

Article

Not peer-reviewed version

SparseTrack: A Physics-Informed Transformer for Real-Time Human Motion Reconstruction from Sparse IMUs

[Adithya Balasubramanyam](#)*, [Suchir M Velpanur](#), [Sushma EJ](#), [Tejasree C J](#), [Prasad B Honnavalli](#), [Gowri Srinivasa](#)

Posted Date: 31 March 2026

doi: 10.20944/preprints202603.2438.v1

Keywords: sparse inertial sensing; human motion reconstruction; transformers; frequency-gated learning; biomechanical constraints; real-time motion capture; digital twins



Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a [Creative Commons CC BY 4.0 license](#), which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

SparseTrack: A Physics-Informed Transformer for Real-Time Human Motion Reconstruction from Sparse IMUs

Adithya Balasubramanyam *, Suchir M Velpanur, Sushma EJ, Tejasree C J, Prasad B Honnavalli and Gowri Srinivasa

PES University, Bengaluru, India
* Correspondence: adithyab@pes.edu

Abstract

Wearable inertial measurement units are widely used for human motion analysis because of their portability and independence from external infrastructure. However, many motion capture systems based on IMUs require dense configurations of sensors, which increase system complexity, cost, and require users to correctly place sensors, resulting in many usability issues in long-term, real-world applications. Reducing the number of wearable sensors is therefore essential for improving accessibility, but sparse sensing introduces challenges related to limited observability and biomechanical consistency. This paper introduces a sparse inertial human motion reconstruction framework using a minimal set of wearable sensors, with a focus on real-time operation and biomechanical plausibility. The framework integrates Movella Xsens DOT IMUs with a learning-based inverse kinematics pipeline and a real-time biomechanical digital twin for motion reconstruction and visualization. The evaluation process consists of two separate phases. The first part of the evaluation process requires the establishment of a real-time motion streaming system which will validate sensor alignment and the consistency of coordinate frames and end-to-end system latency. The second part of the evaluation process involves testing the learning-based sparse inference framework with motion data obtained from the Virginia Tech Natural Motion Dataset which tracks proximal joint orientations using distal inertial measurements collected under limited sensing capabilities. The results show that the established system can accurately reproduce human movements by using only five sensors, achieving a local Mean Per-Joint Position Error (MPJPE) of 5.96 cm. The results of the temporal backbone comparative ablation tests show that Transformer-based temporal modelling achieves better geometric accuracy and temporal smoothness than the recurrent and convolutional baseline models during training with extended temporal data. Further ablation studies show that using physics-informed regularization and hard negative mining is essential for maintaining biomechanical accuracy while also reducing high-frequency motion jitter. The real-time experiments demonstrate that the system maintains its operational performance during interactive latency limits which shows that sparse inertial motion capture technology works well for creating digital twin applications in the field of biomechanics.

Keywords: sparse inertial sensing; human motion reconstruction; transformers; frequency-gated learning; biomechanical constraints; real-time motion capture; digital twins

1. Introduction

1.1. Background and Motivation

Human motion tracking serves as an essential technology that supports multiple fields including biomechanics, sports medicine, physical therapy, ergonomics, safety engineering, and human-computer interaction [1–3]. The growing need for motion sensing technology has increased because digital systems now require sensors that deliver accurate results while being lightweight and scalable

to function in actual environments [4]. For instance, the tracking of patient physical rehabilitation and physical athletic conditioning involves the need for a motion capture system that functions well outside the laboratory [5]. The optical and marker-based systems used in traditional motion capture technologies serve as the primary standard for conducting kinematic studies [6,7]. The systems provide precise location tracking but require fixed multi-camera setups which demand extensive calibration work and ideal lighting conditions and can only operate within specific areas [8]. The financial and infrastructural burden of these systems is enormous therefore small clinics and local sports centres, and educational laboratories face significant challenges when trying to access these systems [9]. The marker-based capture system creates obtrusive tracking which prevents people from moving naturally while monitoring their activity over extended periods [6]. The field of motion analysis needs new sensing methods because current systems cannot perform well outside laboratory environments.

Wearable inertial measurement units (IMUs) have been shown to be a promising substitute for optical motion tracking systems [10]. Wearable IMUs are portable, which is a big advantage, and they don't require any infrastructure, thus giving continuous measurement of orientation, angular velocity, and linear acceleration [11]. Wearable sensors can be used for motion tracking in any setting such as a physiotherapy clinic, sport field, home, and outdoor setting [12]. Nevertheless, for full-body motion tracking with wearable inertial sensors, a dense network of sensors around the body may be needed, ranging in total number from 10 to 17 sensors [13]. Although such systems can achieve high accuracy, they may present usability challenges including user discomfort, longer setup times, higher costs, and reduced adherence, particularly among elderly patients or untrained users undergoing rehabilitation [14].

1.2. Problem Statement

Wearable motion capture technology presents a significant challenge in the accurate reconstruction of full-body human motion with only a small number of inertial sensors. A difficulty arises from the fact that the sensors may not be located on all body segments; as a result, the kinematic chain is only partially visible so that it becomes difficult to accurately determine the position of the joints. As a result, multiple poses can represent the same inertial data from different distances on the body leading to uncertainty in the estimation of the motion. This is particularly important for joints such as the shoulder and hip, which play a critical role in posture, balance, and limb coordination.

In addition to limited observability, other sources of error for inertial sensing systems are sensor noise, bias drift, magnetic disturbances, and sensor placement variation [15]. These sources of error can have a significant effect on system accuracy, especially for systems using sparse sensing configurations. Conventional kinematic models and filter-based approaches often fail to maintain accuracy and stability, especially when using sparse sensing configurations, unless strong assumptions are made or system responsiveness is sacrificed [16].

Such sparse sensor coverage can also result in unintended couplings between distal and proximal joints. In addition, it poses yet another challenge that is often ignored. The learning-based models may incorrectly associate the high-frequency motion of the distal joints (e.g., the wrists performing fast rotation movements) with the motion of the proximal joints, even when they are physically stationary. To address the aforesaid challenges, it is necessary to implement motion inference techniques that can accurately detect spurious correlations as opposed to actual motion.

1.3. Research Objectives

The principal research focus of this study is to establish a learning-based framework for reconstructing biomechanically valid human motion utilizing minimal number of wearable inertial sensors. Specifically, the study aims to assess whether full-body joint kinematics can be estimated from limited inertial measurement data through structured biomechanical framework systems. The study has created a sparse motion reconstruction technique designed as a temporal learning problem which will allow one to capture normal inter-joint coordination patterns from inertial sensor measurements without utilizing direct measurement techniques. The study compares various temporal modelling ar-

chitectures such as Recurrent Neural Networks (RNNs), Convolutional Neural Networks (CNNs), and Transformer-based Architectures under different sparse sensing conditions. Lastly, in order to reduce the effect of spurious distal to proximal joint coupling due to variations in kinematic representations and associated learning strategies, the study will demonstrate how the performance of the resulting reconstructions presented will be validated in real-time through the use of numerical performance metrics as well as through visualizing reconstruction in a biomechanical digital twin.

1.4. Contributions

The key contributions of this work are summarized as follows:

- We present a real-time learning-based sparse inertial motion reconstruction system that can recover unobserved proximal joint articulation from a minimal 5-IMU configuration under observational sparsity.
- We define sparse motion reconstruction as a temporally conditioned inverse kinematics problem and explicitly address the challenges of sparse observability and Kinematic Bleed-Through (KBT), which is an important anomaly in most motion capture systems where distal joint motions are wrongly attributed to proximal joints.
- We introduce a physics-informed frequency-gated Transformer architecture with a multi-task learning objective that enforces soft biomechanical constraints on angular velocities and linear accelerations to ensure temporal stability without any drift.
- We introduce a kinematically informed target prediction based on relative orientations and twist-swing decomposition to allow for the disambiguation of functional limb motions from high variance axial sensor noise.
- We introduce a data mining strategy to specifically address Kinematic Bleed-Through based on distal joint isolation tasks to ensure that the model can explicitly disambiguate distal oscillations from proximal motion priors.
- Quantitative and qualitative evaluations are presented to demonstrate low spatial and angular error in sparse sensing conditions while validating the robustness of our proposed system.

The rest of this paper is structured as follows. Section 2 discusses the related literature on sparse inertial motion capture. Sections 3 to 5 present the novel framework in terms of system design, data processing, and the physics-informed Transformer. Section 6 discusses the experimental design and evaluation metrics. Section 7 presents the experimental results in terms of quantitative evaluation, ablation studies, and real-time performance benchmarking. The rest of the sections 8, 9, 10 present the discussion on the implications, limitations, and future directions.

2. Related Work

Human motion can be captured using optical, depth, and wearable inertial sensors. Each sensor modality (optical, depth and inertial) provides unique trade-offs in terms of accuracy, usability and deployment. A variety of motion capture methods are supported by these sensor modalities, which are extensively used in human motion analysis.

2.1. Dense Motion Capture Systems

Dense motion capture systems based on optical technology have been the dominant source for capturing motion in both biomechanics and animated character production. Optical systems produce a precise spatial reconstruction of 3D joint trajectories due to the nature of tracking 3D spaces with multiple synchronized cameras and reflective markers located on the joints of the body. Optical systems can be used for biomechanical analysis performed in a laboratory environment, gait analysis performed in a clinical environment and animated character production [6].

Optical motion capture systems offer an extremely high level of accuracy; however, the systems have a variety of operational difficulties. The systems require a large number of cameras, each of which must be precisely calibrated. Additionally, the cameras operate under specific lighting conditions, and

the reflective markers utilized on the bodies must also remain in constant view of the cameras. These issues can all lead to decreased tracking accuracy because of poorly controlled lighting, occlusion of markers during complex movements or natural bodily motion, soft tissue artifacts, and a limited available capture area [17]. The requirement for fixed infrastructure for the systems inhibits their ability to be mobile, which leads to the unsuitability of these systems for long period monitoring and outside laboratory use [18]. Researchers have tested a variety of different sensing methods in an effort to solve the above issues and have placed an emphasis on using wearable sensors as well as multiple types of sensors from different manufacturers. A current area of research is developing real-time tracking of human poses using a combination of LiDAR and inertial measurement units (IMUs) to create a method for tracking human movement that does not rely on traditional camera based systems [19]. Other areas of research have explored methods for representing and editing motion and have studied the small movement of joints and how to recreate realistic paths of motion from collected sensor data [20]. These areas of research indicate an increase in the interest in developing flexible motion-capture frameworks that are capable of functioning outside of controlled laboratory environments.

2.2. Wearable Sensor-Based Motion Analysis

Motion capture and analysis using wearable inertial measurement units (IMUs) can be done effectively without using an optical system, because an IMU captures both linear acceleration and angular velocity directly. Sensor fusion algorithms are often used to estimate orientation and quaternions, and in most cases the orientation of the IMU is expressed as the unit quaternion, so that any problem with singularities caused by using Euler angles is avoided [15]. IMUs provide a means to capture motion wherever no other restrictions occur, such as rehabilitation clinics, sports fields, and home settings.

Early inertial motion capture systems that use filtering and kinematic models to estimate joint or segment orientations produced results that were valid over a short period of time. However, they were very sensitive to bias drift, sensor noise, and changes in sensor position which resulted in error accumulation over time [21]. In order to improve the performance, learning-based techniques were developed that attempted to take advantage of the recurring patterns of data in the inertial signals over time. RNN and CNN models achieved increased stability of learned motion references through data-driven learning and utilization of temporal patterns [22].

At the same time, the majority of currently working IMU-based systems for motion capture rely on extensive sensor configurations, which typically consist of 10 - 17 sensors that are placed on key body parts. Although these setups yield extremely accurate results, their accessibility in clinical and real-world applications is limited due to the disadvantages of greater setup complexity, user discomfort, and calibration overhead [23].

2.3. Sparse Sensing Approaches in Biomechanics

The primary goal of sparse inertial motion capture is to preserve biomechanical motion while restoring human movement using the fewest possible wearable sensors.

In earlier sparse sensing techniques, body pose estimation was addressed by formulating a constrained optimization problem that incorporated anthropometric priors and statistical body models to overcome measurement limitations. The Sparse Inertial Poser (SIP) showed the feasibility of reconstructing realistic full-body poses from data gathered by using as few as six inertial measurement units (IMUs), by imposing kinematic and anatomical constraints [22]. Yet, the global optimization of full motion sequences makes these approaches computationally expensive, making them unsuitable for real-time applications.

Data driven approaches have become the preferred means of posture reconstruction. The use of deep learning models has demonstrated that the orientation of proximal joints can be predicted using sparse inertial sensor measurements placed on the body, including distal segments [24]. Sequence to sequence architectures and recurrent networks are capable of achieving a reasonable amount of accuracy across a wide range of normal use; therefore, they are viewed as providing a viable solution across an increasingly large number of real world applications.

Recently, however, Transformer-based models have been explored to model the temporal variable in the inertial motion sequences. The Transformer Inertial Poser (TIP) utilized the self-attention mechanism to model long-range dependencies, thus improving the temporal consistency of the motion sequences over long periods [27]. Alternative hybrid approaches that included more spatial constraints or auxiliary sensors such as ultra-wideband ranging and LiDAR have also been used to reduce drift and improve global positioning systems [28].

Despite recent advances, a number of issues remain unresolved with sparse sensing systems. Several of these persistent issues include limitations of one-way observability leading to ambiguity, accumulation of drift, and unintentional coupled effects from distal-to-proximal joints when dealing with high-frequency, isolated movements of limbs. With fewer sensors, these issues become more challenging to resolve and lead to the need to develop training strategies that specifically consider biomechanical decoupling in sparse observations.

2.4. Research Gap Analysis

Existing motion capture systems have the ability to achieve high accuracy, especially because of the dense distribution of the sensors and large numbers of sensors working together, these same technologies also suffer from a range of limitations with regard to their usability and scalability.

On the other hand, most of the sparse inertial methodologies that can capture motion with fewer inertial sensors and motion capture hardware require very strong prior knowledge, an expensive offline optimization approach or very complex hardware/software designs, thus limiting their ability to function in a real-time environment. Additionally, although the usage of the Transformer architecture has shown promising results for motion reconstruction, it has not been used to overcome the challenges faced by sparse inertial methods. This includes the question of whether long-range temporal attention is a necessity for local kinematics. Numerical accuracy is a prominent aspect that has been considered, but the question of biomechanical plausibility, perceptual stability, and real-time digital twin validation has been given less importance.

It is the large gaps in this arena that are driving the development of the new sparse inertial motion reconstruction methods, where the new methods will: (i) provide for high reliability and usability based on a limited number of sensor configurations, (ii) reduce the distal-to-proximal coupling issues previously experienced in the capture of human motion, and (iii) provide real-time stability that is compatible with the use of these new methods in an interactive biomechanical application environment.

3. System Overview

3.1. Overall Framework Architecture

The proposed system includes a number of interconnected parts that work in concert to achieve sparse inertial motion reconstruction in real time. As depicted in Figure 1, the proposed architecture is divided into three stages. The first stage includes the sensing acquisition layer, where the inertial signals are obtained from a small number of IMUs attached to the body. Orientation, angular velocity, and linear acceleration are the three measurements that are wirelessly transmitted and then processed through a series of steps including calibration and coordinate alignment so that they are similar and compatible across all the sensors. As for the inference layer, it is a temporal learning model that predicts the orientations of the joints which are not measured by making use of the biomechanical dependencies that are present between the body segments that are instrumented and those that are uninstrumented. The joint rotations that are reconstructed are then applied to a humanoid skeleton and rendered in real-time using forward kinematics. This division into separate modules of sensing, inference and visualization gives the system the capability to be easily extended to additional joints or sensing modalities without any need for changes in the core pipeline.

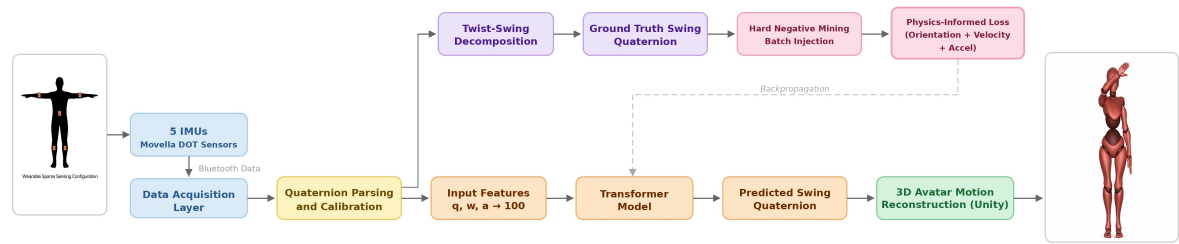


Figure 1. Overview of the proposed sparse inertial motion reconstruction architecture. The pipeline is fundamentally divided into sensing, inference, and visualization layers.

3.2. Sparse Sensor Configuration

The sparse sensing configuration involves a limited amount of inertial sensors to be put on the most informative body parts, i.e., the torso and distal segments of both the upper and lower limbs. As illustrated in Figure 2, five inertial measurement units (IMUs) are placed on the torso, both forearms, and both lower legs. The selection of these positions is intended to collect extensive motion dynamics while at the same time reducing the number of wearable sensing devices needed.

The distal limb segments offer strong kinematic cues and have motion patterns that are coordinated, which makes them dependent on the behaviour of the proximal joint. In the inference framework, dynamics of distal limb movement are used for inferencing. However, proximal joint movement is not measured. Instead, proximal orientation is inferred by learning the relation between distal limb dynamic behavior and proximal joint orientation. This method is based on well-established biomechanical principles for modeling limb segments as part of a kinematic chain. The proposed configuration reduces sensor requirements significantly while maintaining realism and accuracy of biomechanics by leveraging inter-segmental dependencies. This sparse sensing system is scalable and offers real-time capabilities for motion capture outside of a laboratory setting.

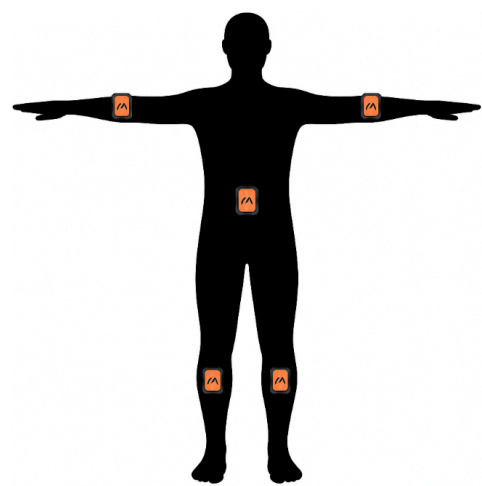


Figure 2. Wearable Sparse Sensing Configuration. Five inertial measurement units (IMUs) are deployed on the distal body segments (wrists and ankles) and the trunk (torso), aiming to capture the human body movements.

3.3. Data Flow and Processing

The data processing pipeline starts with the continuous capture of inertial measurements from the wearable IMUs, which consist of orientation quaternions, angular velocity, and linear acceleration. The data coming from the sensors is synchronized and then sent to the calibration stage, where a reference posture is used to align the coordinate frames of the sensors with those of the virtual skeletal model. Mounting offsets are corrected, and all the orientations are transformed into a common coordinate system that is compatible with animation and learning.

In the case of learning-based reconstruction, the reference motion data is first processed offline to calculate the relative joint orientations between the distal and proximal segments. The relative rotations are then split up into twist and swing components, and the swing rotation, which is responsible for the

functional limb movement, is isolated. At the runtime stage, only the distal IMU measurements are fed into the temporal learning model that predicts the proximal joint swing orientation for the next time step. The predicted orientation is subjected to normalization to validate it on the rotation manifold and is combined with the measured sensor data to get the complete limb pose. The joint rotations that have been reconstructed are applied to a hierarchical skeletal model through forward kinematics, enabling the motion to disperse uniformly through the kinematic chain. This final articulated pose is shown in real time within a Unity-based digital twin, which provides a stable visualization with low latency. This closed-loop system allows for real-time motion reconstruction using minimal sensor data, as well as the evaluation of the quality of the reconstructed motion via both numeric metrics and perceptual analysis.

4. Data Acquisition and Preprocessing

4.1. Sensor Hardware Description

The sparse sensing paradigm presented in the present work utilizes the Movella Xsens DOT (2nd Gen) development kit. The system setup consists of five wireless sensor units placed at strategically located distal landmarks to maximize biomechanical observability. Each sensor unit is a small-sized inertial measurement unit (IMU) with IP68 rating, which includes a high-fidelity 9-axis MEMS module. The detailed specifications of the sensor hardware can be viewed in Table 1 below, as described in the technical specifications of the sensor hardware as per manufacturer’s technical documentation [29].

Table 1. Technical Specifications of the Data Acquisition Hardware [29]

(a) Physical Properties			
Model	Dimensions	Weight	IP Rating
Xsens DOT (2nd Gen)	36.3 × 30.35 × 10.8 mm	11.2 g	IP68 (Water/Dust)
(b) Sensor Specifications			
Module	Full Scale Range	Sensitivity [30]	Rate
Accelerometer	±16 g	2048 LSB/g	800 Hz
Gyroscope	±2000 °/s	16.4 LSB/°/s	800 Hz
Magnetometer	±8 Gauss	0.25 mGauss	60 Hz

4.2. Signal Processing and Transmission Pipeline

Although the internal MEMS components sample at 800 Hz, the onboard processor performs Strap-Down Integration (SDI) to downsample the data to a stable 60 Hz output rate. This ensures that high-frequency motion components are preserved through orientation and velocity increments despite the reduced output rate. In addition, the Movella XKFCore sensor fusion algorithm provides orientation information as a unit quaternion along with the raw calibrated inertial sensor readings. To address the issue of ferromagnetic interference, the magnetometer is only utilized for the initial heading stabilization and is not enabled during the dynamic data collection.

Thereafter, the data transmission is carried out via Bluetooth Low Energy (BLE 5.0), connecting to a central processing workstation running custom bridge software. To facilitate the real-time visualization and debugging (as shown by the digital twin configuration in Figure 1), the hardware interface is managed by a custom-designed MQTT publisher. This middleware utilizes the Movella SDK classes to manage BLE connectivity, battery level monitoring, and the synchronization of the data packets from the five nodes to provide a consistent data stream for the downstream Transformer model.

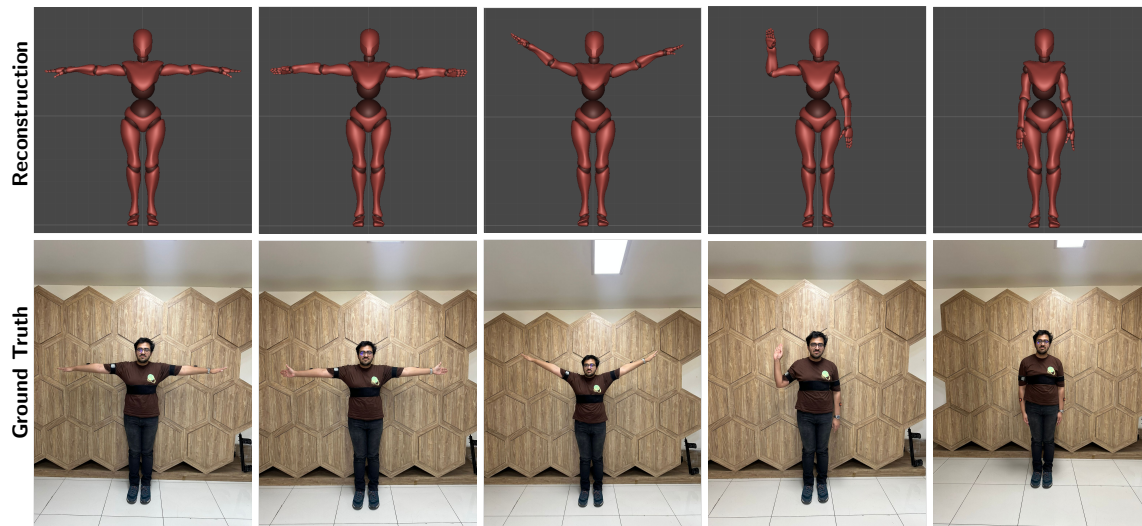


Figure 3. Real-Time Biomechanical Digital Twin. Visualization of the real-time motion streaming pipeline. (Top) The avatar is driven by sensor data in Unity. (Bottom) The corresponding real-world human reference. This configuration is aimed at validating the synchronization of the coordinate frames and the sensor alignments (as discussed in 4.5.1), which are the prerequisites for training the sparse inference model.

4.3. Experimental Data Collection Protocol

All data collection activities with human participants were carried out in accordance with the Declaration of Helsinki [31]. The study participants' involvement was entirely voluntary, with their full consent obtained for the study. Moreover, there was no collection of personally identifiable information. The study did not require institutional review board permission because it posed a minimal risk to the participants and used non-invasive wearable sensors.

During the experiment, a particular data injection procedure known as a 'Hard Negative Mining' was included in the study to mitigate Kinematic Bleed-Through (KBT). Kinematic Bleed-Through is a phenomenon that occurs in sparse IMU motion capture wherein the estimation of a previously unobserved proximal joint configuration is affected by the propagation of distal joint artifacts [26]. This phenomenon leads to a situation where a particular distal pose is related to multiple valid proximal poses. Kinematic Bleed-Through is particularly prominent when the global drift of the root is ignored and local constraints are solely considered [32]. The relationship between the distal joint oscillations and the proximal joint stability leads to Kinematic Bleed-Through, where high-frequency oscillations in the distal sensors are related to the motion of the proximal joints, thereby violating the anatomical constraints of the human kinematic chain [33].

In order to combat this problem, an additional dataset was created to include "Hard Negative" motions. This term refers to motions that are statistically rare in an Activities of Daily Living (ADL) dataset but are still biomechanically valid. Even though an ADL dataset contains a large number of motions that are synergistic in nature, they fundamentally lack the degree of distal isolation required to train a versatile controller. By recording these motions explicitly, the dataset is augmented to include the entire decision boundary of human articulation.

In order to satisfy the distal-proximal decoupling constraint explicitly, distal joint isolation tasks were recorded. Although this was not specifically constrained to the following, the main tasks included several highly constrained motions. To start with, a locked-humerus elbow flexion was recorded. This involved rapid sagittal forearm motions (0.5–2.0 Hz) with the humerus constrained against the ribcage. This motion specifically forces the model to decouple these high-velocity motions from spurious shoulder activity. Next, wrist pronation and supination motions were recorded. This involved high-frequency rotational motions with the elbow constrained to 90°. This specifically eliminates wrist-shoulder bleed-through. Additionally, fixed-thigh knee extensions involving rapid

shank oscillations were performed in a seated posture with a strictly stationary femur, successfully isolating knee dynamics from hip actuation.

In all of these high frequency tasks and other distal isolation tasks within this protocol, participants were also required to physically restrict their proximal joint, specifically through locking their upper arms against their torso, in order to obtain a near-zero proximal angular velocity ($\omega_{\text{proximal}} \approx 0$). All physical trials were subsequently recorded using Movella Xsens DOT sensors, which were then converted into a synchronized CSV format. To ensure the integrity of the labels, stationarity was checked using proximal joint estimates from the full-body reference pipeline. A strict tolerance of $|\omega| < 0.1$ rad/s was applied for all proximal joints, ensuring that the kinematics of the recorded distal joints correctly corresponded to proximal joint stability. This ensures that the Transformer model can correctly decouple high-velocity distal joint signals from any false proximal joint activation.

4.4. Hybrid Dataset Composition

In order for the model to generalize correctly to natural human motion, while also ensuring robustness to infrequent observation ambiguity, a hybrid dataset was constructed from two different sources. First, a primary training corpus was constructed from the Virginia Tech Natural Motion Dataset (VT-NMD) [23], which is currently our standard ground truth for natural human kinematics. The base dataset consists of around 40 hours of unscripted, open-world activities of daily living (ADLs) captured using an industry-standard Xsens MVN Link full-body inertial motion capture suit consisting of 17 sensors. The dataset has 17 subjects, consisting of 13 subjects recorded on a college campus and 4 subjects recorded in a home improvement store, performing a range of activities that include locomotion, object interaction, physical exercises, and household activities [23].

Using the given code for the data analysis [23], the full-body kinematic chain was extracted from the source MVNX data, providing ground truth orientations for the proximal joints (upper arms, thighs). A “Sparse Mask” was applied to the data, retaining only the quaternion and inertial data for the five target locations, i.e., the sternum (torso), the left and right wrists, and the left and right ankles, as shown in Figure 2.

Furthermore, the natural motion dataset was also augmented with the “Hard Negative” injection dataset, which was recorded based on the protocol given in Section 4.3. This dataset targets the data for high-frequency distal joint isolation, recorded as targeted correction data. As for the training, a hard negative injection ratio was used, where the correction data was injected in every training batch with an optimal ratio of $\approx 10\%$ (Section 7.3.3). Lower injection ratios led to the empirical evidence of false positives in the form of persistent proximal activation during rapid distal motion, while higher ratios negatively affected performance on general activities of daily living.

4.5. Computational Preprocessing and Feature Engineering

As a result of the above steps, the raw sensor streams are then subjected to a rigorous computational preprocessing pipeline, which ensures spatial consistency, numerical stability, and kinematic isolation prior to network inference. This pipeline can be broadly categorized as frame calibration, signal sanitization, and geometric feature engineering.

4.5.1. Calibration & Frame Transformation

Prior to every data session, a static T-pose calibration was conducted, which aligns the sensor coordinate frames with the corresponding anatomical body frames. This alignment ensures that the input quaternion q_{input} represents the limb orientation relative to a gravity-aligned global reference frame, invariant to arbitrary sensor mounting placement on the skin. All subsequent transformations are performed deterministically before network inference and are not subjected to learning during training. The frame transformation pipeline implemented by the controller middleware is composed of two mathematical operations.

First, a coordinate system permutation denoted by (T_{coord}) is applied. The raw IMU quaternions, $q_{\text{raw}} = [w, x, y, z]$, are represented in the sensor’s right-handed East-North-Up (ENU) coordinate frame.

To convert this into the global coordinate frame of the application (Unity, left-handed, Y-up), a basis permutation is applied. In accordance with limb-specific mounting convention, the permutation function, $T_{\text{coord}}(q_{\text{raw}})$, is defined as:

$$q_{\text{global}} = \begin{cases} [w, y, -x, z] & \text{for upper body segments} \\ [w, y, -x, -z] & \text{for lower body segments} \end{cases} \quad (1)$$

The sign difference in the z-axis component is meant to account for the sensor mounting orientation relative to the limb's proximal joint axis, ensuring consistent alignment with the virtual world's vertical axis.

Subsequently, the boresight alignment step (T_{calib}) is carried out to account for the sensor mounting variability. The relative boresight quaternion is established during the static calibration step. We define the orientation established during the T-pose at time t_0 as $q_{\text{global}}(t_0)$. The offset for the boresight alignment is computed by taking the inverse of the established orientation:

$$q_{\text{offset}} = q_{\text{global}}(t_0)^{-1} = [w_0, -x_0, -y_0, -z_0] \quad (2)$$

During dynamic motion ($t > 0$), the calibrated orientation is computed by the following equation:

$$q_{\text{calibrated}}(t) = q_{\text{offset}} \otimes q_{\text{global}}(t) \quad (3)$$

This sequence ensures that $q_{\text{calibrated}}(t_0) = [1, 0, 0, 0]$, effectively removing the sensor mounting bias and measuring the true rotation of the limb segment.

4.5.2. Signal Preprocessing and Temporal Segmentation

Raw sensor data is subjected to a rigorous cleaning and formatting pipeline, which ensures numerical stability for the Transformer architecture. First, in order to synchronize the system and reduce computational overhead while preserving high-frequency inertial features, the internal 800 Hz sensor data is downsampled using Strap-Down Integration (SDI) to 60 Hz [29], thereby matching the rendering frame rate of the Unity client. Next, prior to sequence generation, all data streams are subject to a scan for numerical anomalies. In this regard, files containing infinite numbers or extensive NaN sequences, often resulting from Bluetooth packet loss or sensor saturation, are automatically detected and removed from the training pipeline in order to avoid gradient explosion.

Finally, in order to capture temporal features of motion, the continuous time-series data is segmented into fixed-length sequences using a sliding window approach. A sequence length (L) of 120 frames, or 2.0 seconds, is applied, which is sufficient to capture multi-joint coordination while maintaining compatibility with real-time inference constraints. This yields an input tensor of $(B, 120, 10)$ per distal limb, where B is batch size and 10 is input feature dimension per time step. Consequently, all individual limb tensors are concatenated into a unified input stream for the Transformer network.

4.5.3. Feature Representation and Target Reparameterization

A specific set of geometric and inertial features is extracted, which have been chosen with the goal of decoupling the proximal and the distal kinematic chains. The model is provided with the 10-dimensional feature vector ($X_t \in \mathbb{R}^{10}$) at each time step t , representing the state of the distal limb. This feature vector has three parts: the unit quaternion ($q_{\text{distal}} \in \mathbb{R}^4$) representing the orientation of the limb with respect to the global vertical frame, the calibrated angular velocity ($\omega_{\text{distal}} \in \mathbb{R}^3$) and calibrated linear acceleration ($a_{\text{distal}} \in \mathbb{R}^3$).

To ensure numerical stability and the maintenance of proper gradients during the course of the training process, different normalization techniques are applied to different feature groups. The unit quaternions are strictly L2-normalized, thus maintaining the unit norm constraint on the $SO(3)$

manifold. In contrast, global z-normalization has been applied to the inertial features ω_{distal} and a_{distal} . This choice of standardization technique has been motivated by the high kurtosis nature associated with rapid impact accelerations.

In this context, the standard min-max normalization technique has been deemed insufficient. To address this, a standardization technique with outlier clamping has thus been employed:

$$x_{norm} = \text{clamp}\left(\frac{x - \mu}{\sigma + \epsilon}, -5.0, 5.0\right) \quad (4)$$

where μ and σ represent the global mean and standard deviation, respectively, computed over the entire training corpus for each sensor channel. The clamping threshold of $\pm 5\sigma$ was selected to account for rare yet biomechanically valid high-velocity motions that will be used as hard negative samples to train the network. The threshold covers approximately 99.9999% of the distribution, ensuring that extreme yet physically plausible kinematics are retained while removing all non-biological outliers that may result from sensor impact or hardware transients.

Subsequent to this step, the prediction target is reparameterized via a twist-swing decomposition (Figure 4) to decouple anatomically meaningful motion components. The orientations of the proximal and distal limbs are described via unit quaternions $q_{proximal}$ and q_{distal} respectively.

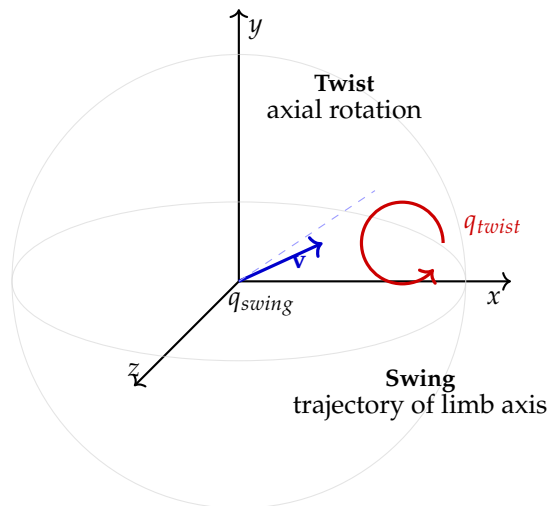


Figure 4. Geometric swing-twist decomposition of relative limb orientation (q_{rel}). Decomposing rotations into swing (q_{swing}) and twist (q_{twist}) components effectively disentangles the limb's spatial trajectory from noisy axial sensor data.

The process starts with computing the relative rotation needed to align the distal frame to the proximal frame to isolate joint articulation from global heading:

$$q_{rel} = q_{proximal} \otimes q_{distal}^{-1} \quad (5)$$

where $\otimes$ denotes the Hamilton quaternion product and q_{distal}^{-1} denotes the quaternion inverse (conjugate) of the distal quaternion.

Next, the twist component is extracted according to the twist-swing decomposition principles described in Motion-Sphere [20]. We define the anatomical twist axis of the limb, $\mathbf{u}$, as the x-axis vector $\mathbf{u} = [1, 0, 0]^T$. Given $q_{rel} = [w, \mathbf{v}]$, where w is the scalar part and $\mathbf{v} = [x, y, z]^T$ is the vector part, we extract the twist component (rotation around the limb axis) by projecting $\mathbf{v}$ onto $\mathbf{u}$:

$$\mathbf{p}_{proj} = (\mathbf{v} \cdot \mathbf{u})\mathbf{u} \quad (6)$$

The non-normalized twist quaternion is constructed as $q_{raw_twist} = [w, \mathbf{p}_{proj}]$. The final unit twist quaternion is:

$$q_{twist} = \frac{q_{raw_twist}}{\|q_{raw_twist}\|} \quad (7)$$

Finally, the swing component, representing the trajectory of the limb axis (the pointing direction) independent of axial rotation, is isolated according to the formulations in Motion-Sphere [20]. Since the total rotation is the composition of swing and twist ($q_{rel} = q_{swing} \otimes q_{twist}$), we isolate the swing component by multiplying by the inverse twist:

$$q_{swing} = q_{rel} \otimes q_{twist}^{-1} \quad (8)$$

The model is trained to minimise the error between the predicted swing $\hat{q}_{swing}$ and its ground truth q_{swing} . This helps to mitigate noise in axial rotation, which is typically seen in wrist-worn sensors during sparse observation tasks.

5. Proposed Sparse Sensing Methodology

5.1. Sparse Observability Problem

The main challenge of this research is to solve a learning-based inverse kinematics reconstruction problem under conditions of extreme observational sparsity. We define the complete kinematic state at time t as $\mathbf{S}_t = \{\mathbf{J}_{proximal}^t, \mathbf{J}_{distal}^t\}$, where each joint state $\mathbf{J}$ contains both orientation and inertial information. The observable input $\mathbf{X}_t$ is defined as a feature embedding derived from $\mathbf{J}_{distal}^t$ and the torso reference.

In a standard Motion Capture (MoCap) setup (e.g., Xsens MVN Link), the observation function $\mathcal{O}_{full}$ is bijective and represents the complete state space:

$$\mathbf{Z}_t = \mathcal{O}_{full}(\mathbf{S}_t) \approx \mathbf{S}_t \quad (9)$$

However, in the proposed framework, the observation function $\mathcal{O}_{sparse}$ plays the role of a masking operator, allowing only the distal states (end effectors) and the root reference (torso) to be visible, while making the proximal states latent. The absence of secondary sensors on the proximal limb segments and global position information (e.g., UWB anchors and optical cameras) leads to a fundamental ill-posed problem. We formally define this condition as *Sparse Observability*. Under this condition, a single instantaneous distal observation (e.g., wrist orientation) is associated with a continuous manifold of valid proximal configurations (e.g., varying elbow swivel angles). As illustrated in Figure 5, this motion duality renders the inverse kinematic problem unsolvable at any single time step.



Figure 5. Visualization of the Sparse Observability Problem. A single distal observation (wrist orientation) corresponds to multiple valid proximal configurations. Without temporal context, a sparse solver collapses these distinct states into a single ambiguous prediction.

To resolve this, we formalize the reconstruction task as learning a mapping function $\mathcal{F}_\theta$ (parameterized by the Transformer) that estimates the conditional probability of the proximal state given a temporal history of distal observations:

$$P(\mathbf{J}_{\text{proximal}}^t \mid \mathbf{X}_{t-L:t}) \quad (10)$$

In practice, this conditional distribution is approximated via deterministic regression using an encoder-only Transformer. The network is trained to minimize the expected reconstruction error over the proximal joint space by mapping the input tensor sequence $\mathbf{X}_{t-L:t}$ (a temporal context of length L containing features from the five observed IMUs) directly to the target state vector $\mathbf{J}_{\text{proximal}}^t$ for the uninstrumented proximal joints (e.g., shoulders and thighs) at time t .

5.2. Sensor Placement Strategy: The “Outside-In” Paradigm

The 5-sensor configuration (illustrated in Figure 2) represents a near-minimal subset sufficient to constrain the primary kinematic chains of the human body.

Typically, in conventional inverse kinematics, the focus is placed on proximal limb segments (e.g., the upper arm) to drive the kinematic chain forward. In contrast, we adopt a different strategy, which we term the “outside-in” tracking paradigm, where emphasis is placed on distal limbs, namely the wrists and ankles. This approach is based on biomechanical principle that distal limb segments act as *kinematic integrators*, accumulating the composite motion of the entire upstream chain. Consequently, placing sensors on distal limb segments implicitly captures latent information from intermediate segments through kinematic constraints. This is not the case for proximal limb sensors, which lack observability of downstream end-effector dynamics.

This configuration achieves two important purposes: first, it maximizes the information density in the sparse input vector, directly addressing the observability constraints as discussed in Section 5.1 and secondly, it captures the high-frequency dynamics in the end-effectors, which are a prerequisite for the proposed frequency-dependent gating mechanism in the Transformer, which in turn enables the attention heads to mitigate kinematic bleed-through by associating high-frequency distal joint oscillation with induced stationarity in the proximal joint, as discussed in Section 5.4.

5.3. Temporal Modeling of Motion Dynamics

In addressing the physics observability and motion duality challenges, wherein an instantaneous distal joint configuration has multiple corresponding valid configurations of the proximal joint, we formulate the reconstruction problem as a Sequence-to-Vector regression task. In this way, the system can leverage the temporal history of dynamics in distal joint configurations to resolve the ambiguity in proximal joint configurations.

5.3.1. Input Phase Space Formulation

The input sequence $\mathbf{X} \in \mathbb{R}^{L \times D_{in}}$ is defined as a sliding window over last L frames of the sequence history. Through extensive empirical ablation study (as discussed in Section 7.3.4), an optimal context window length $L = 120$ frames, or 2.0 seconds at 60 Hz, has been chosen. This time window is critical in addressing the multi-scale aspects of human biomechanics, as it must be long enough to capture the low-frequency preparation phases of the movement (e.g., the subtle torso lean preceding a reach) yet also short enough to capture the high-frequency inertial content of the movement necessary to detect rapid oscillations in the distal joint.

The dimension of the input sequence D_{in} is set to 10, and the feature dimension at each time step t is given by:

$$\mathbf{x}_t = [\mathbf{q}_t, \boldsymbol{\omega}_t, \mathbf{a}_t]^T \quad (11)$$

In this equation, $\mathbf{q}_t \in \mathbb{R}^4$ is the unit quaternion describing the orientation of the distal limb with respect to the global vertical frame, and $\boldsymbol{\omega}_t \in \mathbb{R}^3$ and $\mathbf{a}_t \in \mathbb{R}^3$ are the robustly normalized angular velocity and linear acceleration of the limb, respectively. Crucially, these inertial derivatives are not merely

auxiliary; they provide the high-frequency jitter signals that the Transformer's attention mechanism utilizes to explicitly identify the relationship between distal joint chaos and proximal joint stability.

5.3.2. Temporal Injection via Positional Encoding

The permutation-invariance property of the basic Transformer architecture necessitates the injection of explicit temporal ordering into the input embeddings. Standard sinusoidal Positional Encodings (PE) [38] are directly added to the input embeddings to preserve the sequential nature of the motion. For a position pos and dimension $2i$ or $2i + 1$:

$$PE_{(pos,2i)} = \sin\left(\frac{pos}{10000^{2i/d_{model}}}\right) \quad (12)$$

$$PE_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{2i/d_{model}}}\right) \quad (13)$$

The continuous nature of this PE allows the model to attend to differences in temporal context. This allows for the differentiation between immediate dynamic events (e.g, an impact jerk at time t) and causal history in the trajectory (e.g, velocity initialization at time $t - 50$).

5.4. Learning Architecture

A custom encoder-only variant of the Transformer architecture is employed for high-fidelity signal reconstruction. In contrast to recurrent architectures (e.g, LSTM and GRU), which exhibit limitations in capturing long-range temporal dependencies over extended contexts ($L > 100$), the Transformer's global receptive field facilitates the learning of long-range dependencies [38].

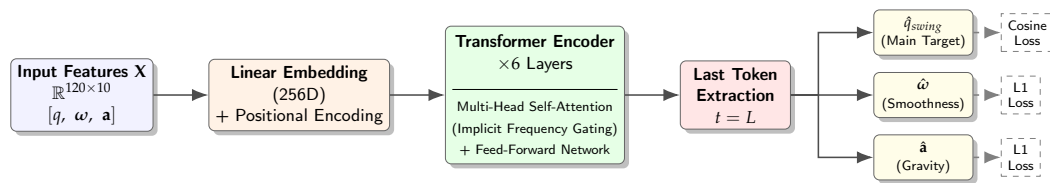


Figure 6. Schematic of the attention-guided Transformer-based architecture for biomechanically constrained sparse motion reconstruction. It employs an encoder-only Transformer to capture distal–proximal synergies via implicit frequency gating, enabling the learning of latent states of uninstrumented joints over a temporal window ($L = 120$). A multi-task head performs regression of orientation, smoothness, and gravity consistency, optimized through a physics-informed objective.

5.4.1. Network Topology

As depicted in Figure 6, the architectural backbone consists of $N = 6$ identical layers of the Transformer encoder with a latent dimension $d_{model} = 256$ and $h = 8$ attention heads. This architecture results in a total of 4,743,946 trainable parameters ($\approx 4.74M$), being sufficient to represent the complex kinematic dependencies between the inputs while keeping the architecture computationally efficient enough to be used for real-time processing at 60 Hz. Each layer of the encoder uses a Pre-LayerNorm configuration to improve the training stability of deeper layers. Each encoder layer consists of two components: (i) Multi-Head Self-Attention (MSA) followed by (ii) a position-wise Feed-Forward Network (FFN).

Multi-Head Self-Attention (MSA) module aggregates value vectors through attention weights computed from query–key interactions. Mathematically, for a single attention head, the attention operation can be written as:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right) \mathbf{V} \quad (14)$$

where d_k represents the dimensionality of each attention head. This mechanism inherently learns a frequency-dependent gating mechanism to help mitigate the Kinematic Bleed-Through (KBT) problem.

The proposed architecture learns a specific synergy between high-frequency oscillatory features of the distal input (ω_{distal} , e.g., rapid wrist shaking) and predicts the stationarity of the proximal input. Effectively, chaos in the lower arm is attended to as a signal for stability in the upper arm, allowing the model to gate the propagation of distal noise and prevent temporally inconsistent high-frequency inputs from corrupting proximal joint predictions.

Position-wise Feed-Forward Network (FFN) applies a fully connected nonlinear transformation to each position independently, with an expansion dimension of $d_{ff} = 1024$ and GELU activation:

$$\text{FFN}(x) = \text{GELU}(x\mathbf{W}_1 + \mathbf{b}_1)\mathbf{W}_2 + \mathbf{b}_2 \quad (15)$$

5.4.2. Multitask Projection Head

The final latent representation of the last time step (z_L) is projected via parallel linear heads to reconstruct the full kinematic state, treating the reconstruction as a Multi-Task Learning (MTL) problem. Specifically, the network predicts the swing quaternion ($\hat{q}_{swing} \in \mathbb{R}^4$) as the primary orientation target decoupled from axial twist, the angular velocity ($\hat{\omega} \in \mathbb{R}^3$) to enforce temporal smoothness constraints, and the linear acceleration ($\hat{\mathbf{a}} \in \mathbb{R}^3$) to ensure consistency with gravity vectors.

5.4.3. Physics-Informed Multi-Task Learning (MTL) Objective

To enforce biomechanical plausibility and model a general-purpose controller capable of handling both ADLs and high-frequency distal joint isolation tasks, we minimize a composite MTL loss function:

$$\mathcal{L}_{MTL} = \lambda_q \mathcal{L}_{quat} + \lambda_\omega \mathcal{L}_{vel} + \lambda_a \mathcal{L}_{acc} \quad (16)$$

This objective consists of three distinct components.

Manifold-Aware Orientation Loss ($\mathcal{L}_{quat}$) utilizes a cosine similarity loss formulation,

$$\mathcal{L}_{quat} = 1 - |\hat{q} \cdot q_{gt}| \quad (\lambda_q = 1.0) \quad (17)$$

to address the antipodal symmetry (double cover) property of the unit quaternion hypersphere $\mathbb{S}^3$, where q and $-q$ represent identical rotations in $SO(3)$. Unlike Euclidean metrics (e.g., MSE), this formulation ensures invariance of the loss to global sign ambiguity ($q \equiv -q$).

Stationarity soft constraint ($\mathcal{L}_{vel}$) acts as the driver for our hard negative mining strategy. We apply a penalty ($\lambda_\omega = 0.1$) to angular velocity errors specifically during hard negative batches (selected via hyperparameter ablation in Section 7.3.2). In a sparse inertial setup lacking global position data (e.g., cameras or UWB), enforcing hard zero-velocity constraints leads to integral drift accumulation. KBT is therefore mitigated as a *physics-guided soft constraint*, forcing the model to minimize drift while allowing the micro-adjustments required to maintain kinematic continuity without requiring a secondary sensor on the proximal limb.

Auxiliary regularization ($\mathcal{L}_{acc}$) introduces a minor loss term ($\lambda_a = 0.01$) (Section 7.3.2) applied to linear acceleration predictions to encourage trajectory smoothness and gravity vector consistency.

5.5. Post-Processing & Kinematic Constraints

While the Transformer effectively models the temporal dynamics of human motion, raw neural network predictions can occasionally violate physiological limits, particularly in unseen scenarios. To further enforce biomechanical validity and constrain the motion duality manifold (Section 5.1), we implement a lightweight post-processing stack at inference.

First, conservative anatomical constraints are applied. Since the model predicts relative orientations, we enforce kinematic constraints by mapping the relative quaternion into a local joint-aligned Euler parameterization, apply bounded constraints, and re-project the result back onto $SO(3)$. This step acts as a deterministic complement to the learned frequency-dependent gating (Section 5.4). While the Transformer learns to minimize Kinematic Bleed-Through (KBT) via soft constraints, this hard clamping layer ensures that residual bleed-through does not result in biomechanically invalid poses.

Specifically, the elbow is modeled as a hinge joint with relative flexion θ_{elbow} constrained to $[5^\circ, 150^\circ]$ to prevent hyperextension or physically impossible flexion, while the glenohumeral (shoulder) joint is restricted within a bounded 3D cone (e.g., Pitch/Yaw/Roll axes: $[-45^\circ, 150^\circ]$) to prevent the arm from clipping through the torso. These conservative limits are chosen to ensure physiological plausibility without over-restricting the natural range of motion.

Following this, temporal smoothing is performed using Kalman filter to eliminate high-frequency jitter, which is commonly observed during frame-by-frame regression. We implement a recursive Linear Kalman Filter directly on the output quaternions, where components are modeled as the state vector $x_k \in \mathbb{R}^4$. The filter dynamically blends the Transformer's prediction with the prior estimate using a constant process noise ($\sigma_p^2 = 10^{-6}$) and measurement noise variance ($\sigma_m^2 = 10^{-3}$):

$$K_k = \frac{P_{k-1} + \sigma_p^2}{P_{k-1} + \sigma_p^2 + \sigma_m^2} \quad (18)$$

$$\hat{x}_k = \hat{x}_{k-1} + K_k(z_k - \hat{x}_{k-1}) \quad (19)$$

The quaternion $\hat{x}_k$ is then re-normalized after each update to maintain unit norm. This acts as an adaptive low-pass filter, smoothing motion while maintaining lower latency than standard moving-average techniques.

5.6. Real-Time Inference & Calibration Mechanism

The system architecture decouples the deep learning inference engine from the Unity-based visualization client through an asynchronous MQTT bridge, enabling modularity and fault tolerance.

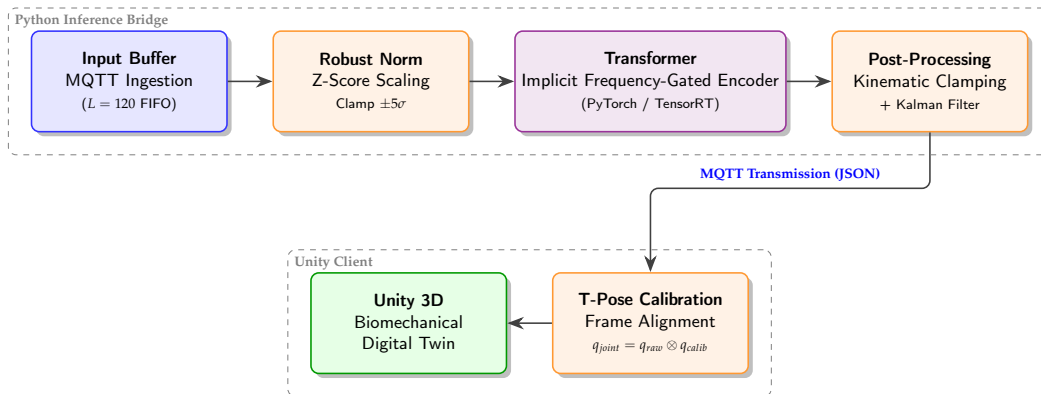


Figure 7. Pipeline for real-time motion inference and visualization. Our system separates the real-time inference engine from the Unity client using an asynchronous MQTT bridge. Sensor data is processed through buffering, normalization, and Transformer-based regression, followed by post-processing and T-pose calibration for visualization. This modular approach provides a stable and low-latency motion reconstruction suitable for real-time interaction (60 Hz).

The real-time inference loop is managed by an inference bridge that maintains a sliding window buffer of length $L = 120$ (Figure 7), where distal sensor packets are ingested via MQTT and inserted into a first-in-first-out (FIFO) queue for processing. Once the buffer has enough data, normalization statistics, as described in Section 4.5.3, are calculated immediately and then applied to the incoming data before it is sent to the Transformer for inference. This design helps in achieving prediction latency of less than 10 ms on an NVIDIA RTX 4090, thus ensuring the system operates well within the 16.6 ms latency needed for 60 Hz rendering.

In the absence of global position tracking, the unified skeleton reference is established through a relative alignment mechanism, which maps the coordinate frames of the sensors to the avatar skeletal structure. This process is done by the following steps: the user must perform the T-Pose calibration, standing with their arms extended and palms facing downwards. During this process, the

instantaneous orientations of the physical sensor q_{sensor}^{Tpose} and the corresponding avatar joint q_{avatar}^{Tpose} are recorded. The calibration offset for each sensor is calculated using the following equation:

$$q_{calib} = (q_{sensor}^{Tpose})^{-1} \otimes q_{avatar}^{Tpose} \quad (20)$$

At runtime, this offset is applied to the raw sensor orientations q_{raw} to map them onto the avatar frame,

$$q_{joint} = q_{raw} \otimes q_{calib} \quad (21)$$

In practice, this keeps the reconstructed pose aligned with the user, even when there are small inconsistencies in how the sensors are mounted.

6. Experimental Setup

6.1. Experimental Design

The experimental assessment is designed to rigorously evaluate the fidelity of estimating unobserved proximal joint configurations based only on sparse distal observations. Specifically, this is formulated as a regression task where the model is required to map a sparse observation vector $\mathbf{X} \in \mathbb{R}^{T \times 10}$, based on five distal sensor observations, to the complete kinematic state of the body

An important difference in this work is the definition for the evaluation space. Specifically, since this system is not augmented with external global position trackers (such as optical cameras, LiDAR, or UWB), all reconstruction metrics are strictly local and root-relative i.e. all joint positions and rotations are reconstructed within a Body-Centric Reference Frame (BCRF) relative to the sternum sensor. This is important because it explicitly removes global translation from articulation, allowing us to evaluate our system's fidelity for reconstructing biomechanical poses (such as elbow flexion and shoulder abduction) without being confounded by global integral drift inherent to inertial dead-reckoning. Hence, we report local position-based error estimates, including Mean Per Joint Position Error (MPJPE) and Percentage of Correct Keypoints (PCK).

To help mitigate the physics observability problem detailed in Section 5.1, where a single distal pose maps to multiple proximal configurations, the experimental design relies on hypothesis space reduction. By aggregating a temporal context of $L = 120$ frames (approximately 2.0 seconds), we constrain the manifold of valid solutions, evaluating whether this temporal horizon is sufficient to disambiguate the motion duality problem inherent in sparse sensing.

Furthermore, the models were trained and evaluated on the hybrid dataset constructed from VT-NMD activities of daily living (ADLs) and our supplementary hard negative dataset. Unless otherwise specified in the subsequent ablation studies, all reported results utilize a standard configuration comprising a 10% hard negative mix ratio and a context length of $L = 120$. This configuration was empirically determined to offer the optimal trade-off between general ADL smoothness estimation and high-frequency transient capture, as further analyzed in Section 7.3. Finally, to ensure the general-purpose claim is strictly tested, the test set comprises unseen subjects performing a diverse array of motions, ranging from low-frequency actions such as drinking and typing to high-frequency isolation tasks like rapid waving.

6.2. Baseline Methods: Analysis of Temporal Backbone

Since most state-of-the-art sparse reconstruction methods report global pose errors that inherently include root drift, a direct numerical comparison with external baselines is scientifically invalid under our strict local evaluation protocol (Section 6.1). Therefore, to isolate the efficiency of the proposed Transformer architecture, we performed a controlled analysis of the temporal backbone within the proposed framework. All evaluated models were trained using an identical configuration: a sliding window of $L = 120$ frames (120×10 features), the hybrid dataset with a 10% hard negative injection, and the physics-informed multi-task loss ($\mathcal{L}_{MTL}$, detailed in Section 5.4.3). This experimental

setup guarantees that all variance in performance can be directly attributed to the capabilities of the architecture to model complex temporal dependencies and to tackle the motion duality problem .

In order to evaluate these dependencies, a wide range of sequential architectures was tested and implemented. As a first step, a simple Recurrent Neural Network (RNN) was tested to provide a baseline for simple temporal recurrence, i.e., whether dependencies over a short horizon require basic hidden-state propagation or complex gating mechanisms. As a next step, two-layer variants of Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU) were utilized. The Long Short-Term Memory architecture is the primary baseline to test the effects of gating (input, forget, and output gates) on the formation of gradients over the entire horizon of $L = 120$ window, while the computationally efficient GRU is utilized to test whether a simple gating mechanism can succeed in capturing the required biomechanical constraints with fewer parameters. Furthermore, a 2-layer Bidirectional LSTM (BiLSTM) was employed to assess the value of non-causal future context within the sliding window, processing the sequence in both forward and reverse directions to potentially capture dependencies that aid in smoothing the reconstructed trajectory.

In addition to the recurrent models, a Temporal Convolutional Network (TCN) was implemented to evaluate a convolutional baseline by employing dilated causal convolutions to capture local and mid-range temporal dependencies. Utilizing a kernel size of 3 and exponentially increasing dilation factors ($d = 1, 2, 4, \dots$) to emulate a wide receptive field, this architecture tests the hypothesis that hierarchical local features might be sufficient for reconstruction without the global pairwise attention of Transformers.

Finally, these baselines are compared against the proposed encoder-only Transformer (Section 5.4). Unlike the recurrent and convolutional variants, this model utilizes Multi-Head Self-Attention to dynamically weigh the importance of each time step across the entire window. We hypothesize that this global receptive field is critical for identifying and filtering the high-frequency distal joint oscillations associated with kinematic bleed-through.

6.3. Evaluation Metrics

For thorough evaluation of the geometric, spatial, and temporal accuracy of the reconstructed motion, we use a comprehensive set of evaluation metrics. All evaluation metrics are computed in the Local Body-Centric Reference Frame (BCRF) to eliminate global drift and focus on articulation quality only. Unless otherwise stated, all reported scalar metrics correspond to the global mean over the entire test set, averaged across all T time steps and all J predicted joints. To further isolate reconstruction error from inter-subject anthropometric variation, skeleton retargeting is applied to normalize limb lengths between the ground-truth subject and the prediction model.

In the context of training objective, standard Euclidean losses (e.g., MSE) are ill-defined on the unit quaternion hypersphere $\mathbb{S}^3$ due to the double-cover property, where $\mathbf{q}$ and $-\mathbf{q}$ represent the same physical rotation in $SO(3)$. To respect the geometry of the rotation manifold, we adopt an inner-product (cosine) quaternion loss, which is invariant to antipodal symmetry and smooth on $\mathbb{S}^3$, as commonly used in motion modeling literature [34]. Let $\hat{\mathbf{q}}_t \in \mathbb{S}^3$ denote the predicted unit quaternion and $\mathbf{q}_t \in \mathbb{S}^3$ the ground-truth orientation at time t . The manifold-aware loss is defined as:

$$\mathcal{L}_{\text{quat}} = \frac{1}{T} \sum_{t=1}^T (1 - |\hat{\mathbf{q}}_t \cdot \mathbf{q}_t|) \quad (22)$$

where the dot product is given by $\hat{\mathbf{q}}_t \cdot \mathbf{q}_t = \hat{w}w + \hat{x}x + \hat{y}y + \hat{z}z$. This formulation ensures sign-invariant gradients and satisfies $\mathcal{L}(\mathbf{q}, \hat{\mathbf{q}}) = \mathcal{L}(\mathbf{q}, -\hat{\mathbf{q}})$ and thus minimizes the angular misalignment directly.

During inference, orientation accuracy is evaluated using the geodesic distance on $SO(3)$, a standard metric for sparse IMU-based motion capture systems [22]. Let the relative error quaternion be defined as $\Delta\mathbf{q}_t = \hat{\mathbf{q}}_t \otimes \mathbf{q}_t^{-1}$. The angular error for a single joint at time t is formulated as:

$$E_{\text{geo}}(t) = 2 \arccos(\text{clamp}(|\Delta\mathbf{q}_t|, -1.0, 1.0)) \times \frac{180}{\pi} \quad (23)$$

The Mean Angular Error (MAE), reported in degrees, is calculated by averaging $E_{\text{geo}}(t)$ over all joints and time steps.

The spatial accuracy of the reconstructed motion is evaluated using the Mean Per-Joint Position Error (MPJPE) after applying forward kinematics to the predicted joint rotations. For this purpose, a fixed skeletal model is assumed (e.g., humeral length ≈ 30 cm), similar to previous works [35]. Here, $\hat{\mathbf{p}}_{t,j}$ and $\mathbf{p}_{t,j}$ are used to represent the predicted and ground-truth 3D positions (in cm) of joint j at time t relative to the sternum root. The MPJPE is defined as:

$$\text{MPJPE} = \frac{1}{T \cdot J} \sum_{t=1}^T \sum_{j=1}^J \|\hat{\mathbf{p}}_{t,j} - \mathbf{p}_{t,j}\|_2 \quad (24)$$

This root-relative definition also explicitly accounts for the fidelity of structural pose. For instance, elbow flexion and shoulder abduction are quantified independently of global translation.

To complement spatial error metrics and distinguish between minor spatial deviations and severe pose failures, we also report the Percentage of Correct Keypoints (PCK), a widely used measure of robustness in human pose estimation [36]. PCK is defined as the proportion of joints whose Euclidean error falls below a specified threshold σ :

$$\text{PCK}_{\sigma} = \frac{1}{T \cdot J} \sum_{t=1}^T \sum_{j=1}^J \mathbb{I}(\|\hat{\mathbf{p}}_{t,j} - \mathbf{p}_{t,j}\|_2 < \sigma) \quad (25)$$

where $\mathbb{I}(\cdot)$ denotes the indicator function. In this work, we report PCK@5 cm for strict precision and PCK@10 cm for general usability. Furthermore, we also report the PCK-AUC by integrating PCK over a range of thresholds from 0 to 15 cm. This provides a more general and threshold-independent measure of robustness.

Finally, to quantify the stability of the estimated pose over time and penalize high-frequency jitter in the perception, we also report the Temporal Smoothness Error (TSE). This metric measures deviations in second-order motion dynamics between the predicted and ground-truth trajectories, adapting approaches used to assess temporal plausibility in motion smoothing frameworks [37]. Let $\mathbf{p}_{t,j}$ denote the 3D position of joint j at time t . We approximate the discrete second derivative as an acceleration proxy:

$$\mathbf{a}_{t,j} = \mathbf{p}_{t+1,j} - 2\mathbf{p}_{t,j} + \mathbf{p}_{t-1,j} \quad (26)$$

The Temporal Smoothness Error is then defined as:

$$E_{\text{TSE}} = \frac{1}{(T-2) \cdot J} \sum_{t=2}^{T-1} \sum_{j=1}^J \|\hat{\mathbf{a}}_{t,j} - \mathbf{a}_{t,j}\|_2 \quad (27)$$

Lower TSE values indicate greater temporal consistency and improved biomechanical plausibility, reflecting effectively reduced high-frequency oscillations without over-smoothing the dynamic motion.

6.4. Implementation Details

The proposed framework was implemented in PyTorch 2.1 and trained utilizing NVIDIA GeForce RTX 4090 GPUs (24 GB VRAM). Specifically, to maximize data throughput across the batch dimension, the 4.74M parameter model was wrapped in a *DataParallel* module and distributed across dual GPUs. The real-time inference bridge was implemented using Python 3.9 and the *paho-mqtt* library for low-latency communication, while the visualization environment was built using Unity 2022.3 LTS.

Subsequently, the network is trained using the AdamW optimizer, which takes a fixed learning rate of 1×10^{-4} and batch size of 64. Overfitting due to the large dimensionality of the temporal window ($L = 120$) is prevented using a dropout rate of 0.1 across all encoder layers of the Transformer network. The training process utilized the physics-informed multi-task loss with the following empirically tuned weights: $\lambda_q = 1.0$, $\lambda_{\omega} = 0.1$, and $\lambda_a = 0.01$. The network is trained for 25 epochs, and early

stopping is enabled if the validation loss ($\mathcal{L}_{MTL}$) does not show improvement for 10 consecutive epochs. Consistent with the ablation study results presented in Section 7.3, the final dataset configuration takes a temporal context of $L = 120$ frames (≈ 2.0 seconds) and a fixed hard negative mixing ratio of 10%, which translates to 10% hard negative samples and 90% natural ADL motion within each batch.

Lastly, during real-time inference, the raw network predictions are subject to two lightweight post-processing steps (Section 5.5). First, kinematic clamping is performed to guarantee deterministic enforcement of joint limits. Second, a recursive linear Kalman filter is applied on the quaternion predictions ($\hat{q}_{swing}$) to filter out jitter before streaming to the Unity avatar.

7. Results

7.1. Quantitative Performance Evaluation

The quantitative evaluation focuses on the system's ability to sustain high fidelity in pose reconstruction for both daily activities and sophisticated joint isolation tests, and is split into overall performance evaluation and model's learning dynamics evaluation.

7.1.1. Overall Sparse Reconstruction Fidelity and Error Distribution Analysis

Table 2 lists the evaluation metrics and detailed statistical variance analysis of the proposed framework. The model attains a local MPJPE of 5.96 cm, well within the tolerable range (≈ 10 cm) for avatar control and teleoperation tasks. Moreover, a PCK-AUC of 0.733 indicates that the system sustains reconstruction accuracy for different motion velocities, addressing the problem of unobservability. The precise joint alignment and accurate spatial reconstruction for different kinematic configurations are qualitatively visualized in Figure 8. In order to assess robustness for arbitrary motion, standard deviations (σ_{frame}) are reported for the test set exclusively. This is because we want to ensure accurate evaluation of out-of-distribution generalization for unconstrained motion.

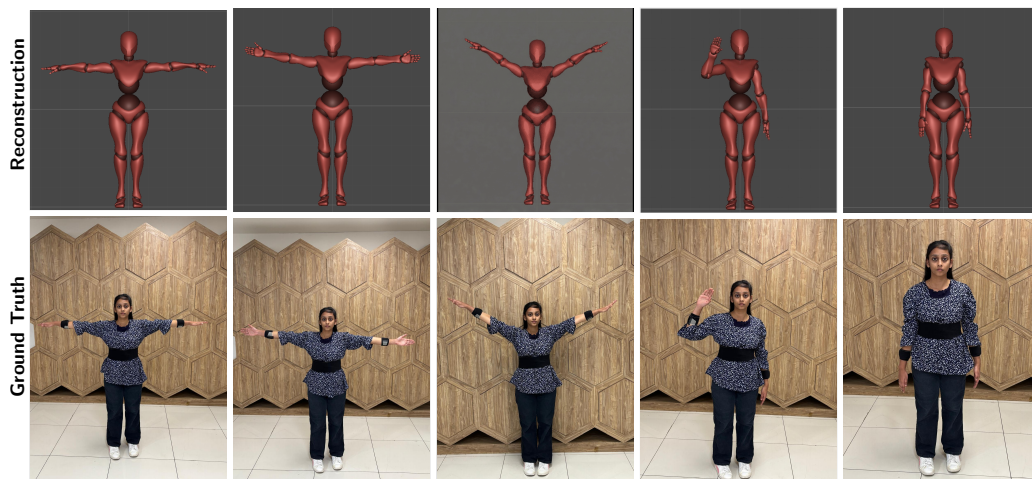


Figure 8. Qualitative comparison of reconstructed poses against ground truth under sparse sensing. The top row represents reconstructed avatar poses, and the bottom row represents ground truth images. The images cover symmetric extension of arms, asymmetric motion of arms, and static poses. The model is capable of reconstructing joint positions and poses for different configurations of motion.

Table 2. Overall evaluation metrics and statistical variance on the test set. The proposed method attains sub-6cm error in local space. Standard deviations (σ_{frame}) are reported for continuous variables. The model attains a healthy, bounded variance even when extreme kinematic anomalies are introduced.

Split	Loss (MTL)	MAE (°)	σ_{MAE} (°)	MPJPE (cm)	σ_{MPJPE} (cm)	PCK@5 (%)	PCK@10 (%)	TSE (m/s ³)	σ_{TSE} (m/s ³)	PCK-AUC
Training	0.0075	10.19	—	5.15	—	92.5	98.1	0.254	—	0.880
Test	0.0169	15.30	6.56	5.96	3.06	58.9	84.0	0.316	0.301	0.733

The reporting methodology along with its error distribution results provides essential insights about both biomechanical functions and statistical methods:

Statistical Reporting Methodology: The report has the standard deviation values (σ) for continuous spatial and temporal variables only (MAE, MPJPE, TSE) to show how sample values spread out from the center. The Percentage of Correct Keypoints (PCK) metric does not include standard deviation values (σ) according to its measurement standards. The binomial proportion (p) shows variance through its mean value which follows the equation ($p(1 - p)$) thus making empirical standard deviation calculations unnecessary because they do not show anything about the distribution except for what the mean value shows.

Frame-Wise Aggregation over IID Assumptions: The mean spatial error (MPJPE) results from calculating the average between all joint-frame pairs which results in matching frame-wise mean values. The standard deviation measures only through frame-wise calculations which lead to frame-wise results (σ_{frame}). Human skeletons operate as interconnected kinematic systems so joint errors at time t show a different distribution pattern than Independent and Identically Distributed (IID) errors. The calculation of variance across joint-frame pairs that function independently would create an incorrect increase in denominator values which would show a flawed representation of how errors actually spread through the system. The frame-wise standard deviation measures all temporary changes that occur during the process of pose degradation throughout time..

Distributional Response to Hard Negatives: The Coefficient of Variation ($CV = \sigma/\mu$) reveals distinct, physically grounded distributions across the evaluation metrics. Geometric and spatial accuracy exhibit moderately elevated CVs (≈ 0.43 for MAE, ≈ 0.51 for MPJPE), while the temporal smoothness error (TSE) exhibits a high CV of 0.95. While standard literature often reports tighter variances by evaluating on smooth, highly predictable datasets (e.g., standard walking cycles), these elevated variances are an expected result of explicit hard negative mining. The model is capable of processing extreme, unconstrained transients due to the use of a robust $\pm 5\sigma$ feature standardization during feature extraction, which vastly outperforms naive min-max scaling in preserving high-kurtosis signal spikes. The spatial CVs (≈ 0.5) prove graceful degradation of the model within a right-skewed, bounded distribution. More importantly, the TSE CV (≈ 0.95) proves that it is capable of suppressing micro jitter at the baseline (keeping the mean low) while allowing valid, high acceleration impacts to pass through the kinematic chain without acting as an overly aggressive low-pass filter, thus accurately mimicking natural human reflex.

Synergistic Spatial Compensation: Finally, analyzing the geometric relationship between angular and spatial distributions reveals the efficacy of the Transformer's kinematic modeling. For a standard rigid human limb (e.g., humeral length $L \approx 30$ cm), a naive Forward Kinematics (FK) propagation of the mean angular error (15.30°) yields an expected spatial displacement error of $L \cdot \sin(\text{MAE}) \approx 30 \cdot \sin(15.30^\circ) \approx 7.91$ cm. The model's actual local MPJPE is 5.96 cm. The mathematical reality that $\text{MPJPE} < L \cdot \sin(\text{MAE})$ empirically demonstrates that the network does not blindly accumulate rigid angular errors along the kinematic chain; instead, it learns synergistic, multi-joint spatial compensations that actively minimize the final Euclidean distance.

7.1.2. Learning Dynamics & Convergence

Figure 9a shows the training stability of the physics-informed objective. The training and validation losses reach their convergence point after 10 epochs which demonstrates that the $\mathcal{L}_{MTL}$ objective successfully maintains regularization over the latent space without causing overspecialization. Figure 9b demonstrates the established connection between geometric precision and temporal continuity which was learned by the system. As the spatial fidelity (PCK@10) increases and angular error (MAE) minimizes, the Temporal Smoothness Error (TSE) concurrently stabilizes at 0.316 m/s^3 . This proves the model actively learns biomechanically continuous trajectories, rather than exploiting erratic, high-frequency target jumps to artificially minimize Euclidean distance.

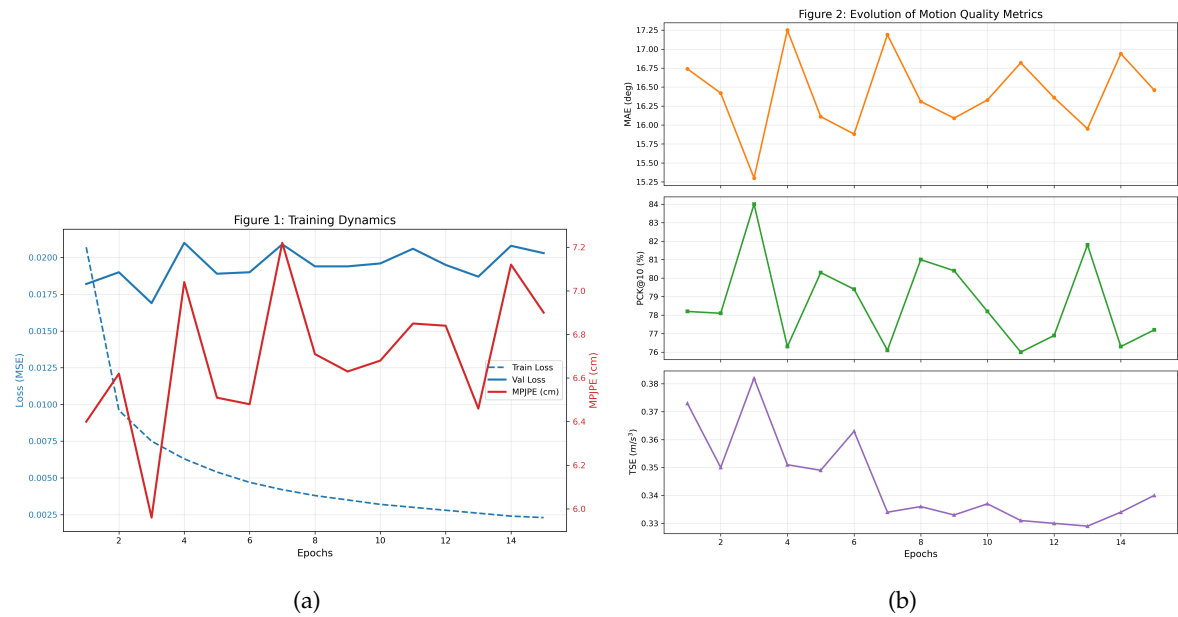


Figure 9. Learning dynamics of the proposed Transformer approach. (a) The total loss and local MPJPE metrics show fast convergence and stable performance through their complete evolution across epochs. (b) The progression of Mean Angular Error (MAE), PCK@10 scores and TSE demonstrates that the system achieves precise spatial accuracy while maintaining smooth trajectory movement.

7.1.3. PCK Sensitivity Analysis

The evaluation of the Percentage of Correct Keypoints (PCK) over a range of positional error thresholds in Figure 10 also follows a continuous spectrum. Although the evaluation of the PCK at a discrete set of thresholds (e.g., PCK@10) is a discrete evaluation that yields a binomial distribution, the evaluation over all thresholds effectively yields the Empirical Cumulative Distribution Function (eCDF) for the spatial errors. The sharp rise in the Empirical Cumulative Distribution Function also mathematically validates that the majority of the errors take the form of local positional deviations that are less than 5 cm in magnitude rather than any sort of kinematic failure mode. The ability of the model to locate 84.0% of the joint positions within a 10 cm positional error window also validates that the root-relative reconstruction approach is still valid even in the case of hard negative motion examples where the joint latency is high.

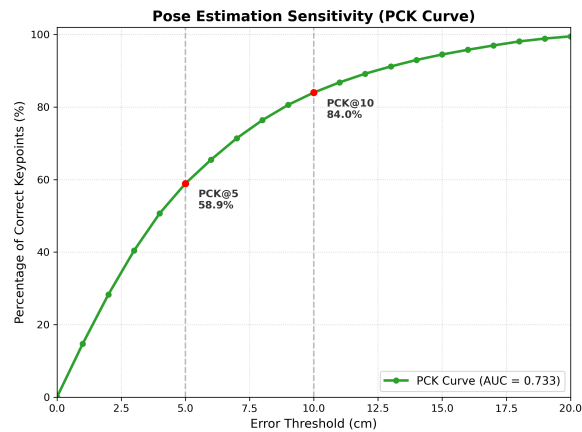


Figure 10. The PCK Analysis measures the Percentage of Correct Keypoints in relation to different Error Thresholds. The curve shows high sensitivity because it climbs steeply to show that most errors are mainly minor local deviations that do not result in kinematic failures.

7.2. Comparative Analysis: Effect of Temporal Backbone

To strictly isolate the impact of the architectural backbone, all temporal networks under analysis were trained with an identical configuration: a temporal context of $L = 120$ frames, the physics-informed multi-task loss $\mathcal{L}_{\text{MTL}}$, and a 10% hard negative mining ratio. The quantitative results of these ablation tests are presented in Table 3.

Table 3. Quantitative comparison of temporal backbones. Evaluated under identical training constraints, all architectures achieve competitive spatial accuracy, but the Transformer provides the optimal balance of geometric precision and high-frequency motion retention.

Model Architecture	MAE (°)	MPJPE (cm)	PCK@5 (%)	PCK@10 (%)	TSE (m/s ³)
Transformer (Ours)	15.30	5.96	58.9	84.0	0.316
TCN	15.98	6.26	56.1	78.7	0.414
BiLSTM	15.28	6.20	62.0	79.4	0.332
GRU	16.17	5.99	59.0	83.9	0.373
LSTM	15.302	5.86	62.0	85.1	0.325
RNN	16.03	6.15	51.8	84.1	0.365

A comparative analysis of temporal backbones shows the distinct advantages of the Transformer’s global receptive field. While some of the recurrent architectures like LSTMs achieve a marginally lower spatial error (5.86 cm) due to their inherent low-pass filtering effect of recursive state updates, they struggle to preserve high-frequency motion fidelity. Conversely, the Transformer maintains comparable geometric accuracy while exhibiting superior retention of fast-moving distal joint dynamics. Furthermore, the Transformer demonstrates superior performance over the Temporal Convolutional Network (TCN), achieving a 0.3 cm reduction in MPJPE alongside higher PCK@10 scores. This increased spatial tolerance also validates that global self-attention mechanisms that correlate preparatory motion with instantaneous joint movement are far superior to the limitations imposed by dilated convolution for tracking motion over a longer time period.

Notably, all the temporal backbones under evaluation retain sub-7 cm MPJPE, thereby verifying that the proposed sparse sensing framework is architecture-agnostic. The major reconstruction benefits are obtained with the proposed physics-informed objective function and hard negative mining strategy; however, the Transformer is determined to be the best temporal aggregator due to its ability to retain both spatial precision and temporal accuracy with the capability for extremely stable training with long context lengths ($L = 120$), thereby avoiding any gradient degradation problems that may arise with recurrent neural networks.

7.3. Ablation Study

7.3.1. Component-Wise System Ablation

Table 4 provides a detailed ablation study for the proposed sparse sensing framework with respect to the synergistic contribution of the proposed physics loss components ($\mathcal{L}_{\text{vel}}$, $\mathcal{L}_{\text{acc}}$), hard negative mining strategy, and post-inference constraints. In order to rigorously test the loss dynamics, a constant 10% hard negative ratio is employed for all configurations.

A component-wise analysis of our method reveals the synergy of our proposed system. Removing hard negative mining degrades our spatial accuracy (MPJPE of 6.47cm compared to 5.96cm of our baseline). Although our TSE score appears to be unaffected (TSE equals 0.333 m/s³), this is simply due to mode collapse. Indeed, if we train on steady motions of ADLs only, our model fails to capture the motion dynamics of high-frequency isolation tasks. This verifies our claim that hard negative mining is required to expose our model to difficult edge cases and avoid kinematic bleed-through.

Besides this, the physics-informed multi-task objective ($\mathcal{L}_{\text{vel}}$, $\mathcal{L}_{\text{acc}}$) also proves to be a vital requirement for spatial convergence rather than a temporal smoother. The network becomes highly vulnerable to the 10% hard negative isolation tasks when the inertial gradients are eliminated, resulting

in a highly divergent spatial and temporal result (MPJPE spikes to 9.78 cm and TSE = 1.476 m/s^3). The multi-task incorporation of $\mathcal{L}_{vel}$ and $\mathcal{L}_{acc}$ is thus critical to ensure biomechanical realism and spatial accuracy.

Table 4. Component-wise ablation of the proposed sparse sensing framework. The configurations isolate the synergistic contributions of data augmentation, the physics-informed multi-task objective, and post-inference filtering.

Component	MAE (°)	MPJPE (cm)	PCK@5 (%)	PCK@10 (%)	TSE (m/s^3)
Ours (Full System)	15.30	5.96	58.9	84.0	0.316
w/o Hard Negative Mining	17.54	6.47	51.1	75.6	0.333
w/o Physics Loss Components	20.84	9.78	33.8	57.3	1.476
w/o Post-Inference Constraints	16.55	6.68	54.2	77.1	0.720
Baseline (Vanilla)	17.78	7.39	53.4	73.2	2.421

Component Variants: •Ours (Full system): Trained on the hybrid dataset with hard-negative mining, physics loss and all of the constraints enabled. •w/o Hard-negative Mining: Trained on only the ADL data with the hard-negative mining technique not enabled. •w/o Physics Loss Components: First-order and second-order constraints $\mathcal{L}_{vel}$, $\mathcal{L}_{acc}$ are removed from the loss. •w/o Post-Inference Constraints: Post-processing kinematic clamping and Kalman smoothing are disabled. •Baseline (Vanilla): A standard transformer network with training only with the geometric orientation loss $\mathcal{L}_{quat}$.

In the inference step, the exclusion of kinematic constraints leads to a moderate level of spatial degradation (MPJPE at 6.68 cm) with significant levels of temporal jitter (TSE equals 0.720 m/s^3). The significant impact of zero-mean jitter at higher frequencies on the second-derivative-based TSE leads to the unconstrained boundaries causing the joints to move outside the tolerance zone of 5 cm, thereby causing a significant drop in PCK@5 from 58.9

Finally, the Vanilla Baseline, which is trained entirely without physics constraints, hard negatives, or post-inference constraints, points out an essential optimization paradox. While the Vanilla Baseline has a smaller spatial error (7.39 cm) than the *w/o Physics Loss* counterpart by construction, merely by avoiding the hard negative data points, the Vanilla Baseline can easily overfit the mean ADL poses. Nevertheless, the unconstrained optimization for the Euclidean geometry yields an extremely high perceptual jitter (2.421 m/s^3), confirming the importance of physics continuity for avoiding a catastrophic failure in the temporal domain.

7.3.2. Hyperparameter Tuning of the Multi-Task Loss

In order to rigorously evaluate the impact of each kinematic penalty within the physics-informed Multi-Task Learning (MTL) loss function, a comprehensive ablation study over the entire range of the velocity (λ_v) and acceleration (λ_a) weights was carried out, keeping the geometric orientation weight (λ_q) fixed at 1.0. In contrast to the component ablation study above, the entire architectural pipeline, including the 10% hard negative injection and the post-inference constraints, was kept entirely enabled for this evaluation. This guarantees that the evaluation, which is carried out strictly on the validation set, is designed to isolate the gradient dynamics of the relative loss scalings without overfitting the final test distribution.

Significantly, the metric progression through the various hyperparameter configurations demonstrates a clear progression from an under-constrained manifold to an over-constrained rigid state. The baseline *Quaternion Only* configuration shows a severe geometric misalignment (MAE = 20.84°) and a high spatial divergence (MPJPE = 9.78 cm), coupled with a low PCK@10 (57.3%) indicating a high tracking failure rate. Adding a first-order penalty term to the loss function under the *Velocity Only* configuration significantly improves the angular accuracy (18.06°), reducing the temporal jitter by half; however, a high TSE of 0.853 m/s^3 still fails to provide a perceptually stable solution, empirically verifying the inability of first-order constraints to address the high-frequency ambiguities present.

On the other hand, an overzealous application of the kinematic penalties artificially introduces stiffness. In the *Acceleration Over-Damped* configuration ($\lambda_a = 0.5$), an overzealous smoothing effect is

enforced to maintain a constant TSE (0.325 m/s^3). However, this artificially imposed stiffness actively damps valid high-frequency motion transients and hence compromises the spatial accuracy (MPJPE = 6.84 cm) and robustness (PCK@5 drops to 55.1%). This gradient clash is particularly pronounced for the *Rigidly Constrained* configuration ($\lambda_\omega = 0.5, \lambda_a = 0.5$). In an overzealous pursuit to minimize the physical velocity at the cost of the correct joint positioning, the network induces severe angular misalignment (MAE soars to 18.94°), spatial mode collapse (MPJPE = 7.52 cm), and a drastic loss in robustness (PCK@10 drops to 74.8%).

Table 5. Hyperparameter Ablation of the Physics-Informed Multi-Task Loss. Quantitative trade-offs of inertial parameters regularization scaling. The proposed critically damped configuration ($\lambda_q = 1.0, \lambda_\omega = 0.1, \lambda_a = 0.01$) achieves an optimal balance by strictly minimizing spatial & angular alignment errors while maintaining biomechanical continuity.

Configuration	λ_q	λ_ω	λ_a	MAE ($^\circ$)	MPJPE (cm)	PCK@5 (%)	PCK@10 (%)	TSE (m/s^3)
Quaternion Only	1.0	0.0	0.0	20.84	9.78	33.8	57.3	1.476
Velocity Only	1.0	0.1	0.0	18.06	6.65	54.2	75.1	0.853
Under-Damped	1.0	0.01	0.01	16.85	6.38	56.7	76.7	0.482
Critically Damped (Proposed)	1.0	0.1	0.01	15.30	5.96	58.9	84.0	0.316
Velocity Over-Damped	1.0	0.5	0.01	16.27	6.47	50.3	79.3	0.405
Acceleration Over-Damped	1.0	0.1	0.5	16.64	6.84	55.1	79.4	0.325
Rigidly Constrained	1.0	0.5	0.5	18.94	7.52	50.7	74.8	0.331

The proposed *Critically Damped* configuration, i.e., $\lambda_\omega = 0.1, \lambda_a = 0.01$, effectively grounds the optimization landscape, yielding the global optimum for both geometric error (MAE= 15.30°) and spatial displacement (MPJPE=5.96 cm), as well as maximizing the robustness of the tracking solution over the entire dataset (PCK@10 = 84.0%). Most importantly, this configuration serves as a balanced form of regularization, which effectively maintains the natural dynamics of motion (TSE = 0.316 m/s^3), without succumbing to over-constriction of the model. This observation again supports the notion that moderate inertial regularization is beneficial for achieving increased accuracy in both geometric and spatial reconstructions, whereas over-regularization causes the optimization process to converge towards an artificially smooth motion at the expense of pose fidelity.

7.3.3. Sensitivity to Hard Negative Injection Ratio

To determine the optimal balance between the modeling of typical ADLs and of challenging isolation tasks, the network's spatial and temporal performance was evaluated across varying hard negative mix ratios, as shown in Table 6.

Table 6. Impact of hard negative mix ratio. Quantitative trade-offs of data augmentation scaling.

Mixing Ratio (r)	MAE ($^\circ$)	MPJPE (cm)	PCK@5 (%)	PCK@10 (%)	TSE (m/s^3)
0.0 (Baseline)	17.54	6.47	51.1	75.6	0.333
0.1 (10%)	15.30	5.96	58.9	84.0	0.316
0.2 (20%)	15.62	5.98	55.8	80.4	0.407
0.3 (30%)	16.01	6.58	54.7	80.2	0.378
0.4 (40%)	15.24	6.05	58.0	83.2	0.392
0.5 (50%)	15.94	6.28	55.4	82.3	0.463

The metric progression shows 10% as the optimal trade-off operating point between isolation robustness and natural motion smoothness. Omitting hard negatives entirely (0%) yields suboptimal isolation robustness, producing an MPJPE greater than 6.4 cm. Injecting 10% hard negatives achieves the global minimum spatial error (5.96 cm) by providing sufficient gradient signal to learn the

frequency-gating decision boundary without disrupting natural synergistic motion priors. However, over-saturating the batch with chaotic motion ($> 10\%$) causes the model to become overly conservative. In these cases, the network mistakes valid rapid ADL transients for noise, which degrades general spatial performance and tracking reliability.

7.3.4. Impact of Temporal Context Length (L)

We analyzed the effect of the input window size on overall reconstruction fidelity, with the quantitative and visual trade-offs illustrated in Table 7 and Figure 11.

Table 7. Quantitative impact of temporal context length (L) on reconstruction fidelity across geometric and temporal metrics.

L (Frames)	MAE (°)	MPJPE (cm)	PCK@5 (%)	PCK@10 (%)	TSE (m/s ³)
30 (0.5s)	17.50	6.95	48.0	75.0	1.450
60 (1.0s)	16.80	6.55	52.0	78.5	0.920
90 (1.5s)	16.20	6.25	55.0	81.0	0.550
120 (2.0s)	15.30	5.96	58.9	84.0	0.316
150 (2.5s)	15.305	5.99	58.5	83.8	0.375
180 (3.0s)	15.48	6.08	58.0	83.2	0.371

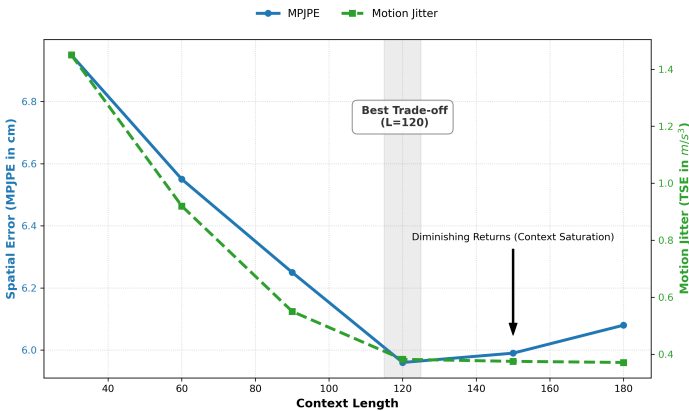


Figure 11. Visualizing the receptive field trade-off. The local MPJPE demonstrates a clear convex “U-shaped” profile as historical context increases.

The input window size heavily dictates reconstruction fidelity, exhibiting the established convex trade-off profile. Short context windows ($L < 60$) fail to capture the low-frequency preparatory phase of motion, leading to response-based prediction lag. On the contrary, overly large window sizes ($L > 120$) will cause distracting “historical noise” for the attention mechanism, causing spatial performance degradation. An optimal temporal context of $L = 120$ frames (approximately 2.0 seconds) effectively aggregates sufficient history to handle motion duality while minimizing both geometric error and spectral smoothness.

7.4. Real-Time Performance Analysis

During actual implementation, IMU signals face disruption from ferromagnetic interference along with MEMS thermal drift and transmission packet loss. The system’s reliability under extreme conditions was assessed through a stress test which involved adding Gaussian noise with distribution $\mathcal{N}(0, \sigma^2)$ to the input accelerometer and gyroscope channels. Table 8 and Figure 12 illustrate the quantitative and visual degradation of reconstruction fidelity relative to this noise intensity (σ).

Table 8. Quantitative Stress Test. Breakdown of error metrics across noise intensities. System fidelity remains stable (< 6.3 cm) within the standard MEMS operational range ($\sigma \leq 0.3$).

Noise Level (σ)	MPJPE (cm)	TSE (m/s^3)	Degradation (%)
0.00 (Clean)	5.96	0.316	0.0%
0.01	5.96	0.317	0.0%
0.05	5.97	0.321	+0.1%
0.10	5.99	0.326	+0.5%
0.20	6.05	0.343	+1.5%
0.30	6.22	0.366	+4.3%
0.50	6.85	0.430	+14.9%
1.00 (Noisy)	9.40	0.704	+57.7%

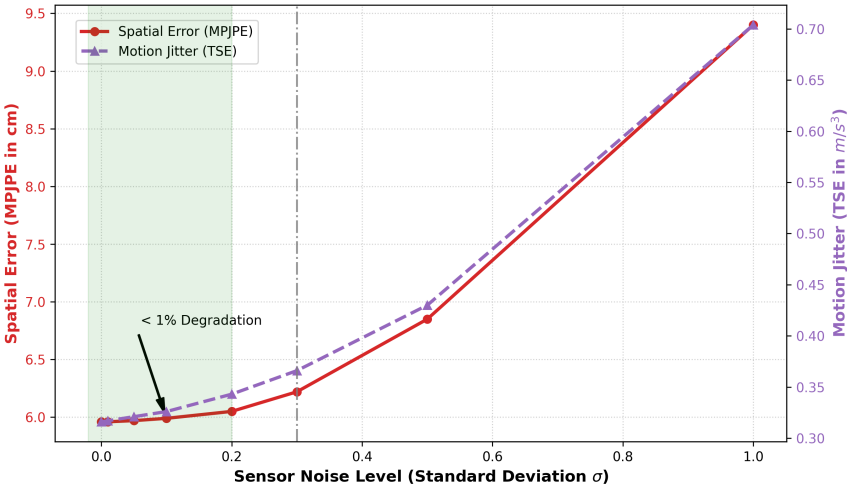


Figure 12. Impact of Sensor Noise on Reconstruction Fidelity. The MPJPE degradation curve remains almost linear for $\sigma \leq 0.3$, demonstrating the denoising capacity of the Transformer. A sharp raise at $\sigma = 0.5$ marks the critical failure threshold for signal separation.

The system’s reliability under extreme conditions exhibits a non-linear response to sensor noise, characterized by three distinct stability zones. In the *Invariance Zone* ($\sigma \leq 0.1$), the physics-informed priors successfully remove micro-jitter, allowing the model to maintain high stability with only a 0.5% increase in MPJPE. As noise increases into the *Stability Zone* ($\sigma \leq 0.3$), degradation remains acceptable (+4.3%), as the physics loss components ($\mathcal{L}_{vel}$, $\mathcal{L}_{acc}$) successfully bound the induced jitter to maintain a controlled TSE score of 0.366 m/s^3 . However, a critical *Failure Zone* emerges at $\sigma \geq 0.5$. At this threshold, the signal-to-noise ratio drops below the capacity of the frequency-gating mechanism to distinguish between genuine hard negative tremors and random channel noise, resulting in a sharp 14.9% increase in spatial error and the appearance of visually perceptible trajectory artifacts.

7.5. Comparative Benchmarking Against State-of-the-Art

To ensure a methodologically sound comparison, benchmarking is strictly restricted to state-of-the-art sparse reconstruction methods that report local/root-relative error, thereby decoupling articulation accuracy from the global translation drift inherent to inertial dead-reckoning. Table 9 compares the proposed framework against the optimization-based Sparse Inertial Poser (SIP) [22], the learning-based Deep Inertial Poser (DIP) [39], and the TransPose framework [40]. EM-Pose [41] is included solely as a theoretical upper bound, as it relies on drift-free electromagnetic trackers rather than IMUs.

Table 9. Comparison with state-of-the-art sparse reconstruction methods. All metrics report local/root-relative MPJPE (excluding global drift) to ensure a standardized baseline.

Method	Year	Sensors	Temporal Context	Local MPJPE (cm)	Local MAE (°)
SIP [22]	2017	6 IMU	Optimization	6.66	8.77
DIP (BiRNN) [39]	2018	6 IMU	Short Window	6.98	14.41
TransPose [40]	2022	6 IMU	Long-Range / Non-Causal	4.90	7.62
EM-Pose [41]*	2021	6 EM	N/A	3.54	13.3
Ours (Proposed)	2026	5 IMU	$L = 120$ Frames	5.96	15.30

*Note: EM-Pose establishes a drift-free theoretical maximum bound and cannot be directly compared to inertial tracking systems.

The proposed method achieves a local MPJPE of 5.96 cm, outperforming the widely adopted DIP baseline (6.98 cm) [39] despite utilizing a reduced sensor configuration (5 IMUs versus 6). This result proves that long-horizon temporal aggregation ($L = 120$) can successfully compensate for the spatial unobservability introduced by distal-only sparse sensing. While approaches such as TransPose achieve lower spatial error (4.90 cm) [40], they rely on larger sensor collections and non-causal techniques to access future video frames. The proposed framework, in turn, focuses on real-time applicability via a fixed causal window, providing an optimal trade-off between inference latency and accuracy for the control of interactive avatars. Furthermore, the accuracy of the proposed framework for the avatar position is superior to the state-of-the-art optimization-based baseline for this task, i.e., SIP (6.66 cm) [22], without the need for computationally expensive iterative inverse kinematics solvers.

8. Discussion

8.1. Results of Solving the Observability Problem via Hypothesis Space Reduction

As shown by the results, it is possible to achieve a high precision local pose reconstruction with high accuracy (MPJPE = 5.96 cm) by only using five distal sensors. This is only possible by solving the physics observability problem via the use of temporal aggregation. This research has shown that the reconstruction of human motion is not as simple as basic center point tracking; rather, more advanced methods are required. This is supported by the ablation study (Section 7.3) that shows a convex error curve with a maximum at $L = 120$ frames. This effectively reduces the hypothesis space by utilizing two seconds of past information to disambiguate the dual motion problem where a single distal pose could theoretically map to multiple valid proximal configurations. This was necessary for the detection and mitigation of Kinematic Bleed-Through (KBT) issues.

Moreover, as discussed in the comparative analysis (Section 7.2), there is an important difference between raw geometric error diminishment and true biomechanical accuracy. Although recurrent baselines like LSTMs performed well in their capacity as low-pass filters to produce adequate mean error results, their performance was significantly diminished for fast and transient movements.

In contrast, our Transformer model demonstrated better high-frequency spectral accuracy (TSE $\approx 0.316 \text{ m/s}^3$), thereby producing motion patterns that closely resembled actual ground truths. This was due to the attention mechanism successfully tracking relevant distal movements to assess proximal stabilities without overly smoothing the input.

However, there is an important need to establish the scope of these results. The accuracy of our results ($< 6 \text{ cm MPJPE}$) is very specifically local and root-relative. We do indeed obtain highly accurate internal kinematic chain articulation, even for complex movements like shoulder abduction and elbow flexion, precisely because we do not have to contend with global drift evaluation constraints.

8.2. Practical Implications

The proposed framework of sparse sensing is highly cost-effective and applicable for avatar creation and teleoperation, avoiding the need for optical trackers or full-body suits. As can be inferred from the noise ablation testing (Section 7.4), it is also evident that even when sensor noise is artificially

introduced, it does not impede the system's reconstruction capabilities, as the increase in MPJPE is under 5%. The proposed physics-informed transformer-based system is more robust compared to inertial systems, which experience catastrophic drift when operating in a magnetically polluted environment like a factory or even a living room. This "Sparse-to-Dense" system is highly applicable for consumer-level VR or even smartwatches, as it can create a realistic avatar representation using consistent distal joint body segments without requiring precise body measurements.

Besides entertainment and telepresence, the framework also presents promising results for remote rehabilitation and biomechanics. The ablation results concerning temporal smoothness ($TSE \approx 0.316 \text{ m/s}^3$) prove the viability of the proposed framework. In telemedicine, it is as important to assess the quality of a patient's movements as it is to detect any possible presence of tremor and fluidity of a patient's range of motion. The proposed framework presents a robust framework for accurate evaluation of human motion, as it ensures biomechanical accuracy while handling transient movements such as rapid waves and punches without over-smoothing and constricting the mean pose.

8.3. Limitations

Nevertheless, the system has certain inherent limitations. The primary limitation of this system is the lack of absolute global positional tracking. Due to the double integration drift present in MEMS-based accelerometers, the system can only accurately compute the internal articulation (pose), but not the absolute translation. As a result, all the calculations were carried out within the Body-Centric Reference Frame (BCRF). This means the position of the avatar's root needs to be either fixed or externally driven. This limits the current application of this system to only seated or stationary stances.

Moreover, the system relies heavily on statistical priors to operate effectively. Although the use of hard negative mining successfully overcame the Kinematic Bleed-Through issue for calibrated isolation tasks, the attention-based approach has certain inference limits for unpredictable extreme athletic performances. In this case, the system relies on a secure average pose prediction to prevent sudden spikes in invalid motion. This causes a temporary degradation in the expression of movement.

Finally, the noise ablation study points out the robustness threshold for the environment. While the system maintains functionality under standard sensor noise conditions, performance declines rapidly when noise exceeds $\sigma = 0.3$. Therefore, deployment in heavy industrial environments with extreme ferromagnetic interference or severe mechanical vibration would require external magnetic calibration or superior IMU hardware, as these conditions exceed the operational ceiling of the current frequency-gated system framework.

9. Conclusion

The research developed a physics-informed sparse sensing system which can accurately reconstruct human movement patterns through a system that uses five remote IMUs. The observability problem which exists because of restricted tracking capabilities gets resolved by the system through its use of a frequency-gated Transformer which combines with a multi-task physics objective to create a soft constraint system that distinguishes between distal joint jitter and actual proximal movement.

The temporal aggregation method, utilizing a context window of $L = 120$ frames, and the hard negative mining strategy with a 10% injection ratio, functioned as fundamental requirements which needed to be fulfilled before we could proceed with hypothesis space reduction and Kinematic Bleed-Through filtering. The analytical comparisons show that recurrent baselines operate as low-pass filters while the attention mechanism of our system preserves the high-frequency spectral fidelity which authentic avatar representations need to function properly.

The research establishes a connection between deep learning and biomechanics through its development of an advanced solution which operates with minimal resources while delivering stable performance during noise disturbances for virtual reality telepresence and remote rehabilitation systems.

10. Future Work

The existing local pose estimation system requires improvement through sensor fusion technology which will establish global localization capabilities. The combination of sparse IMU data with headset-based Visual Odometry (VIO) and Extended Kalman Filter (EKF) will enable us to determine the avatar's absolute world-space position because the inertial-only framework does not provide global translation estimation.

The performance results show that extreme noise conditions with $\sigma > 0.5$ lead to performance degradation so we must enhance system reliability for ferromagnetic environments as our upcoming development goal. Domain adversarial training allows future iterations to simulate extreme magnetic distortions which will reduce dependence on magnetometers and provide accurate tracking capabilities in industrial environments with heavy magnetic interference.

The current system can deliver low-latency performance through discrete GPUs but requires specific optimization for use on standalone VR headsets that operate with Snapdragon XR2. The next research phase needs to examine post-training quantization (PTQ) for converting to INT8 formats which will help decrease inference system requirements. Low-Rank Adaptation (LoRA) and QLoRA techniques offer effective methods for personalized device use because they enable users to customize their devices using their own personal data. The model requires only a small number of parameter adjustments to optimize itself for the user because the edge device handles unique biomechanics of the user through parameter changes which reduce processing costs associated with full network retraining.

Funding: The author(s) received no specific funding for this study.

Data Availability Statement: The original data are available from corresponding authors upon reasonable request.

Acknowledgments: This project was carried out as an internally funded initiative and was carried out with resources available in the CAVE Labs, Center for IoT, PES University, Bengaluru. Department of Computer Science and Engineering provided us with the necessary administrative and academic support.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Winter, D.A. *Biomechanics and Motor Control of Human Movement*, 4th ed.; John Wiley & Sons: Hoboken, NJ, USA, 2009.
2. Cassell, J. Embodied conversational agents: Representation and intelligence in user interfaces. *Commun. ACM* **2000**, *43*, 70–78.
3. Bartlett, R.M. *Introduction to Sports Biomechanics: Analysing Human Movement Patterns*, 2nd ed.; Routledge: London, UK, 2007.
4. Moeslund, T.B.; Hilton, A.; Krüger, V. A survey of advances in vision-based human motion capture. *Comput. Vis. Image Underst.* **2006**, *104*, 90–126.
5. Del Din, S.; Godfrey, A.; Mazzà, C.; Lord, S.; Rochester, L. Free-living gait characteristics in ageing and Parkinson's disease: Effects of environment and ambulatory monitoring. *Gait Posture* **2016**, *49*, 415–422.
6. Cappozzo, A.; Della Croce, U.; Leardini, A.; Chiari, L. Human movement analysis using stereophotogrammetry: Part 1: theoretical background. *Gait Posture* **2005**, *21*, 186–196.
7. Whittle, M.W. *Gait Analysis: An Introduction*, 4th ed.; Butterworth-Heinemann: Oxford, UK, 2007.
8. Richards, J. *Clinical Biomechanics: Assessment of Human Movement*, 2nd ed.; Elsevier: Amsterdam, The Netherlands, 2018.
9. Das, K.; et al. Comparison of markerless and marker-based motion capture systems for gait analysis. *Sci. Rep.* **2023**, *13*, 1–12.
10. Zhou, H.; Hu, H. Human motion tracking for rehabilitation—A survey. *Biomed. Signal Process. Control* **2008**, *3*, 1–18.
11. Sabatini, A.M. Quaternion-based extended Kalman filter for determining orientation by inertial and magnetic sensing. *IEEE Trans. Biomed. Eng.* **2006**, *53*, 1346–1356.

12. Cutti, A.G.; Giovanardi, A.; Rocchi, L.; Davalli, A.; Sacchetti, R. Ambulatory measurement of shoulder and elbow kinematics through inertial and magnetic sensors. *Med. Biol. Eng. Comput.* **2008**, *46*, 169–178.
13. Xsens Technologies. *MVN Awiinda User Manual*; Xsens Technologies: Enschede, The Netherlands, 2021.
14. Huang, X.; et al. Sensor-based wearable systems for monitoring human motion: A review. *Sensors* **2023**, *23*, 1521.
15. Sabatini, A.M. Estimating three-dimensional orientation of human body parts by inertial and magnetic sensing. *Sensors* **2011**, *11*, 1489–1525.
16. Roetenberg, S.; Slycke, P.; Veltink, P. Ambulatory position and orientation tracking in a large volume with calibrated inertial and magnetic sensors. *IEEE Trans. Biomed. Eng.* **2007**, *54*, 883–890.
17. Leardini, A.; Chiari, L.; Della Croce, U.; Cappozzo, A. Human movement analysis using stereophotogrammetry: Part 3. Soft tissue artefact assessment and compensation. *J. Biomech.* **2005**, *38*, 1–11.
18. Windolf, M.; Götzen, N.; Morlock, M. Systematic accuracy analysis of video-based motion capture systems—exemplified on the Vicon-460 system. *Gait Posture* **2006**, *24*, 532–540.
19. Patil, A.K.; Balasubramanyam, A.; Ryu, J.Y.; Bn, P.K.; Chakravarthi, B.; Chai, Y.H. Fusion of Multiple Lidars and Inertial Sensors for the Real-Time Pose Tracking of Human Motion. *Sensors* **2020**, *20*, 5342.
20. Balasubramanyam, A.; Patil, A.K.; Chakravarthi, B.; Ryu, J.Y.; Chai, Y.H. Motion-Sphere: Visual Representation of the Subtle Motion of Human Joints. *Appl. Sci.* **2020**, *10*, 6462.
21. Chakravarthi, B.; Patil, A.K.; Ryu, J.Y.; Balasubramanyam, A.; Chai, Y.H. Scenario-Based Sensed Human Motion Editing and Validation through the Motion-Sphere. *IEEE Access* **2022**, *10*, 28295–28307.
22. von Marcard, T.; Rosenhahn, B.; Black, M.J.; Pons-Moll, G. Sparse Inertial Poser: Automatic 3D Human Pose Estimation from Sparse IMUs. *Comput. Graph. Forum* **2017**, *36*, 349–360.
23. Geissinger, J.H.; Asbeck, A.T. Motion Inference Using Sparse Inertial Sensors, Self-Supervised Learning, and a New Dataset of Unscripted Human Motion. *Sensors* **2020**, *20*, 6330.
24. Zheng, Z.; Ma, H.; Yan, W.; Liu, H.; Yang, Z. Training Data Selection and Optimal Sensor Placement for Deep-Learning-Based Sparse Inertial Sensor Human Posture Reconstruction. *Entropy* **2021**, *23*, 588.
25. Ribeiro, P.M.S.; Matos, A.C.; Santos, P.H.; Cardoso, J.S. Machine Learning Improvements to Human Motion Tracking with IMUs. *Sensors* **2020**, *20*, 6383.
26. Liao, X.; Dong, J.; Song, K.; Xiao, J. Three-Dimensional Human Pose Estimation from Sparse IMUs through Temporal Encoder and Regression Decoder. *Sensors* **2023**, *23*, 3547.
27. Jiang, Y.; Ye, Y.; Gopinath, D.; Won, J.; Winkler, A.W.; Liu, C.K. Transformer Inertial Poser: Real-Time Human Motion Reconstruction from Sparse IMUs with Simultaneous Terrain Generation. *ACM Trans. Graph.* **2022**, *41*, 1–15.
28. Armani, R.; Qian, C.; Jiang, J.; Holz, C. Ultra Inertial Poser: Scalable Motion Capture and Tracking from Sparse Inertial Sensors and Ultra-Wideband Ranging. *ACM Trans. Graph.* **2024**, *43*, 1–16.
29. Movella. *Xsens DOT: Precision Wearable Sensor Platform*; Movella: Henderson, NV, USA, 2023. Available online: <https://www.xsens.com/wearables/xsens-dot> (accessed on 20 March 2026).
30. Movella. *Xsens DOT User Manual*; Document XD0502P, Rev. 1; Movella: Henderson, NV, USA, 2023. Available online: [https://www.movella.com/hubfs/Xsens DOT User Manual.pdf](https://www.movella.com/hubfs/Xsens%20DOT%20User%20Manual.pdf) (accessed on 20 March 2026).
31. World Medical Association. World Medical Association Declaration of Helsinki: Ethical Principles for Medical Research Involving Human Subjects. *JAMA* **2013**, *310*, 2191–2194.
32. Xu, F.; Yi, X.; Liu, S. Improving Global Motion Estimation in Sparse IMU-based Motion Capture. *arXiv* **2025**, arXiv:2505.05010.
33. Nagaraja, D.; Dobinson, R.; Werth, D. Towards Kinematically Constrained Real-Time Human Pose Estimation using Sparse IMUs. In *CEUR Workshop Proceedings*, Aachen, Germany, 2020; Volume 2846, pp. 1–10.
34. Pavllo, D.; Feichtenhofer, C.; Grangier, D.; Auli, M. QuaterNet: A Quaternion-Based Recurrent Model for Human Motion. In *Proceedings of the British Machine Vision Conference (BMVC)*, Newcastle, UK, 3–6 September 2018; p. 299.
35. Martinez, J.; Black, M.J.; Romero, J. On Human Motion Prediction Using Recurrent Neural Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 21–26 July 2017; pp. 2897–2906.
36. Ionescu, C.; Papava, D.; Olaru, V.; Sminchisescu, C. Human3.6M: Large-Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *36*, 1325–1339.

37. Zeng, A.; Sun, X.; Huang, F.; Liu, M.; Lin, Q.; Xu, L. SmoothNet: A Plug-and-Play Network for Refining Human Poses in Videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, QC, Canada, 11–17 October 2021; pp. 11651–11661.
38. Vaswani, A.; Shazeer, N.; Parmar, J.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. In *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*; Curran Associates, Inc.: Red Hook, NY, USA, 2017; pp. 5998–6008.
39. Huang, Y.; Kaufmann, M.; Emmerl, E.; Hilliges, O.; Black, M.J. Deep Inertial Poser: Learning to Reconstruct Human Pose from Sparse Inertial Measurements in Real Time. *ACM Trans. Graph.* **2018**, *37*, 185:1–185:15.
40. Li, J.; Wang, Y.; Zhang, Y.; Xu, C. TransPose: Real-Time 3D Human Pose Estimation from Sparse Inertial Sensors. *IEEE Trans. Vis. Comput. Graph.* **2022**, *28*, 4624–4635.
41. Kaufmann, M.; Zhao, J.; Tang, C.; Tao, L.; Twigg, C.; Song, J.; Wang, R.; Hilliges, O. EM-Pose: 3D Human Pose Estimation from Sparse Electromagnetic Trackers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, QC, Canada, 11–17 October 2021; pp. 11510–11520.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.