# ProbSIP: Probabilistic Modeling for Ambiguity-Reduced Sparse Inertial Poser

Shanyan Guan[1] Yunbo Wang[1*] Xintao Lv[1] Yanhao Ge[2] Xiaokang Yang[1]

[1] MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University
[2] vivo Mobile Communication Co. Ltd

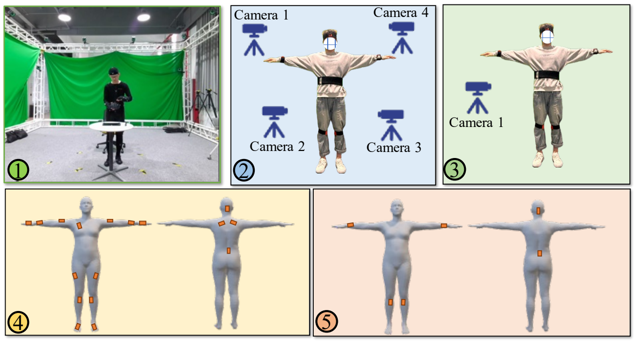
## Abstract

Human motion capture is critical for applications like VR/AR, gaming, and motion detection, requiring accuracy, convenience, and user comfort. Traditional methods, like marker-based optical systems, offer accuracy but lack convenience and comfort. Inertial Measurement Units (IMUs) provide a solution, but their dense placement on the body is cumbersome. This work introduces ProSIP, a sparse IMU-based approach using only six IMUs for motion capture. ProSIP uses a probabilistic modeling approach inspired by Variational Autoencoders to address the ambiguity that sparse IMU input might match different body motions. The core idea of ProSIP is to align the representation between sparse IMU data and full-body motion, which thus can significantly reduce ambiguity. As verified on DIP-IMU and TotalCapture datasets, ProSIP can accurately capture complex motions, such as 'sitting and grasping'.

## Introduction

Human motion capture, which aims to capture 3D human movements, is crucial for many applications, such as VR/AR (Vidal et al. 2018), gaming (Stoeve et al. 2021), and motion detection (Alarfaj, Qian, and Liu 2021). These applications impose three requirements on human motion capture: (1) it must **accurately** capture 3D body movement, (2) it should be **flexible**, enabling capture anywhere, and (3) it must be **nonintrusive** for users.

In Fig. 1, we have a brief illustration of the motion capture system that is commonly used. The most accurate way to capture human motion is through marker-based optical motion capture systems like Vicon (vic 2014). However, these systems are limited to controlled studios where users must wear special clothing and attach multiple markers to their clothing. Marker-less multi-camera systems (Kocabas, Karagoz, and Akbas 2019) offer decent accuracy, but they require camera calibration, time synchronization, and are not easily portable. Single-camera motion capture systems (Guan et al. 2021; Bogo et al. 2016; Kanazawa et al. 2018) are more flexible and widely applicable, yet they still lack accuracy. Most importantly, vision-based systems, whether optical or multi-/single-camera motion capture systems, perform poorly when visibility is limited due to occlusion, dim light, or motion blur. By contrast, mounting inertial measurement units (IMUs) directly on the user's body overcomes the need for line-of-sight. IMUs can capture the orientation and acceleration of limbs, which makes them ideal for capturing body movements. Commercial IMU-based motion capture systems (Schepers et al. 2018) place multiple IMUs on each limb to comprehensively capture the whole body's movements. However, this dense placement of IMUs (usually more than 17) is inflexible and results in a poor user experience because of the need for tight-fitting and unnonintrusive clothing.



|  | Category | Accuracy | Flexible | Nonintrusive |
|---|---|---|---|---|
| Vision-based | ① Marker-based | ✓ | ✗ | ✗ |
|  | ② Single-camera | ✗ | ✓ | ✓ |
|  | ③ Multi-camera | ✓ | ✗ | ✗ |
| IMU-based | ④ Dense IMUs | ✓ | ✓ | ✗ |
|  | ⑤ Six IMUs | ✓ | ✓ | ✓ |

Figure 1: Comparisons between vision-based and IMU-based motion capture systems.

Some recent works (Marcard et al. 2017; Huang et al. 2018) have proposed the task of **sparse inertial poser**, which only binds the IMU at six key points, including the lower leg, wrist, head, and pelvis, to realize accurate, flexible, and nonintrusive motion capture. Reducing the number of IMUs required for motion capture technology makes it more practical, but tracking a few limbs introduces ambiguity in the inference process, called **one-to-many mapping**

[*] Corresponding Author

**ambiguity**: a single segment of IMU data can correspond to multiple possible 3D motions. Existing works (Marcard et al. 2017; Huang et al. 2018; Yi, Zhou, and Xu 2021) use the temporal context of sparse IMU data to reduce ambiguity, but this approach is not effective for complex actions. The reason is that sparse IMU data cannot fully describe body movement, even when fed into the network temporally.

The main focus of this work is to address the problem of one-to-many mapping ambiguity in sparse inertial motion inference caused by input data sparsity. To reduce this ambiguity, we propose a probabilistic modeling paradigm called ProSIP. The key to reducing ambiguity in sparse human motion estimation is to minimize the difference between the representation of sparse IMU data and that of full-body motion. To do this, we draw inspiration from the Variational Autoencoder (VAE) and use ProSIP to estimate the prior distribution of full-body motion representations based on sparse IMU data. We then estimate the posterior distribution of motion using actual full-body motion data (such as SMPL posture parameters) as input. Finally, we minimize the distance between the two distributions to achieve representation alignment between sparse IMU data and full-body motion. This method differs from traditional VAEs in that the prior distribution changes with the IMU data, allowing for more accurate modeling of real-world dynamics. We evaluated ProSIP on two benchmarks: the DIP-IMU dataset and the TotalCapture dataset. The results show that ProSIP consistently outperforms existing approaches and demonstrates the effectiveness of representation alignment.

## Related Work

**Inertial Motion Capture.** Attaching an Inertial Measurement Unit (IMU) to a limb can capture the acceleration and direction of that limb's movement (Foxlin 1996; Bachmann et al. 2001; Roetenberg et al. 2005; Del Rosario et al. 2018; Vitali, McGinnis, and Perkins 2020). However, commercial inertial motion capture systems like Xsens (Schepers et al. 2018) use multiple IMUs to track the user's full-body movement. This can be intrusive, so reducing the number of IMUs is preferable. However, sparse IMUs make reconstructing human posture challenging. Early methods (Slyper and Hodgins 2008; Tautges et al. 2011; Riaz et al. 2015; Schwarz, Mateus, and Navab 2009) relied on a Lazy Learning strategy (Aha 1997) to retrieve the most similar action from a human motion database based on the similarity of acceleration as the prediction result, which limited the accuracy of the motion capture effects. Recent research (Marcard et al. 2017; Huang et al. 2018; Yi, Zhou, and Xu 2021; Jiang et al. 2022) has demonstrated that better human motion capture can be achieved by inputting both acceleration and direction. However, existing methods ignore the ambiguity of sparse IMU data and rely solely on data-driven approaches, which can lead to imprecision inference for untracked limb attitudes. To address this issue, this work proposes a probabilistic modeling paradigm that makes the motion representation of sparse IMU data closer to the representation of whole-body motion, resulting in more accurate inference results.

**Representation Learning with Generative Models** Representation learning researches aim to automatically learn effective data representation for tasks. Early methods rely on hand-designed features (Lowe 1999; Sivic and Zisserman 2008). Generative models (Goodfellow et al. 2014; Kingma and Welling 2014) make progress in this field. Variational Autoencoder (VAE) (Kingma and Welling 2014) is used for representation learning. It encodes data into latent code and decodes them to reconstruct the original data. Reconstruction loss and KL divergence loss ensure the quality of data reconstruction and continuity of the latent space. VAEs have been widely used to model data representations, *e.g.*, image distribution (Higgins et al. 2016; Denton and Fergus 2018) and skeletal sequence representations (Ling et al. 2020).

## Methodology

Similar to previous literature (Huang et al. 2018), the sparse inertial poser system places six IMUs on the legs, arms, pelvis, and head, as shown in Fig. 3. The six IMUs capture and utilize the acceleration $\boldsymbol{a} \in \mathbb{R}^3$ and orientation $\boldsymbol{R} \in \mathbb{R}^3$ of each IMU as its fundamental input. The overall input is defined as $\boldsymbol{X}_t = [\boldsymbol{a}_t^{1:6}, \boldsymbol{R}_t^{1:6}]$. The goal of sparse inertial poser is to estimate complete body posture $\boldsymbol{Y}_t \in \mathbb{R}^{J\times 3}$, where $J$ is the number of skeleton joints. The problem is that many limbs of the body are not directly observable, resulting in *one-to-many mappings* from the observed sparse IMU data $\boldsymbol{X}_t$ to $\boldsymbol{Y}_t$.

To tackle the one-to-many mapping ambiguity in the sparse inertial poser, we propose a probabilistic modeling paradigm, named **ProSIP**. ProSIP maps $\boldsymbol{X}$ and $\boldsymbol{Y}$ to the motion distribution space for representation, and then reduces the mapping ambiguity by minimizing the distance between the two representations. The fundamental concept of ProSIP is that whether the data is in the form of sparse IMU readings represented by $\boldsymbol{X}$ or a complete body posture represented by $\boldsymbol{Y}$, both provide different perspectives of body movement. Therefore, in the representation space of body movement, the features of both types of data should be consistent with each other. Next, we introduce the overall framework and then describe the training objectives.

### Overall Framework Design

The overall framework of ProSIP is shown in Fig. 2, mainly composed of three modules: (1) Prior Encoding Module, (2) Posterior Encoding Module, (3) Dynamic Decoding Module. ProSIP uses $\boldsymbol{z}_t$ as the representation of the motion distribution.

**Prior Encoding Module** This module takes IMU measurement data $\boldsymbol{X}_t$ as input and estimates the prior distribution of $\boldsymbol{z}_t$, $p_\phi(\boldsymbol{z}_t|\boldsymbol{X}_t)$. The formal definition of the encoding module is:

$$\boldsymbol{z}_t \sim p_\phi(\boldsymbol{z}_t|\boldsymbol{X}_t). \tag{1}$$

The prior encoding module includes an encoder $\mathcal{E}_\phi$ that encodes IMU data $\boldsymbol{X}_t$ into hidden space, a bidirectional LSTM $\mathcal{M}_\phi$ that captures long-distance temporal dependencies, and a fully-connected layer $\mathcal{F}_\phi$ that outputs the mean $\mu_\phi(t)$ and variance $\sigma_\phi(t)$ of $q_\phi(\boldsymbol{z}_t|\boldsymbol{X}_t, \boldsymbol{Y}_t)$.

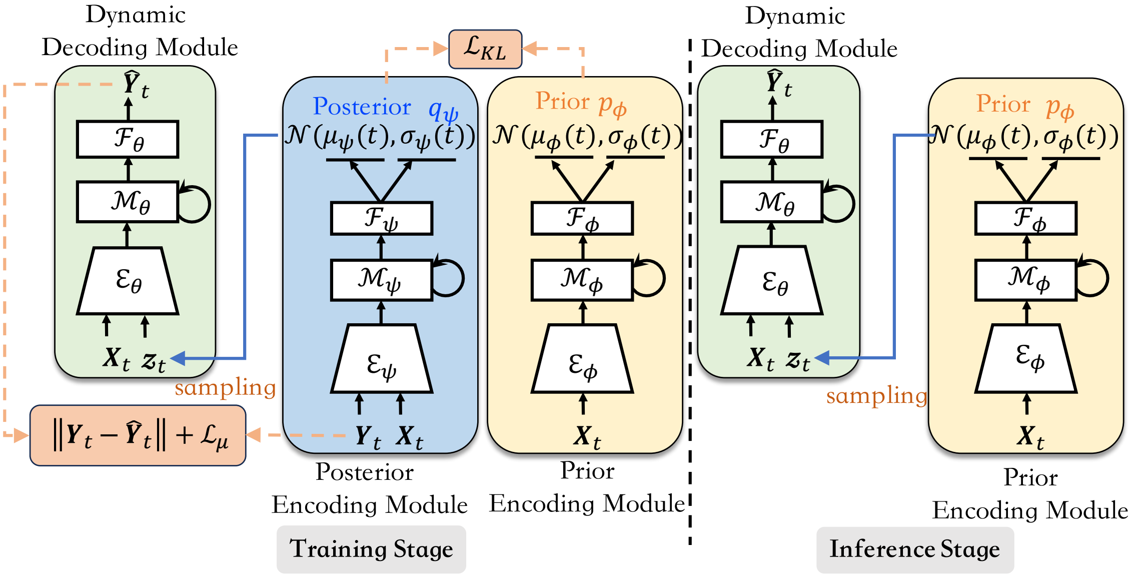


Figure 2: The overall framework of the proposed method. $\boldsymbol{X}_t$ is the IMU measurements, and $\boldsymbol{Y}_t$ is the ground-truth annotation of full body posture.

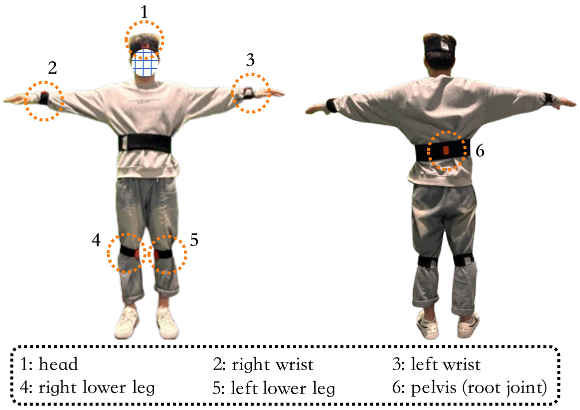


Figure 3: Illustration of the IMU placement.

**Posterior Encoding Module** This module takes human posture $\boldsymbol{Y}_t$ and IMU measurement data $\boldsymbol{X}_t$ as input, and calculates the posterior distribution of $\boldsymbol{z}_t$, $q_\psi(\boldsymbol{z}_t|\boldsymbol{X}_t, \boldsymbol{Y}_t)$. Following common practices in variational inference, we model $q_\psi(\boldsymbol{z}_t|\boldsymbol{X}_t, \boldsymbol{Y}_t)$ with a Gaussian distribution $\mathcal{N}(\mu_\psi, \sigma_\psi)$. The formal definition of the representation module is as follows:

$$\boldsymbol{z}_t \sim q_\psi(\boldsymbol{X}_t, \boldsymbol{Y}_t). \tag{2}$$

In this module, an encoder $\mathcal{E}_\psi$, which maps IMU data $\boldsymbol{X}_t$ and body posture $\boldsymbol{Y}_t$ to hidden space, then a bidirectional LSTM module $\mathcal{M}_\psi$ to capture long-distance temporal dependencies, and finally a fully-connected layer $\mathcal{F}_\psi$ to output the mean $\mu_\psi(t)$ and variance $\sigma_\psi(t)$ of $q_\psi(\boldsymbol{z}_t|\boldsymbol{X}_t, \boldsymbol{Y}_t)$.

**Dynamic Decoding Module** This module uses IMU measurement data $\boldsymbol{X}_{1:T}$ and hidden variables $\boldsymbol{z}_{1:T}$ as input to estimate whole-body motion $\hat{\boldsymbol{Y}}_{1:T}$. The formal definition is:

$$\hat{\boldsymbol{Y}}_{1:T} = p_\theta(\boldsymbol{X}_{1:T}, \boldsymbol{z}_{1:T}). \tag{3}$$

During training, $\boldsymbol{z}_t$ is sampled from the posterior distribution $q_\psi$, while during testing, it is drawn from the prior distribution $p_\phi$. The dynamic module first uses an encoder $\mathcal{E}_\theta$ to map $\boldsymbol{X}_{1:T}$ and $\boldsymbol{z}_{1:T}$ to the same hidden space, then uses a bidirectional LSTM $\mathcal{M}_\theta$ for temporal feature extraction, and finally uses a multi-layer perception (MLP) $\mathcal{F}_\theta$ to estimate the pose parameters $\hat{\boldsymbol{Y}}_{1:T}$ of SMPL.

## Training Objectives

In order to minimize the mapping ambiguity between sparse IMU data and whole body motion data, this paper constraints the distance between the posterior distribution $q_\psi$ and the prior distribution $p_\phi$ of $\boldsymbol{z}_t$. Similar to training VAE, the distance between probability distributions is minimized by minimizing the following constraint function:

$$\begin{aligned}\mathcal{L}_{dist} = & \frac{1}{T}\|\boldsymbol{Y}_{1:T} - p_\theta(\boldsymbol{X}_{1:T}, \boldsymbol{z}_{1:T})\| \qquad (4)\\ & + \lambda_{KL}\frac{1}{T}\sum_{t=1}^{T}\mathcal{L}_{KL}(q_\psi(\boldsymbol{z}_t|\boldsymbol{X}_t, \boldsymbol{Y}_t)\|p_\phi(\boldsymbol{z}_t|\boldsymbol{X}_t)), \qquad (5)\end{aligned}$$

where $\lambda_{KL}$ is the weight of the KL divergence $\mathcal{L}_{KL}$. We use the reparameterization trick (Kingma and Welling 2014) for training. In addition, since our task is a regression task rather than a generative task, we constraint the human motion estimated by the mean $\mu_\psi$ of the posterior distribution $q_\psi$ to be the same as the actual motion, that is:

$$\mathcal{L}_{exp} = \|\boldsymbol{Y}_{1:T} - p_\theta\left(\boldsymbol{\mu_\psi(t)}\right)\| \tag{6}$$

Therefore, the final training function of the model is:

$$\mathcal{L}_{dist} + \lambda_{exp}\mathcal{L}_{exp} \tag{7}$$

where $\lambda_{exp}$ is the weight of $\mathcal{L}_{exp}$.

**Model Implementation** Consistent with the previous work (Huang et al. 2018; Yi, Zhou, and Xu 2021), this paper ignores the rotation of SMPL's ankles, wrists, soles, and palms. In the entire framework, the joint rotation of SMPL and the direction measured by the IMU are both converted to the coordinate system with the SMPL root node (i.e., the pelvis, Pelvis) as the origin, expressed as a $3 \times 3$ rotation matrix. In the representation module, the encoder $\mathcal{E}_{\psi}$ is a fully connected layer stack with residual connection, and the output feature dimension is 256. The bidirectional LSTM $\mathcal{M}_{\psi}$ is an LSTM with two layers, with an output dimension of 256. The output dimension of the fully connected layer $\mathcal{F}_{\psi}$ is 2048, and the lengths of the mean $\mu_{\psi}$ and variance $\sigma_{\psi}$ are both 1024. The structure of the representation module is like the representation module, except that the input of the encoder $\mathcal{E}_{\phi}$ is only the IMU data $\boldsymbol{X} \in \mathbb{R}^{5\times(9+3)}$. For the dynamic module, $\mathcal{E}_{\theta}$ maps the IMU data and $\boldsymbol{z}_t \in \mathcal{R}^{1024}$ to a feature of length 256. The bidirectional LSTM module $\mathcal{M}_{\theta}$ contains two layers of LSTM, with an output dimension of 256. The multilayer perceptron is a three-layer fully connected layer with residual connections, outputting SMPL's pose parameters $\boldsymbol{Y} \in \mathbb{R}^{15\times 9}$. In the inference stage, this paper uses the mean value of the prior distribution as the input to the dynamic module to infer the overall posture.

# Experiment

## Experimental Setup

**Dataset** Same as the previous methods (Huang et al. 2018; Yi, Zhou, and Xu 2021), this method trains the model on the training subset of DIP-IMU and AMASS, and then tests the model performance on the testing subset of DIP-IMU and TotalCapture. The introduction of the datasets is as follows:

- *DIP-IMU (Huang et al. 2018)* is a dataset captured in real scenarios, collecting IMU data of 10 actors (9 males and 1 female) wearing Xsens. The total action time is approximately 90 minutes, including simple actions like walking, raising hands, squatting, grabbing, etc. DIP-IMU uses SIP (Marcard et al. 2017) to fit Xsens's 17 IMU time-series data to obtain SMPL annotations.
- *AMASS (Mahmood et al. 2019)* is a synthetic large-scale human motion dataset. AMASS collects Mocap raw data in different formats from multiple existing raw datasets, then obtains SMPL annotations that fit Mocap data through MoSh (Loper, Mahmood, and Black 2014). AMASS contains motion data of over 300 actors with a total time of over 40 hours. As the DIP-IMU dataset is relatively small, we adopt AMASS to co-train ProSIP. Specifically, virtual IMU sensors are placed on the SMPL model to calculate IMU's direction and acceleration.
- *TotalCapture (Trumble et al. 2017)* collects multimodal data of 5 actors in a studio, including IMU data, multi-view video, and Mocap data of Vicon. All these data have been synchronized over time. The SMPL annotations are also obtained by using SIP (Marcard et al. 2017).

**Single-stage Training Strategy** Taking into account that the synthesized IMU data on AMASS cannot simulate measurement noises, IMU location drifts, etc. that appear in real situations, previous methods often used a two-stage method for training: pre-training on AMASS to obtain a better initial model, and then finetuning on DIP-IMU to adapt the model to the real IMU data. This two-stage training method is very complex and requires a fine adjustment of the training times to avoid underfitting AMASS to cause the initialization model to not learn enough, or overfitting DIP-IMU causing a catastrophic forgetting of prior knowledge. For the problem that two stages are difficult to control the learning strength, this paper directly trains on DIP-IMU and AMASS at the same time and controls the problem of overfitting by controlling the ratio of them in the same batch of data (Batch). In the experiments, the data of AMASS occupies $80\%$, and the data of DIP-IMU occupies $20\%$. The ablation experiment proves that within a certain range of ratios, the model's performance does not fluctuate greatly. The method uses Adamax (Kingma and Ba 2015) as the optimizer, the learning rate is set to $0.001$, $\beta_1$ and $\beta_2$ are 0.9 and 0.999 respectively, and the weight decay coefficient is $0.0001$. The batch size (Batchsize) is set to 128. The weights of the loss functions $\mathcal{L}_{KL}$ and $\mathcal{L}_{exp}$ are $0.001$ and $1.0$ respectively. This paper uses the early stopping method (Early Stopping) to control the training rounds of the model, and the model is trained for about 10 hours on the 3080 GPU.

**Comparison Method and Evaluation Metrics** This paper selects DIP (Huang et al. 2018) and TransPose (Yi, Zhou, and Xu 2021) as comparison methods. DIP is the first to use deep neural networks to solve the problem of sparse inertia-based human motion inference, and TransPose is the best-performing model at that time, using a multi-stage regression framework: firstly regressing the skeletal position according to the IMU data, and then regressing the body posture according to the skeletal position. This paper quantitatively evaluates the model's performance from three aspects:

- The predictive accuracy of the joint angle (Joints Angular Error, JAE), measuring the average rotation error of all body critical points in the SMPL coordinate system.
- Predictive accuracy of joint positions (Joint Position Error, JPE), measuring the average positional error of all body keypoints in the root node coordinate system.
- Predictive accuracy of mesh vertex positions (Vertex Position Error, VPE), measuring the average positional error of all vertices on the body mesh in the root node coordinate system. Compared to JPE, VPE can reflect the accuracy of the skeleton self-spin.

## Quantitative Evaluation

Quantitative experiments are conducted with this paper using the online testing setup, and the results are shown in Tab. 1. In the online test setting, the test data arrives frame by frame, which is more in line with most practical application scenarios. Consistent with the comparative methods (Huang et al. 2018; Yi, Zhou, and Xu 2021), this paper uses 20 frames of historical data, one frame of current data

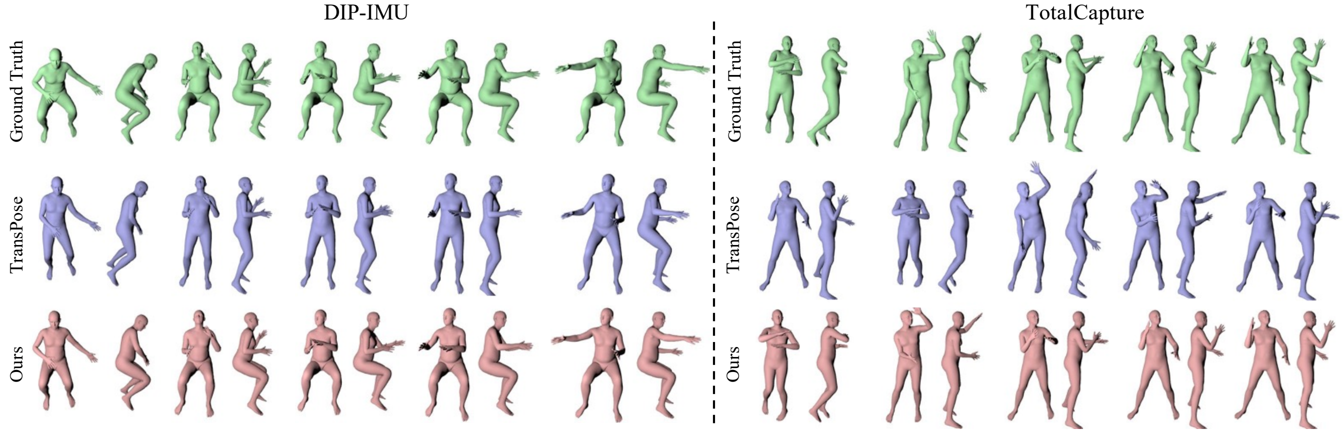


Figure 4: Qualitative comparisons between our method and TransPose on DIP-IMU and TotalCapture. Both font-view and side-view results are shown.

Table 1: Quantitative results on DIP-IMU and TotalCapture under the online setting.

| | DIP-IMU | | | TotalCapture | | |
|---|---|---|---|---|---|---|
| Method | JAE (deg) | JPE (cm) | VPE (cm) | JAE (deg) | JPE (cm) | VPE (cm) |
| DIP (Huang et al. 2018) | 15.2 (±8.5) | 7.3 (±4.2) | 9.0 (±5.0) | 17.5 (±10.1) | 9.6 (±6.0) | 11.4 (±6.9) |
| TransPose (Yi, Zhou, and Xu 2021) | 8.9 (±4.8) | 6.0 (±3.7) | 7.1 (±4.2) | 12.9 (±6.2) | 6.6 (±3.9) | 7.5 (±4.4) |
| Ours | **7.7 (±2.0)** | **5.3 (±1.7)** | **6.2 (±1.9)** | **12.8 (±4.7)** | **6.6 (±2.5)** | **7.4 (±3.0)** |

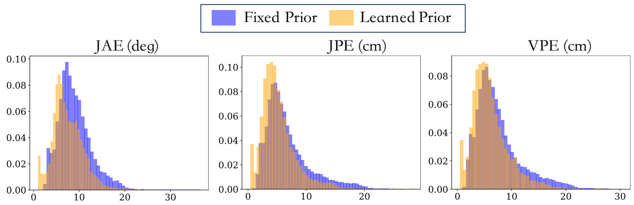


Figure 5: Learned prior (ours) *vs.* fixed prior (standard VAE). Y-axis: ratio of samples (0-1).

and 5 frames of future data. This paper selects DIP (Huang et al. 2018) and TransPose (Yi, Zhou, and Xu 2021) as the comparative methods and conducts a quantitative comparison on the benchmark test sets DIP-IMU and TotalCapture. Two observations can be made from Tab. 1: 1. On DIP-IMU, this method significantly outperforms the comparison methods in all evaluation indexes, for example, compared with TransPose (the best performing method currently), the average JAE of this method is reduced by 1.2 degrees (deg), the performance is improved by 13.8%, which is a large improvement in this field. On TotalCapture, the average error of this method in all evaluation indexes is close to TransPose, but the variance of all indexes is much smaller than that of TransPose, for example, the variance of JAE is reduced by 1.5 degrees, which indicates that the stability of this method is better than that of TransPose, this is a very important advantage for practical applications. In the two benchmark test sets, the average error and variance of this method are far lower than that of DIP. The results of the quantitative experiment verify the advantages of the information enhancement method based on Bayesian posterior constraint, this method can make the motion representation obtained from sparse IMU data more similar to the whole body motion, thereby improving the rationality and accuracy of the inference results.

## Qualitative Evaluation

**Qualitative Comparison on DIP-IMU** This section first qualitatively compares this method with TransPose on the DIP-IMU dataset using the online testing setup, and the experimental results are as shown in Fig. 4. The first example shown is ”sitting and grabbing.” From Figure 4 (top), it can be observed that TransPose did not correctly predict the posture of the thigh (no IMU tracking), resulting in its action being closer to ”standing and grabbing.” However, this method is able to accurately infer the posture of the thigh due to the enhancement of the motion representation of sparse IMU data with Bayesian posterior constraint. In addition, from the figure, it can be seen that the arm movement of TransPose is also not as accurate as this method. The second example is ”raising and waving both arms”, which is also a commonly performed action in daily life. From Fig. 4 (bottom), it can be seen that TransPose estimated the posture of the arm incorrectly, resulting in the action of the hands being significantly inconsistent with the actual movement. As observed from the front view and side view, the method of

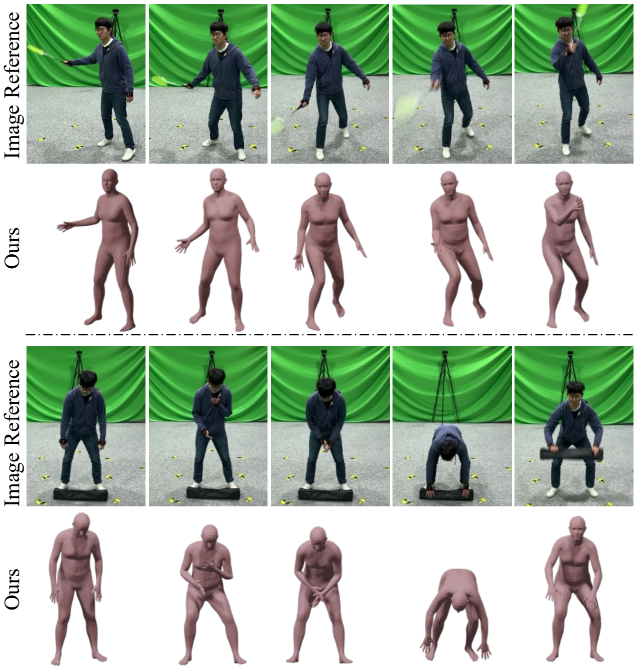


Figure 6: Live demo of our method

this paper is very close to the actual movement.

**Qualitative Comparison on TotalCapture** Then, this section conducted a qualitative comparison between this method and TransPose on TotalCapture, and the results are shown in Fig. 4. As can be seen from the figures (especially the third column), for skeletons with IMU tracking, TransPose can accurately predict their posture, but for skeletons without IMU tracking, such as the upper arm, its results are significantly worse than this method. This also verifies that, with the help of Bayesian posterior constraint for representation enhancement, this method can more accurately infer human motion on datasets that have not been pre-trained.

**Live Demo** As shown in Fig. 6, this paper evaluates the effectiveness of this method in a real environment (based on an online testing setup). The current benchmark tests are mainly simple actions, for example, DIP-IMU is mainly waving hands, lifting legs and walking, and TotalCapture is also mainly walking and running. In this experiment, this paper selects two actions with a large body movement range, ”playing badminton” and ”weightlifting”, for testing. This paper uses Xsens’s IMU sensors to collect inertia data, but it needs to be emphasized that this paper still maintains the ”sparse tracking” setting, that is, only data from 6 sensors are used. As can be seen from the figure, even for fast movements (playing badminton) and bending over to pick up, which can easily cause sensor offset and complex limb dynamics, this method can still accurately infer the three-dimensional movement of the body. This again validates the effectiveness of information-enhanced inference: enhancing the motion representation of sparse IMU based on Bayesian posterior constraint can effectively improve the ability of sparse inertia-based motion capture models to infer real and complex movements.

## Ablation Study

**Analysis of the Randomness in the Prior Distribution** As previously mentioned, this paper maps sparse IMU data to a prior distribution of motion (Gaussian distribution), using its mean as the input to infer the full-body posture for the dynamic module. Unlike prediction tasks, inertial human motion capture is a deterministic regression task where only one actual scenario occurs. Therefore, excessive randomness in the prior distribution is not desirable for this study. To verify the applicability of this method for regression tasks, this paper conducts random sampling with the prior distribution and explores its randomness through visualized results. The results of random sampling are shown in Fig. 7. The first column displays the actual human postures that occurred, while the second to sixth columns are the results randomly sampled from the prior distribution. To better compare the differences in random results, this paper merges and renders them in the same coordinate system (as seen in the last column). An analysis of the results in the last column reveals that the postures in the randomly sampled outcomes are distinct from each other, as can be clearly seen in the postures of the right hand and right leg. However, from the second to the sixth column, it can be observed that the randomly sampled results are very close to the actual actions. This indicates that the prior distribution can model the uncertainty in reasoning inertial human motion with sparse data, while also ensuring that the outcomes of random sampling do not significantly differ from each other.

**The Impact of Dynamic Prior** Fig. 5 compares the online test results on DIP-IMU using two different methods of constructing the prior distribution. As mentioned before, the prior distribution dynamically changes with different IMU data inputs. In contrast, traditional VAE assumes a fixed normal distribution for the prior, $\mathcal{N}(\mathbf{0}, \mathbf{1})$. From Fig. 5, it is observable that after replacing the fixed prior with the dynamic prior distribution used in this method, the histogram overall shifts towards reduced error, leading to more accurate estimation of body postures. This is because the fixed prior overlooks the variations in IMU observations, essentially assuming that all human motions are sampled and mapped from a normal distribution, which is inconsistent with the characteristics of real-world movements, thereby increasing the difficulty of learning the distribution of full-body motion. In contrast, the prior distribution in this method changes according to different sparse IMU data, more accurately modeling the distribution of movements in the real world, hence resulting in smaller errors.

**The Impact of DIP-IMU’s Proportion in Training Data** As mentioned earlier, this paper employs a single-stage training strategy, conducting training simultaneously on DIP-IMU and AMASS. AMASS is a large synthetic dataset, crucial for the model to learn the mapping from IMU data to human posture. DIP-IMU is a real dataset, significantly smaller in scale compared to AMASS, but it provides noise interference encountered by IMUs in real scenarios, such

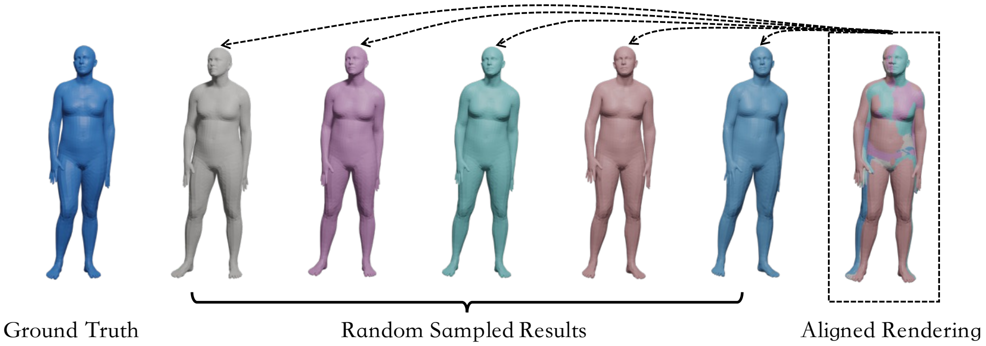


Figure 7: Analysis of the randomness of prior distributions.

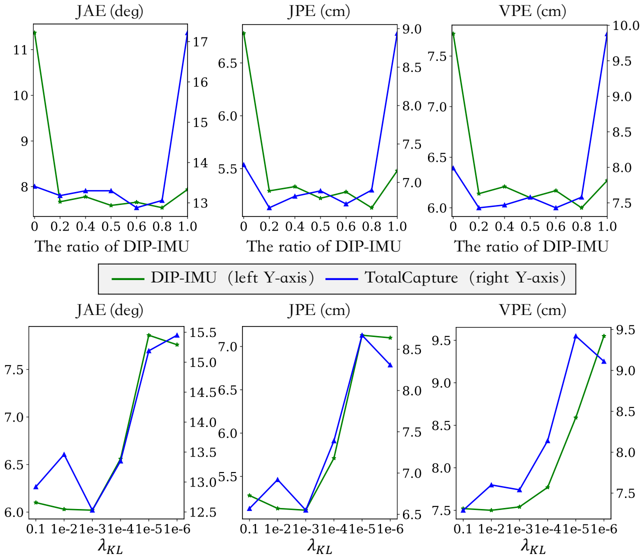


Figure 8: The impact of different ratio of DIP-IMU in the training data (Upper Part) and $\lambda_{KL}$. Evaluations were carried out on both the DIP-IMU test set (green line, left Y-axis) and TotalCapture (blue line, right Y-axis) in terms of JAE, JPE, and VPE.

as location drift, electromagnetic environment interference, etc. Both datasets complement each other, which is also validated by the experimental results in Fig. 8 (top). From Fig. 8 (Top), we can observe that the error curve exhibits a "U-shaped" trend: the model performs poorly when the proportion of DIP is either 0% or 100%, and performs better when both DIP and AMASS are used in training. Additionally, it can be seen that between 0.2 and 0.8, the model's performance is relatively similar, indicating that this method is robust to the relative proportions of the two datasets.

**The Impact of $\lambda_{KL}$** The lower part of Fig. 8 shows the impact of different values of $\lambda_{KL}$ on ProSIP. The results are reported under the online setting on the DIP-IMU test set and TotalCapture. From Fig. 8, we can observe that as $\lambda_{KL}$ decreases, the errors (JAE, JPE, VPE) initially start to decline, then begin to increase gradually after $\lambda_{KL}$ drops below $1e-3$. This is because $\lambda_{KL}$ controls the balance between the reconstruction constraint (the first term of Eq. (5)) and the KL constraint (the second term of Eq. (5)). The goal of the reconstruction constraint is to enable the model to reconstruct human postures, while the KL constraint is used to enhance the expressive capability of the motion representation in sparse IMU data. When $\lambda_{KL}$ is small, the reconstruction constraint dominates, allowing the model to better reconstruct bones tracked by IMU but failing to accurately infer untracked bones. Conversely, when $\lambda_{KL}$ is large, despite the posterior constraint for full-body motion, the weaker reconstruction constraint leads to the model's inability to effectively learn motion representation from sparse IMU data. Therefore, it is necessary to adjust the size of $\lambda_{KL}$ so that the model can simultaneously learn to reconstruct full-body motion (by the reconstruction constraint) and eliminate ambiguities in sparse IMU data (by the KL constraint).

## Conclusion

Sparse Inertial Poser suffers from mapping ambiguity from sparse IMUs data to full-body motion. To tackle this problem, we introduce a Bayesian posterior constraint method that aligns the distribution of IMU data with that of full-body motion data, making the IMU data representation closer to the actual full-body motion. In order to model the dynamics of real motion, the distribution estimated from the IMU input changes with the data, rather than being fixed. This method has been tested on authoritative benchmark datasets DIP-IMU and TotalCapture, and it achieved the best results at that time in three evaluation metrics: Joint Angle Error (JAE), Joint Position Error (JPE), and Vertex Position Error (VPE) of the SMPL model.

## Acknowledgments

This work was supported by NSFC (62250062, 62106144), Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102), and Shanghai Sailing Program (21Z510202133) from the Science and Technology Commission of Shanghai Municipality.

## References

2014. *VICON Mocap System.*

Aha, D. 1997. *Lazy Learning*.

Alarfaj, M.; Qian, Y.; and Liu, H. 2021. Detection of Human Body Movement Patterns Using IMU and Barometer. In *2020 International Conference on Communications, Signal Processing, and their Applications (ICCSPA)*, 1–6. IEEE.

Bachmann, E. R.; McGhee, R. B.; Yun, X.; and Zyda, M. J. 2001. Inertial and Magnetic Posture Tracking for Inserting Humans into Networked Virtual Environments. In *VRST*, 9–16.

Bogo, F.; Kanazawa, A.; Lassner, C.; Gehler, P.; Romero, J.; and Black, M. J. 2016. Keep it SMPL: Automatic Estimation of 3D Human Pose and Shape from a Single Image. In *ECCV*, 561–578.

Del Rosario, M. B.; Khamis, H.; Ngo, P.; Lovell, N. H.; and Redmond, S. J. 2018. Computationally Efficient Adaptive Error-State Kalman Filter for Attitude Estimation. *IEEE Sensors Journal*, 9332–9342.

Denton, E.; and Fergus, R. 2018. Stochastic video generation with a learned prior. In *ICML*, 1174–1183.

Foxlin, E. 1996. Inertial Head-tracker Sensor Fusion by A Complementary Separate-bias Kalman Filter. In *VRAIS*, 185–194.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative Adversarial Nets. In *NeurIPS*, 2672–2680.

Guan, S.; Xu, J.; Wang, Y.; Ni, B.; and Yang, X. 2021. Bilevel Online Adaptation for Out-of-Domain Human Mesh Reconstruction. In *CVPR*, 10472–10481.

Higgins, I.; Matthey, L.; Pal, A.; Burgess, C.; Glorot, X.; Botvinick, M.; Mohamed, S.; and Lerchner, A. 2016. beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. In *ICLR*.

Huang, Y.; Kaufmann, M.; Aksan, E.; Black, M.; Hilliges, O.; and Pons-Moll, G. 2018. Deep Inertial Poser: Learning to Reconstruct Human Pose from Sparse Inertial Measurements in Real Time. *ACM Transactions on Graphics*, 1–15.

Jiang, Y.; Ye, Y.; Gopinath, D.; Won, J.; Winkler, A. W.; and Liu, C. K. 2022. Transformer Inertial Poser: Real-time Human Motion Reconstruction from Sparse IMUs with Simultaneous Terrain Generation. In *SIGGRAPH Asia*, 1–9.

Kanazawa, A.; Black, M. J.; Jacobs, D. W.; and Malik, J. 2018. End-to-end Recovery of Human Shape and Pose. In *CVPR*, 7122–7131.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In *ICLR*.

Kingma, D. P.; and Welling, M. 2014. Auto-Encoding Variational Bayes. In *ICLR*.

Kocabas, M.; Karagoz, S.; and Akbas, E. 2019. Self-Supervised Learning of 3D Human Pose using Multi-view Geometry. In *CVPR*, 1077–1086.

Ling, H. Y.; Zinno, F.; Cheng, G.; and Van De Panne, M. 2020. Character Controllers using Motion Vaes. *ACM Transactions on Graphics*, 1–40.

Loper, M. M.; Mahmood, N.; and Black, M. J. 2014. MoSh: Motion and Shape Capture from Sparse Markers. *ACM Transactions on Graphics*, 220:1–220:13.

Lowe, D. G. 1999. Object Recognition from Local Scale-Invariant Features. In *ICCV*, 1150–1157.

Mahmood, N.; Ghorbani, N.; Troje, N. F.; Pons-Moll, G.; and Black, M. J. 2019. AMASS: Archive of Motion Capture as Surface Shapes. In *ICCV*, 5442–5451.

Marcard, T.; Rosenhahn, B.; Black, M.; and Pons-Moll, G. 2017. Sparse Inertial Poser: Automatic 3D Human Pose Estimation from Sparse IMUs. *Computer Graphics Forum*, 349–360.

Riaz, Q.; Tao, G.; Krüger, B.; and Weber, A. 2015. Motion Reconstruction using very Few Accelerometers and Ground Contacts. *Graphical Models*, 23–38.

Roetenberg, D.; Luinge, H.; Baten, C.; and Veltink, P. 2005. Compensation of Magnetic Disturbances Improves Inertial and Magnetic Sensing of Human Body Segment Orientation. *IEEE Trans. Neural Syst. Rehab. Eng*, 395 – 405.

Schepers, M.; Giuberti, M.; Bellusci, G.; et al. 2018. Xsens MVN: Consistent Tracking of Human Motion Using Inertial Sensing. *Xsens Technol*, 1–8.

Schwarz, L.; Mateus, D.; and Navab, N. 2009. Discriminative Human Full-Body Pose Estimation from Wearable Inertial Sensor Data. 159–172.

Sivic, J.; and Zisserman, A. 2008. Efficient Visual Search of Videos Cast as Text Retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 591–606.

Slyper, R.; and Hodgins, J. 2008. Action Capture with Accelerometers. 193–199.

Stoeve, M.; Schuldhaus, D.; Gamp, A.; Zwick, C.; and Eskofier, B. M. 2021. From the laboratory to the field: IMU-based shot and pass detection in football training and game scenarios using deep learning. *Sensors*, 21(9): 3071.

Tautges, J.; Zinke, A.; Krüger, B.; Baumann, J.; Weber, A.; Helten, T.; Müller, M.; Seidel, H.-P.; and Eberhardt, B. 2011. Motion Reconstruction Using Sparse Accelerometer Data. *ACM Transactions on Graphics*, 18.

Trumble, M.; Gilbert, A.; Malleson, C.; Hilton, A.; and Collomosse, J. 2017. Total Capture: 3D Human Pose Estimation Fusing Video and Inertial Sensors. In *BMVC*, 1–13.

Vidal, A. R.; Rebecq, H.; Horstschaefer, T.; and Scaramuzza, D. 2018. Ultimate SLAM? Combining events, images, and IMU for robust visual SLAM in HDR and high-speed scenarios. *IEEE Robotics and Automation Letters*, 3(2): 994–1001.

Vitali, R.; McGinnis, R.; and Perkins, N. 2020. Robust Error-State Kalman Filter for Estimating IMU Orientation. *IEEE Sensors Journal*, 3561–3569.

Yi, X.; Zhou, Y.; and Xu, F. 2021. TransPose: Real-time 3D Human Translation and Pose Estimation with Six Inertial Sensors. *ACM Transactions on Graphics*, 1–13.