Improving Sparse IMU-based Motion Capture with Motion Label Smoothing

Zhaorui Meng[1], Lu Yin[1], Yangqing Hou[1], Anjun Chen[1]*, Shihui Guo[1], Yipeng Qin[2]

[1]Xiamen University
[2]Cardiff University
mengzhaorui@stu.xmu.edu.cn, yinlu@stu.xmu.edu.cn, 38120241150235@stu.xmu.edu.cn, anjunchen@xmu.edu.cn, guoshihui@xmu.edu.cn, qiny16@cardiff.ac.uk

Abstract

Sparse Inertial Measurement Units (IMUs) based human motion capture has gained significant momentum, driven by the adaptation of fundamental AI tools such as recurrent neural networks (RNNs) and transformers that are tailored for temporal and spatial modeling. Despite these achievements, current research predominantly focuses on pipeline and architectural designs, with comparatively little attention given to regularization methods, highlighting a critical gap in developing a comprehensive AI toolkit for this task. To bridge this gap, we propose *motion label smoothing*, a novel method that adapts the classic label smoothing strategy from classification to the sparse IMU-based motion capture task. Specifically, we first demonstrate that a naive adaptation of label smoothing, including simply blending a uniform vector or a "uniform" motion representation (e.g., dataset-average motion or a canonical T-pose), is suboptimal; and argue that a proper adaptation requires increasing the *entropy* of the smoothed labels. Second, we conduct a thorough analysis of human motion labels, identifying three critical properties: 1) Temporal Smoothness, 2) Joint Correlation, and 3) Low-Frequency Dominance, and show that conventional approaches to entropy enhancement (e.g., blending Gaussian noise) are ineffective as they disrupt these properties. Finally, we propose the blend of a novel skeleton-based Perlin noise for motion label smoothing, designed to raise label entropy while satisfying motion properties. Extensive experiments applying our motion label smoothing to three state-of-the-art methods across four real-world IMU datasets demonstrate its effectiveness and robust generalization (plug-and-play) capability.

Introduction

Human motion capture plays a critical role in diverse domains, including film production (Menache 2000), interactive gaming (Geng and Yu 2003), and medical rehabilitation (Mousavi Hondori and Khademi 2014). Recently, sparse Inertial Measurement Units (IMUs) based motion capture systems have emerged as a lightweight yet promising alternative.

These systems achieve real-time human motion reconstruction using only six IMUs strategically positioned on the wrists, ankles, head, and hips. This minimal configuration offers compelling advantages in portability, affordability, and resilience to occlusions or lighting variations, making them highly suitable for ubiquitous motion capture scenarios.

Figure 1: A live comparison between the state-of-the-art sparse IMU-based motion capture system, GlobalPose (Yi, Pan, and Xu 2025) (left, red), and its improved variant enhanced with our motion label smoothing technique (right, green) clearly illustrates the effectiveness of our method.

Fueled by recent advances in AI, sparse IMU-based motion capture has made remarkable progress. Early methods (Huang et al. 2018; Yi, Zhou, and Xu 2021) lever-

*Corresponding author.

aged RNNs to reconstruct human motion; TIP (Jiang et al. 2022) introduced transformer architectures to improve accuracy; PIP (Yi et al. 2022) enhanced RNNs with hidden-state initialization to disambiguate complex motions; PNP (Yi, Zhou, and Xu 2024) calibrated acceleration signals via an autoregressive MLP. While effective, these approaches largely focus on adapting core neural architectures for temporal and spatial modeling to the task. In contrast, regularization, an equally critical component of deep learning, remains largely unexplored, revealing a key gap in building a more comprehensive AI toolkit for sparse IMU-based motion capture.

In this paper, we address this gap by introducing *motion label smoothing*, an adaptation of the classic label smoothing technique (Szegedy et al. 2016) tailored for sparse IMU-based motion capture. While it may appear straightforward, this adaptation poses significant challenges. Specifically, a naive adaptation of label smoothing from classification tasks (Szegedy et al. 2016; Müller, Kornblith, and Hinton 2019), such as blending the ground-truth label with a uniform label vector or a "uniform" motion representation (e.g., dataset-average motion or a canonical T-pose), proves suboptimal. We argue that this stems from a fundamental misinterpretation of "smoothness" in label smoothing: it is meant to increase label entropy, not merely enforce uniformity across label vectors. In classification, incorporating a uniform vector supports this objective; in motion capture, however, a "uniform" motion collapses into a static pose (e.g., a T-pose), paradoxically reducing entropy rather than enhancing it. Therefore, to properly adapt label smoothing for sparse IMU-based motion capture, we first conduct a rigorous analysis of motion labels, identifying their three key properties: (1) Temporal Smoothness: motion evolves continuously over time; (2) Joint Correlation: rotations of adjacent joints within a kinematic chain are inherently linked; and (3) Low-Frequency Dominance: joint rotation signals in Euclidean space are dominated by low-frequency components. Building on these properties, we demonstrate that naive entropy-enhancement strategies, such as blending Gaussian or uniform noise, are ineffective as they inevitably disrupt these intrinsic characteristics. To address this challenge, we propose a novel skeleton-based Perlin noise method for motion label smoothing. Specifically, we first map joint rotations from the SO(3) manifold to a tractable Euclidean R6D representation (Zhou et al. 2019), where these motion properties can be explicitly modeled. In this space, we construct a structured Perlin noise field whose spatial distribution encodes joint correlations (via kinematic chains) and whose temporal continuity preserves smoothness while still raising label entropy. Overlaying this skeleton-based noise onto motion labels yields smoothed labels that adhere to the principles of label smoothing while respecting the properties of human motion. We conduct comparison experiments across three state-of-the-art sparse IMU-based motion capture models and four real-world IMU datasets. Empirically, we also compare our method to naive adaptations of label smoothing and other label modification strategies. Extensive experimental results demonstrate that our method consistently improves motion capture performance and outperforms competing strategies.

In summary, our contributions are:

- We propose *motion label smoothing*, a novel adaptation of the classic label smoothing technique specifically tailored for sparse IMU-based motion capture, featuring the blending of a skeleton-based Perlin noise to motion data.
- To justify its necessity, we identify why a naive adaptation of label smoothing from classification is suboptimal, showing that they misinterpret "smoothness" as uniformity rather than entropy enhancement, leading to the misuse of static, low-entropy motions instead of meaningful regularization.
- In addition, we conduct the first rigorous analysis of motion labels for sparse IMU-based capture, identifying their three key properties: (1) Temporal Smoothness, (2) Joint Correlation, and (3) Low-frequency Dominance; and demonstrate that naive entropy-enhancement strategies (e.g., Gaussian or uniform noise) are also ineffective as they inevitably disrupt these properties.
- Extensive experiments on three state-of-the-art sparse IMU-based motion capture models and four real-world datasets demonstrate both the effectiveness and strong generalization capability of our method.

Related Work

Sparse IMU-based Motion Capture

Sparse IMU-based motion capture reconstructs human poses by estimating the local rotations of the 24 SMPL (Loper et al. 2023) joints through inverse kinematics, using signals from six IMUs placed on the left forearm, right forearm, left lower leg, right lower leg, head, and hips, respectively. However, this task remains highly challenging due to the sparsity of IMU data and its inherent noise. To address these challenges, prior works have primarily focused on adapting fundamental AI tools to enrich the toolkit for this task. Specifically, early solutions leveraged Recurrent Neural Networks (RNNs) (Huang et al. 2018) to perform end-to-end pose estimation. Subsequent work by (Yi, Zhou, and Xu 2021) broke the task into multiple prediction stages, enhancing pose accuracy and enabling global motion tracking with additional RNNs. Studies by (Jiang et al. 2022; Wu et al. 2024) explored the Transformer (Vaswani et al. 2017) framework as an additional tool for inertial motion capture, while (Yi et al. 2022) integrated physical optimization to improve the physical plausibility of predicted motions. Additionally, (Yi, Zhou, and Xu 2024) mitigated the impact of non-inertial acceleration at the root joint during motion by adapting a self-supervised MLP network. Other efforts have targeted IMU drift and calibration errors, (Zuo et al. 2025) introduced a real-time IMU calibration technique using a Transformer-based calibrator network, and (Shao et al. 2025) addressed magnetic interference in sparse IMU-based motion capture systems with an auxiliary LSTM to correct orientation errors leveraging human motion priors. Most recently, (Yi, Pan, and Xu 2025) achieved translation estimation in the full 3D space using physical optimization and refined pose estimation, making it state-of-the-art in sparse-IMU motion tracking.

Nevertheless, while these methods have collectively advanced the AI toolkit for sparse IMU-based motion capture, they have predominantly focused on pipeline and architectural designs, with comparatively little attention given to *regularization* methods, revealing a critical gap in developing a comprehensive AI toolkit for this task.

Label Smoothing

Label smoothing (Szegedy et al. 2016) is a widely used regularization technique for improving the performance of deep learning models, with applications spanning diverse domains such as image classification, machine translation, and speech recognition (Chorowski and Jaitly 2016; Real et al. 2019; Huang et al. 2019; Wei et al. 2021; Zhou et al. 2021, 2025). In practice, label smoothing operates in classification tasks by replacing one-hot labels with a softened target distribution formed by blending the ground truth label with a uniform label vector, which is commonly understood as a regularization technique designed to prevent models from becoming overly confident in their predictions and to improve their generalization capabilities (Pereyra et al. 2017; Müller, Kornblith, and Hinton 2019). Additionally, (Lukasik et al. 2020) has highlighted its effectiveness in coping with label noise; (Yuan et al. 2020) further notes that knowledge distillation represents a form of learned label smoothing regularization, wherein label smoothing regularization serves as a virtual teacher model for knowledge distillation; (Zhang et al. 2021) proposes generating soft labels based on statistical predictions of the model for the target class, thereby implementing label smoothing. Furthermore, (Lienen and Hüllermeier 2021) introduces label relaxation, where the target is represented as a set of probabilities defined by an upper probability distribution. (Keriven 2022) explores smoothing in graph neural networks.

Nevertheless, despite extensive research on label smoothing, its potential in sparse IMU-based motion capture remains unexplored.

Preliminaries

Sparse IMU-based Motion Capture. Our task involves predicting human motion based on real-time measurements from six IMUs attached to six body locations:

$$R = \texttt{Poser}(A_S, O_S, \omega_S^{root}) \tag{1}$$

where `Poser` is the pose estimation network; acceleration $A_S \in \mathbb{R}^3$, orientation $O_S \in \mathrm{SO}(3)$, and angular velocity of the root joint $\omega_S^{root} \in \mathbb{R}^3$ are the input measured by the six IMUs; the output R comprises the rotations of the 24 joints of the SMPL (Loper et al. 2023) human model (body pose).

Label Smoothing. As defined in (Szegedy et al. 2016), given a ground-truth label y, label smoothing modifies y into y', which comprises a mixture of a vector u and y weighted by $1-\epsilon$ and ϵ, respectively:

$$y' = (1-\epsilon)y + \epsilon u \tag{2}$$

where u is a uniform vector of value $\frac{1}{K}$, where K represents the number of classes in a classification task.

Motion Label Smoothing. We adapt label smoothing (Eq. 2) to our task (Eq. 1) by replacing the y in Eq. 2 with the ground truth rotation R in Eq. 1:

$$R' = (1-\epsilon)R + \epsilon u \tag{3}$$

Nevertheless, this adaptation is challenging due to the choice of u, which will be discussed in detail below.

Analysis

Naive Adaptation of Label Smoothing is Ineffective

Despite its effectiveness, the interpretation of label smoothing remains ambiguous, and we argue that the naive adaptation of it would undermine its potential as such an adaptation stems from a misinterpretation.

Interpretation Ambiguity of Label Smoothing. Despite its clear definition (Eq. 2), label smoothing has two seemingly equally valid interpretations regarding the choice of u:

- *Uniform u (naive interpretation).* As indicated by the use of the symbol "u", label smoothing works by increasing the *uniformity* of y' using a high-uniformity u which distributes label probability uniformly across all classes.
- *High-entropy u (our interpretation).* Unlike the naive interpretation, we argue that label smoothing works by increasing the entropy of y' using a high-entropy u.

Thus far, it is difficult to assess the validity of these two interpretations, as they result in the same u for classification tasks, i.e., a uniform vector of value $\frac{1}{K}$ as mentioned above.

Naive Adaptation is Ineffective. Nonetheless, these two interpretations would lead to *different* u when adapting label smoothing to sparse IMU-based motion capture:

- *Naive adaptation.* To ensure uniformity, u should represent an "average" motion label, e.g., the T-Pose or the mean motion derived from the training dataset.
- *Our adaptation.* To increase entropy, u should be a noise vector tailored for motion representations.

Empirically, through extensive experiments (Table 3) and accompanying discussions, we demonstrate that (1) the naive adaptation is ineffective, revealing that the naive interpretation is indeed a misinterpretation; (2) our interpretation is valid, offering new insights into the underlying mechanisms of label smoothing.

Naive Entropy-Enhancement Strategies are Ineffective

As mentioned above, our adaptation is non-trivial as it requires the design of noise vectors tailored to motion representations. To address this, we first conduct a rigorous analysis of motion labels to identify three key properties, and then demonstrate that naive noise designs (e.g., Gaussian) are ineffective as they disrupt these properties.

Property 1 (Temporal Smoothness)**.** *Human motion is constrained by muscle strength, skeletal structure, joint range of motion, and inertia, which inherently precludes abrupt*

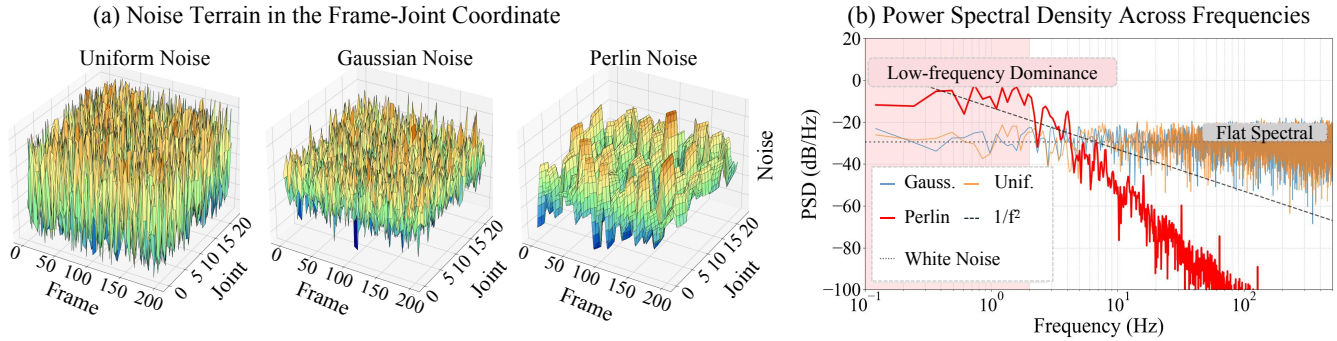

Figure 2: (a) Noise terrain of three types of noise in the frame-joint coordinate system, where only Perlin noise exhibits continuity across frames and correlation across joints. (b) Power spectral density (PSD) of the three types of noise, where only Perlin noise is dominated by low-frequency components.

changes in joint rotations R over short time intervals, i.e., $\|\omega(t)\|$ is bounded:

$$\omega(t) \approx \frac{R(t+\Delta t) - R(t)}{\Delta t}, \quad \|\omega(t)\| \leq M, \tag{4}$$

where t is time; $\omega(t)$ is the first derivative of R over t; $\|\cdot\|$ is a vector norm (e.g., L2-norm). M represents a physiological upper bound derived from biomechanical constraints.

Property 2 (Joint Correlation)**.** *The human body is a highly coupled system of rigid skeletal chains, where proximal joint motion constrains distal joints (e.g., elbow rotation influences the shoulder). This structure can naturally be modeled as a skeleton tree, with the waist node as the root and edges representing joint dependencies and biomechanical constraints. Then, for any parent-child joint pair (j_{parent}, j_{child}), the rotation range of j_{child} is constrained by the rotation angle of j_{parent} (e.g., the knee range of motion narrows when the hip is flexed):*

$$R_{child}(t) \in A(R_{parent}(t)) \tag{5}$$

where the admissible set A is defined as:

$$A = \{R_{child} | \phi_{\min}(R_{parent}) \leq R_{child} \leq \phi_{\max}(R_{parent})\} \tag{6}$$

where ϕ is a Lipschitz continuous joint correlation function; $\phi_{\min}, \phi_{\max}$ are its minimum and maximum values.

Property 3 (Low-Frequency Dominance)**.** *Human motion data exhibits a predominance of low-frequency signals corresponding to normal movements, with high-frequency components (e.g., intense jitter) being relatively rare. Let $R(t) \in \mathbb{R}^d$ denote the d-dimensional motion label at time t. We define low-frequency dominance as: $\exists$ a frequency threshold $f_c \ll f_{\max}$ for all channels $c \in \{1, 2, ..., d\}$ and a ratio threshold $\alpha \gg 0$ such that:*

$$\frac{\sum_{c=1}^{d} \int_0^{f_c} P_c(f) df}{\sum_{c=1}^{d} \int_0^{f_{\max}} P_c(f) df} \geq \alpha, \tag{7}$$

where $f_{\max}$ is the Nyquist frequency ($f_{\max} = \frac{1}{2\Delta t}$ and Δt is the IMU sampling interval); the power spectral density (PSD) $P_c(f)$ for each channel c is computed as:

$$P_c(f) = \left| \int_0^T R_c(t) e^{-i2\pi f t} dt \right|^2 \tag{8}$$

where T denotes the number of frames sampled by IMU. In our experiments, we have $f_c = 5Hz$ when $\alpha = 0.7$.

Invalidity of Naive Strategies. Given the above properties, we show that naive entropy-enhancing strategies, such as implementing u (Eq. 3) as a Gaussian or uniform noise, are ineffective as they disrupt these properties. Specifically, since Gaussian and uniform noise are *independent and identically distributed (i.i.d.)*, they have:

- **An inherent trade-off between noise amplitude, and temporal smoothness together with joint correlation** (Properties 1, 2). Sine an *i.i.d.* distribution lacks dependencies across both temporal and joint dimensions, increasing the noise amplitude (necessary for effective regularization) inevitably amplifies discrepancies along these dimensions and violates Properties 1 and 2.
- **Flat power spectral densities (PSDs)** containing significant high-frequency components, contradicting Prop. 3.

Method

As mentioned above, adapting label smoothing to our task (Eq. 3) is challenging, as its u must satisfy Properties 1, 2, 3, as well as having a sufficiently large noise amplitude for effective regularization. To address this challenge, we propose a novel design that implements u as a *skeleton-based Perlin noise* as follows, ensuring continuity and smoothness while eliminating sharp discontinuities.

Skeleton-based Perlin Noise

Specifically, we define our skeleton-based Perlin noise as:

$$u = \texttt{sk-Perlin}(JC, \mathcal{H}, size) \tag{9}$$

where JC represents the six joint chains defined from the SMPL (Loper et al. 2023) skeleton (*i.e.* , left leg, right leg, left arm, right arm, torso, and head), with the six IMUs attached to their terminal joints; $\mathcal{H} = \{S_b, S_t, S_s, p, oct, l\}$ denotes the basic parameters in constructing the Perlin nosie (Perlin 1985), including base scale S_b, time scale S_t, space scale (joint scale) S_s, persistence p, octaves oct and lacunarity l; and $size$ denotes the dimensional extent of each spatio-temporal axis, which we set to same dimension as the ground truth label.

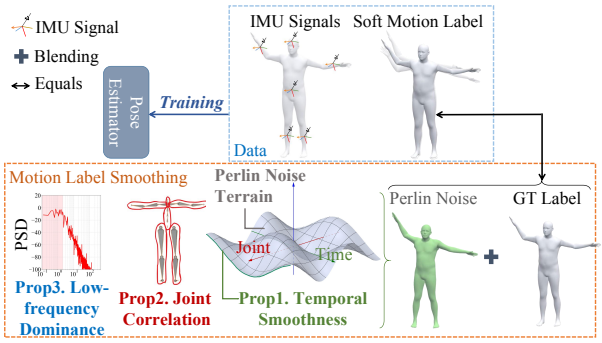

Figure 3: Overview of our *Motion Label Smoothing* method. The ground truth motion label R is blended with a carefully-constructed skeleton-based Perlin noise u, which satisfy Properties 1, 2, 3, as well as having a sufficiently large noise amplitude for effective regularization.

Amplitude-Decoupling and Properties-Satisfying Design

Perlin noise (Perlin 1985) creates smooth, natural textures and is widely utilized in computer graphics and natural phenomenon simulation. Its construction process involves dividing the three-dimensional space ($time$, $joint$, and $channel$ in our `sk-Perlin`) into a grid of equal intervals, representing coordinates along the three dimensions of a given size, segmented by points (x, y, z). At each grid point, a random unit gradient vector $\mathbf{g}_{x,y,z}$ (with length 1 and random direction) is assigned. The influence at point (x, y, z) is computed as the dot product of the gradient vectors from the surrounding eight grid points. Subsequently, an interpolation function smooths the values across these grid points. Enhanced detail is achieved by superimposing multiple noise layers of varying frequencies and amplitudes, known as octaves oct.

Amplitude-Decoupling. The amplitude of the base noise, controlled by S_b, directly determines the overall magnitude of our `sk-Perlin`, while the interpolation function governs the smoothness of transitions between adjacent grid points. Thus, the amplitude and smoothness of Perlin noise are decoupled and controlled by distinct parameters. In other words, the interpolated nature of Perlin noise (unlike the *i.i.d.* Gaussian or uniform noises) ensures smoothness while maintaining an effective noise amplitude S_b.

Satisfaction of Properties 1 and 2. As outlined above, interpolation along the x-direction ensures smoothness and continuity in the temporal dimension, with the temporal scaling factor S_t (temporal frequency) stretching the x-axis to enhance smoothness. Additionally, as illustrated in Fig. 3, our skeleton-based Perlin approach initially applies a base noise to each joint chain, followed by a single-octave noise to each joint within the chain, stripping all high-frequency details to produce ultra-smooth offsets. The final Perlin noise is synthesized by combining and scaling the base noise with the offsets. This method guarantees that the noise within each joint chain exhibits correlation (derived from the base noise) while distinguishing individual joints, thereby satisfying Property 2. We plot the different noises in the $time - joint$ coordinate in Fig. 2 (a), clearly showing that our method, based on interpolation, ensures continuity in both temporal and spatial (joint) dimensions.

Satisfaction of Property 3. The low-frequency dominance of Perlin noise arises from the low-resolution gradient field of the base grid and the attenuation of high-frequency components. The amplitude of octave superposition decays exponentially with frequency ($\frac{1}{2^i}$), while the persistence p regulates the contribution of high frequencies. We set a low octave count $oct = 5$ to reduce the number of high-frequency layers and adjust $p = 0.5$ to attenuate high-frequency effects, ensuring that low-frequency components predominate. This effect is validated through power spectral density (PSD) analysis, as shown in Figure 2 (b). Visually, while Gaussian and uniform noise display similar spectral densities across all frequencies, Perlin noise exhibits significantly higher density in low-frequency bands, with a monotonic decrease as frequency increases.

Experiments

Implementation Details

Training Setup and Datasets. Our method is designed as a plug-and-play training tool that requires no modification to model architectures. Existing approaches that innovate through model architecture have established a standardized training pipeline, utilizing the AMASS (Mahmood et al. 2019) dataset and synthesized IMU data as the training set, followed by fine-tuning with real IMU datasets (Huang et al. 2018; Trumble et al. 2017). Our experiments adhere to this same setup. Since our method modifies only the joint rotation data, fine-tuning is performed exclusively on the pose estimation networks, while parameters of other networks, such as those predicting joint velocities or foot-ground contact probabilities, are retained using publicly available pre-trained weights. All training parameters and details strictly follow the original implementations.

Our test set is selected from four real IMU datasets, including TotalCapture (Trumble et al. 2017), ANDY (Maurice et al. 2019), CIP (Palermo et al. 2022) and DIP-IMU (Huang et al. 2018).

Baseline Methods We evaluate the effectiveness of our method on three representative baseline pose estimation algorithms:

- TransPose (Yi, Zhou, and Xu 2021), the first real-time algorithm for global human motion tracking using only six IMUs;
- PIP (Yi et al. 2022), the first method to incorporate physical constraints through optimization, whose framework has been adopted by many follow-up works; and
- GlobalPose (Yi, Pan, and Xu 2025), the most recent and state-of-the-art algorithm in this domain.

Evaluation Metrics. We employ four standard evaluation metrics used in exsiting works to assess the effectiveness of

Method	TotalCapture			ANDY		
	SIP Err	**Joint Err**	**Mesh Err**	**SIP Err**	**Joint Err**	**Mesh Err**
TransPose	14.28	5.31	5.89	31.91	13.77	18.96
TP+Ours	12.49 (↓ 12.54%)	5.00 (↓ 5.84%)	5.55 (↓ 5.77%)	31.20 (↓ 2.23%)	13.42 (↓ 2.54%)	18.46 (↓ 2.64%)
PIP	11.16	4.55	5.26	29.59	13.56	18.79
PIP+Ours	10.54 (↓ 5.56%)	4.38 (↓ 3.74%)	5.07 (↓ 3.61%)	29.04 (↓ 1.86%)	13.49 (↓ 0.52%)	18.67 (↓ 0.64%)
GlobalPose	9.85	3.96	4.35	39.58	17.66	23.45
GP+Ours	7.84 (↓ 20.41%)	3.26 (↓ 17.68%)	3.75 (↓ 13.79%)	36.71 (↓ 7.25%)	16.92 (↓ 4.19%)	22.22 (↓ 5.25%)
Method	**CIP**			**DIP-IMU**		
	SIP Err	**Joint Err**	**Mesh Err**	**SIP Err**	**Joint Err**	**Mesh Err**
TransPose	28.46	10.82	11.91	14.04	4.86	5.80
TP+Ours	26.19 (↓ 7.97%)	10.65 (↓ 1.57%)	11.71 (↓ 1.68%)	13.57 (↓ 3.35%)	4.64 (↓ 4.53%)	5.50 (↓ 5.17%)
PIP	25.57	9.02	10.76	12.08	4.33	5.06
PIP+Ours	23.87 (↓ 6.65%)	8.60 (↓ 4.66%)	10.27 (↓ 4.55%)	11.62 (↓ 3.81%)	4.18 (↓ 3.46%)	4.88 (↓ 3.56%)
GlobalPose	23.04	6.98	8.13	13.77	4.36	5.08
GP+Ours	22.32 (↓ 3.12%)	6.47 (↓ 7.31%)	7.70 (↓ 5.29%)	13.50 (↓ 1.96%)	4.27 (↓ 2.06%)	4.98 (↓ 1.97%)

Table 1: Quantitative comparisons between baseline methods and those augmented with our motion label smoothing method with error percentage reduction (↓ percentage%).

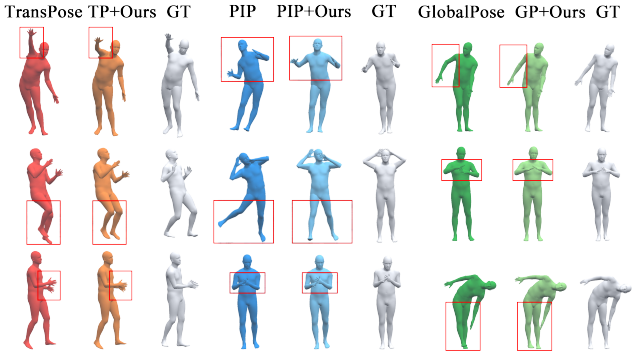

Figure 4: Qualitative comparisons with baseline methods. Examples are from the TotalCapture and CIP datasets.

the methods. *SIP Error (°)* (Von Marcard et al. 2017) measures mean global rotation error of hips and shoulders; *Angular Error (°)* measures mean global rotation error of all joints; *Positional Error (cm)* measures mean position error of all joints; *Mesh Error (cm)* measures mean vertex error of the posed SMPL meshes. Lower values indicate higher motion capture accuracy.

Comparisons

Quantitative Comparisons. We compare the performance of baseline methods using conventional supervision with those fine-tuned using our approach across four test datasets. Table. 1 reports the percentage reduction in error metrics for each of the three baseline algorithms after applying our method, showing consistent improvements across all settings. We attribute this to our method being the first regularization strategy specifically designed for sparse IMU-based motion capture. Its plug-and-play nature enables integration into a wide range of pose estimation pipelines, thereby expanding the current sparse-IMU based motion capture toolkit. Notably, we have advanced the state-of-the-art accuracy of GlobalPose, achieving a significant 20.41% reduction in SIP Error on the TotalCapture dataset.

Qualitative Comparisons. Furthermore, we provide qualitative comparisons on the test data in Fig. 4, revealing noticeable enhancements in actual motion quality when integrating our training method. For instance, in the third example of TransPose (bottom-left panel of Fig. 4), our method yield improvements in the arm movements, transitioning hand positions from being misaligned across different levels to being aligned at the same level, bringing them closer to the ground truth compared to the original results. In the first example of GlobalPose (Fig. 4, top-right), the baseline method produces a bendy right arm that misaligns with the actual pose, whereas our method yields a straight arm that aligns accurately.

Evaluation

Ablation Studies. We conduct ablation studies on the core properties of human motion addressed by our proposed method, following the logic outlined below. Using the baseline method as a control, we initially apply Gaussian noise as a naive entropy-enhancement label smoothing approach (*Baseline w/ Label Smoothing*). Subsequently, guided by the properties we have defined, we refine the noise to align with human motion properties by sequentially incorporating: 1) Temporal Smoothness: We apply Gaussian smoothing to the noise in *Baseline w/ Label Smoothing* (*Baseline + T*); 2) Joint Correlation: We leverage the same joint chain constraints as detailed in the Method section to further enhance the noise's alignment with human motion properties (*Baseline + T + J*); 3) Low-Frequency Dominance: We replace the Gaussian noise from step 2 with a Perlin noise field, designed with specific scale and persistence parameters to

reflect low-frequency dominance (*Baseline + T + J + L (ours)*). As shown in Table. 2, the naive label smoothing method provides a moderate improvement in the baseline model's performance. However, the sequential integration of the three defined properties further reduces errors, underscoring the significance of our proposed skeleton-based Perlin noise tailored to these properties.

Alternative Design. In Table. 3, we substantiate the necessity of our proposed motion label smoothing method by comparing it against alternative potential solutions, including the configurations below:

1. **Naive Adaptation**: To evaluate the rationality of our approach of increasing label entropy, following the method outlined in the Naive Adaptation section, we replace the uniform vector u with two stationary motion vectors: the T-Pose and the average motion derived from the AMASS dataset (Mahmood et al. 2019). This method maintains formal consistency with Eq. 2, and we adjust ϵ to 0.1 to align the magnitude of the added term with our noise blending scheme.
2. **Entropy-Enhancement**: We compare our method with other approaches to increasing label entropy, including the addition of uniform noise and Gaussian noise with consistent intensity.
3. **Alternatives**: Other potential label smoothing methods:
   (a) *Temporal Smoothing*: A naive temporal smoothing approach by directly applying Gaussian smoothing to the original labels.
   (b) *Knowledge Distillation*: Widely employed to enhance model performance, the soft labels produced by the teacher model in knowledge distillation act as an implicit form of label smoothing regularization (LSR) (Yuan et al. 2020). We utilize poses optimized with a physics-based module (proposed in (Yi et al. 2022) and refined in (Yi, Pan, and Xu 2025)), which incorporates physical information, as the applied distribution in Eq. 2. This distribution is overlaid onto the ground truth to form a distillation-based LSR.

Our method outperforms all the aforementioned solutions, a result we attribute to its motion-property-awareness. We analyze or mathematically demonstrate why these methods fall short of our approach, drawing on the human motion properties we have identified.

Limitations

While our method enhances existing sparse-IMU systems, it also inherits the inherent limitations of these methods. For instance, prior work relies on the template SMPL body model, overlooking the impact of body shape on IMU data, which hinders generalization to individuals with diverse body types, such as children or exceptionally tall subjects. Moreover, existing methods leverage public datasets such as AMASS for training; although extensive, these datasets are still limited in the range of motion types, thereby posing challenges in reconstructing complex motions like slipping and street dance.

Method	SIP	Ang	Joint	Mesh
Baseline (GlobalPose)	9.85	9.55	3.96	4.35
w/ Label Smoothing				
Baseline	8.82	8.65	3.82	4.43
Baseline+*T*	8.59	8.30	3.77	4.37
Baseline+*T*+*J*	8.22	8.02	3.52	4.12
Baseline+*T*+*J*+*L*(ours)	**7.84**	**7.87**	**3.26**	**3.75**

Table 2: We examine the effectiveness of incorporating the proposed motion properties (Temporal Smoothness (*T*), Joint Correlation (*J*), and Low-frequency Dominance (*L*)).

Method	SIP	Ang	Joint	Mesh
Naive Adaptation				
T-pose Vector	8.97	8.97	3.96	4.73
AMASS Mean Vector	8.75	9.17	3.87	4.60
Entropy-Enhancement				
Uniform Noise	8.72	8.61	3.82	4.44
Gaussian Noise	8.82	8.65	3.82	4.43
Alternatives				
Temporal Smoothing	8.23	8.07	3.57	4.15
Distillation	8.46	8.25	3.59	4.17
Ours	**7.84**	**7.87**	**3.26**	**3.75**

Table 3: Quantitative comparison with alternative strategies.

Conclusion

In this work, we introduce the first regularization tool for the sparse-IMU based motion capture AI toolkit. We initially show that a naive adaptation of traditional label smoothing is insufficient. Subsequently, we conduct a systematic study, identifying three inherent properties of human motion data that constrain the effectiveness of naive entropy-enhancement label smoothing techniques. Finally, we propose a novel task-specific motion label smoothing method, achieved by blending skeleton-based Perlin noise with ground-truth labels, and show its alignment with the identified properties. Extensive experiments confirm the effectiveness of our motion label smoothing method.

Acknowledgments

This work is supported by National Natural Science Foundation of China (62472364, 62072383), the Public Technology Service Platform Project of Xiamen City (No.3502Z20231043), Xiaomi Young Talents Program / Xiaomi Foundation and the Fundamental Research Funds for the Central Universities (20720240058), "Young Eagle Plan" Top Talents of Fujian Province. Anjun Chen is the corresponding author.

References

Chorowski, J.; and Jaitly, N. 2016. Towards better decoding and language model integration in sequence to sequence models. *arXiv preprint arXiv:1612.02695*.

Geng, W.; and Yu, G. 2003. Reuse of motion capture data in animation: A review. In *International Conference on Computational Science and Its Applications*, 620–629. Springer.

Huang, Y.; Cheng, Y.; Bapna, A.; Firat, O.; Chen, D.; Chen, M.; Lee, H.; Ngiam, J.; Le, Q. V.; Wu, Y.; et al. 2019. Gpipe: Efficient training of giant neural networks using pipeline parallelism. *Advances in neural information processing systems*, 32.

Huang, Y.; Kaufmann, M.; Aksan, E.; Black, M. J.; 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 (TOG)*, 37(6): 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 2022 Conference Papers*, 1–9.

Keriven, N. 2022. Not too little, not too much: a theoretical analysis of graph (over) smoothing. *Advances in Neural Information Processing Systems*, 35: 2268–2281.

Lienen, J.; and Hüllermeier, E. 2021. From label smoothing to label relaxation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 8583–8591.

Loper, M.; Mahmood, N.; Romero, J.; Pons-Moll, G.; and Black, M. J. 2023. SMPL: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, 851–866.

Lukasik, M.; Bhojanapalli, S.; Menon, A.; and Kumar, S. 2020. Does label smoothing mitigate label noise? In *International Conference on Machine Learning*, 6448–6458. PMLR.

Mahmood, N.; Ghorbani, N.; Troje, N. F.; Pons-Moll, G.; and Black, M. J. 2019. AMASS: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF international conference on computer vision*, 5442–5451.

Maurice, P.; Malaisé, A.; Amiot, C.; Paris, N.; Richard, G.-J.; Rochel, O.; and Ivaldi, S. 2019. Human movement and ergonomics: An industry-oriented dataset for collaborative robotics. *The International Journal of Robotics Research*, 38(14): 1529–1537.

Menache, A. 2000. *Understanding motion capture for computer animation and video games*. Morgan kaufmann.

Mousavi Hondori, H.; and Khademi, M. 2014. A review on technical and clinical impact of microsoft kinect on physical therapy and rehabilitation. *Journal of medical engineering*, 2014(1): 846514.

Müller, R.; Kornblith, S.; and Hinton, G. E. 2019. When does label smoothing help? *Advances in neural information processing systems*, 32.

Palermo, M.; Cerqueira, S. M.; André, J.; Pereira, A.; and Santos, C. P. 2022. From raw measurements to human pose-a dataset with low-cost and high-end inertial-magnetic sensor data. *Scientific Data*, 9(1): 591.

Pereyra, G.; Tucker, G.; Chorowski, J.; Kaiser, Ł.; and Hinton, G. 2017. Regularizing neural networks by penalizing confident output distributions. *arXiv preprint arXiv:1701.06548*.

Perlin, K. 1985. An image synthesizer. *ACM Siggraph Computer Graphics*, 19(3): 287–296.

Real, E.; Aggarwal, A.; Huang, Y.; and Le, Q. V. 2019. Regularized evolution for image classifier architecture search. In *Proceedings of the aaai conference on artificial intelligence*, volume 33, 4780–4789.

Shao, Y.; Yi, X.; Yin, L.; Guo, S.; Yong, J.; and Xu, F. 2025. MagShield: Towards Better Robustness in Sparse Inertial Motion Capture Under Magnetic Disturbances. *arXiv preprint arXiv:2506.22907*.

Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2818–2826.

Trumble, M.; Gilbert, A.; Malleson, C.; Hilton, A.; and Collomosse, J. 2017. Total capture: 3d human pose estimation fusing video and inertial sensors. In *Proceedings of 28th British Machine Vision Conference*, 1–13.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Von Marcard, T.; Rosenhahn, B.; Black, M. J.; and Pons-Moll, G. 2017. Sparse inertial poser: Automatic 3d human pose estimation from sparse imus. In *Computer graphics forum*, volume 36, 349–360. Wiley Online Library.

Wei, J.; Liu, H.; Liu, T.; Niu, G.; Sugiyama, M.; and Liu, Y. 2021. To smooth or not? when label smoothing meets noisy labels. *arXiv preprint arXiv:2106.04149*.

Wu, Y.; Yin, L.; Guo, S.; Qin, Y.; et al. 2024. Accurate and steady inertial pose estimation through sequence structure learning and modulation. *Advances in Neural Information Processing Systems*, 37: 42468–42493.

Yi, X.; Pan, S.; and Xu, F. 2025. Improving Global Motion Estimation in Sparse IMU-based Motion Capture with Physics. *arXiv preprint arXiv:2505.05010*.

Yi, X.; Zhou, Y.; Habermann, M.; Shimada, S.; Golyanik, V.; Theobalt, C.; and Xu, F. 2022. Physical inertial poser (pip): Physics-aware real-time human motion tracking from sparse inertial sensors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13167–13178.

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 (TOG)*, 40(4): 1–13.

Yi, X.; Zhou, Y.; and Xu, F. 2024. Physical Non-inertial Poser (PNP): Modeling Non-inertial Effects in Sparse-inertial Human Motion Capture. In *SIGGRAPH 2024 Conference Papers*.

Yuan, L.; Tay, F. E.; Li, G.; Wang, T.; and Feng, J. 2020. Revisiting knowledge distillation via label smoothing regularization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3903–3911.

Zhang, C.-B.; Jiang, P.-T.; Hou, Q.; Wei, Y.; Han, Q.; Li, Z.; and Cheng, M.-M. 2021. Delving deep into label smoothing. *IEEE Transactions on Image Processing*, 30: 5984–5996.

Zhou, H.; Song, L.; Chen, J.; Zhou, Y.; Wang, G.; Yuan, J.; and Zhang, Q. 2021. Rethinking soft labels for knowledge distillation: A bias-variance tradeoff perspective. *arXiv preprint arXiv:2102.00650*.

Zhou, Y.; Barnes, C.; Lu, J.; Yang, J.; and Li, H. 2019. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5745–5753.

Zhou, Z.; Wei, S.; Zhang, X.; Dou, W.; Qu, M.; and Cai, Y. 2025. Training Deep Neural Networks with Virtual Smoothing Classes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 23036–23044.

Zuo, C.; Huang, J.; Jiang, X.; Yao, Y.; Shi, X.; Cao, R.; Yi, X.; Xu, F.; Guo, S.; and Qin, Y. 2025. Transformer IMU calibrator: Dynamic on-body IMU calibration for inertial motion capture. *arXiv preprint arXiv:2506.10580*.