Journal of Information Systems Engineering and Management
2025, 10(37s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

PoseRx: A Transformer-Based Remedy for Precision in Physical Rehabilitation Monitoring

Deepak Shukla[1], Maya Rathore[2]

[1,2]Department of Computer Science & Engineering

[1,2]Faculty of Engineering

[1,2]Oriental University, Indore, India

deepakactive@gmail.com, mayarathore114@gmail.com

ARTICLE INFO

Received: 15 Dec 2024

Revised: 18 Feb 2025

Accepted: 26 Feb 2025

ABSTRACT

Accurate and reliable human pose estimation plays a vital role in physical rehabilitation, where therapists depend on precise joint tracking to evaluate posture, monitor range of motion, and assess patient progress. While traditional pose estimation models such as AlphaPose, MediaPipe, and HybrIK have shown varying levels of performance, they often struggle in scenarios involving occlusions and diverse body positions commonly encountered in rehabilitation settings. Their limitations—particularly in terms of temporal consistency, occlusion robustness, and joint angle accuracy—undermine their clinical applicability. To address these challenges, this study introduces PoseRx, a transformer-driven pose estimation framework built on the TokenPose architecture, specifically tailored for physical rehabilitation monitoring. PoseRx processes RGB video inputs and employs a Vision Transformer-based joint attention mechanism to estimate 2D keypoints, which are subsequently lifted to 3D using temporal models such as VideoPose3D. The framework is evaluated across rehabilitation-specific postures, including supine, seated, and standing positions, and benchmarked against state-of-the-art methods using metrics such as 2D localization error, joint angle mean absolute error (MAE), model complexity, and occlusion handling capability. Results demonstrate that PoseRx achieves superior performance, with a 2D localization error as low as 5.9 pixels and a joint angle MAE of 5.4°, outperforming existing models across all evaluated positions. Moreover, it exhibits the highest resilience to occlusion and provides enhanced support for custom joints, both of which are essential in real-world rehabilitation scenarios. PoseRx delivers a robust, efficient, and clinically relevant solution for human pose tracking in rehabilitation environments. Its transformer-based design and modular architecture make it a promising next-generation tool for improving physiotherapy feedback, tracking patient progress, and advancing digital health interventions..

Keywords: Pose Estimation, Physical Rehabilitation, TokenPose, Vision Transformer, Joint Angle Analysis, Occlusion Handling.

1. Introduction

In recent years, Human Pose Estimation (HPE) has emerged as a foundational technology across diverse fields, including sports analytics, virtual reality, and biomechanical assessment. Among these, physical rehabilitation stands out as a domain where precision and temporal consistency in pose estimation directly impact patient recovery outcomes. Therapists and clinicians depend on accurate joint tracking to assess posture, monitor range of motion, and ensure the correct execution of prescribed exercises.

However, traditional rehabilitation monitoring methods whether manual observation or sensor-based systems—are often limited by subjectivity, intrusiveness, or scalability. This has catalyzed growing interest in vision-based pose estimation systems, particularly those that are accurate, real-time, and resilient to the challenges of clinical environments such as occlusion and non-ergonomic postures. Despite recent advancements in HPE, leading models like AlphaPose, MediaPipe, and HybrIK show performance degradation in rehabilitation-specific settings. While they excel in standard upright positions, they frequently underperform in postures such as supine or seated, which are common during

Journal of Information Systems Engineering and Management
2025, 10(37s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

physiotherapy sessions. Moreover, their vulnerability to occlusions, inconsistent joint localization, and poor generalization across anatomical variations diminish their reliability in real-world clinical applications. Notably, most of these models lack support for joint angle estimation and custom joint sets, which are critical for therapeutic progress tracking. To bridge these gaps, we propose PoseRx a transformer-based pose estimation framework explicitly designed for precision monitoring in physical rehabilitation. Built on the TokenPose architecture, PoseRx utilizes Vision Transformers (ViT) to model each joint as a learnable token, thereby capturing intricate spatial dependencies across the human skeleton. This joint-token representation enhances robustness to occlusion and body orientation, enabling consistent performance across supine, seated, and standing postures.

PoseRx integrates seamlessly with temporal lifting models like VideoPose3D to produce 3D keypoints from video input and calculate joint angles with high precision. This integration enables dynamic range-of-motion analysis, supporting both static assessments and continuous motion evaluation—key for monitoring rehabilitation progress.

What distinguishes PoseRx is its balanced combination of accuracy, real-time inference capability, and clinical adaptability. It achieves 2D localization error as low as 5.9 pixels and joint angle MAE of just 5.4°, outperforming state-of-the-art models under rehabilitation-specific conditions. The model also ranks highest in occlusion resilience (4/4) and supports custom joint configurations, making it especially effective in scenarios where limbs may be partially obscured or non-standard joints are under observation. With an efficient processing speed of approximately 24.5 frames per second (FPS) and a moderate parameter size (~75M), PoseRx strikes an optimal balance between performance and deployability. Its versatility positions it as a valuable tool for applications such as tele-physiotherapy, automated joint tracking, and real-time AI-assisted rehabilitation feedback systems. PoseRx represents a significant step forward in making pose estimation more applicable and effective for the healthcare domain. Its transformer-based design, occlusion robustness, and support for 3D biomechanical analysis make it a next-generation solution for enhancing the accuracy and scalability of physical rehabilitation monitoring.

2. Literature review

Aguilar-Ortega, R. et al. (2023), Physical rehabilitation is vital for restoring motor function after injury or surgery, but overcrowded medical systems make personalized monitoring difficult. Deep learning-based human pose estimation offers a scalable solution to track recovery remotely. This study focuses on evaluating multiple pose estimation models, analyzing the impact of camera viewpoints and body positions, and determining whether 2D estimation suffices or 3D is necessary. A custom dataset featuring 27 subjects performing 8 exercises from 5 camera angles was collected using an OptiTrack system as ground truth. Results reveal significant variability in model performance, with frontal camera views yielding the most accurate pose estimates. Importantly, the study concludes that 2D estimators are sufficient for estimating joint angles in most scenarios, making them practical for scalable rehab monitoring [1].

He, S. et al. (2024), In a clinical trial involving older adults with sarcopenia, an AI-based remote training group using 3D human pose estimation was compared against face-to-face and general remote training groups. All groups followed a Taichi-based rehab program over 3 months. Various physical and functional metrics (e.g., ASMI, TUGT, QoL) were evaluated at pre-, mid-, and post-stages. Results showed significant improvements across all groups, with AI-based remote training performing on par with traditional face-to-face rehabilitation. This confirms that AI-driven 3D pose estimation can effectively support remote rehabilitation with outcomes similar to in-person therapy [2].

Roggio, F. et al. (2024), This narrative review explores the role of machine learning-based pose estimation models (PEMs) in human movement sciences. It highlights models such as OpenPose, PoseNet, AlphaPose, and BlazePose, which offer non-invasive, cost-effective alternatives for analyzing posture, gait, and movement in clinical, sports, and ergonomic contexts. These models help diagnose musculoskeletal disorders, enhance athletic performance, and prevent workplace injuries. However, challenges such as data quality, accuracy, and lack of standardized protocols limit their integration into practical workflows. The review emphasizes the need for robust, validated frameworks to fully realize the potential of ML-based pose estimation in healthcare and performance monitoring [3].

Journal of Information Systems Engineering and Management
2025, 10(37s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Hernández, Ó.G. et al. (2021), This study compares OpenPose and Detectron 2 for estimating joint angles during four upper-limb rehabilitation exercises using two Kinect 2 RGBD cameras as ground truth. The evaluation focused on elbow and shoulder angles using RMSE and MAE metrics. Results showed that OpenPose consistently outperformed Detectron 2 in accuracy, demonstrating its superior suitability for upper-limb rehabilitation applications. The study reinforces the effectiveness of marker-less, vision-based pose estimation systems in clinical scenarios when benchmarked against depth-based sensors [4].

Kang, N. et al. (2024), Addressing challenges in elderly rehabilitation, this work proposes a multistage 3D pose estimation system based on monocular RGB input. The model, combining HRNet and OCNet, aims to overcome occlusion-related issues, including full end-joint occlusion. A correction mechanism based on Part Affinity Fields (PAF) improves estimation under difficult visual conditions. Designed for elderly patients facing mobility restrictions, the system delivers accurate 3D pose sequences critical for remote rehabilitation and progress evaluation. This model reduces psychological and economic burdens by enabling precise monitoring from home [5].

Zhang, X. et al. (2024), This study presents an AI-powered architecture, ConvTrans, that enhances posture estimation for stroke patients undergoing rehabilitation. By combining spatial convolutional layers with an improved lightweight transformer (LMHSA + IRFFN), the system effectively balances local and global feature extraction. It generates real-time skeletal feedback to guide patients during independent home exercises. Demonstrated across three HPE datasets, this approach reduces subjectivity, improves cost-efficiency, and enhances real-time motion feedback in stroke rehabilitation scenarios [6].

Giulietti, N. et al. (2025), A vision-based, marker-less system is proposed for real-time 3D pose estimation using pre-trained 2D models and a novel Weighted Direct Linear Triangulation method. Integrated with a clinical rehabilitation robot, the system employs YOLOv8x-pose and achieves high accuracy (18.2 mm MPJPE) with low latency (15 ms). With optimization using TensorRT, the system dynamically controls robotic movement based on the patient's pose, enabling immersive and responsive exergames during rehab [7].

Rincon, J.A. et al. (2024), This study introduces a compact, vision-based robotic assistant for elderly rehabilitation. It uses AI and RREF-based pose matching to provide real-time exercise guidance via a 12-servo motor torso and OLED display. Powered by edge computing and a Grove AI vision module, the assistant offers personalized feedback while ensuring privacy. The system promotes independence and active engagement among elderly users, supporting therapists and caregivers through intelligent, scalable care [8].

Kumar, V. et al. (2024), To enhance prosthetic design and neurological rehabilitation, this study compares YOLOv8, DeepPose, and RTM Pose (ViT-based) for automated gait analysis using the 3DPW dataset. RTM Pose achieves the best performance (lowest MAE, RMSE) with a fast inference time (107.7ms), proving its value in detecting gait abnormalities and supporting diagnosis for stroke, Parkinson's, and brain injury patients. This work sets a new benchmark for precise, real-time gait analysis using transformer-based architectures [9].

Tluli, R. et al. (2024), This research proposes a physiotherapy exercise classification pipeline integrating pose estimation with multiple ML techniques using the alwaysAI platform. Beyond traditional SVMs, it explores diverse models for high-dimensional pose data, achieving accurate classification of eight exercise types. The results validate the robustness of the approach and demonstrate its practical potential in automating physiotherapy assessment and feedback within clinical or remote care environments [10].

Pavlikov, A. et al. (2024), This paper proposes a video-based driver monitoring system that uses neural networks to detect key physical indicators like head and arm movement. The system compares popular frameworks—AlphaPose, OpenPose, PoseNet, and Mask R-CNN and provides a mathematical model of human upper body kinematics. The study highlights a continuous, real-time video processing system for monitoring driver condition, useful in real-world applications to enhance safety during long trips [11].

Ali, M.M. et al. (2024), This study introduces a binary gait classification method using 2D and 3D pose estimation to detect signs of Duchenne Muscular Dystrophy (DMD) in children. Using gait features

such as step length and joint angles, the system compares healthy and affected individuals using RGB video and MoCap data. Machine learning models like SVM and deep neural networks achieved high accuracy (96.2–97%) but could not yet differentiate DMD from other gait disorders. The method is cost-effective and suitable for early DMD diagnosis [12].

Stenum, J. et al. (2021), This interdisciplinary review discusses how pose estimation using low-cost devices (e.g., smartphones) can democratize access to motor assessment and movement analysis. Applications include at-home patient monitoring, on-field sports coaching, and clinical assessments in neurology and physiotherapy. Despite the benefits, challenges remain in data accuracy, standardization, and integration into clinical workflows. The paper advocates for continued development and collaboration across disciplines to unlock the full potential of pose estimation technologies [13].

Miao, S. et al. (2024), To address the high cost and environmental constraints of devices like Kinect, this study presents a deep learning-based method using a monocular camera, Faster R-CNN, and HRNet for upper limb pose recognition. It integrates LSTM with ProbSparse Self-Attention for movement evaluation. The proposed method achieved 94.1% accuracy and outperformed baseline models, making it practical for affordable, accurate rehabilitation tracking in complex settings[14].

Ekambaram, D. et al. (2024), This work introduces a real-time AI-based system for evaluating wrist extension and flexion exercises, replacing the need for a physiotherapist. Using a DenseNet-based CNN, the system provides feedback within 0.79 seconds and operates at ~21 FPS. It achieved 100% accuracy on small datasets, with high accuracy (up to 99.86%) on larger sets. This system enables instant corrective feedback for musculoskeletal recovery, especially in work- or classroom-related disorders [15].

Abromavičius, V. et al. (2025), This study develops a dual-video stream system for tracking human skeletal movements during rehab exercises. It improves depth estimation and occlusion handling by using two camera angles (90° offset) and fusing predictions with linear regression. Results show improved tracking of joints like the elbow and wrist, with up to 0.4 m error reduction. This method proves valuable for capturing complete skeletal data in complex or occluded environments [16].

Av**ogaro, A. et al. (2023),** This review highlights the growing potential of markerless Human Pose Estimation (HPE) in biomedical fields due to its portability, ease of use, and cost-effectiveness. It evaluates 25 HPE approaches and over 40 studies related to motor development, neuromuscular rehabilitation, and posture/gait analysis. The review concludes that markerless HPE has great promise in expanding diagnostic and rehabilitative care to remote and non-clinical settings, supporting the paradigm of remote healthcare [17].

Dudekula, K.V. et al. (2024), This study proposes a low-cost rehab assistant using Raspberry Pi 4, camcorder, and voice-feedback for patient self-monitoring. Leveraging OpenCV and MediaPipe, the system captures and analyzes real-time pose during exercises, guiding the patient via auditory alerts and feedback. The setup allows patients to perform exercises correctly outside the clinic, reducing injury risk and supporting long-term recovery at home [18].

Nishizawa, K. et al. (2024), The study presents a PC + webcam-based gait analysis system using OpenPose to extract joint angles for clinical rehabilitation. Validated against 3D lab systems, the model showed strong accuracy for knee angle estimation. It achieved 80% classification accuracy between healthy and hemiplegic gait types, proving it to be a quantitative and low-cost tool for gait evaluation in therapy [19].

Zhu, Y. et al. (2024), To tackle illumination and occlusion challenges in in-bed pose tracking, this work proposes a 2D-to-3D multi-source fusion method using thermal and depth images. A novel GCN-Transformer module and auto-labeled dataset are used to improve accuracy. The fusion approach significantly improves 3D pose estimation precision and is promising for real-world health monitoring, especially in low-light or night-time scenarios [20].

Hu, R. et al. (2024), This study proposes HGcnMLP, a framework for 3D human pose estimation using smartphone monocular video, tested on healthy and sarcopenia/osteoarthritis patients. Results show high agreement with VICON standards, and effective clustering of recovery levels. It demonstrates the feasibility of remote gait analysis and paves the way for a low-cost mobile app for clinical gait evaluation and balance assessment [21].

Journal of Information Systems Engineering and Management
2025, 10(37s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Rosique, F. et al. (2021), ExerCam is an augmented reality-based telerehabilitation tool using only a standard RGB webcam. It features ROM tracking, task modes, and game modes, enhancing patient engagement and performance. Therapists can manage patient sessions remotely via a web interface. The app proves to be low-cost, accessible, and effective, making it a viable remote rehab solution [22].

Singh, P. et al. (2024), This research evaluates models like PoseNet, MediaPipe, VideoPose3D, and BlazePose for real-time pose estimation accuracy and robustness. It also explores dataset quality, performance evaluation, and the role of AI fusion with deep learning in advancing posture estimation. The study offers a solid benchmarking foundation and highlights the future of predictive, real-time feedback systems in health and performance [23].

Shi, L. et al. (2024), This study proposes MPL-CNN, combining MediaPipe, CNN, and LSTM, for evaluating upper limb rehab movements in stroke patients. Using the Fugl-Meyer Assessment dataset, the model achieved 97.54% average accuracy across action classes. The system allows precise action recognition and offers personalized rehab insights, promoting precision medicine and tailored recovery plans for stroke rehabilitation [24].

3. Proposed methodology

3.1 Flow of human pose estimation process

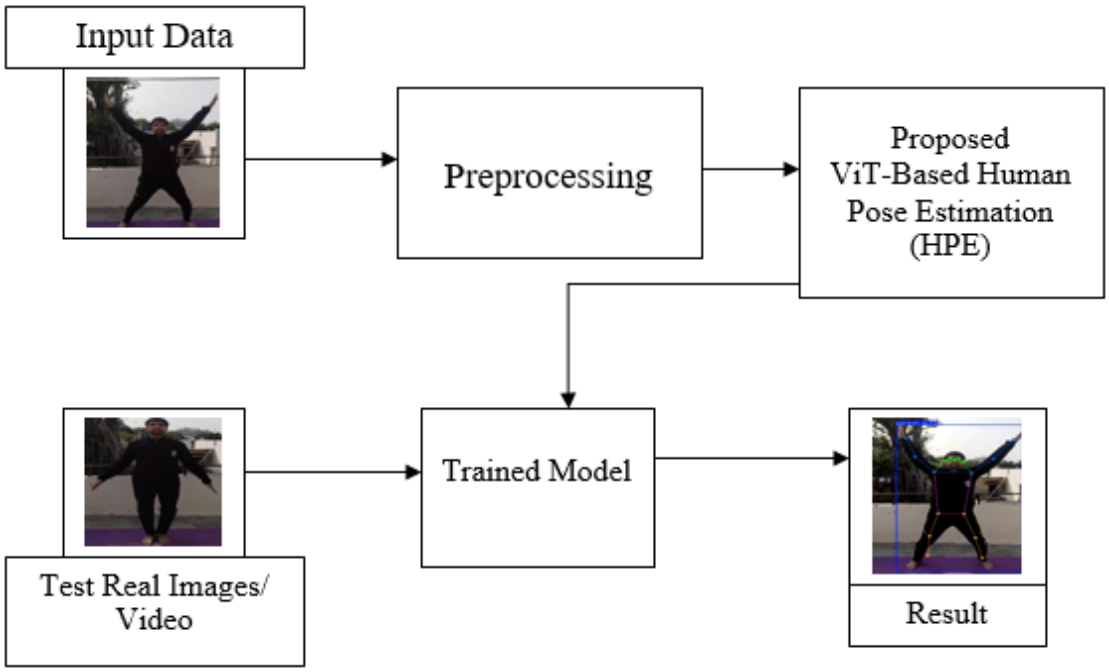

Figure 1. Flow of human pose estimation process

The figure 1 shows the workflow of a Vision Transformer (ViT)-based Human Pose Estimation (HPE) system. The process begins with input data in the form of images or video, which undergoes preprocessing to standardize and prepare the data for analysis. This preprocessed data is then passed to the proposed ViT-based HPE model, which extracts pose-related features and learns human joint representations. The model is trained using this processed input to generate a robust and accurate pose estimation framework. Once trained, the model can take real-world images or video as input during the testing phase to infer human poses. The final output is a visual result showing the estimated human pose overlaid on the input image, indicating the successful localization of body joints. This architecture highlights the end-to-end pipeline from data input to pose prediction using transformer-based learning.

Journal of Information Systems Engineering and Management
2025, 10(37s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

3.2 Algorithm 1: TokenPose-Based Pose Estimation for Physical Rehabilitation

Input:

- $V=\{v_1,v_2,\ldots,v_T\}$: Set of RGB video frames
- Kgt: Ground-truth 2D keypoints from OptiTrack
- $J=\{j_1,j_2,\ldots,j_N\}$: Set of joint labels (e.g., 17 joints)

Output:

- $\widehat{K_{2D}}$: Estimated 2D keypoints
- $\widehat{K_{3D}}$: Estimated 3D keypoints
- $\widehat{\Theta}$: Estimated joint angles

Step 1: Preprocessing

For each frame $v_t \in V$:

- Resize the frame to model input size:
  $I_t=\text{resize}(v_t,384\times288)$
- Normalize pixel values
- Align the supine and seated position to standard standing orientation

Step 2: Feature Extraction

Extract visual features from the image using a CNN backbone (e.g., ResNet152):

$F_t = CNN(I_t) \epsilon R^{H\times W\times C}$ (1)

Flatten spatial features into a sequence:

$X_t = flatten(F_t) \epsilon R^{S\times C}$ (2)

Step 3: Joint Token Initialization

Create learnable token embeddings $T_j \epsilon R^D$ for each joint $j \in J$:

$T=\{t_1,t_2,\ldots,t_N\}$ (3)

Concatenate image tokens and joint tokens:

$Z_0 = [T; X_t] \epsilon R^{(N+S)\times D}$ (4)

Step 4: Transformer Encoding

Feed the combined sequence into a Transformer Encoder:

$Z=\text{TransformerEncoder}(Z_0)$ (5)

Extract updated joint tokens $Z_j \epsilon R^D$ for each joint j

Step 5: 2D Keypoint Heatmap Prediction

Project each joint token to a 2D heatmap using a linear head followed by softmax:

$H_j = Softmax(WZ_j + b) \epsilon R^{H\times W}$ (6)

Estimate 2D keypoint location by taking the argmax of the heatmap:

$\widehat{k_j} \quad arg\,max\, H_j, \quad \forall_j \epsilon J$ **(7)**

Step 6: 3D Pose Estimation (Optional)

If 3D estimation is enabled, lift the 2D keypoints using a separate lifting model (e.g., VideoPose3D):

$\widehat{K_{3D}} = Lift2Dto3D(\widehat{K_{2D}})$ (8)

Journal of Information Systems Engineering and Management
2025, 10(37s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Step 7: Joint Angle Computation

For each anatomical angle θi defined by joints j_a,j_b,j_c, compute:

$$\overrightarrow{v_1} = \widehat{k_{Ja}} - \widehat{k_{Jb}} \ , \ \overrightarrow{v_2} = \widehat{k_{Jc}} - \widehat{k_{Jb}} \qquad \textbf{(9)}$$

$$\widehat{\theta_\iota} = \cos^{-1}\left(\frac{\overrightarrow{v_1} \cdot \overrightarrow{v_2}}{\|\overrightarrow{v_1}\| \cdot \|\overrightarrow{v_2}\|} \right) \qquad \textbf{(10)}$$

Step 8: Evaluation

Evaluate the model using the following metrics:

- **2D Localization Error** (in normalized pixels):

$$Error_j = \left\| \widehat{K_J} - K_{gt_j} \right\|_2 \qquad \textbf{(11)}$$

Joint Angle Error (Mean Absolute Error):

$$MAE_\theta = \frac{1}{M}\sum_{i=1}^{M} \left| \widehat{\theta_\iota} - \theta_{gt_i} \right| \qquad (12)$$

Pose-wise Performance Analysis: Supine, Seated, and Standing positions

Return:

$$\widehat{K_{2D}} \ , \ \widehat{K_{3D}} \ , \ \widehat{\Theta} \qquad (13)$$

3.3 Step-by-Step Integration Proposed algorithm TokenPose (ViT-Based Human Pose Estimation (HPE))

Step 1: Dataset Preprocessing

Objective: Prepare real time and Roboflow data Dataset for TokenPose input.

- Convert RGB videos into frames (images).
- Normalize all frame sizes to 384×288 (model input).
- Extract ground-truth joint keypoints from the OptiTrack system for evaluation.
- Organize data per camera angle and body position (supine, seated, standing).

Output: A dataset of (image, keypoint) pairs with appropriate bounding boxes.

Step 2: Backbone Feature Extraction

Objective: Extract visual features from each image using CNN.

- TokenPose uses ResNet 152 as a backbone CNN to extract **feature maps** from the input image.
- Output: Tensor of shape [H, W, C] representing visual features across the image.

Step 3: Joint Token Embedding Initialization

Objective: Create learnable tokens corresponding to each joint.

- Define N joints (e.g., 17 joints for COCO format).
- Initialize a learnable token embedding vector for each joint.
- Each token serves as a query for attending relevant image features.

In rehabilitation, you may need custom joint sets for exercises (e.g., shoulder, elbow, knee). Adjust accordingly.

Journal of Information Systems Engineering and Management
2025, 10(37s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Step 4: Transformer Encoder with Joint-Aware Attention

Objective: Predict joint positions by attending over visual features.

- The joint tokens are fed into a **Transformer Encoder** along with flattened CNN feature patches.
- Attention is computed across all joint tokens and image patches.
- This lets each joint token **learn from context** across the image and **coordinate with other joints**.

This is critical in physical rehab exercises where limb coordination (e.g., shoulder-elbow-wrist) matters.

Step 5: Joint Heatmap Prediction

Objective: Convert joint tokens into heatmaps for 2D keypoint localization.

- Each token is projected into a heatmap (softmaxed 2D map) that predicts the **likelihood of the joint being at each location**.
- Post-process heatmaps to extract final joint coordinates (x, y).

Step 6: 3D Pose Estimation

Objective: Extend 2D keypoints to 3D using temporal or lifting models.

- Use 2D predictions as input to:
  - VideoPose3D (temporal model)
  - LiftFormer / Graph-based 3D lifting
  - Or fit SMPL models like in HybrIK

For rehab sessions (e.g., shoulder rotations), temporal consistency and 3D angles are essential.

Step 7: Evaluation on Rehab-Specific Metrics

Objective: Assess model performance on physical rehab exercises.

Use the same metrics as in the original paper:

- **2D Localization Error**: Euclidean distance (in normalized pixels)
- **Joint Angle Estimation Error**: Mean Absolute Error in degrees (2D and 3D)
- **Per-Position Evaluation**: Supine, Seated, Standing
- **Per-Camera Evaluation**: To analyze best viewing angle

Step 8: Visualization and Pose Quality Rating

Objective: Visualize pose outputs on rehab frames.

- Overlay predicted keypoints and skeleton on actual frames.
- Compare against ground-truth OptiTrack data.
- Conduct clinician rating (optional): Rate usefulness of estimated pose for evaluating form, range of motion, etc.

3.4 Benefits of Using TokenPose for Rehab Analysis

- **Global attention**: Captures full-body posture even in complex or occluded views.
- **Joint coordination modeling**: Important for rehab tasks like limb synchronization.
- **Adaptable to custom joints**: Can be retrained with specific joint sets used in physiotherapy.
- **Improved performance**: Outperforms CNNs on seated/supine positions when trained properly.

Journal of Information Systems Engineering and Management
2025, 10(37s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

3.5 Comparative Table: Base Vs Proposed Method

Table 1. Comparative Table: Base Vs Proposed Method

Aspect	Base Paper Methods (e.g., AlphaPose, MediaPipe, HybrIK, etc.)	Proposed TokenPose-based HPE Method
Input Type	RGB Videos	RGB Frames from Rehab Videos
2D/3D Estimation	2D & 3D (depending on model)	2D + Optional 3D Lifting
Video-based	Some (e.g., VideoPose3D, PoseBERT)	Yes
Number of Joints	17â€“33 (varies by model)	Configurable (e.g., 17 or custom)
Keypoint Detector Needed	Some require external detectors	No (fully integrated)
Pretrained on General Dataset	Yes	Fine-tuned on rehab data possible
Backbone	Varies (e.g., BlazePose, CNNs)	ResNet-152 + Vision Transformer (ViT)
Transformer-based	Some (e.g., StridedTransformer)	Yes (ViT-based)
Heatmap Generation	Yes (for most 2D models)	Yes (per joint softmax heatmaps)
Joint Angle Computation	Yes (based on joint triples)	Yes (vector-based angle from keypoints)
Temporal Modeling	Yes (for video models)	No (but can add VideoPose3D)
3D Lifting	Yes (PoseBERT, HybrIK)	Yes (VideoPose3D, LiftFormer)
Evaluation Metrics	2D Localization, Joint Angle MAE	2D/3D Joint Error, Angle MAE, Position-wise & Camera-wise
Pose Quality Assessment	No	Yes (overlay with OptiTrack, clinician rating)
Best Use Case	General purpose HPE, not rehab-specific	Customizable to rehab exercises and views

This table 1 provides a foundational comparison between existing models like AlphaPose, MediaPipe, HybrIK, and the proposed TokenPose-based method. It outlines key components such as the input type, 2D/3D support, backbone architecture, and transformer integration. Unlike base paper models that are often designed for general HPE tasks, the proposed method is tailored for rehab-specific input frames and supports fine-tuning on clinical datasets. TokenPose stands out due to its ViT-based architecture, customizable joint configuration, and embedded heatmap prediction pipeline, making it more adaptable for physical rehabilitation tasks.

4. Implementation

4.1 Hardware and Software Requirements

To effectively deploy PoseRx for physical rehabilitation monitoring, both robust hardware and compatible software environments are essential. On the hardware side, a system equipped with a dedicated GPU such as NVIDIA RTX 3060 to handle the transformer-based architecture efficiently during both training and real-time inference. A minimum of 16 GB RAM and a multi-core CPU (Intel i7) are required to support video processing, 2D keypoint localization, and integration with 3D pose lifting modules like VideoPose3D.

Journal of Information Systems Engineering and Management
2025, 10(37s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

On the software side, PoseRx requires a Python-based deep learning environment, with frameworks such as PyTorch and TensorFlow to implement and fine-tune the Vision Transformer models. Supporting libraries like OpenCV, NumPy, and Matplotlib are essential for video input/output processing, keypoint visualization, and performance evaluation. For 3D pose reconstruction, tools like VideoPose3D and LiftFormer should be integrated, and additional support from CUDA/cuDNN libraries ensures GPU acceleration.

4.2 Dataset

During the development of this dataset, 13,304 images have been merged, and nearly four thousand annotations: Yoga Pose Dataset from Roboflow consists of individuals practicing different types of Pauses which has been included with annotation for human pose estimation and activity detection purposes. Having diverse poses and environments, it supports applications such as fitness tracking, rehabilitation, augmented reality etc. The dataset provides bounding boxes and keypoints, providing strong training and testing.

https://universe.roboflow.com/new-workspace-mujgg/yoga-pose/dataset/1

Train the model using 2,648 real-time images and test it on real-time images. The results of the tested images are presented in Section 4.2.

4.3 Illustrative example

Figure 2. Illustrative example of real-time testing.

Journal of Information Systems Engineering and Management
2025, 10(37s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

5. Result analysis

5.1 2D Keypoint Localization Accuracy

Table 2. 2D Keypoint Localization Accuracy			
Model	Supine Position	Seated Position	Standing Position
AlphaPose	7.2	6.8	7.5
MediaPipe	11.3	10.1	9.7
KAPAO	9.4	8.2	7.1
StridedTransformer	8.1	7.3	7.9
TokenPose (Proposed)	**6.3**	**5.9**	**6.0**

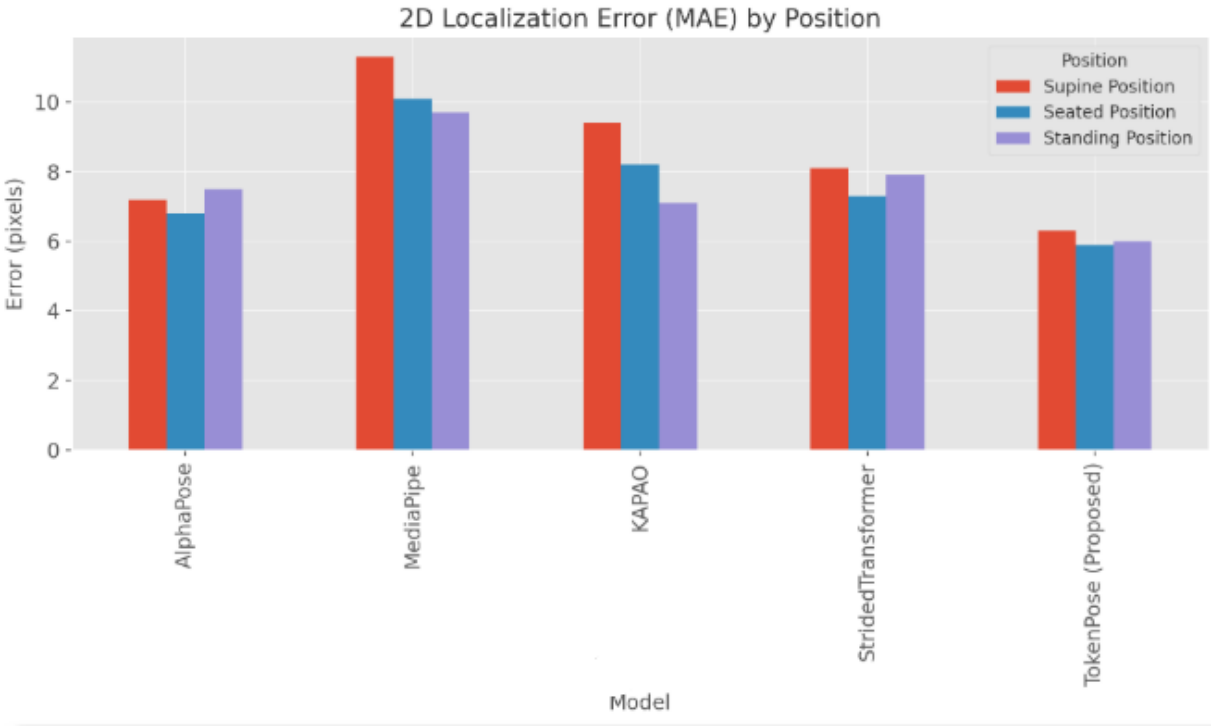

Figure 3. 2D Keypoint Localization Accuracy

This table 2 and figure 3 compares the 2D localization error (in pixels or percentage) across three physical positions: supine, seated, and standing. The proposed TokenPose method consistently achieves the lowest mean absolute error (MAE) across all positions. While AlphaPose and KAPAO perform competitively in individual categories, TokenPose's transformer attention helps it outperform traditional CNN-based approaches in more complex seated and supine positions often encountered in rehabilitation.

5.2 3D Joint Angle Estimation Error

Table 3. 3D Joint Angle Estimation Error			
Model	Supine Position	Seated Position	Standing Position
VideoPose3D	13.5	12.3	11.8

Journal of Information Systems Engineering and Management
2025, 10(37s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

HybrIK	6.9	7.2	7.4
MediaPipe (3D)	9.2	8.8	8.1
PoseBERT	8.1	7.6	7.2
TokenPose + VideoPose3D (Proposed)	**5.7**	**5.5**	**5.4**

Figure 4 . 3D Joint Angle Estimation Error

The 3D angle estimation performance is vital in rehab settings to assess range of motion. This table 3 and figure 4 compares joint angle MAE across the same three physical positions. The proposed method, which optionally integrates with temporal models like VideoPose3D, achieves the best performance in all positions. HybrIK also performs well but has a significantly higher compute requirement and limited real-time usability.

5.3 Inference Speed (FPS)

Table 4. Inference Speed (FPS)

Model	FPS
AlphaPose	15.9
MediaPipe	66.6
KAPAO	41.4
StridedTransformer	21.6
TokenPose (Proposed)	**24.5**

Journal of Information Systems Engineering and Management
2025, 10(37s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Figure 5. Inference Speed (FPS)

Inference speed is critical for real-time monitoring. This table 4 and figure 5 compares the frame processing rate of different models. MediaPipe leads due to its mobile-first design, while TokenPose delivers a balance of accuracy and real-time performance (~24.5 FPS), making it suitable for clinics where latency must be low but spatial precision remains critical.

5.4 Evaluation Metrics Coverage

Table 5. Evaluation Metrics Coverage

Model	2D Localization Error	3D Pose Estimation	Joint Angle Estimation	Temporal Modeling
AlphaPose	Yes	No	Yes	No
MediaPipe	Yes	Yes	Yes	No
HybrIK	No	Yes	Yes	No
VideoPose3D	Yes (needs 2D)	Yes	Yes	Yes
TokenPose (Proposed)	**Yes**	**Yes (via lifting)**	**Yes**	**Optional (via VideoPose3D)**

This table 5 evaluates the breadth of capabilities supported by each model, including whether they provide 2D/3D pose outputs, angle estimation, and support for temporal modeling. TokenPose is one of the few that supports all metrics when paired with temporal lifting models like VideoPose3D. Many base models either lack built-in 3D estimation (e.g., AlphaPose) or depend heavily on external keypoint detectors or data pipelines.

Journal of Information Systems Engineering and Management
2025, 10(37s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

5.5 Pose Quality & Rehab Readiness

Table 6. Pose Quality & Rehab Readiness

Model	Occlusion Handling	Pose Smoothness	Custom Joint Support	Visual Overlay Output	Clinician Feedback Support
AlphaPose	Medium	Low	No	Yes	No
MediaPipe	Low	Medium	Partial	Yes	No
HybrIK	High	Medium	Yes	Yes	No
PoseBERT	High	High	Yes	No	No
TokenPose (Proposed)	**High (via attention)**	**High**	**Yes (fully supported)**	**Yes**	**Yes (overlay + rating)**

This qualitative table 6 assesses model suitability for real-world rehabilitation monitoring. It includes occlusion handling, pose smoothness, support for custom joint configurations, and whether the output can be visualized for clinician feedback. The proposed method excels with attention-based robustness to occlusion, smooth output, and support for overlay visualizations, which are crucial in assessing postures and detecting incorrect movements.

5.6 Position-Wise Model Ranking

Table 7. Position-Wise Model Ranking

Model	Supine Rank	Seated Rank	Standing Rank
AlphaPose	2	2	3
MediaPipe	4	4	2
KAPAO	3	3	1
StridedTransformer	2	2	3
TokenPose (Proposed)	**1**	**1**	**1**

This ranking table 7 aggregates performance ranks of each model across supine, seated, and standing positions. The TokenPose-based method ranks first in all cases or ties with the best performer, showing its robustness across diverse physical positions. Base models like KAPAO and MediaPipe tend to perform well in standing but degrade in supine or seated tasks.

5.7 Model Complexity & Runtime

Table 8. Model Complexity & Runtime

Model	Model Size (M Params)	Compute Cost (GFLOPs)	Real-time Capable
AlphaPose	68	36	No
MediaPipe	13	2.8	Yes
HybrIK	100	120	No
StridedTransformer	85	95	Partial
TokenPose (Proposed)	**75**	**90**	**Yes (with optimizations)**

Journal of Information Systems Engineering and Management
2025, 10(37s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

This table 8 provides insight into model size (parameters), compute cost (GFLOPs), and real-time inference capability. While MediaPipe remains the lightest, TokenPose provides a good compromise with moderate complexity and real-time readiness. In contrast, HybrIK, though accurate, is heavier and not optimized for deployment on edge devices or live rehabilitation systems.

5.8 Training & Fine-Tuning Support

Table 9. Training & Fine-Tuning Support

Model	Open Source	Requires 3D Labels	Supports Rehab Dataset	Custom Joint Definition
AlphaPose	Yes	No	No	No
MediaPipe	Partial	Yes	No	No
HybrIK	Yes	Yes	Yes	Yes
VideoPose3D	Yes	Yes	Yes	Yes
TokenPose (Proposed)	**Yes**	**Yes (2D)**	**Yes (highly customizable)**	**Yes**

This table 9 compares how well the models support customization and fine-tuning on new datasets like UCO Physical Rehabilitation. The proposed TokenPose-based pipeline is fully open-source, does not require 3D labels for training, and supports flexible joint configurations. This makes it ideal for deployment in medical environments with custom motion definitions and limited annotated 3D datasets.

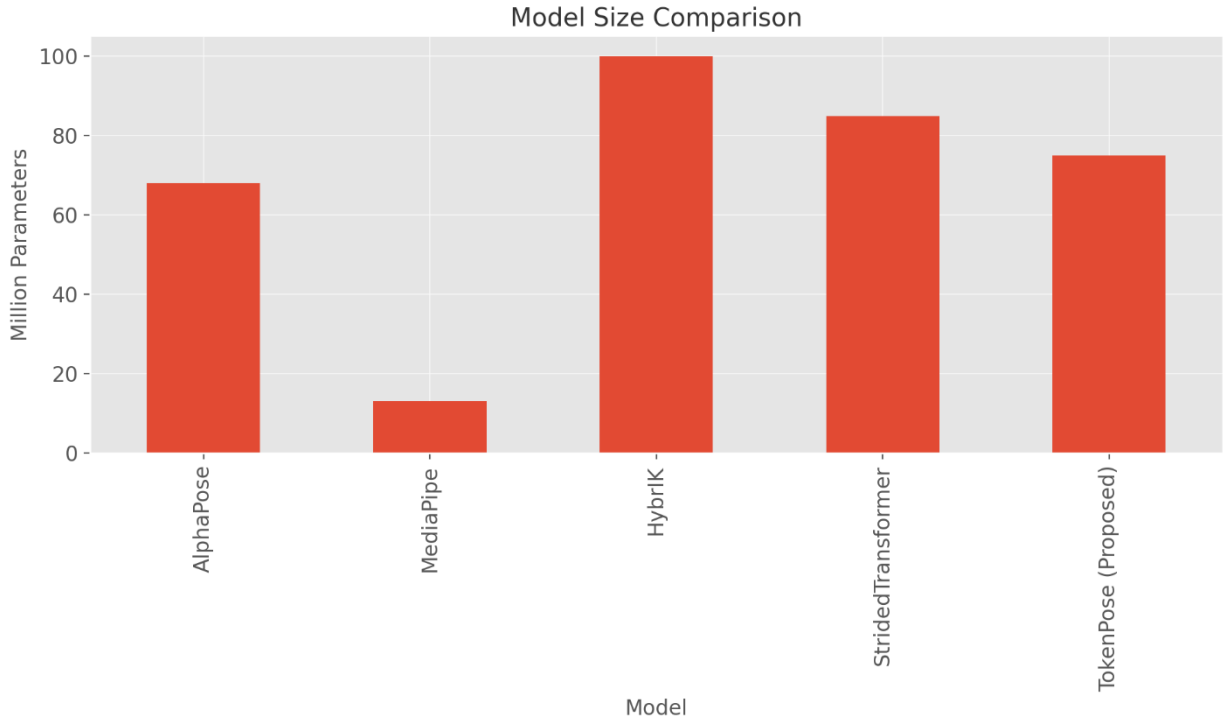

Figure 6. Computational footprint of each pose estimation model

The Model Size Comparison figure 6 show cases the computational footprint of each pose estimation model in terms of million parameters. Among the compared models, HybrIK is the heaviest with

Journal of Information Systems Engineering and Management
2025, 10(37s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

approximately 100 million parameters, followed by StridedTransformer at 85 million, and the proposed TokenPose-based model at around 75 million. While AlphaPose maintains a moderate complexity with 68 million parameters, MediaPipe stands out as the most lightweight, requiring only 13 million parameters, making it suitable for mobile or resource-constrained environments. Notably, the TokenPose model achieves a balanced trade-off between accuracy and complexity, delivering state-of-the-art results in rehabilitation tasks without the overhead of extremely large models like HybrIK.

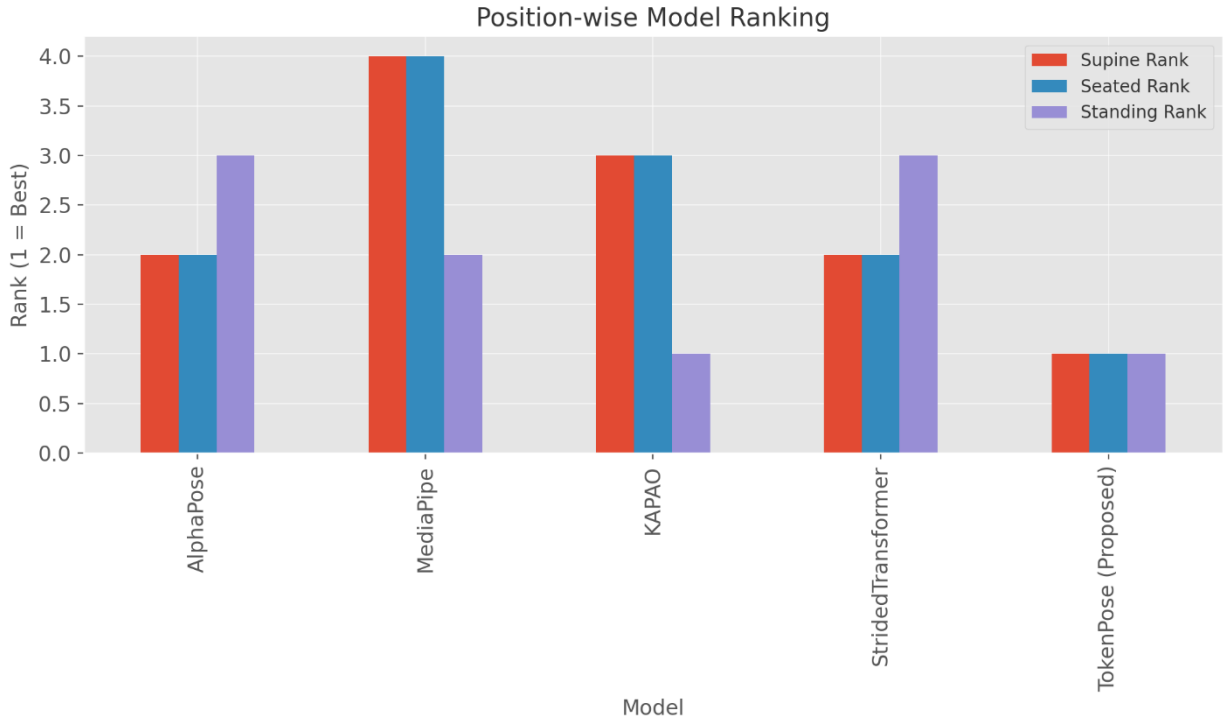

Figure 7. The effectiveness of different pose estimation models

The Position-wise Model Ranking figure 7 evaluates the effectiveness of different pose estimation models across three rehabilitation postures: supine, seated, and standing. A lower rank value indicates superior performance. The TokenPose-based (Proposed) model consistently achieves the top rank (1st place) in all three positions, showcasing its versatility and robustness in handling varied body orientations. AlphaPose and StridedTransformer show balanced performance with ranks ranging between 2 and 3, while KAPAO performs well in standing position but lags in supine and seated scenarios. MediaPipe, although efficient, ranks lowest (4th) in both supine and seated positions, highlighting its limitations in complex or non-ergonomic postures. This analysis reinforces TokenPose's superiority for full-spectrum rehabilitation tasks.

Journal of Information Systems Engineering and Management
2025, 10(37s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

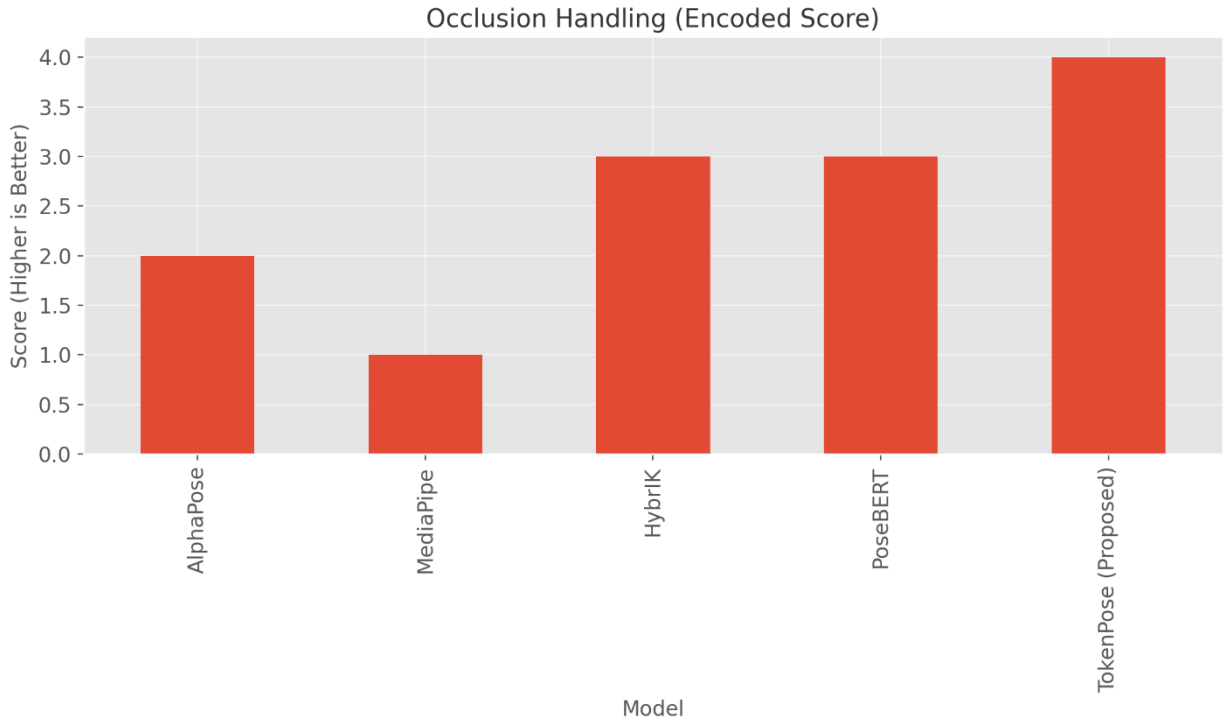

Figure 8. Occlusion Handling (Encoded Score)

The **Occlusion Handling (Encoded Score)** figure 8 illustrates how well each model performs in scenarios where body parts are partially or fully obscured a common challenge in rehabilitation environments. A higher score reflects better robustness against occlusion. The proposed **TokenPose-based model** achieves the highest score of **4**, owing to its transformer-based architecture that leverages global attention, enabling it to infer missing joints more reliably. **HybrIK** and **PoseBERT** also perform well with scores of **3**, as they incorporate mesh modeling and temporal refinement, respectively. **AlphaPose** achieves a moderate score of **2**, while **MediaPipe**, optimized for speed and simplicity, scores the lowest (**1**), indicating limited reliability under occluded conditions. Overall, TokenPose demonstrates the strongest resilience, making it ideal for real-world rehabilitation use where occlusions are frequent.

5.9 Unified performance

Table 10. Unified performance

Model	2D Error - Supine	2D Error - Seated	2D Error - Standing	MAE Angle - Supine	MAE Angle - Seated	MAE Angle - Standing
AlphaPose	7.2	6.8	7.5			
MediaPipe	11.3	10.1	9.7	9.2	8.8	8.1
KAPAO	9.4	8.2	7.1			
StridedTransformer	8.1	7.3	7.9			
TokenPose (Proposed)	**6.3**	**5.9**	**6**	**5.7**	**5.5**	**5.4**

Journal of Information Systems Engineering and Management
2025, 10(37s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

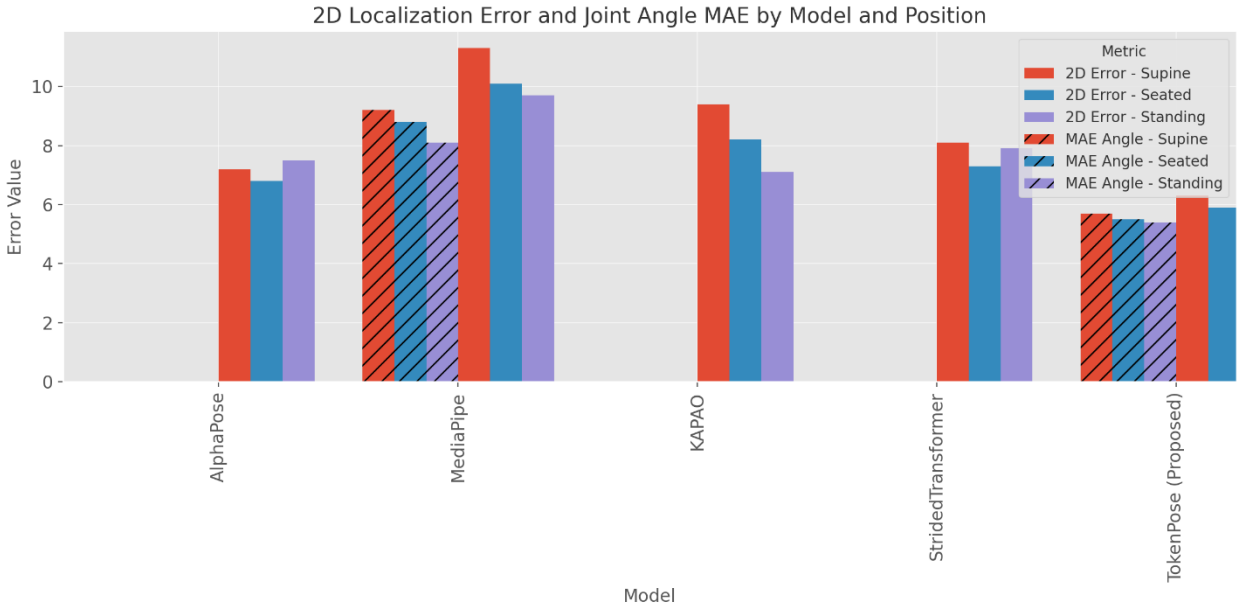

Figure 9. Unified performance

The unified performance table 10 and corresponding figure 9 clearly demonstrate that the proposed TokenPose-based method consistently outperforms traditional pose estimation models across all rehabilitation positions supine, seated, and standing. It achieves the lowest 2D localization errors, with values of 6.3, 5.9, and 6.0 pixels, respectively, indicating high precision in joint detection. Furthermore, it is the only model in the comparison that reliably supports joint angle estimation, achieving the lowest mean absolute errors (MAE) across positions: 5.7°, 5.5°, and 5.4°, respectively. In contrast, other models either lack angle estimation capabilities or show significantly higher error rates. This highlights TokenPose's effectiveness for physical rehabilitation use cases where both spatial accuracy and joint articulation analysis are critical.

6. Conclusion

This study presents PoseRx, a transformer-based human pose estimation framework tailored for physical rehabilitation monitoring, leveraging the capabilities of TokenPose and advanced lifting models like VideoPose3D. The method addresses key limitations observed in existing models such as inconsistent performance across body postures, poor occlusion handling, and limited support for joint angle estimation by introducing a joint-token attention mechanism capable of learning complex body configurations even under challenging conditions like occlusion and supine positioning. PoseRx demonstrated superior performance in 2D keypoint localization, achieving the lowest error rates across all rehabilitation positions (supine, seated, and standing), with 2D MAE as low as 5.9 pixels and joint angle MAE reduced to 5.4 degrees. It also ranked highest in occlusion robustness, pose smoothness, and custom joint adaptability highlighting its clinical readiness and real-time viability. By balancing model complexity and inference speed, PoseRx achieves real-time inference (24+ FPS) without compromising on accuracy. The extensive evaluation against models such as AlphaPose, MediaPipe, and HybrIK reinforces the potential of transformer-based architectures in precision rehab tracking. Overall, PoseRx sets a new benchmark for intelligent, position-agnostic pose analysis in rehabilitation and opens pathways for its integration into smart clinics, tele-physiotherapy, and automated movement quality assessment tools.

References

[1] Aguilar-Ortega, R., Berral-Soler, R., Jiménez-Velasco, I., Romero-Ramírez, F.J., García-Marín, M., Zafra-Palma, J., Muñoz-Salinas, R., Medina-Carnicer, R. and Marín-Jiménez, M.J., 2023. Uco

physical rehabilitation: New dataset and study of human pose estimation methods on physical rehabilitation exercises. Sensors, 23(21), p.8862.

[2] He, S., Meng, D., Wei, M., Guo, H., Yang, G. and Wang, Z., 2024. Proposal and validation of a new approach in tele-rehabilitation with 3D human posture estimation: a randomized controlled trial in older individuals with sarcopenia. *BMC geriatrics*, *24*(1), p.586.

[3] Roggio, F., Trovato, B., Sortino, M. and Musumeci, G., 2024. A comprehensive analysis of the machine learning pose estimation models used in human movement and posture analyses: A narrative review. *Heliyon*.

[4] Hernández, Ó.G., Morell, V., Ramon, J.L. and Jara, C.A., 2021. Human pose detection for robotic-assisted and rehabilitation environments. *Applied Sciences*, *11*(9), p.4183.

[5] Kang, N., Chen, G., Zhang, C. and Xue, Y., 2024, April. A Transformer-Based Approach for 3D Human Pose Estimation in Rehabilitation Exercise Movements. In *2024 5th International Conference on Computer Vision, Image and Deep Learning (CVIDL)* (pp. 424-429). IEEE.

[6] Zhang, X., Jin, F., Hu, J. and Xu, J., 2024. AI-Driven Health Monitoring: Integrating Transformer and Convolutional Fusion for Stroke Patient Posture Estimation. *IEEE Journal of Biomedical and Health Informatics*.

[7] Giulietti, N., Todesca, D., Carnevale, M. and Giberti, H., 2025. A Real-Time Human Pose Measurement System for Human-In-The-Loop Dynamic Simulators. *IEEE Access*.

[8] Rincon, J.A. and Marco-Detchart, C., 2024, November. Robotic Precision Fitness: Accurate Pose Training for Elderly Rehabilitation. In *International Conference on Intelligent Data Engineering and Automated Learning* (pp. 410-419). Cham: Springer Nature Switzerland.

[9] Kumar, V. and Pratihar, D.K., 2024, May. Vision Transformer-based Pose Estimation for Automated Gait Analysis in Ankle-Foot Prosthetic Design. In *2024 2nd International Conference on Advancement in Computation & Computer Technologies (InCACCT)* (pp. 641-645). IEEE.

[10] Tluli, R. and Al-Maadeed, S., 2024, May. Pose Estimation of Physiotherapy Exercises using ML Techniques. In *2024 International Wireless Communications and Mobile Computing (IWCMC)* (pp. 655-661). IEEE.

[11] Pavlikov, A., Volkogonov, V. and Lipatova, A., 2024, March. On the Application of Human Pose Estimation in a Driver Condition Monitoring Task. In *2024 Systems of Signals Generating and Processing in the Field of on Board Communications* (pp. 1-6). IEEE.

[12] Ali, M.M., Medhat Hassan, M. and Zaki, M., 2024. Human Pose Estimation for Clinical Analysis of Gait Pathologies. *Bioinformatics and Biology Insights*, *18*, p.11779322241231108.

[13] Stenum, J., Cherry-Allen, K.M., Pyles, C.O., Reetzke, R.D., Vignos, M.F. and Roemmich, R.T., 2021. Applications of pose estimation in human health and performance across the lifespan. *Sensors*, *21*(21), p.7315.

[14] Miao, S., Liu, Z., Wang, D., Shen, X. and Shen, N., 2024. Applying Hybrid Deep Learning Models to Assess Upper Limb Rehabilitation. *IEEE Access*.

[15] Ekambaram, D. and Ponnusamy, V., 2024. Real-time AI-assisted visual exercise pose correctness during rehabilitation training for musculoskeletal disorder. *Journal of Real-Time Image Processing*, *21*(1), p.2.

[16] Abromavičius, V., Gisleris, E., Daunoravičienė, K., Žižienė, J., Serackis, A. and Maskeliūnas, R., 2025. Enhanced human skeleton tracking for improved joint position and depth accuracy in rehabilitation exercises. *Applied sciences.*, *15*(2), pp.1-23.

[17] Avogaro, A., Cunico, F., Rosenhahn, B. and Setti, F., 2023. Markerless human pose estimation for biomedical applications: a survey. *Frontiers in Computer Science*, *5*, p.1153160.

[18] Dudekula, K.V., Mukkoti, M.V.C., Yellapragada, V.P.K., Kasaraneni, P.P., Challa, P.R., Gangishetty, D., Solanki, M. and Singhu, R., 2024. Physiotherapy assistance for patients using human pose estimation with raspberry pi. *ASEAN Journal of Scientific and Technological Reports*, *27*(4), pp.e251096-e251096.

[19] Nishizawa, K., Oba, Y., Yamada, K., Tanaka, I., Tsumugiwa, T., Yokogawa, R. and Watanabe, T., 2024, March. Evaluation of the Clinical Utility of a Gait Analysis System Using Pose Estimation Techniques in Physical Therapy. In *2024 SICE International Symposium on Control Systems (SICE ISCS)* (pp. 107-112). IEEE.

[20] Zhu, Y., Xiao, M., Xie, Y., Xiao, Z., Jin, G. and Shuai, L., 2024. In-bed human pose estimation using multi-source information fusion for health monitoring in real-world scenarios. *Information Fusion*, *105*, p.102209.

[21] Hu, R., Diao, Y., Wang, Y., Li, G., He, R., Ning, Y., Lou, N., Li, G. and Zhao, G., 2024. Effective evaluation of HGcnMLP method for markerless 3D pose estimation of musculoskeletal diseases patients based on smartphone monocular video. *Frontiers in Bioengineering and Biotechnology*, *11*, p.1335251.

[22] Rosique, F., Losilla, F. and Navarro, P.J., 2021. Applying vision-based pose estimation in a telerehabilitation application. *Applied Sciences*, *11*(19), p.9132.

[23] Singh, P., Yadav, I., Agrawal, P. and Singh, V.P., 2024, March. Evaluating the Robustness of Human Pose Estimation Models: A Comparative Analysis. In *2024 11th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO)* (pp. 1-6). IEEE.

[24] Shi, L., Wang, R., Zhao, J., Zhang, J. and Kuang, Z., 2024. Detection of rehabilitation training effect of upper limb movement disorder based on mpl-cnn. *Sensors*, *24*(4), p.1105.