Spatial-Related Sensors Matters: 3D Human Motion Reconstruction Assisted with Textual Semantics

Xueyuan Yang [1,2], Chao Yao[1,2*], Xiaojuan Ban[1,2,3,4*]

[1]Beijing Advanced Innovation Center for Materials Genome Engineering, Beijing 100083, China.
[2]University of Science and Technology Beijing, Beijing 100083, China.
[3]Key Laboratory of Intelligent Bionic Unmanned Systems, Ministry of Education, Beijing 100083, China.
[4]Institute of Materials Intelligent Technology, Liaoning Academy of Materials, Shenyang 110004, China.
m202210673@xs.ustb.edu.cn,{yaochao, banxj}@ustb.edu.cn

Abstract

Leveraging wearable devices for motion reconstruction has emerged as an economical and viable technique. Certain methodologies employ sparse Inertial Measurement Units (IMUs) on the human body and harness data-driven strategies to model human poses. However, the reconstruction of motion based solely on sparse IMU data is inherently fraught with ambiguity, a consequence of numerous identical IMU readings corresponding to different poses. In this paper, we explore the spatial importance of sparse sensors, supervised by text that describes specific actions. Specifically, uncertainty is introduced to derive weighted features for each IMU. We also design a Hierarchical Temporal Transformer (HTT) and apply contrastive learning to achieve precise temporal and feature alignment of sensor data with textual semantics. Experimental results demonstrate our proposed approach achieves significant improvements in multiple metrics compared to existing methods. Notably, with textual supervision, our method not only differentiates between ambiguous actions such as sitting and standing but also produces more precise and natural motion.

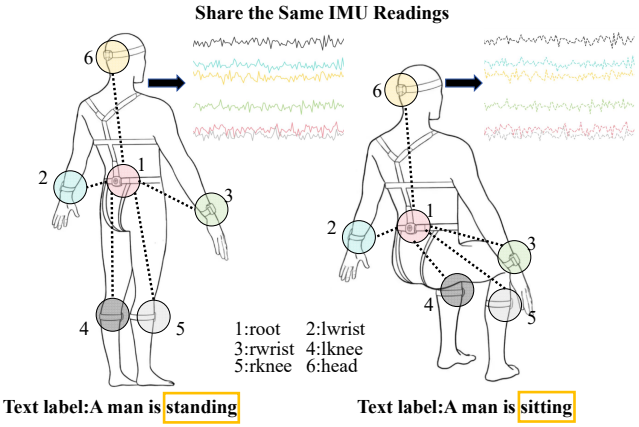

Figure 1: Considering specific postures such as standing and sitting, the rotational data and acceleration output by the sensors are largely invariant. Incorporating additional information such as text can help to address this challenge.

Introduction

Human motion reconstruction is a pivotal technique for accurately capturing 3D human body kinematics, with critical applications in gaming, sports, healthcare, and film production. One of the prevalent methods in motion reconstruction is the optical-based approach, which involves analyzing images of individuals to ascertain their respective poses (Chen et al. 2020; Sengupta, Budvytis, and Cipolla 2023; Cao et al. 2017). With the rapid progression of wearable techniques, various sensor devices have also been used to reconstruct human motion. For example, Xsens (Schepers et al. 2018) system employs 17 densely positioned IMUs to facilitate the reconstruction of human body poses. Compared to optical methods, IMUs offer robustness against variable lighting conditions and occlusions, allow for unrestrained movement in both indoor and outdoor environments, and enable the generation of naturalistic human motion. However, the dense placement of wearable IMUs on the body can be intrusive and costly.

To address this issue, some methods (Huang et al. 2018; Yi, Zhou, and Xu 2021; Jiang et al. 2022b; Von Marcard et al. 2017; Yi et al. 2022) have deployed sparse IMUs on the body and analyzed temporal signals to model human body poses. These approaches not only reduce the number and cost of IMUs but also enhance the wearability and minimize invasiveness. Nevertheless, it should be noted that there are still some limitations that restrict the utilization of sparse sensors. Specifically, motion reconstruction using sparse inertial sensors constitutes an under-constrained problem: distinct postures can yield identical sensor outputs. As illustrated in Fig. 1, the sensors generate similar rotation matrices and acceleration outputs when the subject is sitting and standing, making accurate differentiation between these postures challenging. Besides, the inherent distinction of spatial relations between IMUs, has rarely been used in previous methods, thereby revealing opportunities for potential enhancements.

In this paper, we introduce a novel framework for sensor-based 3D human motion reconstruction, leveraging spatial relationships and textual supervision to accurately generate naturalistic human body poses. Sparse sensors are designed to capture the motion characteristics of different body

*Corresponding author.

parts. Considering that the correlations among these features contain a crucial priori knowledge about the human body's skeletal structure, our method employs intra-frame spatial attention to model the correlation between IMUs, allowing the model to concentrate on the distinct characteristics of different body regions at one point in time. Moreover, in response to the inherent potential instability of IMU readings, the concept of sensor uncertainty is introduced. This allows for the optimization of sensor outputs and the adaptive adjustment of each sensor's relative contribution. However, relying solely on sensor data is insufficient for resolving the problems of ambiguity. Text, with its rich motion information, can aid the model in identifying human motion states and resolving issues of ambiguity. Finally, to facilitate better modality fusion, we propose unique modules to align sensor features with text features in both temporal and semantic dimensions.

In the realm of sensor data and text fusion, IMU2CLIP (Moon et al. 2022) bears resemblance to our work, aligning images and IMU sensor data with corresponding text using the CLIP (Radford et al. 2021). The methodologies diverge in several key respects. IMU2CLIP is designed for modality transitivity, facilitating text-based IMU retrieval, IMU-based video retrieval, and natural language reasoning tasks with motion data. In contrast, our approach underscores the synergistic potential of multimodal information, using text to resolve ambiguities inherent in sparse sensor data. Furthermore, to achieve enhanced modality fusion, the Hierarchical Temporal Transformer module was designed, and contrastive learning was employed to ensure temporal and semantic synchronization between the textual and sensor data. Cross-attention mechanisms were then utilized to merge features from both modalities. Experimental results show that our proposed framework achieves state-of-the-art performance compared with some classical methods, both in quantitative and qualitative measurements.

In summary, our work makes the following contributions:

- We present a sensor-based approach to 3D human motion reconstruction that is augmented with textual supervision. This method leverages the rich semantic information contained within the text to enhance the naturalness and precision of the modeled human poses.
- We introduce a spatial-relation representation model which computes the correlations between sensors within a frame while also taking into account the uncertainty of each IMU.
- We design a Hierarchical Temporal Transformer module to achieve temporal alignment between sensor features and textual semantics. A contrastive learning mechanism is also adopted to optimize the alignment between the two modalities in high-dimensional space.

Related Work

Sensor-based Human Motion Reconstruction

Full-body sensor-based motion reconstruction is a widely utilized technique in commercial motion capture systems. A prominent example is the Xsens system (Schepers et al. 2018), which achieves detailed reconstruction of human movements by equipping the body with 17 strategically placed IMUs. However, this method presents drawbacks, primarily its invasive impact on human movement due to the intensive IMUs placement, as well as its substantial cost.

Efforts have been made to implement motion reconstruction using sparse IMUs, thereby enhancing the usability of inertial sensor-based motion reconstruction, albeit at the expense of some degree of accuracy. For instance, studies (Slyper and Hodgins 2008; Tautges et al. 2011) have achieved human motion reconstruction with as few as four to five accelerometers, by retrieving pre-recorded postures with analogous accelerations from motion reconstruction databases. (Von Marcard et al. 2017) developed an offline system that operates with only six IMUs, optimizing the parameters of the SMPL body model (Loper et al. 2015) to fit sparse sensor inputs. With the advent of the deep learning era, (Huang et al. 2018) synthesized inertial data from an extensive human motion dataset to train a deep neural network model based on a Bidirectional Recurrent Neural Network that directly mapped IMU inputs to body postures. (Yi, Zhou, and Xu 2021) decomposed body posture estimation into a multi-stage task to improve the accuracy of posture regression through the use of joint locations as an intermediate representation. Moreover, recent methodologies such as (Dittadi et al. 2021) and AvatarPoser (Jiang et al. 2022a) estimated full-body posture using only head and hand sensors, yielding promising results.

However, reconstructing human motion from a set of sparse IMUs presents an under-constrained problem, where similar sensor readings may correspond to different postures. Some approaches have sought to address this issue to a certain extent through unique network designs. For instance, Physical Inertial Poser (Yi et al. 2022) approximated the under-constrained problem as a binary classification task between standing and sitting, proposing a novel RNN initialization strategy to replace zero initialization. It then distinguished between standing and sitting based on instantaneous acceleration. Transformer Inertial Poser (Jiang et al. 2022b) introduced past history outputs as inputs to differentiate ambiguous actions. Other methodologies have explored the integration of multimodal information to impose additional constraints on the model, enhancing the generation of precise poses. For instance, studies (Von Marcard et al. 2018; Malleson et al. 2017; Von Marcard, Pons-Moll, and Rosenhahn 2016) have significantly improved estimation accuracy by combining inertial sensors with video data, although challenges such as occlusion, lighting issues, and mobility restrictions still persist. Fusion Poser (Kim and Lee 2022) incorporates head height information from a head tracker into the model's input.

Textual Semantics in Human Motion Field

In the burgeoning field of multimodal processing, text, with its rich semantic information and ease of annotation, is increasingly utilized in the human motion domain. Studies such as (Guo et al. 2022; Zhang et al. 2022; Tevet et al. 2022) can generate high-quality 3D human motions from textual descriptions. These findings affirm that texts encap-

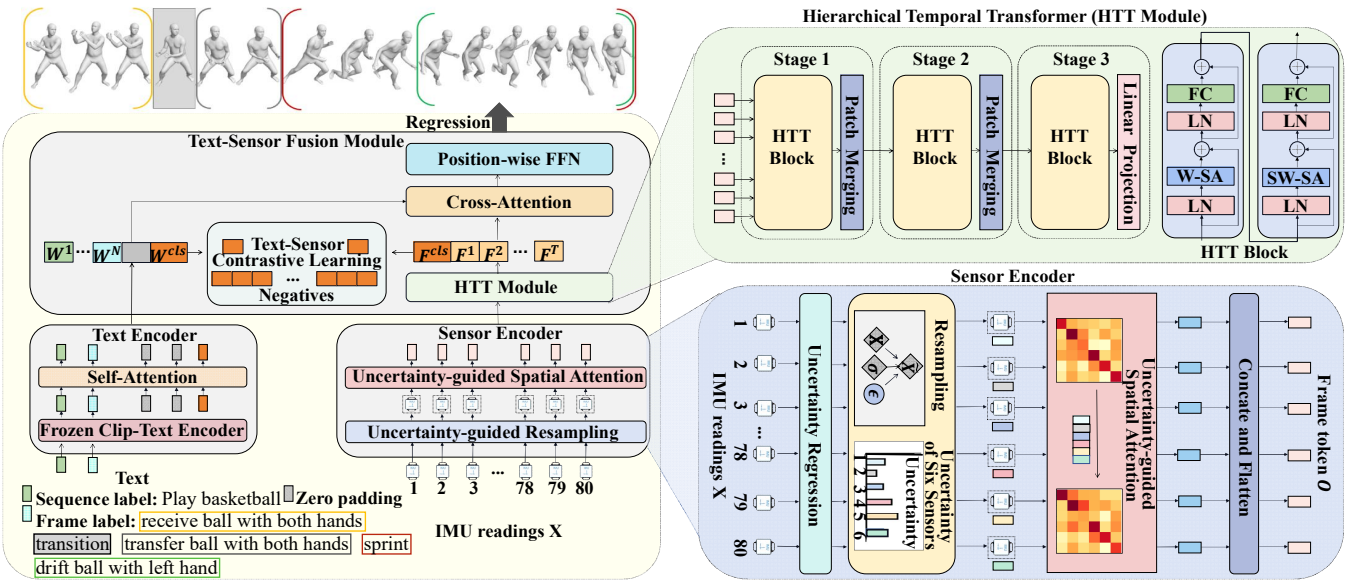

Figure 2: Overview of our method. Our model encapsulates three distinct encoders: a Text Encoder, a Sensor Encoder, and a Text-Sensor Fusion Module. The details of the Sensor Encoder and the Hierarchical Temporal Transformer (HTT) module are illustrated on the right. The schematic of the model output is adapted from (Punnakkal et al. 2021).

sulate rich motion information. We posit that text supervision could disambiguate actions, thereby enhancing the naturalness and precision of generated motions.

Method

Our primary target is to reconstruct accurate human poses using data from 6 IMUs placed on the legs, wrists, head, and pelvis (root), coupled with textual supervision. The sensors provide inputs in the form of tri-axial acceleration, $a \in \mathbb{R}^3$, and rotation matrices, $R \in \mathbb{R}^{3\times3}$. As illustrated in Fig. 2, our framework consists of a Text Encoder, a Sensor Encoder, and a Text-Sensor Fusion module. The Text Encoder converts the input text W such as ["receive ball with both hands", ..., "transition"] into a sequence of embeddings: $\{W^{cls}, W^1, ..., W^N\}$, where W^{cls} represents the embedding of the [CLS] token, and N denotes the number of text labels. For the Sensor Encoder, a motion sequence composed of sensor data frames $X^t = [(a^t_{root}, R^t_{root}), ..., (a^t_{head}, R^t_{head})], t \in [1, T]$ is encoded into a sequence of embeddings that contain intra-frame spatial relations: $\{O^1, ..., O^T\}$. Within the Text-Sensor Fusion module, these spatial embeddings are then processed by the Hierarchical Temporal Transformer to extract a unified spatio-temporal fusion representation $\{F^{cls}, F^1, ..., F^T\}$, where F^{cls} denotes the embedding of the [CLS] token. Before applying cross-attention in the fusion process, Text-Sensor contrastive learning is strategically implemented to refine the alignment between the unimodal representations of the two modalities. Finally, a simple regression head is employed to derive human pose rotational data $q \in \mathbb{R}^{j\times6}$ for j key points (with each rotation encoded by a 6D vector (Zhou et al. 2019)), the corresponding three-dimensional position $p \in \mathbb{R}^{j\times3}$, and the root's speed data $s \in \mathbb{R}^3$.

Text Encoder

We utilize the first 4 layers of the frozen CLIP (Radford et al. 2021) VIT/B32 text encoder, augmented with two additional transformer layers, to form our Text Encoder. Specifically, given a text label sequence W, it is initially tokenized and mapped into a sequence of tokens $\widetilde{W}$ using CLIP, with a randomly initialized tensor prepended as the [CLS] token. It is important to note that W provides two kinds of semantic labels: sequence-level and frame-level labels, as defined in the dataset configuration. For frame-level labels, despite each frame having its own text description, they are largely repetitive. For example, the label "walk" might apply continuously over a series of frames. To mitigate computational load, only non-repetitive frame-level texts are chronologically ordered as inputs. For sequence labels, if the total number is less than the threshold M, we use all sequence labels as input. Otherwise, one-third of the labels are selected based on their temporal information, specifically choosing those that best match the sensor subsequence. To differentiate between sequence-level and frame-level labels, two learnable group position embeddings G are developed for each. Additionally, Sinusoidal Position Embeddings (Vaswani et al. 2017) P are utilized, with time information computed independently for both the sequence and frame levels, accommodating their unique characteristics.

$$\overline{W}^i = \widetilde{W}^i + P^i + G^i, \quad for \ \ i \in [1, N] \tag{1}$$

Then the processed features $\overline{W}$ and the [CLS] token are fed into the self-attention layers to better extract textual semantics.

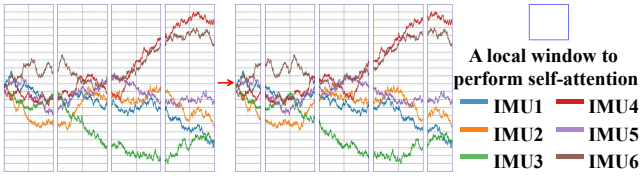

Figure 3: An illustration of window self-attention (left) and shifted window self-attention (right).

Sensor Encoder

The Sensor Encoder captures the intricate relations within sparse sensors via spatial modeling. It includes a resampling strategy and a spatial attention mechanism, both guided by estimated uncertainty for each IMU.

Uncertainty Estimation: First, we estimate the uncertainty for each IMU reading, where the original IMU readings, denoted by $X^t \in \mathbb{R}^{6\times(3+3\times3)}$ are fed into an uncertainty regression head, yielding uncertainty $\sigma^t \in \mathbb{R}^{72}$ for each channel.

Uncertainty-guided Resampling: Rather than directly using the original readings X^t, we resample IMU readings denoted as $\widetilde{X}^t$ from a Gaussian distribution $\mathcal{N}(X^t, \sigma^t)$, with X^t as the mean and predicted uncertainty σ^t as the variance. This resampling method ensures that the values with low uncertainty remain largely unchanged, while the values with high uncertainty are resampled, thereby optimizing the sensor data. Notably, the resampling procedure is only employed during the training. During inference, the uncertainty is simply regressed for each channel, and the original sensor readings X^t are utilized as $\widetilde{X}^t$. We apply the reparameterization trick (Kingma and Welling 2022) for efficient gradient descent by sampling $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{1})$ to compute $\widetilde{X}^t$ as follows: $\widetilde{X}^t = X^t + \sigma^t \cdot \epsilon$.

Uncertainty-guided Spatial Attention (UGSA): After sampling the IMU readings for the t-th frame $\widetilde{X}^t$ with corresponding uncertainty σ^t, we map $\widetilde{X}^t$ to a $6 \times c$ feature embedding Z^t, where 6 represents the number of sensors and c signifies the dimension of spatial features. We then conduct self-attention (Vaswani et al. 2017) on Z^t. It is noted that for the computation of t-th frame's attention between two sensors, denoted as j and k, the uncertainty $\sigma_k^t \in \mathbb{R}^{12}$ (summed over its 12 channels) of sensor k is taken into account by dividing the attention score by it.

$$A_{j,k}^t = \frac{\left(Z_j^t P^Q\right)\left(Z_k^t P^K\right)^T}{\sqrt{c} \cdot \sum \sigma_k^t} \tag{2}$$

where $P^Q, P^K \in \mathbb{R}^{c\times c}$ are the Query and Key projection matrices. This unique alteration ensures that sensors with high uncertainty contribute less when computing spatial correlations. The output of the UGSA module for the t-th frame, O^t, matches the input dimensions $Z^t \in \mathbb{R}^{6\times c}$. After flattening O^t to $\mathbb{R}^{1\times(C=6\times c)}$, we concatenate the output vectors from T frames to form $O \in \mathbb{R}^{T\times C}$.

Text-Sensor Fusion Module

The Text-Sensor Fusion Module aligns and fuses bimodal features. Specifically, we employ a Hierarchical Temporal Transformer to acquire spatiotemporally fused sensor features for temporal synchronization with text features. Subsequently, contrastive learning is used to align the multimodal features in a high-dimensional space, followed by the application of cross-attention for feature fusion.

Hierarchical Temporal Transformer (HTT): The HTT module is utilized for temporal alignment between sensor features and textual semantics. We hypothesized that information derived from adjacent frames is pivotal for the estimation of the current frame pose. In response to this hypothesis, window self-attention (W-SA) and shifted window self-attention (SW-SA) mechanisms are incorporated to constrain the scope of attention computation, introducing a convolution-like locality to the process. Furthermore, to integrate information from distant frames and thus extend the receptive field, a patch merge operation is implemented. These approaches facilitate the extraction of sensor features at diverse granularity levels and concurrently reduce the computational complexity of the transformer from a quadratic to a linear relationship with the sequence length.

Given a window size of I, a sensor sequence of length L is divided into $\frac{L}{I}$ non-overlapping subintervals. Local window attention computations are first performed within these subintervals. To create interconnections between these non-overlapping segments, we adopt a shifted window attention module inspired by (Liu et al. 2021), enabling a new partitioning method that enhances self-attention across segments. The W-SA and SW-SA always appear alternately, constituting a Hierarchical Temporal Transformer Block as shown in the top-right corner of Fig. 2.

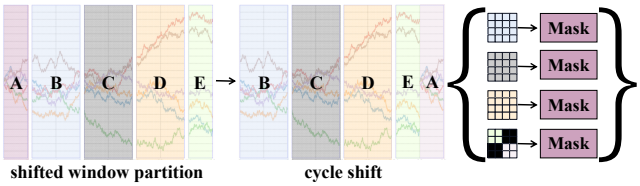

Figure 4: An efficient methodology for batch computation of self-attention within the context of shifted window partitioning.

When applying shifted window attention to temporal sequences, the window count increases from $\frac{L}{I}$ to $\frac{L}{I}+1$, resulting in some windows being smaller than I. To address this, we introduce a batch computation with a leftward cyclic shift, depicted in Fig. 4. This shift can produce windows with non-contiguous sub-windows. We tackle this by designing a masking mechanism that restricts self-attention to within each sub-window, maintaining the number of batched windows and ensuring computational efficiency. After the computation, the original sequence order is restored.

In the patch merge operation, each procedure consolidates two adjacent tokens into one, effectively halving the token count and doubling each token's dimensionality. These

transformed tokens are then fed into the subsequent stages. Within the final stage, the patch merge is omitted, and tokens are restored to their original count and dimensions through linear projection and reshaping. Within a sensor sequence, we map the output features $F \in \mathbb{R}^{T\times C}$ to a feature with dimensions $1 \times C$, serving as the [CLS] token. This [CLS] token, in conjunction with F, forms the cumulative output $\{F^{cls}, F^1, ..., F^T\}$, which encompasses the spatio-temporal features.

Feature Fusion: Given the sensor features set $\{F^{cls}, F^1, ..., F^T\}$ and the text features set $\{W^{cls}, W^1, ..., W^N\}$, we apply contrastive learning to align these features in a high-dimensional joint space, utilizing the [CLS] tokens as anchors. Subsequently, the sensor features are fused with textual features through cross-attention. Corresponding group embeddings and temporal position embeddings are designed for both textual and sensor features.

Losses

We train our model with three objectives: uncertainty learning on the Sensor Encoder, Text-Sensor contrastive learning on the unimodal encoders and recon loss on the Text-Sensor Fusion module. The relevant equations are presented below. The parameters δ, γ, λ, α, β are used to balance the different loss weights.

Uncertainty Loss: We aim to estimate the uncertainty of the input IMU data. Inspired by (Kendall and Gal 2017), we set our uncertainty estimation loss as:

$$\mathcal{L}_\sigma = \frac{\delta}{T}\sum_{t=1}^{T}\left(\left\|\frac{(q^t-\hat{q}^t)}{\sum_{j=1}^{6}\sigma_j^t}\right\|^2 + \left\|\frac{(p^t-\hat{p}^t)}{\sum_{j=1}^{6}\sigma_j^t}\right\|^2 + \sum_{j=1}^{6}\left\|\sigma_j^t\right\|^2\right) \tag{3}$$

The term σ_j^t denotes the uncertainty of the j-th sensor at the t-th frame. The terms $||q^t-\hat{q}^t||^2$ and $||p^t-\hat{p}^t||^2$ represent the squared discrepancies between the predicted and true values of the joint rotation angles and the joint positions for the t-th frame, respectively.

Contrastive Loss: We use text-sensor contrastive learning to learn better unimodal representations before fusion. Given a batch of B text-sensor pairs, the model learns to maximize the similarity between a sensor sequence and its corresponding text while minimizing the similarity with the other $B-1$ texts in the batch, and vice versa.

$$\mathcal{L}_{\text{contrastive}} = -\frac{\gamma}{2B}\sum_{i=1}^{B}(H1+H2) \tag{4}$$

where

$$H1 = \log\frac{e^{s_{i,i}/\tau}}{\sum_{j=1}^{B}e^{s_{i,j}/\tau}},\ H2 \quad = \log\frac{e^{s_{i,i}/\tau}}{\sum_{j=1}^{B}e^{s_{j,i}/\tau}} \tag{5}$$

The $s_{i,j}$ represents the similarity calculated by cosine similarity between the i-th sensor sequence and the j-th text, and τ is a learnable temperature parameter that controls the concentration of the distribution.

Recon Loss: Our model is optimized to capture motion characteristics by minimizing the L_2 losses on joint orientations q, joint locations p, and root speed s, as shown in Equations (6) and (7).

$$\mathcal{L}_{\text{recon}} = \lambda\cdot D(q,\hat{q}) + \beta\cdot D(p,\hat{p}) + \alpha\cdot D(s,\hat{s}) \tag{6}$$

where

$$D(x,\hat{x}) = \frac{1}{T}\sum_{t=1}^{T}\left|x^t-\hat{x}^t\right|^2 \tag{7}$$

calculates the mean discrepancy between the model's predicted values x and the true values $\hat{x}$.

The full objective of our model is:

$$\mathcal{L} = \mathcal{L}_\sigma + \mathcal{L}_{\text{contrastive}} + \mathcal{L}_{\text{recon}} \tag{8}$$

Experiment

Dataset Setting

Our experiment employed two types of data: sensor data captured during human motion and the corresponding textual annotations.

We utilized the Babel dataset (Punnakkal et al. 2021) for semantic annotations, which provides two levels of text labels for around 43 hours of AMASS mocap sequences (Mahmood et al. 2019): sequence labels describe the overall actions, while frame labels detail each action per frame. For the DIP-IMU dataset (Huang et al. 2018), which lacks Babel's semantic annotations, we manually added sequence-level labels, albeit less comprehensive.

Regarding the motion data, given the scarcity of real datasets and the extensive data requirements inherent in deep learning, we followed the previous method (Jiang et al. 2022b) and synthesized more diverse inertial data from the extensive AMASS dataset. This enriched synthesized data, combined with real data, was used for training. The configuration details of the motion datasets are as follows:

AMASS: The AMASS dataset unifies various motion reconstruction datasets. We synthesized a subset of AMASS, incorporating the CMU, Eyes Japan, KIT, ACCAD, DFaust 67, HumanEva, MPI Limits, MPI mosh, and SFU datasets.

DIP-IMU: The DIP-IMU dataset comprises IMU readings and pose parameters from approximately 90 minutes of activity by 10 subjects. We reserved Subjects 9 and 10 exclusively for evaluation and utilized the rest for training.

Totalcapture: The Totalcapture dataset (Trumble et al. 2017) comprises 50 minutes of motion captured from 5 subjects. Following previous works, we used real IMU data for evaluation, but ground truth and synthesized IMU readings were still integrated into the training set. Due to missing semantic annotations from Babel in some sequences, only 27 fully annotated sequences were utilized.

Metric

For a fair comparison, we adhered to the evaluation methodology previously used for (Yi, Zhou, and Xu 2021). We used five metrics for pose evaluation: 1) SIP error, which measures the average global rotation error of the limbs in degrees; 2) Angular error, the average global rotation error of

	Totalcapture					DIP-IMU				
Method	SIP Err	Ang Err	Pos Err	Mesh Err	Jitter	SIP Err	Ang Err	Pos Err	Mesh Err	Jitter
SIP	–	–	–	–	–	21.02/9.61	8.77/4.38	6.66/3.33	7.71/3.80	3.86/6.32
DIP	–	–	–	–	–	16.36/8.60	14.41/7.90	6.98/3.89	8.56/4.65	23.37 /23.84
Transpose	12.30/5.90	11.34/4.84	4.85/2.63	5.54/2.89	**1.31/2.43**	13.97/6.77	**7.62 /4.01**	4.90/2.75	5.83/3.21	**1.19/1.76**
Ours	**7.92/4.38**	**9.35/4.10**	**3.70/2.03**	**4.32/2.29**	1.74/1.55	**13.34/6.71**	8.33/4.70	**4.71/2.72**	**5.75/3.29**	1.81/1.72

Table 1: In offline settings, our method is evaluated against SIP, DIP, and Transpose on the Totalcapture and DIP-IMU datasets, focusing on the assessment of body poses. The mean values, followed by the standard deviations, are presented in the report. Bold numbers indicate the best performing entries.

	Totalcapture					DIP-IMU				
Method	SIP Err	Ang Err	Pos Err	Mesh Errr	Jitter	SIP Err	Ang Err	Pos Err	Mesh Err	Jitter
DIP	–	–	–	–	–	17.10/9.59	15.16/8.53	7.33/4.23	8.96/5.01	30.13/28.76
PIP	–	–	–	–	–	15.02	8.73	5.04	5.95	**2.4**
TIP	11.74/6.75	11.57/5.12	5.26/3.00	6.10/3.44	9.69/6.68	15.33/8.44	8.89/5.04	5.22/3.32	6.28/3.89	10.84/6.87
Transpose	13.65/7.83	11.84/5.36	5.64/3.42	6.35/3.70	**8.05/11.70**	16.68/8.68	8.85/4.82	5.95/3.65	7.09/4.24	6.11/7.92
Ours	**9.67/5.12**	**10.49/4.55**	**4.36/2.37**	**5.05/2.69**	13.30/16.86	**14.18/7.14**	**8.25/4.45**	**4.76/2.76**	**5.80/3.26**	14.41/17.18

Table 2: In online settings, our method is evaluated against DIP, PIP, TIP, and Transpose on the Totalcapture and DIP-IMU datasets, focusing on the assessment of body poses. Bold numbers indicate the best performing entries.

all body joints, also in degrees; 3) Positional error, the average Euclidean distance error of all joints, with the spine aligned, measured in centimeters; 4) Mesh error, the average Euclidean distance error of the body mesh vertices, with the spine aligned, also in centimeters; 5) Jitter error, defined as the average jerk of all body joints in predicted motion and measured in $10^2 m/s^3$, reflects motion smoothness.

Training Details

The entire training and evaluation regimen was conducted on a system equipped with 1 Intel(R) Xeon(R) Silver 4110 CPU and 1 NVIDIA GeForce RTX 2080 Ti GPU. Our model was developed using PyTorch 1.13.0, further accelerated by CUDA 11.6. Our model configuration sets the input sequence length T at 80 frames, with a window and shifted size of 20 and 10 frames, respectively, and a threshold of M being 15. The training process, utilizing a batch size of 40, incorporates the Adam optimizer (Kingma and Ba 2017) initialized with a learning rate of $2e^{-5}$. To balance the magnitude of the loss, we set λ and α to 1, β to 10, δ to 0.1, and γ to 0.01. We focus on regressing information for the 15 major joints as defined in the SMPL model, instead of all joints. Additionally, we apply a moving average with a window size of 15 to the model's output, enhancing the smoothness of the predicted poses.

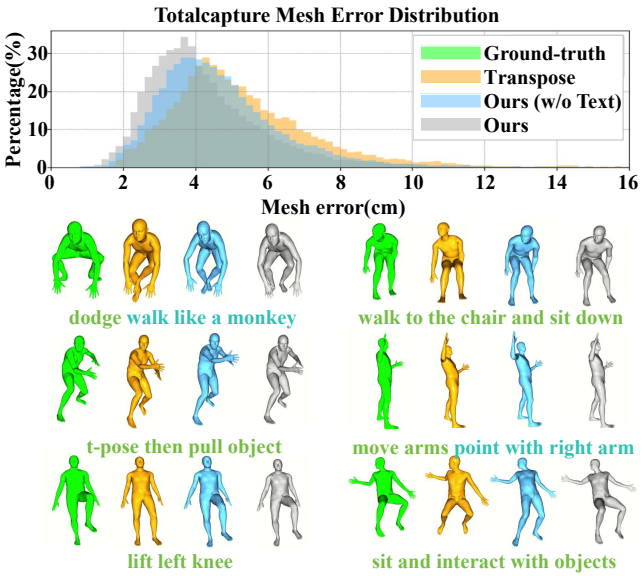

Figure 5: Mesh error distribution and qualitative comparisons between our method (with/without text) and Transpose. Text labels are below, with sequence labels in green and frame labels in blue.

Comparisons

We conducted quantitative and qualitative comparisons with SIP (Von Marcard et al. 2017), DIP (Huang et al. 2018), Transpose (Yi, Zhou, and Xu 2021), PIP (Yi et al. 2022) and TIP (Jiang et al. 2022b) on the Totalcapture and DIP-IMU datasets. In this comparison, we utilized the best-performing models published by the authors. For TIP, the authors employed a human body format different from ours. Therefore, we converted TIP's output into our format before conducting the comparison. The results for the Totalcapture and DIP datasets setting in offline mode are presented in Table 1. Unlike previous methods, our approach does not consider all IMU readings when estimating the current pose. However, our method achieves satisfactory results after integrating semantic information. The performance of our method on the DIP dataset is not as impressive as on the Totalcapture dataset, which can be attributed to the DIP dataset's fewer and less detailed semantic annotations. As shown in Fig 5, our method excels in processing ambiguous actions like standing and sitting, and is adept at capturing finer de-

tails, such as the accurate alignment of hands and feet with the ground truth. This demonstrates a more natural, realistic, and precise performance.

It is worth noting that our full model cannot reconstruct human motion in real-time due to the requirement for semantic annotation. Therefore, we employ only the Sensor Encoder and the HTT module for the evaluation in real-time mode. Our method accesses 70 past frames, 5 current frames, and 5 future frames through a sliding window approach, with a tolerable latency of 83 ms. As shown in Table 2, despite the absence of semantic information, our method still achieved superiority on multiple metrics, thereby validating the effectiveness of our network design.

The performance of our approach on the jitter metric is not as robust as other metrics, primarily owing to a constrained receptive field from the sliding window mechanism and the patch merging operation, which combines adjacent tokens into a single token. However, we posit that jitter, unlike the other four pose-accuracy metrics, isn't as critical. This perspective is based on the observation that visual discrepancies due to jitter are less noticeable when comparing our method with other approaches, while variations in pose precision are notably apparent.

Ablation

We perform three ablations to validate our key design choices: (1) without text semantic information; (2) without the Uncertainty-guided Spatial Attention (UGSA) module; (3) without the Hierarchical Temporal Transformer (HTT) module. Table 3 summarizes the results on the Totalcapture dataset (offline). Ablation experiments underscore the efficacy of our methodological design, with the integration of semantic information being the most salient contribution, followed by the implementation of UGSA and the HTT module.

Method	SIP Err	Ang Err	Pos Err	Mesh Err	Jitter
w/o Text	9.21/4.75	10.30/4.43	4.19/2.23	4.86/2.54	1.87/1.60
w/o UGSA	8.67/4.73	9.94/4.37	4.04/2.22	4.67/2.50	1.70/1.55
w/o HTT	8.35/4.57	9.70/4.29	3.89/2.10	4.52/2.35	**0.44/1.21**
Ours	**7.92/4.38**	**9.35/4.10**	**3.70/2.03**	**4.32/2.29**	1.74/1.55

Table 3: Evaluation of ablation models on the Totalcapture dataset. Bold numbers indicate the best performing entries.

Without semantic information, the model's predictions fluctuate in ambiguous situations, a phenomenon illustrated in Fig. 6 by the erratic alternation between sitting and standing positions. By incorporating a simple semantic annotation like "sitting", our model is able to maintain the desired sitting posture effectively.

Our findings indicate that the absence of Uncertainty-guided Spatial Attention affects the accuracy of the results. Fig. 7 illustrates how uncertainty fluctuates over time. Uncertainty increases across all sensors during complex movements like squatting and crawling, particularly in the hand regions. Conversely, a transition to a standing posture leads to a marked reduction in uncertainty, with the leg sensors showing the lowest levels.

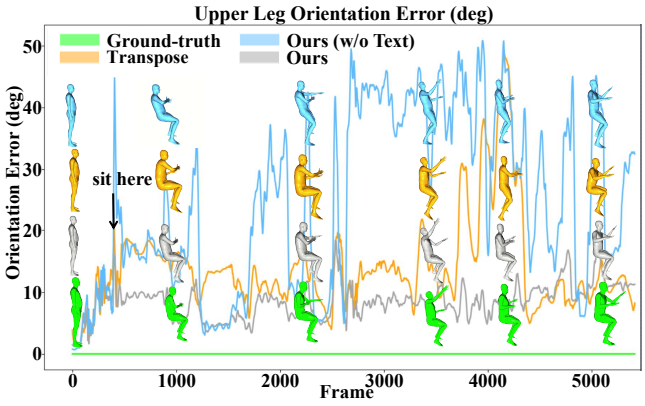

Figure 6: We demonstrated a comparison between our method (with/without text) and Transpose in a sitting situation, focusing on the analysis of upper leg rotation error.

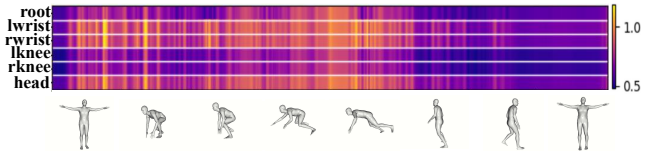

Figure 7: Temporal evolution of uncertainty across six sensors: each row represents a different sensor, with color variations indicating changes in uncertainty.

In examining the Hierarchical Temporal Transformer (HTT), we discern that employing window attention and patch merging within this module, instead of global attention, not only curtails computational needs but also elevates performance in almost all metrics, barring jitter. We consider such a trade-off to be acceptable.

These ablation findings affirm our approach's superior capacity for modeling sensor information and its ability to leverage semantic cues for generating more precise and natural movements.

Conclusion

In this paper, we are dedicated to addressing the ambiguity issues associated with using sparse inertial sensors for motion reconstruction. Our approach involves enhancing the sensor data modeling capabilities and incorporating textual supervision. In the realm of sensor data modeling, we introduced an Uncertainty-guided Spatial Attention Module to model spatial relationships amongst IMUs while considering their respective uncertainty. For the modal fusion, we leverage the Hierarchical Temporal Transformer module for achieving temporal alignment between sensor features and textual semantics. Furthermore, we employ contrastive learning to align features from both modalities in a high-dimensional space before fusion. Experimental results have validated the effectiveness of our method. Looking ahead, we plan to explore the integration of real-time execution capabilities into our framework. This could include the combination of natural language reasoning with motion data.

Acknowledgements

This article is sponsored by National Key R&D Program of China 2022ZD0118001, National Natural Science Foundation of China under Grant 61972028, 62332017, 62303043 and U22A2022, and Guangdong Basic and Applied Basic Research Foundation 2023A1515030177, 2021A1515012285.

References

Cao, Z.; Simon, T.; Wei, S.-E.; and Sheikh, Y. 2017. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 7291–7299.

Chen, L.; Ai, H.; Chen, R.; Zhuang, Z.; and Liu, S. 2020. Cross-view tracking for multi-human 3d pose estimation at over 100 fps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3279–3288.

Dittadi, A.; Dziadzio, S.; Cosker, D.; Lundell, B.; Cashman, T. J.; and Shotton, J. 2021. Full-body motion from a single head-mounted device: Generating smpl poses from partial observations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 11687–11697.

Guo, C.; Zou, S.; Zuo, X.; Wang, S.; Ji, W.; Li, X.; and Cheng, L. 2022. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5152–5161.

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, J.; Streli, P.; Qiu, H.; Fender, A.; Laich, L.; Snape, P.; and Holz, C. 2022a. Avatarposer: Articulated full-body pose tracking from sparse motion sensing. In *Proceedings of the European conference on computer vision (ECCV)*, 443–460. Springer.

Jiang, Y.; Ye, Y.; Gopinath, D.; Won, J.; Winkler, A. W.; and Liu, C. K. 2022b. Transformer Inertial Poser: Real-Time Human Motion Reconstruction from Sparse IMUs with Simultaneous Terrain Generation. In *SIGGRAPH Asia 2022 Conference Papers*, SA '22 Conference Papers.

Kendall, A.; and Gal, Y. 2017. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in Neural Information Processing Systems (NeurIPS)*, 30.

Kim, M.; and Lee, S. 2022. Fusion Poser: 3D Human Pose Estimation Using Sparse IMUs and Head Trackers in Real Time. *Sensors*, 22(13): 4846.

Kingma, D. P.; and Ba, J. 2017. Adam: A Method for Stochastic Optimization. arXiv:1412.6980.

Kingma, D. P.; and Welling, M. 2022. Auto-Encoding Variational Bayes. arXiv:1312.6114.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 10012–10022.

Loper, M.; Mahmood, N.; Romero, J.; Pons-Moll, G.; and Black, M. J. 2015. SMPL: A skinned multi-person linear model. *ACM Transactions on Graphics (TOG)*, 34(6): 1–16.

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 (ICCV)*, 5442–5451.

Malleson, C.; Gilbert, A.; Trumble, M.; Collomosse, J.; Hilton, A.; and Volino, M. 2017. Real-time full-body motion capture from video and imus. In *2017 International Conference on 3D Vision (3DV)*, 449–457. IEEE.

Moon, S.; Madotto, A.; Lin, Z.; Dirafzoon, A.; Saraf, A.; Bearman, A.; and Damavandi, B. 2022. IMU2CLIP: Multimodal Contrastive Learning for IMU Motion Sensors from Egocentric Videos and Text. arXiv:2210.14395.

Punnakkal, A. R.; Chandrasekaran, A.; Athanasiou, N.; Quiros-Ramirez, A.; and Black, M. J. 2021. BABEL: Bodies, Action and Behavior with English Labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 722–731.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 8748–8763. PMLR.

Schepers, M.; Giuberti, M.; Bellusci, G.; et al. 2018. Xsens MVN: Consistent tracking of human motion using inertial sensing. *Xsens Technol*, 1(8): 1–8.

Sengupta, A.; Budvytis, I.; and Cipolla, R. 2023. HuManiFlow: Ancestor-Conditioned Normalising Flows on SO (3) Manifolds for Human Pose and Shape Distribution Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4779–4789.

Slyper, R.; and Hodgins, J. K. 2008. Action capture with accelerometers. In *Proceedings of the 2008 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, 193–199.

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 (TOG)*, 30(3): 1–12.

Tevet, G.; Gordon, B.; Hertz, A.; Bermano, A. H.; and Cohen-Or, D. 2022. Motionclip: Exposing human motion generation to clip space. In *Proceedings of the European conference on computer vision (ECCV)*, 358–374. Springer.

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 (NeurIPS)*, 30.

Von Marcard, T.; Henschel, R.; Black, M. J.; Rosenhahn, B.; and Pons-Moll, G. 2018. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *Proceedings of the European conference on computer vision (ECCV)*, 601–617.

Von Marcard, T.; Pons-Moll, G.; and Rosenhahn, B. 2016. Human pose estimation from video and imus. *IEEE transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 38(8): 1533–1547.

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.

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 (CVPR)*, 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.

Zhang, M.; Cai, Z.; Pan, L.; Hong, F.; Guo, X.; Yang, L.; and Liu, Z. 2022. MotionDiffuse: Text-Driven Human Motion Generation with Diffusion Model. arXiv:2208.15001.

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 (CVPR)*, 5745–5753.