Upper-Limb Pose Prediction using Six-Axis Wrist-Worn IMUs

Abhay Sheel Anand
Computer Science
University of Massachusetts Amherst
Amherst, USA
asanand@umass.edu

Jeongah Lee
Computer Science
University of Massachusetts Amherst
Amherst, USA
jeongahlee@cs.umass.edu

Patrick Do
Computer Science
University of Massachusetts Amherst
Amherst, USA
phuocdo@umass.edu

***Abstract*—Estimating accurate 3D upper-limb joint positions from minimal sensing is critical for scalable motion analysis in health monitoring and rehabilitation. We present a lightweight method for predicting shoulder, elbow, and wrist joint coordinates using only six-axis inertial data from two wrist-worn IMUs. Synthetic inertial signals were generated from the high-resolution motion capture GRAB dataset, a subset of AMASS, using the VirtualIMU pipeline. We propose a ConvTransformer-based sequence-to-sequence model that incorporates biomechanical constraints—joint angle limits and bone length consistency—to produce physically realistic and temporally stable pose estimates. Our model achieves a Mean Per Joint Position Error (MPJPE) of 0.106m and a Mean Joint Velocity Error (MJVE) of 0.002, outperforming LSTM and Bi-LSTM baselines across all conditions. Joint-level analysis shows up to 50% error reduction on challenging joints such as the elbow. Results demonstrate the feasibility of accurate, real-world upper-limb motion tracking using only two wearable sensors. Code available at: *https://github.com/jeongah-jasmine-lee/cs690r*.**

***Index Terms*—Human pose estimation, inertial measurement units (IMU), wearable sensing, upper-limb tracking**

I. Introduction

Accurate estimation of human motion is central to applications in physical rehabilitation, ergonomic design, sports science, and immersive human-computer interaction. While optical motion capture (MoCap) systems provide high-fidelity pose data, their high cost, lack of portability, and dependence on controlled environments hinder large-scale deployment. In contrast, inertial measurement units (IMUs) embedded in consumer wearables offer a lightweight, cost-effective, and privacy-preserving alternative for motion analysis in real-world settings.

Recent advances in full-body pose estimation from arrays of IMUs have demonstrated promising results. However, reconstructing upper-limb kinematics from a single wrist-worn IMU remains an open challenge due to the lack of direct measurements from proximal joints like the elbow and shoulder. Overcoming this under-constrained setting would be transformative for scenarios requiring minimal instrumentation, such as remote physiotherapy, continuous activity monitoring, and embedded assistive systems.

In this study, we investigate whether 3D positions of upper-limb joints—specifically the shoulder, elbow, and wrist—can be accurately predicted using only six-axis inertial data from a single wrist-mounted sensor. We propose a ConvTransformer architecture that combines convolutional feature extraction with temporal attention to model the inverse kinematics mapping. To address ambiguity in sparse sensing, we incorporate biomechanical constraints, including joint angle limits and bone length consistency, to guide the model toward generating physiologically valid and temporally stable pose estimates.

Our training and evaluation pipeline uses a high-quality MoCap dataset, GRAB (GRasping Actions with Bodies), from which we synthesize IMU signals using the VirtualIMU framework (Figure 2). We evaluate model performance using Mean Per Joint Position Error (MPJPE) and Mean Joint Velocity Error (MJVE), and validate outputs qualitatively through animated skeleton overlays.

We aim to answer three core research questions: (1) Which model architectures (e.g., LSTM, BiLSTM, ConvTransformer) best capture upper-limb motion from limited inertial data? (2) How does IMU preprocessing (e.g., filtering, orientation estimation) affect pose prediction accuracy? (3) Can biomechanical constraints enhance the realism and accuracy of predicted poses?

Our findings show that the ConvTransformer consistently outperforms baseline models, achieving the lowest MPJPE and MJVE across all experimental conditions. While both LSTM and ConvTransformer exhibit robustness under biomechanical constraints and preprocessing, Bi-LSTM suffers significant performance degradation, indicating sensitivity to data transformations. Notably, ConvTransformer halves the worst-case joint errors (especially for elbows, joints 6 and 7), reducing error variance from 0.03–0.115 m (LSTM) to 0.02–0.057 m.

In summary, our contributions are threefold. First, we demonstrate that accurate upper-limb pose estimation is feasible using only two wrist-worn IMUs, overcoming sparse input with deep learning and biomechanical priors. Second, we introduce a ConvTransformer-based model tailored to inertial input, validated on synthetic signals from MoCap data. Third, we offer detailed joint-wise analysis and ablation studies revealing how biomechanical constraints and signal preprocessing influence model robustness and generalization.

This work lays the foundation for scalable and unobtrusive upper-limb motion tracking using commodity wearables, with implications for health monitoring, rehabilitation, and context-

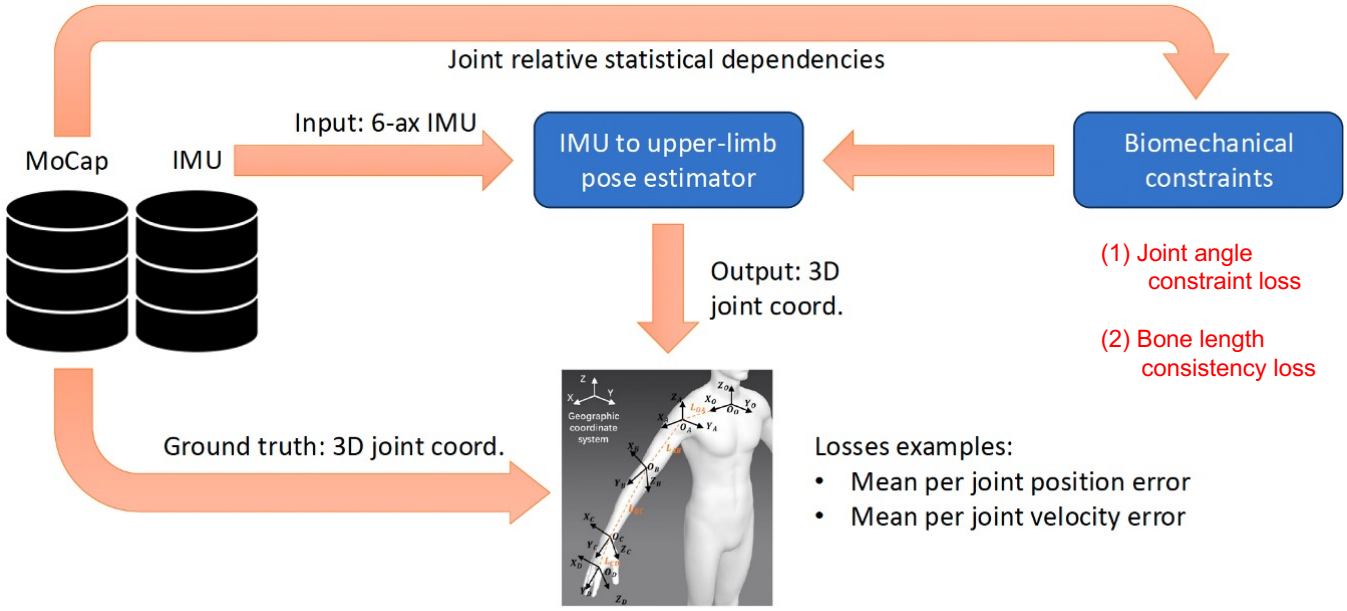

Fig. 1. Overview of the full pose estimation pipeline.

aware applications.

II. Related works

A. Virtual IMU Data Generation for Deep Learning

Recent advancements in virtual IMU synthesis have addressed the critical bottleneck of wearable data scarcity for training deep learning models. Traditional IMU datasets are limited by the cost and effort required for synchronized inertial and motion capture recordings. To overcome this, researchers have proposed methods to simulate IMU data from skeletal representations.

While prior works such as SynthIMU [1] and IMUTube [2] offer pipelines for converting video-derived skeletons into inertial signals, they often assume rigid sensor placements and lack robust data augmentation. In contrast, our own work, VirtualIMU [3], introduces a fully differentiable and physically grounded pipeline for synthesizing six-axis IMU data from both video and motion capture skeletal inputs. Our method systematically accounts for variability in sensor placement and orientation using Monte Carlo simulation and provides augmentation strategies that reflect realistic variability in human morphology.

Unlike earlier pipelines, VirtualIMU enables the creation of population-specific and personalized sensor models, improving the generalizability of downstream pose estimation networks. In this work, we adopt the VirtualIMU framework to generate large-scale synthetic IMU signals from the GRAB motion capture dataset for training our wrist-mounted pose estimation model.

B. Sparse IMU-Based Pose Estimation

Upper-limb pose estimation using wearable sensors has gained significant attention in recent years, particularly with the emergence of sparse IMU-based and deep learning approaches.

Ultra Inertial Poser [4] demonstrated that even sparse IMU and ultra-wideband (UWB) signals could be fused to estimate full-body pose using a graph-based spatio-temporal model. While their method combines multiple sensing modalities, our approach uses only wrist-worn IMUs, making the sensing setup significantly more minimal. Capturing Upper Body Kinematics [5] proposed a deep sequence-to-sequence model using three IMUs (wrists and pelvis) to predict upper-body motion. In contrast, our work tackles the more challenging setting of predicting full upper-limb kinematics using just two wrist IMUs. IMUPoser [6] addressed real-time full-body pose estimation from IMUs embedded in phones, earbuds, and watches, utilizing a bi-directional LSTM and inverse kinematics refinement. Unlike IMUPoser, which depends on global orientation inputs, our model relies entirely on local sensor readings and incorporates biomechanical priors to ensure anatomical plausibility.

C. Biomechanical Constraints in Pose Estimation

Biomechanical constraints play a vital role in ensuring that predicted poses are not only accurate but also physically and anatomically plausible. Prior research has introduced various constraint-based formulations to regularize deep learning-based pose estimation models.

Van der Steen et al. [7] explored joint-angle coordination to stabilize tool-use motion, highlighting the physiological coupling between limb segments. Similarly, Wang et al. [8] and Hsu et al. [9] applied joint-angle and bone-length constraints to mitigate implausible deformations in 3D human pose estimation.

Recent methods have expanded constraint-based modeling across different domains. Pavllo et al. [10] enforced smooth joint trajectories with bone-length preservation for video-based 3D pose lifting. Wandt et al. [11] introduced a canonical pose space with geometric constraints to resolve scale and depth ambiguity in monocular setups. Zanfir et al. [12] added biomechanical priors to predict temporally coherent poses from monocular images using learned anthropometric models. Kanazawa et al. [13] combined differentiable rendering and SMPL-based constraints to match 3D poses with human shape consistency. Arnab et al. [14] included temporal and kinematic losses, while Xu et al. [15] enforced biomechanical realism using the GHUM model for global and local joint behavior. Taneja et al. [16] proposed PhysCap, an optimization-based method with physics-informed priors, and Rempe et al. [17] modeled contact-aware kinematics for realistic human motion synthesis.

Our work extends this line of research by integrating joint-angle and bone-length constraints into a sequence-to-sequence deep learning model trained solely on six-axis data from a single wrist-worn IMU. Unlike previous studies that rely on multiple sensors or camera inputs, we demonstrate that biomechanical priors can regularize pose predictions even in sparse sensing setups, yielding anatomically consistent results with minimal instrumentation.

III. Methodology

A. Datasets and IMU Synthesis

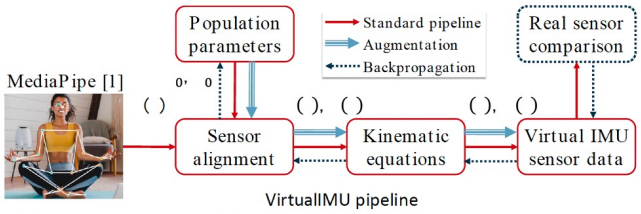

Fig. 2. VirtualIMU pipeline used to synthesize six-axis inertial data (accelerometer and gyroscope) from MoCap sequences.

We utilize GRAB, a high-quality motion capture subdataset of AMASS, to generate synthetic IMU signals for training and evaluation. Amass [18] is a large-scale compilation of over 40 MoCap datasets unified under a common SMPL representation. It includes more than 300 subjects and 40+ hours of recorded motion, covering a wide range of full-body activities such as walking, running, reaching, sitting, and gesturing. The dataset offers high spatial and temporal resolution, making it well-suited for generating physically realistic inertial data.

GRAB (GRasping Actions with Bodies) [19], a subset of AMASS, focuses specifically on detailed hand-object interaction scenarios. It includes 10 subjects interacting with 51 everyday objects in controlled settings. Recorded activities encompass lifting, passing (to someone or from dominant hand to non-dominant hand), and functional usage (e.g., pouring, pressing). The MoCap annotations include upper-body joint positions—shoulders, elbows, wrists—and precise thumb and hand kinematics, enabling fine-grained modeling of forearm dynamics.

Using the VirtualIMU pipeline [3], we synthesize six-axis inertial data—comprising 3-axis acceleration and 3-axis angular velocity—by placing virtual IMUs on the forearms of both arms. The IMU synthesis process leverages the spatial trajectories of relevant joints and simulates sensor measurements within a local coordinate frame. This enables us to replicate realistic sensor readings while maintaining full access to ground-truth joint positions.

B. Preprocessing Pipeline

To enhance signal quality and align coordinate frames, we apply several preprocessing steps. First, filtering is applied to reduce high-frequency noise and sensor drift in both the accelerometer and gyroscope data. We use a combination of low-pass and band-pass filters within the 0.1–20 Hz range to retain only the relevant motion-related signal components.

Next, we estimate the orientation of each virtual sensor using the AQUA algorithm. This orientation estimate is then used to transform the raw IMU measurements, specifically the accelerometer and gyroscope data, into a unified global coordinate system from their local sensor coordinate frames. After this transformation, we subtract the gravitational acceleration component from the Z-axis of the global-frame accelerometer signals, allowing us to isolate motion-related accelerations.

Finally, we segment the continuous time-series data into fixed-length windows to prepare it for sequence modeling. A sliding window approach is used, with each window spanning 1 second and overlapping the previous one by 75%. This segmentation strategy allows the model to learn from temporally coherent input sequences while maintaining high sampling coverage of the original data.

C. Model Architectures

To evaluate the effectiveness of different temporal modeling strategies, we implement and compare three deep learning architectures for upper-limb pose prediction from inertial signals. As a baseline, we use a 2-layer unidirectional Long Short-Term Memory (LSTM) network, which processes input sequences frame by frame. This model is designed to capture short-range motion dynamics efficiently while maintaining low latency and modest computational overhead.

We also explore a 2-layer Bidirectional LSTM (Bi-LSTM), which processes sequences in both forward and backward directions. By leveraging information from both past and future time steps, the Bi-LSTM is capable of capturing richer temporal dependencies across the motion sequence.

Lastly, we introduce the ConvTransformer architecture, which draws inspiration from the work of Shavit et al. [20]. This hybrid model combines convolutional layers for local temporal feature extraction with a Transformer encoder to model long-range spatial-temporal dependencies. As shown in Figure 3, this architecture is designed specifically to support

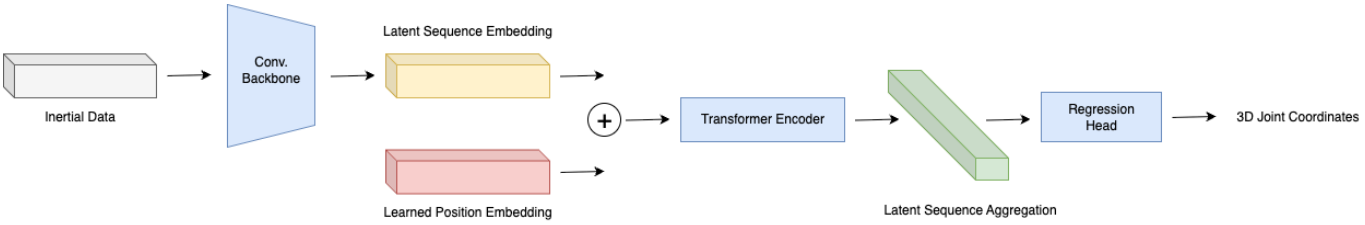

Fig. 3. ConvTransformer architecture.

fine-grained 3D joint prediction using only wrist-mounted IMU data.

All of our models operate on fixed-length windows of synchronized six-axis IMU readings from two devices. Concretely, each input batch is a tensor of shape (B, W, 4, 3)—B batches, W timesteps, and "4 × 3" channels that encode two 3-axis accelerometers and two 3-axis gyroscopes (six axes per sensor flattened as four triaxial channels), which we then reshape to (B, W, 12) per step. For the LSTM and Bi-LSTM baselines, those 12-dimensional vectors feed into a two-layer Bi-LSTM with the hidden size of 24, producing an output sequence of shape (B, W, 24) that we interpret as per-step 3D coordinates of 8 joints (left and right shoulders, elbows, wrists, and fingers). The ConvTransformer likewise accepts (B, W, 12) input, projects it via 1D convolutions into tokens, applies self-attention across timesteps, and through a regression head emits a tensor of shape (B, Wx8x3)—the concatenated 3D coordinates of all eight target joints across the window.

D. Biomechanical Constraints

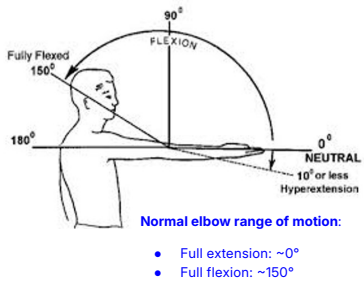

Fig. 4. Joint-angle constraint loss. Predictions falling outside the physiological flexion range (0° to 150°) are penalized to maintain anatomical realism.

To enforce anatomical plausibility in the predicted joint positions, we introduce two biomechanical loss components into our model: joint angle constraint and bone length consistency. The joint angle constraint targets the elbow joint, where the flexion angle is defined by the vector geometry between the shoulder, elbow, and wrist. Predictions are penalized when elbow angles fall outside the physiological range of 0° to 150°, ensuring that the estimated motion adheres to known anatomical limitations (Figure 4).

In addition, the bone length consistency constraint minimizes variance in the estimated lengths of anatomical

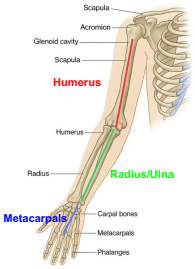

Fig. 5. Visualization of upper-limb bone segments used in the bone length consistency loss. The upper arm (humerus), forearm (radius/ulna), and hand (metacarpals) are tracked to ensure stable, anthropometrically consistent segment lengths across time.

segments—namely the upper arm (humerus), forearm (radius/ulna), and hand (metacarpals)—across time. This helps enforce temporal smoothness and anthropometric realism in the model's output (Figure 5).

The full loss function used during training combines these biomechanical constraints with the standard mean squared error (MSE) loss. When no constraints are applied, the loss is computed as:

$$Loss = MSE$$

When biomechanical terms are incorporated, the overall loss function becomes:

$$Loss = \lambda_1 \cdot MSE + \lambda_2 \cdot Loss_{angle} + \lambda_3 \cdot Loss_{bone}$$

Here, $Loss_{angle}$ penalizes elbow flexion angles outside the valid range, while $Loss_{bone}$ maintains consistent bone segment lengths across frames to reflect realistic body proportions. The scalar weights $\lambda_1, \lambda_2, \lambda_3$ control the relative importance of each term, allowing us to balance prediction accuracy with biomechanical fidelity.

This approach enables the model to minimize spatial prediction error while also generating motion sequences that are biomechanically plausible and temporally stable.

E. Evaluation Metrics and Analysis

Model performance is assessed using two key metrics: Mean Per Joint Position Error (MPJPE) and Mean Joint Velocity Error (MJVE).

The MPJPE metric measures the average Euclidean distance between the predicted and ground-truth joint positions across

all frames and joints. It evaluates the spatial accuracy of pose estimation and is formally defined as:

$$MPJPE = \frac{1}{N_F} \cdot \frac{1}{N_J} \sum_{f,j} \|p_{f,j} - \hat{p}_{f,j}\|_2$$

Here, N_F denotes the number of frames, N_J the number of joints, $p_{f,j}$ the ground truth position of joint j at frame f, and $\hat{p}_{f,j}$ the corresponding predicted position.

The MJVE metric quantifies the difference in temporal smoothness between the predicted and ground-truth joint velocities, thereby evaluating the model's ability to preserve motion continuity. It is defined as:

$$MJVE = \frac{1}{N_F} \cdot \frac{1}{N_J} \sum_{f,j} \|v_{f,j} - \hat{v}_{f,j}\|_2$$

In this expression, $v_{f,j}$ and $\hat{v}_{f,j}$ represent the ground truth and predicted velocities of joint j at frame f, respectively. Like MPJPE, the averaging is performed across all joints and frames to reflect overall performance.

Together, MPJPE and MJVE capture both the spatial precision and temporal stability of the predicted 3D joint trajectories.

IV. Results

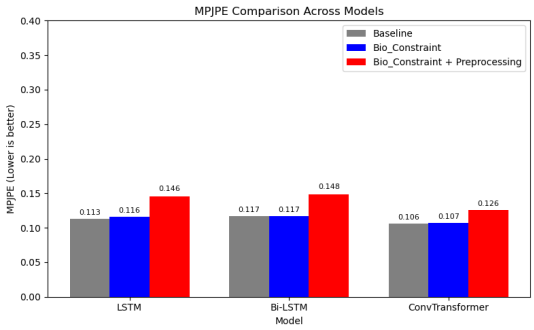

Fig. 6. MPJPE

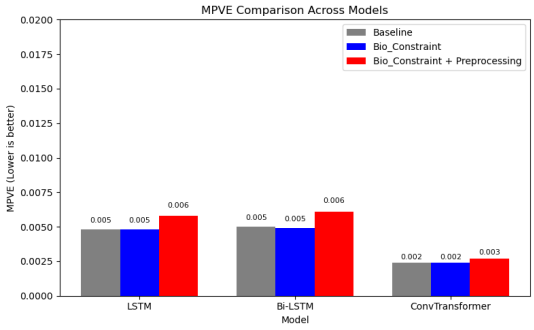

Fig. 7. MPVE

Figure 6 and Figure 7 present a comparative analysis across three model architectures (LSTM, Bi-LSTM, and ConvTransformer) under three experimental conditions: baseline, biomechanical constraint only, and biomechanical constraint with preprocessing.

ConvTransformer achieves the best overall performance, with the lowest MPJPE (0.106 in baseline) and MPVE (0.002 in baseline). Both LSTM and ConvTransformer demonstrate robustness across all three settings, with performance variation contained within a $\pm 3\%$ margin. In contrast, Bi-LSTM shows the most pronounced degradation when both biomechanical constraints and preprocessing are applied, indicating its sensitivity to input transformation.

Introducing biomechanical constraints alone caused only marginal increases in error across all models. However, adding preprocessing steps further degraded performance—particularly for Bi-LSTM and ConvTransformer—likely due to information loss and distributional shift introduced by orientation transformations and smoothing filters.

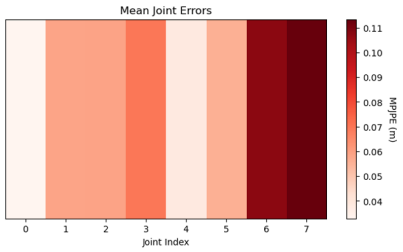

Fig. 8. Per-joint MPJPE for the LSTM model. The model exhibits larger errors and higher variance, particularly for distal joints. Joint indices: 0—Left shoulder, 1—Left elbow, 2—Left wrist, 3—Left finger, 4—Right shoulder, 5—Right elbow, 6—Right wrist, 7—Right finger.

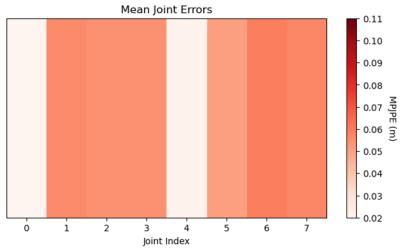

Fig. 9. Per-joint MPJPE for the ConvTransformer model. The model demonstrates low error variance across joints, with notable improvements at the distal joints (Joint 6: Right wrist, Joint 7: Right finger). See joint index mapping in caption of Figure 8.

Joint-level analysis reveals that ConvTransformer significantly reduces worst-case joint errors. Specifically, it halves the error for the most challenging joints (joints 6 and 7, corresponding to elbows) and compresses the MPJPE variance across all joints from 0.03–0.115 m (LSTM) to 0.02–0.057 m (ConvTransformer). This indicates that ConvTransformer not only improves average prediction accuracy but also enhances consistency across joints.

In addition to numerical metrics, we generated animated 3D skeleton overlays to qualitatively validate the predicted trajectories against ground truth MoCap sequences. Visual

inspection confirms that ConvTransformer produces smoother, anatomically plausible motion with reduced limb jitter and better joint coordination than its counterparts.

Moreover, we observe a clear gradient of increasing error along the kinematic chain: the shoulder exhibits the lowest MPJVE, the elbow error is slightly higher, the wrist is higher still, and the fingers show the largest errors. This monotonic rise from proximal to distal joints likely reflects both the accumulation of orientation uncertainty down the chain and the smaller movement amplitudes (and thus lower signal-to-noise) at the finger joints.

V. Discussion

Our experimental results highlight several critical insights into model behavior, data representation, and biomechanical modeling in the context of upper-limb pose estimation using minimal sensor input.

ConvTransformer consistently outperformed other models across both spatial (MPJPE) and temporal (MPVE) metrics. Its hybrid design—combining local temporal convolution with global attention—appears well-suited to capturing both short-term dynamics and long-range joint dependencies inherent in upper-limb motion. LSTM also performed reliably and exhibited minimal sensitivity to constraints or preprocessing, making it an attractive choice for low-power or latency-constrained settings. Bi-LSTM, however, showed notable degradation when biomechanical constraints and preprocessing were applied, suggesting overfitting to unprocessed training distributions or conflict with constraint-induced gradient signals.

Contrary to our initial hypothesis, biomechanical constraints did not consistently improve quantitative metrics. While they introduced anatomical priors into the learning process, they occasionally acted as a form of regularization that limited model expressiveness—especially for models like Bi-LSTM that were not trained to reconcile constraints during optimization. Moreover, the use of fixed joint-angle thresholds and uniform bone-length penalties may have introduced optimization conflicts, particularly when joint configurations near valid boundaries were penalized disproportionately.

Preprocessing IMU signals—such as global frame transformation and gravity compensation—was expected to enhance model generalization. However, our findings suggest that these steps can introduce artifacts or reduce motion fidelity, leading to increased MPJPE and MPVE. This is especially evident for models not retrained on transformed inputs, highlighting a mismatch between training and inference distributions.

Our joint-wise analysis revealed that elbow joints (particularly joints 6 and 7) contributed disproportionately to overall error. This aligns with the inherent difficulty of inferring proximal joint motion from distal sensing, and confirms the value of architectural designs like ConvTransformer that reduce variance and improve stability across challenging joints.

This study relies on synthetic IMU data generated from MoCap using VirtualIMU. Although this setup ensures controlled experimentation and reliable ground truth, the domain gap between synthetic and real-world IMU signals may limit generalization. Additionally, our constraints assume fixed anthropometrics and are not personalized per subject, which could affect accuracy in diverse populations.

VI. Conclusion

In this work, we investigated the feasibility of predicting 3D upper-limb skeletal poses using only six-axis IMU data from two wrist-worn sensors. Our approach combined synthetic inertial data generation with sequence-to-sequence deep learning models and incorporated biomechanical constraints to improve anatomical plausibility.

Among the models evaluated, ConvTransformer demonstrated superior performance in both spatial and temporal accuracy, achieving the lowest MPJPE and MPVE across all experimental conditions. While LSTM maintained robust and consistent predictions, Bi-LSTM exhibited sensitivity to input transformations and constraint integration. Our joint-wise analysis further revealed that ConvTransformer significantly reduced errors for anatomically challenging joints, particularly the elbows.

Interestingly, the inclusion of biomechanical constraints and preprocessing steps did not always yield performance gains, and in some cases, led to degradation due to distribution shift or over-regularization. These findings underscore the importance of aligning model training with the intended inference conditions and carefully balancing constraint design with prediction flexibility.

Overall, our results suggest that accurate, lightweight upper-limb motion tracking is possible with a single-point IMU, provided that architectural and data-driven priors are effectively leveraged. This work serves as a foundation for deploying IMU-based pose estimation in real-world applications such as remote rehabilitation, assistive wearables, and health monitoring systems.

References

[1] J. Huang, A. Kanazawa, and J. Malik, "Synthimu: Learning imu data for human pose estimation," in *European Conference on Computer Vision (ECCV)*. Springer, 2020.

[2] H. B. Kwon, C. Tong, H. Haresamudram, Y. Gao, G. D. Abowd, N. D. Lane, and T. Ploetz, "Imutube: Automatic extraction of virtual on-body accelerometry from video for human activity recognition," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 4, no. 3, pp. 1–29, 2020.

[3] I. Gavier, Y. Liu, and S. I. Lee, "Virtualimu: Generating virtual wearable inertial data from video for deep learning applications," in *IEEE International Conference on Body Sensor Networks (BSN)*. IEEE, 2023.

[4] Z. Chen, S. Xu, H. Rhodin, and L. Jiang, "Ultra inertial poser: Scalable motion capture and tracking from sparse inertial sensors and ultra-wideband ranging," in *ACM SIGGRAPH 2024 Conference Proceedings*, 2024.

[5] C. Dai, T. Liu, and M. Wang, "Capturing upper body kinematics and localization with low-cost wearable sensors," *Sensors*, vol. 22, no. 8, p. 2931, 2022.

[6] A. Gilbert, C. Ladha, and et al., "Imuposer: Full-body pose estimation using imus in phones, watches, and earbuds," in *Proceedings of the IEEE International Conference on Computer Vision*, 2021.

[7] M. van der Steen, R. Bongers, and et al., "Joint-angle coordination patterns ensure stabilization of a body–tool system," *Frontiers in Psychology*, vol. 7, p. 933, 2016.

[8] G. Wang, S. Tang, and Y. Wu, "Motion projection consistency based 3d human pose estimation with virtual bones from monocular videos," *IEEE Transactions on Cognitive and Developmental Systems*, 2022.

[9] P.-W. Hsu, F.-J. Huang, T.-H. Chen, and S.-Y. Kuo, "Blapose: Bone-length-aware learning for 3d human pose estimation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2024, pp. 1234–1243.

[10] D. Pavllo, C. Feichtenhofer, D. Grangier, and M. Auli, "3d human pose estimation in video with temporal convolutions and semi-supervised training," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 7753–7762.

[11] B. Wandt and B. Rosenhahn, "Canonpose: Self-supervised monocular 3d human pose estimation in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 13 294–13 303.

[12] A. Zanfir, A.-I. Popa, M. Zanfir, and C. Sminchisescu, "Monocular 3d pose and shape estimation of multiple people in natural scenes—the importance of multiple scene constraints," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 2148–2157.

[13] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik, "End-to-end recovery of human shape and pose," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7122–7131.

[14] A. Arnab, C. Doersch, and A. Zisserman, "Exploiting temporal context for 3d human pose estimation in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 3395–3404.

[15] W. Xu, E. Bazavan, A. Zanfir, W. T. Freeman, R. Sukthankar, and C. Sminchisescu, "Ghum & ghuml: Generative 3d human shape and articulated pose models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 6184–6193.

[16] A. Taneja, S. Tripathi, G. Lee, and Y. Sheikh, "Physcap: Physically plausible monocular 3d motion capture in real time," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 12 815–12 825.

[17] D. Rempe, N. Mahmood, M. J. Black, G. Pons-Moll, and L. Liu, "Humor: 4d human motion model for realistic animation and simulation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 11 007–11 017.

[18] N. Mahmood, N. Ghorbani, N. F. Troje, G. Pons-Moll, and M. J. Black, "Amass: Archive of motion capture as surface shapes," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 5442–5451.

[19] S. Brahmbhatt, C. Ham, C. C. Kemp, and J. Hays, "ContactDB: Analyzing and predicting grasp contact via thermal imaging," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. [Online]. Available: https://contactdb.cc.gatech.edu

[20] Y. Shavit and I. Klein, "Boosting inertial-based human activity recognition with transformers," *IEEE Access*, vol. 9, pp. 53 540–53 547, 2021.