized with force plates and RGB video, to serve as a challenging new evaluation set. This dataset includes over 13 thousand frames of fencing actions performed by a national-level fencer and a novice. Following standard protocol, we first evaluate SPUD on benchmark HPE datasets 3DPW (von Marcard et al., 2018), RICH (Huang et al., 2022), H36M (Ionescu et al., 2014), then further demonstrate its improved adaptability by evaluating on the faster and more athletic movements from our fencing dataset, as well as for golf swings from the GolfDB dataset (McNally et al., 2019). Lastly, we evaluate our pose prior for motion denoising on HumanEva (Sigal et al., 2010), Transitions (Mahmood et al., 2019), and HPS (Fiche et al., 2024).

In summary, we propose SPUD as the first SKEL-based PHPE method using diffusion, with the following contributions: (i) we introduced two novel diffusion-priors — DForce: a kinetics prior, and (ii) DBone: a SKEL-based pose prior. (iii) We expanded the biomechanics database from (Zhu et al., 2025) for anthropometrics estimation. (iv) We evaluated SPUD for general human motions on 3DPW, RICH, H36M, and also demonstrated improved adaptability by evaluating on two disparate sports — golf from GolfDB and (v) fencing from our new force-annotated mocap video dataset, which will be made publicly available upon acceptance.

## 2 Related Work

**Monocular 3D human pose estimation** (HPE) (Li et al., 2022b;a; Wang et al., 2020; Mehta et al., 2017; Martinez et al., 2017) refers to estimating human motion from a single camera. If the method is purely data-driven, with the human structural information modeled implicitly, the output pose can be inconsistent with human anatomy (*e.g.,* hyperextended joints, erratic bone lengths). This can be mitigated by inferring the parameters for a parametric human model or incorporating pose priors. To further reduce motion artifacts, physics-based HPE (PHPE) (Zhu et al., 2025; Zhang et al., 2024; Gartner et al., 2022a;b; Xie et al., 2021; Shimada et al., 2021; Rempe et al., 2020; Shimada et al., 2020) incorporate the underlying dynamics equations (7) that govern human motion. However, since the aforementioned human models and priors only regulate human poses kinematically, they may not be directly compatible with the dynamics modeling required for PHPE, as discussed next.

**Parametric human shape models** often represent the human body as a triangulated mesh, parameterized by pose and shape parameters learned from 3D scans. Many early works (Zuffi & Black, 2015; Chen et al., 2013; Hasler et al., 2009) are based on SCAPE (Anguelov et al., 2005), the first to factor body deformations into separate shape and pose spaces. SMPL (Loper et al., 2015), the most popular in the current literature, uses a collection of linear functions to map pose, root translation, and shape parameters to a human mesh. Its pose parameter represents rotations for the root and 23 joints, each with 3 DoFs. Its shape parameter $\beta \in \mathbb{R}^m$ denotes the first $m$ principal component analysis coefficients in a low-dimensional shape space. SMPL-X (Pavlakos et al., 2019) learns an expressive model by fusing SMPL with the FLAME head model (Li et al., 2017) and MANO

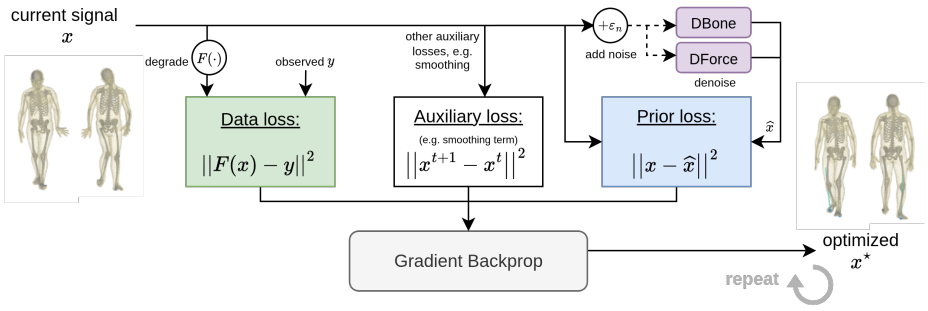


Figure 2: SPUD's optimization framework

hand model (Romero et al., 2017). Other adaptations include the quaternion-based STAR (Osman et al., 2020) and SoftSMPL (Santesteban et al., 2020), which incorporates a nonlinear shape space. Recently, MHR (Ferguson et al., 2025) was introduced to provide more precise decoupled control of pose and shape parameters, Anny (Brégier et al., 2026) incorporates phenotype parameters to broaden the applicability of subject groups, such as infants and elders, and SOMA (Saito et al., 2026) acts as a universal model to unify the above. However, since the focus of these human models is on animation, their kinematic skeletons are often incompatible with common biomechanics standards (Wu et al., 2002; 2005). Another recent model, SKEL (Keller et al., 2023), was introduced to bridge this gap, by integrating a biomechanical skeleton with the SMPL mesh, for more anatomically accurate joint center positions and degrees of freedom (DoFs) (Fig. 1). Hence, we adopt SKEL as a more suitable human model for PHPE.

**Pose priors** can be used to evaluate the plausibility of estimated poses and penalize unlikely ones. Prior pose distributions are often learned from large datasets using generative models, such as with generative adversarial networks in (GANs) (Kanazawa et al., 2018), variational autoencoders (VAEs) (Pavlakos et al., 2019), normalizing flow (Zanfir et al., 2020), neural distance fields (NDFs) (Tiwari et al., 2022). Recently, motion generation approaches (Fiche et al., 2024; Geng et al., 2023) have begun leveraging Vector Quantized VAEs (van den Oord et al., 2017) to tokenize poses into a discrete latent space. Diffusion models (Ho et al., 2020; Sohl-Dickstein et al., 2015) have also been trained as priors (Lu et al., 2025; Müller et al., 2024; Shafir et al., 2024) due to their superior expressiveness in learning complex distributions. In comparison, GANs (Goodfellow et al., 2020) can be limited in diversity and unstable to train due to their adversarial training nature, the expressiveness of VAEs (Kingma & Welling, 2014) can also be bottlenecked by their latent Gaussian prior, flow-based methods (Rezende & Mohamed, 2015) often require specialized architectures to construct reversible transforms, and NDFs (Chibane et al., 2020) can struggle to generalize for pose data. DPoserX (Lu et al., 2025) is a recent whole-body pose prior designed to solve inverse kinematics problems based on variational diffusion sampling (Mardani et al., 2024). Furthermore, unlike most diffusion-based methods, which are tailored for specific generation tasks and trained to condition on certain inputs, DPoserX focuses on flexibility, and is an unconditional prior that can be integrated into various pose-related tasks via test-time optimization. Hence, we hope to extend their framework to inverse dynamics for PHPE. To do so, we trained a novel kinetics prior to regularize ground reaction forces. Furthermore, since existing pose priors are SMPL-based, we also trained a novel SKEL-based pose prior for better biomechanics compatibility.

# 3 SPUD

Given observed and possibly noisy positions, our goal is to recover the underlying human dynamics (DoFs, joint torques, external forces) and anthropometric properties (dimensional, mass, inertial properties) required to produce the motion. We describe our methodology below.

## 3.1 Inverse problem setup

Given some known mapping $f : x \to y$ that maps signal $x$ to observation $y$, our goal is to estimate its inverse $f^{-1}$, so that we can retrieve the original signal $x = f^{-1}(y)$ from the observed $y$.

For example, during *inverse kinematics*, we want to recover DoFs $\mathbf{q}$ and human dimensions $\beta$ from 3D positions $\mathbf{P}$. Let $f_K(\mathbf{q}, \beta) = \mathbf{P}$ denote forward kinematics. In this case, we can write our setup as $f(\cdot) = f_K(\cdot)$, where our goal is to estimate $f_K^{-1}$ to recover signal $x = \{\mathbf{q}, \beta\}$ from the observed $y = \mathbf{P}$. We can extend this to additional tasks such as motion denoising and pose estimation, by degrading $\mathbf{P}$ using some task-specific function $g$, such that $f(\cdot) = g(f_K(\cdot))$. For example, when we have *noisy* inputs, we can write $g(\mathbf{P}) = \mathbf{P} + \epsilon$, with random noise $\epsilon \in \mathcal{N}(0, \sigma^2)$, or during *occlusion*: $g(\mathbf{P}) = \text{Mask}(\mathbf{P})$, or *pose estimation*: $g(\mathbf{P}) = \text{CameraProjection}(\mathbf{P})$, or $g$ can be some combination of the above. As a whole, we can view $f$ as some "degradation" of the original signal.

In our case, we further extend this to recover the underlying external forces $\mathbf{F}$, joint torques $\boldsymbol{\tau}$, and anthropometric properties $\mathcal{A}$ required to produce $\mathbf{q}$, based on a forward dynamics update step $f_d(q^t, F^t, \boldsymbol{\tau}^t, \mathcal{A}) = q^{t+1}$. For time sequence $\mathbf{q} = \{q^0, q^1, ..., q^{T-1}\}$ of length $T$, we denote $f_D(q^0, \mathbf{F}, \boldsymbol{\tau}, \mathcal{A}) = \{q^0, f_d(q^0, F^0, \boldsymbol{\tau}^0, \mathcal{A}), ..., f_d^T(q^0, F^0, \boldsymbol{\tau}^0, \mathcal{A})\} = \mathbf{q}$ as applying $f_d$ recursively $T$ times. Putting it all together, given observed $y$, we want to recover signal $x = \{\mathbf{q}, \beta, \mathbf{F}, \boldsymbol{\tau}, \mathcal{A}\}$, based on the nested relations (1 - 2), which are elaborated in Sec. 3.2.

$$y = g(f_K(\mathbf{q}, \beta)) \tag{1}$$

$$\mathbf{q} = f_D(q^0, \mathbf{F}, \boldsymbol{\tau}, \mathcal{A}) \tag{2}$$

### 3.2 Dynamics computations

**Kinematics.** We use SKEL (3) as our forward kinematics model to infer 3D positions $\mathbf{P}$ from shape parameters $\beta$, 43 joint DoFs $\mathbf{x}_B$, and 6 world DoFs (3 rotational and 3 translational) $\boldsymbol{\Theta}$, for 49 total DoFs $\mathbf{q} \triangleq \{\boldsymbol{\Theta}, \mathbf{x}_B\}$ corresponding to 24 bones/segments (*a pelvis, lumbar spine, thorax, head, and a pair of femurs, tibiae, tali, calcanei, toes, scapulae, humeri, ulnae, radii, hands*). $\mathbf{P}_{skin}$ refers to the 3D positions of the SMPL surface mesh vertices and $\mathbf{P}_{bone}$ for the bones. The global frame is set such that when $\boldsymbol{\Theta} = 0$, the origin coincides with the human's root, and the $X$ direction is frontward of the human, $Y$ is opposite of gravity, and $Z$ is to the right of the human.

$$\mathbf{P}_{skin}, \mathbf{P}_{bone} = \text{SKEL}(\mathbf{q}, \beta) = \text{SKEL}(\boldsymbol{\Theta}, \mathbf{x}_B, \beta) \tag{3}$$

**Anthropometrics.** Let anthropometics features $\mathcal{A} \triangleq \cup_k \{\rho_k, I_{0,k}, CoM_k\}$ be the collection of densities, inertia tensors, and center of masses for each segment $k$. We follow (Zhu et al., 2025) and infer $\mathcal{A}$ from offset parameter $\alpha$ and shape parameter $\beta$ as in (4), where $\bar{\mathcal{A}}$ and $\varsigma$ are mean and range values from a biomechanics database, and the shape scaling factor $s_\beta$ is a function of $\beta$.

$$\mathcal{A} = f_a(\alpha, \beta) = s_\beta(\bar{\mathcal{A}} + \alpha\varsigma) \tag{4}$$

Our approach improves on (Zhu et al., 2025) in the following: (i) while the values $\bar{\mathcal{A}}$ are defined based on bony anatomic landmarks, as shown in Fig. 1, not all of SMPL's keypoints correspond directly to bones, whereas SKEL's is, making this biomechanics database more anatomically suitable for our framework. (ii) Since the kinematic skeleton of SKEL is inherently different from SMPL (*i.e.,* SKEL has scapulae), for compatibility, we added scapula anthropometrics data from (Saul et al., 2015; de Groot & Brand, 2001; Klein Breteler et al., 1999). (iii) For the other segments, we also added data from (Dumas & Wojtusch, 2018; Winter, 2009; De Leva, 1996; Plagenhoef et al., 1983; Hatze, 1980; 1977) to include a wider distribution. (iv) This additional data also contain segment densities $\rho_k$, which we use (along with segment volume from $\mathbf{P}_{skin}$) to compute segment mass, whereas previously Zhu et al. (2025) estimates segment mass from 1D length information only, based on scaling values from (Dumas et al., 2007).

**Ground reaction forces.** Let $\mathbf{F}^{t,k} \in \mathbb{R}^6$ denote ground reaction forces (GRFs) and moments (GRMs) at time $t$ for segment $k$, each with 3 components. For notation simplicity, we remove the $\{t, k\}$ superscript. We infer $\mathbf{F}$ from force parameters $\delta$ and local center of pressures $\boldsymbol{z}$ as

$$\mathbf{F} \triangleq [\text{GRF}, \text{GRM}]^\intercal = f_e(\mathbf{q}, \mathcal{A}, \delta, \boldsymbol{z}) = [\boldsymbol{F}W,\ R\boldsymbol{z}L \times \boldsymbol{F}W]^\intercal \tag{5}$$

where $\boldsymbol{F} = \text{GRF}/W$ denotes the GRFs scaled by body weight $W$, which we inferred from (4). Let $R$ denote the segment's rotation matrix. The $\delta_x$ and $\delta_z$ determines the tangential directions of $\boldsymbol{F}$, while $\delta_y$ ensures that it is within some friction cone (6), where $\boldsymbol{F}_Y$ corresponds to the GRFs scaled by body weight in the vertical direction $Y$ (as defined under (3)).

$$\boldsymbol{F}_X = R\delta_x,\ \boldsymbol{F}_Z = R\delta_z,\ \boldsymbol{F}_Y = \sqrt{\boldsymbol{F}_X^2 + \boldsymbol{F}_Z^2 + \delta_y^2} \tag{6}$$

**Joint torques.** Finally, we infer joint torques via Lagrangian dynamics (7), where mass matrix $\mathcal{M}$, Coriolis term $\mathcal{C}$, and Jacobian $\mathcal{J}$ can all be computed analytically from $\mathbf{q}$ and $\mathcal{A}$. Following (Zhu et al., 2025; Xie et al., 2021), we compute these Lagrangian terms based on (Liu & Jain, 2012).

$$\boldsymbol{\tau} = f_l(\mathbf{q}, \mathcal{A}, \mathbf{F}) = \mathcal{M}\ddot{\mathbf{q}} + \mathcal{C} - \mathcal{J}^\intercal \mathbf{F} \tag{7}$$

### 3.3 Optimization

We follow the TTO framework in (Lu et al., 2025) to optimize parameters $\{\mathbf{q}, \beta, \alpha, \delta, \boldsymbol{z}\}$ by minimizing a sum of three types of losses $\boldsymbol{L}_{total} = \boldsymbol{L}_{data} + \boldsymbol{L}_{aux} + \boldsymbol{L}_{prior}$.

**Data loss** $\boldsymbol{L}_{data} = ||f(x) - y||^2$ is the distance between the observation $y$ and $f(x)$, our current recovered signal degraded by $f$. In practice, we break down $\boldsymbol{L}_{data}$ into two components $L_{kin}$ and $L_{dyn}$, corresponding to the nested equations (1) and (2) respectively. $L_{kin}$ is the motion reconstruction loss from previous methods, and $L_{dyn}$ is equivalent to computing the residual force $\boldsymbol{\tau}_{res}$ (*i.e.* the entries of $\boldsymbol{\tau}$ that correspond to the root, which should be zero).

$$\boldsymbol{L}_{data} = L_{kin} + L_{dyn} \tag{8}$$

$$L_{kin} = ||g(f_K(\mathbf{q}, \mathcal{A})) - y||^2 \tag{9}$$

$$L_{dyn} = ||\boldsymbol{\tau}_{res}||^2 \tag{10}$$

**Prior loss** $\boldsymbol{L}_{prior} = ||x_0 - \hat{x}||^2$ is the distance between current signal $x_0$ and denoised signal $\hat{x}$, where $\hat{x} = \mathcal{D}(x_n)$ is the noisy signal $x_n$ denoised by diffusion model $\mathcal{D}$, following noise schedule $n$. In practice, we apply this prior loss on force features $\mathbf{x}_F$ and bone poses $\mathbf{x}_B$, with diffusion models DForce $\mathcal{D}_F$ and DBone $\mathcal{D}_B$, which we describe in Sec. 3.4. To expedite the process for human pose denoising, we implement the truncated schedule in (Lu et al., 2025).

$$\boldsymbol{L}_{prior} = L_{DBone} + L_{DFone} = ||\mathbf{x}_F - \mathcal{D}_F(\mathbf{x}_{F,n})||^2 + ||\mathbf{x}_B - \mathcal{D}_B(\mathbf{x}_{B,n})||^2 \tag{11}$$

**Auxiliary loss** $\boldsymbol{L}_{aux}$ includes additional auxiliary terms. By default, we set this as a temporal joint acceleration smoothing term with weight $\lambda_{acc} = 20$.

$$\boldsymbol{L}_{aux} = \lambda_{acc}||\ddot{\mathbf{x}}_B||^2 \tag{12}$$

### 3.4 Diffusion priors

**DForce.** We represent our force features as $\mathbf{x}_F = \{\boldsymbol{F}_Y, \boldsymbol{\Psi}, \boldsymbol{\psi}, \boldsymbol{f}, \boldsymbol{z}\}$, where $\boldsymbol{F}_Y$ are the vertical GRFs scaled by body weight. Let $\Psi \subset \mathbf{P}_{skin}$ denote the 3D positions of the root and 10 feet landmarks (*SMPL feet*, *SMPL ankles*, *heels*, *big toes*, *small toes*). Feet features $\boldsymbol{\Psi} = \{\Psi_Y, \dot{\Psi}_Y, \dot{\Psi}_{XZ}\}$ include vertical positions $\Psi_Y$, vertical velocities $\dot{\Psi}_Y$, and tangential velocities $\dot{\Psi}_{XZ}$. We use lower case $\boldsymbol{\psi}, \boldsymbol{f}, \boldsymbol{z}$ to denote the feet landmark positions, GRFs, and CoPs in the root's coordinate system, respectively. We chose to convert $\boldsymbol{f}$ and $\boldsymbol{z}$ to the root's coordinate system rather than the feet because the latter is defined differently between SMPL and SKEL (Fig. 1), and DForce was trained on the MinT dataset, which does not include SKEL labels. On the other hand, since SKEL shares SMPL's mesh, positions $\Psi$ can be extracted from both human models using the same joint regressor.

**DBone.** Bone features $\mathbf{x}_B$ are the joint rotation parameters in the SKEL model (3).

Inspired by (Lu et al., 2025; Ci et al., 2023), both DBone and DForce consist of two fully connected blocks, without conditional input pathways. We trained DBone on the training set of HSMR (Xia et al., 2025) for 1 million steps, with a batch size of 1024. DForce was trained on the MinT dataset (Schneider et al., 2024) for 1 million steps, with a batch size of 128.

## 4 Experiments

**Implementation.** In Sec. 4.1, following standard protocol, we evaluated our PHPE results for general human movements on benchmark datasets: 3DPW (von Marcard et al., 2018), RICH (Huang et al., 2022), and H36M (Ionescu et al., 2014). In Sec. 4.2, to show SPUD's adaptability to new domains, we conducted further ablation experiments on two disparate sport-specific datasets: GolfDB

(McNally et al., 2019) and our new fencing dataset. We compared our results to recent regression-based PHPE methods PhysPT (Zhang et al., 2024) and MusclePose (Zhu et al., 2025). For all methods, WHAM (Shin et al., 2024) was used as the initial kinematic predictor. Since predictions are in world coordinates, we reported Procrustes-aligned joint errors (PJE) in millimeters (mm). Unless otherwise stated, all other distance related metrics will also be in mm. Lastly, in Sec. A.1, we included the motion denoising experiments from Lu et al. (2025) to compare against SOTA pose priors on the HumanEva (Sigal et al., 2010), Transitions (Mahmood et al., 2019), and HPS (Fiche et al., 2024) datasets. Since SKEL does not have a gender-neutral model, SPUD used the male SKEL model for all experiments, following HSMR (Xia et al., 2025). Additional dataset-specific details on implementation and evaluation metrics can be found in their respective subsections.

### 4.1 BENCHMARK DATASETS

**General movements.** Following previous works, we evaluated on the portions of 3DPW with ground-feet contact only (about 70%), for a total of 22024 motion frames at 30 fps. For RICH, we evaluated on the *cooking, lunge, wiping chairs, impro2, spray, eating* actions, for a total of 34193 motion frames at 30 fps. We reported our results in Tab. 1, which includes GRF quantiles, where $F_Y$ denotes the predicted vertical GRF/body weight for one foot, and $\Sigma F_Y$ is the sum of both feet. $\Sigma F_Y$ should usually be close to 1 times body weight. PJEs are computed from the 14 LSP joints (Johnson & Everingham, 2010) for 3DPW and 24 SMPL joints for RICH.

Overall, we see a slightly higher PJE for SPUD than the regression-based methods, which aligns with the discussion in Sec. 1 that regression-based methods dominate in positional accuracy on general large-scale datasets. This gap in positional accuracy may also indicate a bias stemming from the SKEL model (43 joint rotational DoFs) being less flexible than SMPL (69 DoFs), as discussed in (Xia et al., 2025), and also from the mismatch in gender models: the initial kinematic estimates from WHAM, the predictions from PhysPT and MusclePose, and the ground truth labels, are all in SMPL neutral, whereas SPUD implements the male SKEL model.

The two SOTA PHPE methods were compared to because PhysPT has better overall positional accuracy, while MusclePose produces more plausible kinetic predictions. And although we see similar $\Sigma F_Y$ quantiles between SPUD and MusclePose, the predicted forces are still difficult to evaluate for a wide range of motions without force-labels. Hence we assessed the predicted forces, as well as kinematic plausibility, in more detail in the following experiments, on specific actions.

Table 1: Inference results on 3DPW, RICH, and H36M-Walk. Curly brackets indicate the median, 99th, and maximum quantiles. Square brackets indicate the percentage of frames with floating.

| | 3DPW | | RICH | | H36M-Walk | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | PJE↓ | $\Sigma F_Y$ {med, q99, max} | PJE↓ | $\Sigma F_Y$ {med, q99, max} | PJE↓ | GP↓ | FL[% frames]↓ | FS↓ | $F_Y$ {med, q99, max} |
| PhysPT | **35.3** | {0.21, 0.46, 1.47} | **40.3** | {0.25, 0.51, 2.28} | **41.0** | **4.9** | 21.0 [62.9] | 76.6 | {0.01, 1.98, 4.99} |
| MusclePose | 41.0 | {**0.97**, 1.42, 1.88} | 51.0 | {**0.96**, 1.25, 2.44} | 42.9 | 13.2 | **20.4** [**7.2**] | 30.9 | {0.42, 1.01, 1.12} |
| **ours: SPUD** | 35.7 | {**0.99**, 1.33, 2.10} | 42.2 | {**0.99**, 1.32, 2.49} | 49.5 | 9.5 | 25.2 [9.4] | **26.8** | {**0.51**, 1.03, 1.55} |

**Gait.** We assessed human gait on the *walking* action of H36M's test set, at 10 fps, for a total of 5800 frames. Results are reported in Tab. 1, along with additional kinematic plausibility metrics. PJEs are computed from the 14 LSP joints. **Ground penetration** (**GP**) is computed as the average vertical distance to the ground in mm of vertices below the ground. Let $h_{min}$ be the minimum vertex height for a frame. We computed the **floating loss** (**FL**) as the average $h_{min}$ among frames when floating occurs. We defined floating as when $h_{min} > 10$ mm, when at least one foot is in contact with the ground. For H36M-Walk, we assumed that at least one foot is always in contact with the ground (*i.e.,* no floating in the ground truth). **Foot sliding** (**FS**) is computed as the average displacement (mm) in consecutive frames for vertices within 10 mm of the ground. Overall, we see that PhysPT has less ground penetration and more floating, MusclePose is the opposite, while SPUD is in between, and with less foot sliding.

We plotted the median $F_Y$ (dashed), along with a 25-75% quantile band for 892 gait cycles in Fig. 3. Compared to PhysPT (purple), SPUD has more reasonable magnitudes of $F_Y$, peaking consistently around 1 times body weight. Compared to MusclePose (light blue), SPUD also stays around this peak longer ("flatter" top). However, since these benchmark datasets do not contain

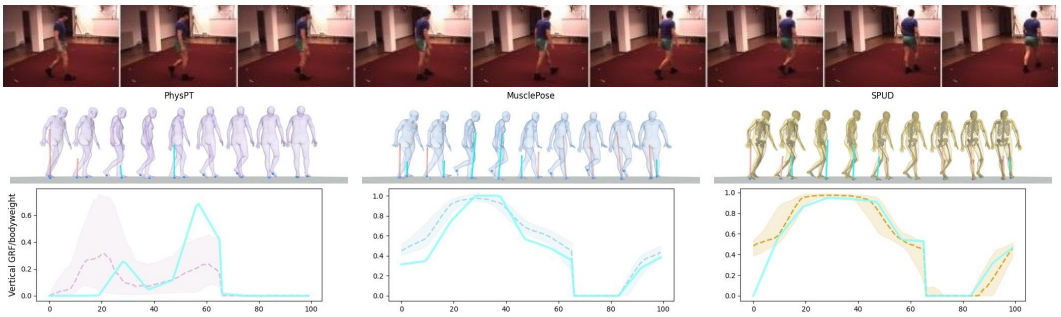


Figure 3: Median $F_Y$ (dashed), along with a 25-75% quantile band for gait cycles (892 steps) from H36M-Walk. The blue solid line corresponds to the right foot of the step in the video example.

force annotations, force predictions from PHPE are only assessed qualitatively, and often focused on gait, as with the case here. Hence, we collected a force-annotated dataset for fencing, to further allow for quantitative evaluations on more athletic movements, which we report next.

### 4.2 *Citius, Altius, Fortius*

**Fencing.** To overcome the lack of force-annotated motion video datasets for athletic human movements, we collected a marker-based mocap dataset for fencing techniques, synchronized with force plates. This evaluation set includes 28 videos for two subjects: a currently active national-level foil fencer (S1), and a novice (S2). Each video is 2.6-5.7 seconds long, and consists of a combination of fencing actions including *step forward/back, jump forward, lunge, recover forward/back, extension hit, point in line, forward reposte, feint, shoulder flick, circle parry, low line parry*. During evaluation, we downsampled the videos to 30 fps, for a total of 1852 frames for S1 and 1226 frames for S2. Our results are reported in Tab. 2, with the PJE computed from 16 keypoints (*glabella, TV10, TV2, mid IPS, hips, knees, ankles, shoulders, knees, ankles*) extracted from the SMPL mesh to align with the mocap markers. We also reported the **vertical force error** (**VFE**) as the mean absolute difference between the predicted $F_Y$ and the experimental force plate data ${}^{\text{FP}}F_Y$. Since the experimental force plate data were not post-processed for noise, higher extreme values can be present, including a maximum ${}^{\text{FP}}F_Y$ of 4.22 times body weight, as shown on the top of Tab. 2. Hence we also reported the median VFE in brackets for reference. However, we see that the maximum value of our predicted $F_Y$ (1.44) is close to the 99th percentile of ${}^{\text{FP}}F_Y$ (1.49), and our median $F_Y$ (0.48) is also closest to the median ${}^{\text{FP}}F_Y$.

Since the back (*ba*) foot of a lunge is an area of interest for fencers, we isolated the $\text{VFE}^{ba}_{lunge}$ values in Tab. 2; and since there were only 19 total lunges, we plotted all of their $F_Y^{ba}$ curves in red in Fig. 4, with force plate ${}^{\text{FP}}{F_Y}^{ba}$ in gray. The bold curves correspond to the RGB video example. Furthermore, after the back foot pushes off during a lunge, the $\text{VFE}^{ba}_{lunge}$ can be potentially inflated, as the force plate ${}^{\text{FP}}{F_Y}^{ba}$ can be lower than expected, due to the back foot sometimes sliding into the gaps between the force plates. This may also explain the lower median $\text{VFE}^{ba}_{lunge}$ for the other methods, as they also produced more floating artifacts (much higher FL in Tab. 2 and pink circles in Fig. 4), leading to more zero $F_Y$ predictions. In summary, we showed improved kinetic results with lower overall VFEs, and improved kinematic plausibility, with low FL and GP. FS was omitted due to the inherent amount of sliding that occurs in fencing. And although the PJE is higher for all methods, the gap in magnitude with PhysPT remains consistent with the results in Tab. 1.

Lastly, the shaded rows of Tab. 2 represent ablation results. To adhere to this section's "*Faster, Higher, Stronger*" moniker, we lowered the weight of the joint angle smoothing term ($\lambda_{acc}$) to 10 in our implementation, to embody more athletic and dynamic movements. We see that as $\lambda_{acc}$ is lowered from 50 to 10, the PJE and median $\text{VFE}^{ba}_{lunge}$ are improved more consistently for S1 (the faster and better fencer) than S2. Furthermore, compared to when the priors were removed (last row), our VFE (row 3) improved by 9.8%, and more noticeably for fast movements that are likely to be more outside the training distribution, as $\text{VFE}^{ba}_{lunge}$ improved by 13.3% during lunges.

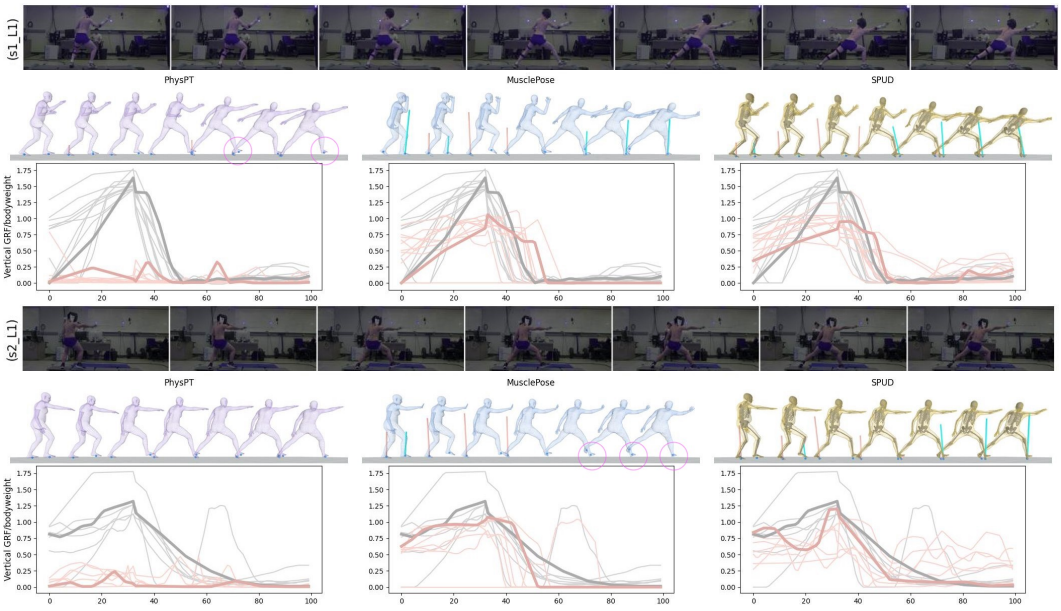


Figure 4: Predicted $F_Y^{ba}$, in red compared to force plate ${}^{\text{FP}}F_Y{}^{ba}$ in gray for the back/left foot of all lunges, normalized in time such that the maximum ${}^{\text{FP}}F_Y{}^{ba}$ are aligned at 30% of the lunge. The bold curves correspond to the video examples, with pink circles indicating when the front foot fails to snap to the ground at the end of the lunge. *Top row*: expert fencer (S1). *Bottom*: novice (S2).

Table 2: Inference results for fencing. $\text{VFE}^{ba}_{lunge}$ is the VFE of the back foot during a lunge. Round brackets indicate median VFEs. Shaded rows are ablations (abs).

| | | | | ${}^{\text{forceplate}}F_Y$: {med:**0.48**, q99:**1.49**, max:4.22} | | | S1: expert | | S2: novice | |
|---|---|---|---|---|---|---|---|---|---|---|
| | PJE ↓ | GP ↓ | FL [% frames] ↓ | VFE (med) ↓ | $\text{VFE}^{ba}_{lunge}$ ↓ | $F_Y$ {med, q99, max} | PJE ↓ | $\text{VFE}^{ba}_{lunge}$ ↓ | PJE ↓ | $\text{VFE}^{ba}_{lunge}$ ↓ |
| PhysPT | **59.33** | 22.4 | 25.1 [85.2] | 0.432 (0.364) | 0.462 (0.208) | {0.05, 0.37, 1.30} | **67.61** | 0.452 (**0.102**) | **46.75** | 0.473 (0.423) |
| MusclePose | 77.16 | **9.9** | 43.0 [37.6] | 0.276 (0.220) | **0.252** (**0.121**) | {0.43, 1.03, 1.28} | 95.54 | 0.299 (0.113) | 49.44 | **0.203** (**0.133**) |
| **ours: SPUD** | **61.37** | 10.3 | **18.6** [6.8] | **0.249** (**0.190**) | **0.255** (0.188) | {**0.48**, 1.11, **1.44**} | **70.42** | 0.249 (**0.147**) | 47.70 | **0.262** (0.218) |
| (abs) $\lambda_{acc} = 20$ | 61.40 | 10.3 | 18.8 [6.4] | 0.255 (0.195) | 0.267 (0.208) | {0.48, 1.12, 2.06} | 70.51 | 0.262 (0.174) | **47.64** | 0.272 (0.219) |
| $\lambda_{acc} = 30$ | 61.51 | 10.3 | 18.6 [6.7] | 0.257 (0.195) | 0.261 (**0.185**) | {0.48, 1.10, 1.85} | 70.67 | **0.247** (0.162) | 47.67 | 0.276 (0.220) |
| $\lambda_{acc} = 50$ | 61.65 | 10.3 | 19.2 [6.4] | 0.254 (0.197) | 0.262 (0.211) | {0.48, 1.11, **1.46**} | 70.86 | 0.252 (0.186) | 47.74 | 0.272 (0.223) |
| w/o prior | 61.70 | 12.4 | 21.8 [8.0] | 0.276 (0.200) | 0.294 (0.199) | {0.49, 1.17, 2.31} | 70.64 | 0.285 (0.182) | 48.19 | 0.304 (**0.208**) |

**Golf.** Next, we evaluated golf swings on the GolfDB dataset (McNally et al., 2019). Compared to the golf swings in PennAction, GolfDB offers more videos at a consistent frame rate (30 fps) with detailed annotations (athlete name, gender, club type, camera view, swing phase segments, YouTube ID). Among the 758 non-slo-mo videos/swings, we removed the ones that included partial slo-mo, and kept 742 for evaluation, for 48790 total frames. Unlike in fencing or most movements, we know that during a golf swing, certain keypoints on the feet (back foot big toe, front foot small toe and heel) should exhibit minimal sliding, and that the wrists should be close together, with the front (*fr*) elbow extended. Taking advantage of the flexibility of TTO, we incorporated this golf information $L_{golf}$ into the auxiliary loss $\boldsymbol{L}_{aux}$. We reported our results in Tab. 3. Since GolfDB does not include ground truth force nor position labels, VFEs and PJEs were not computed. To assess sliding, we computed **total foot sliding** (**TFS**) as the sum of tangential distances traveled between adjacent frames for each swing, then averaged over all swings.

Furthermore, taking advantage of the annotations in GolfDB, we aligned our results in Fig. 5 by swing phases, with the downswing exaggerated to better visualize the athlete's weight transfer from their back foot $F_Y^{ba}$ to front $F_Y^{fr}$. From both Fig. 5 and the last three columns of Tab. 3 ($F_Y^{fr} - F_Y^{ba}$), we can see that SPUD resulted in a cleaner weight transfer (*i.e.,* where the red and blue curves intersect) during the downswing, compared to MusclePose, which seemed to occur near impact; and both methods produced more realistic magnitudes compared to PhysPT.

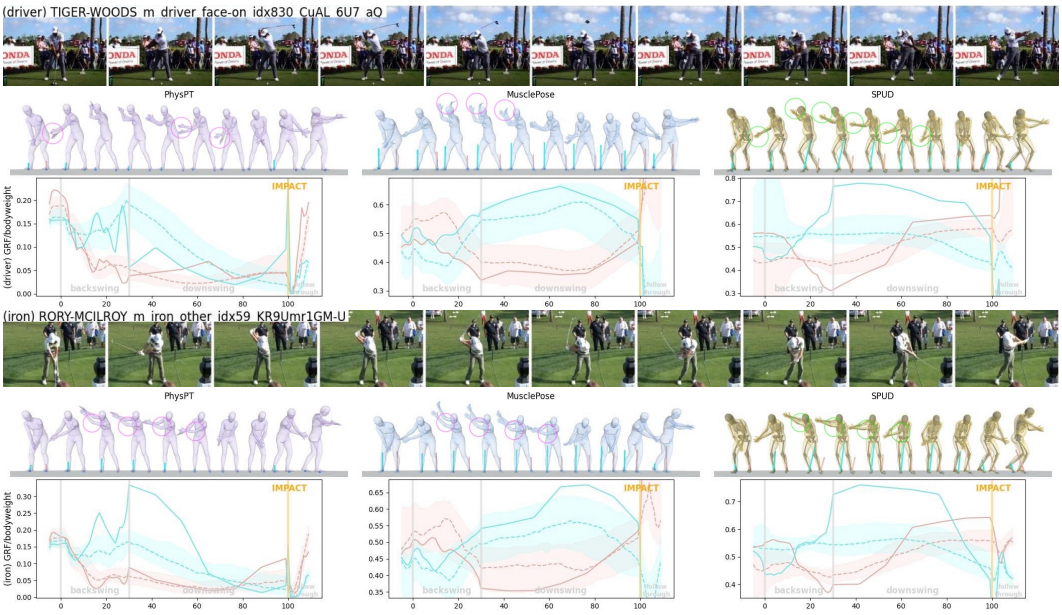


Figure 5: Median $F_Y$ (dashed) with 25-75% quantile band for golf swings from GolfDB, normalized in time to align swing phases, with the downswing zoomed in (70%) to better visualize the weight transfer from the back foot (blue) to front (red). The solid curves correspond to the video examples. *Top row*: 495 driver swings, with pink circles indicating when the wrists struggle to stay close. *Bottom*: 128 iron swings, pink circles indicate when the front elbow struggles to extend.

Table 3: Inference results on GolfDB. "Wrists mm" is the average distance between the wrists. "*fr*Elbow deg" is the average degree of the front elbow. "BS" indicates the backswing phase, "DS" for downswing, "FT" for follow-through. Round brackets indicate median values.

| | GP↓ | FL [% frames]↓ | TFS mm↓ *ba*BigToe,*fr*SmallToe,*fr*Heel | | | Wrists mm↓ BS | Wrists mm↓ DS | *fr*Elbow deg↑ BS | *fr*Elbow deg↑ DS | $F_Y$ {med, q99, max} | $F_Y^{fr} - F_Y^{ba}$ BS (med) | $F_Y^{fr} - F_Y^{ba}$ DS (med) | $F_Y^{fr} - F_Y^{ba}$ FT (med) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| PhysPT | **4.5** | 22.9 [7.5] | 424 | 417 | 387 | 156 | 160 | 144 | 124 | {0.08, 0.27, 1.11} | -0.054 (-0.049) | -0.049 (-0.035) | 0.054 (0.050) |
| MusclePose | 10.90 | **16.2** [**2.5**] | 737 | 778 | 777 | 174 | 245 | 142 | 123 | {0.48, 0.93, 1.44} | 0.005 (0.044) | -0.194 (-0.137) | 0.449 (0.321) |
| **ours: SPUD** | 13.3 | 18.6 [2.9] | **303** | **342** | **273** | **62** | **56** | **160** | **152** | {0.50, 1.00, 1.32} | -0.205 (-0.107) | -0.023 (-0.009) | 0.224 (0.122) |
| w/o $L_{golf}$ | 5.9 | 20.3 [13.7] | 786 | 707 | 722 | 164 | 133 | 144 | 123 | {0.50, 1.04, 1.75} | -0.347 (-0.146) | -0.277 (-0.137) | 0.241 (0.121) |

Kinematically, although GP and FL are slightly higher for SPUD, the output motion is more "golf-like", with less sliding (lower TFS), closer wrists, and a more extended elbow (closer to 180 degrees), as reported in columns 3-9 of Tab. 3, where the shaded row serves as an ablation to show the contributions of $L_{golf}$. Accordingly, from the pink circles in row 1 of Fig. 5, we see cases when MusclePose struggle to keep the wrists together during the top of the swing, and PhysPT during the bottom of the swing. And in row 2, the pink circles indicate when these two methods struggle to keep the front elbow extended during the swing.

## 5 Conclusion

We developed SPUD as a diffusion-based TTO framework for PHPE. SPUD extends the success of current methods that solve inverse kinematics tasks via variational diffusion sampling to inverse dynamics tasks, regularized by its novel kinetics prior. Additionally, SPUD is the first PHPE method that infers human dynamics for the SKEL model, making it more compatible for downstream biomechanics analysis. Through comprehensive experiments, we demonstrated SPUD's versatility on various inverse dynamics and motion denoising tasks, as well as its adaptability to disparate movements from different sports. Lastly, we also presented a new fencing dataset, to address the lack of publicly available force-annotated mocap video datasets for athletic human movements, and hope it can facilitate more robust evaluations for future PHPE research.

## AI USE STATEMENT

We **have not used** generative AI tools for any of the tasks with **required** disclosure, including: *generate synthetic data sets, help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, assist with translation, clean and reformat dataset, support qualitative and thematic data analysis, interpret results.*

Additionally, we **used** generative AI tools for the following tasks with **recommended** disclosure: *summarize or analyse existing literature, edit a research paper to improve readability, identify relevant literature, format references.*

We have reviewed all AI-assisted work. We have proofread all text generated and edited by LLMs for the tasks listed above. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

## REFERENCES

Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. Scape: Shape completion and animation of people. *ACM Transactions on Graphics*, 24 (3):408–416, 2005. doi: 10.1145/1073204.1073207.

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 *European Conference on Computer Vision (ECCV)*, volume 9909, pp. 561–578. Springer International Publishing, 2016. doi: 10.1007/978-3-319-46454-1_34.

Romain Brégier, Guénolé Fiche, Laura Bravo-Sánchez, Thomas Lucas, Matthieu Armando, Philippe Weinzaepfel, Grégory Rogez, and Fabien Baradel. Human mesh modeling for Anny body. In *Computer Vision – ECCV 2026*, pp. 170–188, Cham, 2026. Springer Nature Switzerland.

Yinpeng Chen, Zicheng Liu, and Zhengyou Zhang. Tensor-based human body modeling. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2013. doi: 10.1109/CVPR.2013.21.

Julian Chibane, Aymen Mir, and Gerard Pons-Moll. Neural unsigned distance fields for implicit function learning. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, pp. 21638–21652. Curran Associates, Inc., 2020.

Hai Ci, Mingdong Wu, Wentao Zhu, Xiaoxuan Ma, Hao Dong, Fangwei Zhong, and Yizhou Wang. GFPose: Learning 3d human pose prior with gradient fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. doi: 10.1109/CVPR52729.2023.00465.

Jurriaan H. de Groot and Ronald Brand. A three-dimensional regression model of the shoulder rhythm. *Clinical Biomechanics*, 16(9), 2001. ISSN 0268-0033. doi: 10.1016/S0268-0033(01)00065-1.

Paolo De Leva. Adjustments to zatsiorsky-seluyanov's segment inertia parameters. *Journal of Biomechanics*, 29(9), 1996. ISSN 0021-9290. doi: 10.1016/0021-9290(95)00178-6.

Raphaël Dumas and Janis Wojtusch. Estimation of the body segment inertial parameters for the rigid body biomechanical models used in motion analysis. In *Handbook of Human Motion*, volume 1–3. 2018. doi: 10.1007/978-3-319-14418-4_147.

Raphaël Dumas, Laurence Chèze, and Jean-Pierre Verriest. Adjustments to mcconville et al. and young et al. body segment inertial parameters. *Journal of Biomechanics*, 40(3):543–553, 2007. doi: 10.1016/j.jbiomech.2006.02.013.

Aaron Ferguson, Ahmed A. A. Osman, Berta Bescos, Carsten Stoll, Chris Twigg, Christoph Lassner, David Otte, Eric Vignola, Fabian Prada, Federica Bogo, et al. MHR: Momentum human rig, 2025. URL https://arxiv.org/abs/2511.15586.

Guénolé Fiche, Simon Leglaive, Xavier Alameda-Pineda, Antonio Agudo, and Francesc Moreno-Noguer. VQ-HPS: Human pose and shape estimation in a vector-quantized latent space. In *European Conference on Computer Vision (ECCV)*, pp. 471–490, 2024. doi: 10.1007/978-3-031-72943-0_27.

Erik Gartner, Mykhaylo Andriluka, Erwin Coumans, and Cristian Sminchisescu. Differentiable dynamics for articulated 3d human motion reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022a. doi: 10.1109/CVPR52688.2022.01284.

Erik Gartner, Mykhaylo Andriluka, Hongyi Xu, and Cristian Sminchisescu. Trajectory optimization for physics-based reconstruction of 3d human pose from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022b. doi: 10.1109/CVPR52688.2022.01276.

Zigang Geng, Chunyu Wang, Yixuan Wei, Ze Liu, Houqiang Li, and Han Hu. Human pose as compositional tokens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. doi: 10.1109/CVPR52729.2023.00071.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11), 2020. ISSN 1557-7317. doi: 10.1145/3422622.

Vladimir Guzov, Aymen Mir, Torsten Sattler, and Gerard Pons-Moll. Human POSEitioning system (HPS): 3d human pose estimation and self-localization in large scenes from body-mounted sensors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. doi: 10.1109/CVPR46437.2021.00430.

Nils Hasler, Carsten Stoll, Martin Sunkel, Bodo Rosenhahn, and Hans-Peter Seidel. A statistical model of human pose and body shape. *Computer Graphics Forum*, 28(2), 2009. ISSN 1467-8659. doi: 10.1111/j.1467-8659.2009.01373.x.

Herbert Hatze. A complete set of control equations for the human musculo-skeletal system. *Journal of Biomechanics*, 10(11–12), 1977. ISSN 0021-9290. doi: 10.1016/0021-9290(77)90094-X.

Herbert Hatze. A mathematical model for the computational determination of parameter values of anthropomorphic segments. *Journal of Biomechanics*, 13(10), 1980. ISSN 0021-9290. doi: 10.1016/0021-9290(80)90171-2.

Jonathan Ho, Ajay N. Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, 2020.

Chun Hao P. Huang, Hongwei Yi, Markus Hoschle, Matvey Safroshkin, Tsvetelina Alexiadis, Senya Polikovsky, Daniel Scharstein, and Michael J. Black. Capturing and inferring dense full-body human-scene contact. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. doi: 10.1109/CVPR52688.2022.01292.

Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6M: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7), 2014. ISSN 0162-8828. doi: 10.1109/TPAMI.2013.248.

Zhongyu Jiang, Zhuoran Zhou, Lei Li, Wenhao Chai, Cheng-Yen Yang, and Jenq-Neng Hwang. Back to optimization: Diffusion-based zero-shot 3d human pose estimation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2024. doi: 10.1109/WACV57701.2024.00603.

Sam Johnson and Mark Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *British Machine Vision Conference*, 2010.

Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018. doi: 10.1109/CVPR.2018.00744.

Marilyn Keller, Keenon Werling, Soyong Shin, Scott Delp, Sergi Pujades, C. Karen Liu, and Michael J. Black. From skin to skeleton: Towards biomechanically accurate 3d digital humans. *ACM Transactions on Graphics*, 42(6), 2023. ISSN 1557-7368. doi: 10.1145/3618381.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations (ICLR 2014)*, 2014.

Mary D. Klein Breteler, Cornelis W. Spoor, and Frans C. T. Van Der Helm. Measuring muscle and joint geometry parameters of a shoulder for modeling purposes. *Journal of Biomechanics*, 32, 1999. doi: 10.1016/S0021-9290(99)00122-0.

Tianye Li, Timo Bolkart, Michael J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4d scans. *ACM Transactions on Graphics*, 36(6), 2017. doi: 10.1145/3130800.3130813.

Wenhao Li, Hong Liu, Hao Tang, Pichao Wang, and Luc Van Gool. MHFormer: Multi-hypothesis transformer for 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022a. doi: 10.1109/CVPR52688.2022.01280.

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 *European Conference on Computer Vision (ECCV)*, 2022b.

Karen Liu and Sumit Jain. A quick tutorial on multibody dynamics, 2012.

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. In *ACM Transactions on Graphics*, volume 34, 2015. doi: 10.1145/2816795.2818013.

Junzhe Lu, Jing Lin, Hongkun Dou, Ailing Zeng, Yue Deng, Xian Liu, Zhongang Cai, Lei Yang, Yulun Zhang, Haoqian Wang, et al. DPoser-X: Diffusion model as robust 3d whole-body human pose prior. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9988–9997, 2025. doi: 10.1109/ICCV51701.2025.00931.

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*, 2019. doi: 10.1109/ICCV.2019.00554.

Morteza Mardani, Jiaming Song, Jan Kautz, and Arash Vahdat. A variational perspective on solving inverse problems with diffusion models. In *International Conference on Learning Representations (ICLR)*, 2024.

Julieta Martinez, Rayat Hossain, Javier Romero, and James J. Little. A simple yet effective baseline for 3d human pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017. doi: 10.1109/ICCV.2017.288.

William McNally, Kanav Vats, Tyler Pinto, Chris Dulhanty, John McPhee, and Alexander Wong. GolfDB: A video database for golf swing sequencing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019. doi: 10.1109/CVPRW.2019.00311.

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 *2017 International Conference on 3D Vision (3DV)*, 2017. doi: 10.1109/3DV.2017.00064.

Lea Müller, Vickie Ye, Georgios Pavlakos, Michael J. Black, and Angjoo Kanazawa. Generative proxemics: A prior for 3d social interaction from images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. doi: 10.1109/CVPR52733.2024.00925.

Ahmed A. A. Osman, Timo Bolkart, and Michael J. Black. STAR: Sparse trained articulated human body regressor. In *European Conference on Computer Vision (ECCV)*, pp. 598–613, 2020. doi: 10.1007/978-3-030-58539-6_36.

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. doi: 10.1109/CVPR.2019.01123.

Stanley Plagenhoef, F. Gaynor Evans, and Thomas Abdelnour. Anatomical data for analyzing human motion. *Research Quarterly for Exercise and Sport*, 54(2), 1983. ISSN 2168-3824. doi: 10.1080/02701367.1983.10605290.

Davis Rempe, Leonidas J. Guibas, Aaron Hertzmann, Bryan Russell, Ruben Villegas, and Jimei Yang. Contact and human dynamics from monocular video. In *19th ACM SIGGRAPH/Eurographics Symposium on Computer Animation (SCA)*, 2020.

Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J. Guibas. HuMoR: 3d human motion model for robust pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. doi: 10.1109/ICCV48922.2021.01129.

Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International Conference on Machine Learning (ICML)*, volume 2, 2015.

Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics*, 36, 2017. doi: 10.1145/3130800.3130883.

Jun Saito, Jiefeng Li, Michael de Ruyter, Miguel Guerrero, Edy Lim, Ehsan Hassani, Roger Blanco Ribera, Hyejin Moon, Magdalena Dadela, Marco Di Lucca, et al. SOMA: Unifying parametric human body models, 2026. URL https://arxiv.org/abs/2603.16858.

Igor Santesteban, Elena Garces, Miguel A. Otaduy, and Dan Casas. SoftSMPL: Data-driven modeling of nonlinear soft-tissue dynamics for parametric humans. *Computer Graphics Forum*, 39(2), 2020. ISSN 1467-8659. doi: 10.1111/cgf.13912.

Katherine R. Saul, Xiao Hu, Craig M. Goehler, Meghan E. Vidt, Melissa Daly, Anca Velisar, and Wendy M. Murray. Benchmarking of dynamic simulation predictions in two software platforms using an upper limb musculoskeletal model. *Computer Methods in Biomechanics and Biomedical Engineering*, 18(13), 2015. ISSN 1476-8259. doi: 10.1080/10255842.2014.916698.

David Schneider, Simon Reiß, Marco Kugler, Alexander Jaus, Kunyu Peng, Susanne Sutschet, M. Saquib Sarfraz, Sven Matthiesen, and Rainer Stiefelhagen. Muscles in time: Learning to understand human motion by simulating muscle activations. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-2147.

Yonatan Shafir, Guy Tevet, Roy Kapon, and Amit H. Bermano. Human motion diffusion as a generative prior. In *International Conference on Learning Representations (ICLR)*, 2024.

Soshi Shimada, Vladislav Golyanik, Weipeng Xu, and Christian Theobalt. PhysCap. *ACM Transactions on Graphics*, 39(6), 2020. doi: 10.1145/3414685.3417877.

Soshi Shimada, Vladislav Golyanik, Weipeng Xu, Patrick Pérez, and Christian Theobalt. Neural monocular 3d human motion capture with physical awareness. *ACM Transactions on Graphics*, 40(4), 2021. doi: 10.1145/3450626.3459825.

Soyong Shin, Juyong Kim, Eni Halilaj, and Michael J. Black. WHAM: Reconstructing world-grounded humans with accurate 3d motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. doi: 10.1109/CVPR52733.2024.00202.

Leonid Sigal, Alexandru O. Balan, and Michael J. Black. HumanEva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International Journal of Computer Vision*, 87(1–2), 2010. ISSN 0920-5691. doi: 10.1007/s11263-009-0273-6.

Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning (ICML)*, volume 3, 2015.

Garvita Tiwari, Dimitrije Antic, Jan Eric Lenssen, Nikolaos Sarafianos, Tony Tung, and Gerard Pons-Moll. Pose-NDF: Modeling human pose manifolds with neural distance fields. In *European Conference on Computer Vision (ECCV)*, pp. 572–589. Springer, 2022. doi: 10.1007/978-3-031-20065-6_33.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *Advances in Neural Information Processing Systems*, 2017.

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 *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 614–631, 2018. doi: 10.1007/978-3-030-01249-6_37.

Jingbo Wang, Sijie Yan, Yuanjun Xiong, and Dahua Lin. Motion guided 3d pose estimation from videos. In *European Conference on Computer Vision (ECCV)*, pp. 764–780, 2020. doi: 10.1007/978-3-030-58586-0_45.

David A. Winter. *Biomechanics and Motor Control of Human Movement*. 4 edition, 2009. doi: 10.1002/9780470549148.

Ge Wu, Sorin Siegler, Paul Allard, Chris Kirtley, Alberto Leardini, Dieter Rosenbaum, Mike Whittle, Darryl D. D'Lima, Luca Cristofolini, Hartmut Witte, et al. Isb recommendation on definitions of joint coordinate system of various joints for the reporting of human joint motion – part i: Ankle, hip, and spine. *Journal of Biomechanics*, 35(4):543–548, 2002. doi: 10.1016/S0021-9290(01)00222-6.

Ge Wu, Frans C. T. Van der Helm, H. E. J. Veeger, Mohsen Makhsous, Peter Van Roy, Carolyn Anglin, Jochem Nagels, Andrew R. Karduna, Kevin McQuade, Xuguang Wang, et al. Isb recommendation on definitions of joint coordinate systems of various joints for the reporting of human joint motion – part ii: Shoulder, elbow, wrist and hand. *Journal of Biomechanics*, 38(5):981–992, 2005. doi: 10.1016/j.jbiomech.2004.05.042.

Yan Xia, Xiaowei Zhou, Etienne Vouga, Qixing Huang, and Georgios Pavlakos. Reconstructing humans with a biomechanically accurate skeleton. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. doi: 10.1109/CVPR52734.2025.00504.

Kevin Xie, Tingwu Wang, Umar Iqbal, Yunrong Guo, Sanja Fidler, and Florian Shkurti. Physics-based human motion estimation and synthesis from videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. doi: 10.1109/ICCV48922.2021.01133.

Andrei Zanfir, Eduard Gabriel Bazavan, Hongyi Xu, William T. Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Weakly supervised 3d human pose and shape reconstruction with normalizing flows. In *European Conference on Computer Vision (ECCV)*, pp. 465–481, 2020.

Yufei Zhang, Jeffrey O. Kephart, Zijun Cui, and Qiang Ji. PhysPT: Physics-aware pretrained transformer for estimating human dynamics from monocular videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2305–2317, 2024.

Kevin Zhu, AliAsghar MohammadiNasrabadi, Alexander Wong, and John McPhee. 3d human pose estimation with muscles. In *Advances in Neural Information Processing Systems*, 2025.

Silvia Zuffi and Michael J. Black. The stitched puppet: A graphical model of 3d human shape and pose. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2015. doi: 10.1109/CVPR.2015.7298976.

# A APPENDIX

## A.1 MOTION DENOISING

Lastly, we followed the same motion denoising experiments in (Lu et al., 2025) to compare DBone's inverse kinematics results to SMPL-based pose priors Pose-NDF (Tiwari et al., 2022) and DPoserX (Lu et al., 2025). Specifically, given noisy or occluded SMPL joints ($y = g(\mathbf{P}_{SMPL})$), we aim to recover SKEL body pose parameters $\mathbf{x}_B$, whereas the other methods recover SMPL's body pose.

For **noisy inputs** ($y = \mathbf{P}_{SMPL} + \epsilon$) we added random noise, with means of 20, 40, 100 mm, to the SMPL joint positions of the test split (Rempe et al., 2021) of AMASS (Mahmood et al., 2019), which includes HumanEva (Sigal et al., 2010), Transitions (Mahmood et al., 2019)) and HPS (Guzov et al., 2021), and reported denoising results in Tab. 4.

For **partial/occluded inputs** ($y = Mask(\mathbf{P}_{SMPL})$) we masked the positions of the SMPL joints in the legs and left arm of the test split of AMASS, and reported denoising results in Tab. 5.

Table 4: MJE (in mm) of motion denoising results for different noise scales on the AMASS test set and HPS.

| | AMASS test | | | HPS | | |
|---|---|---|---|---|---|---|
| (noise scale mm) | 20 | 40 | 100 | 20 | 40 | 100 |
| w/o prior | 15.3 | 24.2 | 51.5 | 16.3 | 23.7 | 50.9 |
| Pose-NDF | 13.8 | 22.1 | 46.1 | 15.6 | 21.6 | 47.5 |
| DPoserX | **13.6** | 19.9 | **33.2** | **13.5** | 20.5 | 35.3 |
| **ours: SPUD** | 15.1 | **18.9** | 33.3 | 13.8 | **14.7** | **25.2** |

Table 5: MJE (in mm) of motion denoising results with different body part occlusion on the AMASS test set. Vis: visible joints. Occ: occluded joints. MVE: mean vertex error in mm.

| | legs | | | | left arm | | | |
|---|---|---|---|---|---|---|---|---|
| | Vis. | Occ. | All | MVE | Vis. | Occ. | All | MVE |
| w/o prior | 2.6 | 147.2 | 55.2 | 54.5 | 2.6 | 248.7 | 47.4 | 99.1 |
| Pose-NDF | **2.5** | 157.1 | 58.7 | 56.4 | **2.5** | 177.0 | 34.2 | 78.6 |
| DPoserX | 2.8 | **122.4** | **46.3** | **36.5** | 2.7 | **78.0** | **16.4** | **38.1** |
| **ours: SPUD** | 5.0 | 182.2 | 69.4 | 57.3 | 4.5 | 110.4 | 23.8 | 61.7 |

Tab. 4 shows that SPUD's denoising results are on-par with SOTA kinematic pose priors, with competitive results on AMASS and outperforming DPoserX on HPS. However, from Tab. 5, SPUD's performance seems to suffer with increased occlusion. Our overall worse performance on the AMASS test split could also be partially attributed to DBone being trained on the HSMR training set, which largely contains individual images, whereas Pose-NDF and DPoserX were both trained on the motion frames of the AMASS train split. Although seemingly statistically significant, the $\sim$ 2mm increase in error for the visible joints may not be practically significant, and could stem from the potential fitting error between SKEL and SMPL discussed in Sec. 4.1.

## A.2 ADDITIONAL RESULTS

Additional output examples are included in the supplementary material. This includes a fencing video in which the individual in the background was manually censored in post to preserve anonymity for submission (the participating subject wore a mask that obscured identifiable facial features). The remaining video examples are from public datasets.