Accuracy and Reliability of Monocular 3D Pose Estimation for Lower-Limb Movement Analysis

Guangjun Wang[1]*, Jiaojiao Chen[2], Mengxia Hu[3], Cezhong Luo[4], Kaibo Li[5], Xiaohui Chen[6]
School of Artificial Intelligence and Computer Science, Anqing Normal University, Anqing, Anhui 246133, China[1, 2, 3, 4, 5]
School of Economics, Hefei University of Economics, Hefei, Anhui 230000, China[3]
Department of Electronic Information, Anqing Vocational and Technical College, Anqing, Anhui 246003, China[6]

***Abstract*—Optical motion capture provides accurate joint kinematics, but its cost and setup complexity limit routine use in rehabilitation and home monitoring. Monocular RGB video is a low-cost alternative, yet many 3D pose estimation studies report overall errors on public benchmarks without distinguishing lower-limb movements or individual joints. This study evaluated the spatial and frame-to-frame displacement errors of a two-stage pipeline combining BlazePose 2D detection with a dilated temporal convolutional network for 3D lifting. We constructed a paired dataset of monocular RGB video and optical motion capture from 10 healthy participants performing six standardized lower-limb movements, yielding 450 samples and approximately 202,500 frames. Under the within-participant recording-set evaluation, the fine-tuned pipeline achieved an MPJPE of 32.03 mm, compared with 121.70 mm for direct BlazePose 3D output. Errors varied across movements, from 11.21 mm for quiet standing to 62.83 mm for seated knee extension, with distinct patterns at the hip, knee, and ankle. At a noise standard deviation of 5.4 pixels, independent Gaussian perturbations of 2D keypoints increased MPJPE by approximately 0.9% but increased mean per-joint velocity error (MPJVE), measured here as frame-to-frame displacement error, by about 234%. Selected joint-angle measurements also showed movement-dependent agreement with optical motion capture. These findings provide movement- and joint-specific evidence on position accuracy, temporal sensitivity, and selected kinematic agreement under controlled laboratory conditions.**

***Keywords*—3D pose estimation; monocular RGB; lower-limb analysis; accuracy; noise robustness; temporal error analysis**

I. Introduction

This work was supported by the Key Project of Anhui Provincial Social Science Planning "Five Cultures" Research on "Digital-Intelligent Protection and Dissemination of Huangmei Opera" (AHSKWHZ2025D003); the Open Project of State Key Laboratory of Virtual Reality Technology and Systems, Beihang University (VRLAB2026B04); and the "AI + Education" Course Program of Anhui Provincial Department of Education (2024aijy210).

Quantitative assessment of hip, knee, and ankle motion supports rehabilitation, sports performance assessment, and home monitoring. Although optical motion capture provides accurate measurements, its equipment, space, and marker-placement requirements limit use outside the laboratory. Advances in markerless human pose estimation [1], [2] have made consumer-grade RGB video a lower-cost option for movement analysis, including smartphone-based kinematic reconstruction and related biomechanical applications [3], [4]. Single-camera setups have modest hardware requirements, although depth ambiguity makes their accuracy sensitive to movement configuration and viewing angle.

Previous studies have identified depth ambiguity as a source of monocular reconstruction error and reported variation across movements, joints, and camera orientations [5]–[7]. Such variation is particularly relevant to lower-limb analysis. Because the hip, knee, and ankle form a kinematic chain, distal-joint estimates may be affected by upstream localization errors, occlusion, and 2D keypoint jitter. In addition, low frame-wise position error does not necessarily imply low frame-to-frame displacement error. Overall mean per-joint position error (MPJPE) alone therefore cannot characterize every movement or derived metric.

Research on monocular 3D pose estimation has improved coordinate accuracy on public benchmarks using temporal convolutions [8], Transformers [9], graph convolutions [10], and large-scale motion representation learning [11]. Application-oriented studies have also examined pose estimation for gait [12], clinical movements [5], and sports tasks [6]. In rehabilitation, smartphone-based monocular estimates have been compared with optical motion capture [13], while VideoPose3D-type pipelines have been applied to gait kinematics [14]. These developments motivate task-specific evaluation. The joint effects of movement type, anatomical location, and 2D input perturbations on spatial and temporal metrics nevertheless remain less well characterized under a consistent protocol.

To address these gaps, we constructed a paired dataset of monocular RGB video and optical motion capture covering six standardized lower-limb movements performed by 10 healthy participants. We analyzed the BlazePose-plus-temporal-convolution pipeline across five dimensions: movement-specific error, joint-level error, alignment sensitivity, frame-to-frame displacement error, and sensitivity to synthetic 2D keypoint noise. Under the tested synthetic perturbations, frame-to-frame displacement error increased more sharply than position error, highlighting the value of evaluating temporal measures alongside position accuracy.

Fig. 1 summarizes the acquisition and frame-matching workflow, the 2D-to-3D estimation pipeline, and the multidimensional error analysis.

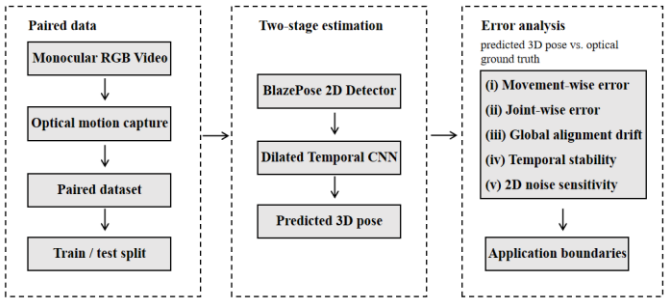

Fig. 1. Overview of the study framework.

Section II reviews relevant methods and applications. Section III describes the dataset and evaluation procedure, followed by the results in Section IV and their interpretation in Sections V and VI.

II. Related Work

A. Monocular 3D Human Pose Estimation

Monocular 3D human pose can be estimated directly from RGB images or recovered by lifting detected 2D keypoints. Direct-regression methods include coarse-to-fine volumetric prediction [15], structured pose prediction [16], pose grammar networks that encode body structure [17], and compositional pose regression [18]. In 2D-to-3D lifting, keypoint detection is separated from depth recovery, making it possible to use detectors such as OpenPose [19] or lighter real-time alternatives. Martinez et al. [20] showed that a fully connected residual network can provide a competitive 2D-to-3D mapping. Pavllo et al. [8] used dilated temporal convolutions to incorporate pose information from neighboring frames and address ambiguity in single-frame estimates. BlazePose [21] provides efficient on-device 2D keypoint detection and was selected as the front end in this study.

Spatiotemporal modeling has since broadened beyond convolutional networks. PoseFormer [9] captures spatial and temporal dependencies with Transformers, while MixSTE [22] combines spatial and temporal encoding and PoseFormerV2 [23] incorporates frequency-domain information. MotionAGFormer [24] integrates Transformer and graph-convolution components; MotionBERT [11] uses large-scale motion representation pretraining; and PoseMamba [25] explores state-space models for long sequences. Self-supervised and weakly supervised methods can reduce reliance on 3D annotations [26], [27]. Although these approaches have improved benchmark accuracy, reported errors depend on the 2D inputs, training data, and temporal-window settings, making direct comparisons difficult.

B. Validation for Lower-Limb and Rehabilitation Applications

In rehabilitation, gait, and sports analysis, evaluation must account for both 3D coordinate accuracy and movement-dependent errors. OpenCap combines consumer-grade video with musculoskeletal modeling for kinematic estimation [3]. OpenCapBench emphasizes the need to separate global rigid-body error from local joint error [4]. Lima et al. [28] validated OpenCap during squatting, hopping, jumping, and side-stepping and observed differences across movements and joints. Rode et al. [5] compared several monocular pose estimators on a large clinical movement dataset and reported MPJPE values of about 146–249 mm. Studies using AthletePose3D [6] and the sports-motion analysis of Fukushima et al. [7] likewise reported performance differences across movements and camera orientations.

Clinical and rehabilitation studies offer complementary evidence. Hu et al. [13] compared smartphone monocular video with a Vicon system in 15 healthy adults and 12 patients with musculoskeletal disorders and reported clinical validity using intraclass correlation coefficients. Natraj et al. [14] examined VideoPose3D for gait kinematics after spinal cord injury. Dill et al. [29] evaluated monocular and stereo-fusion accuracy for physical exercises using a MediaPipe-based front end, while Maskeliūnas et al. [30] applied 2D-to-3D lifting to tele-rehabilitation. Medrano-Paredes et al. [31] compared monocular estimates with inertial-sensor measurements and found movement- and joint-dependent errors. Building on this evidence, the present study examines joint- and movement-specific accuracy, alignment sensitivity, frame-to-frame displacement error, and synthetic 2D noise within the same paired-data evaluation.

C. Error Metrics and Robustness

MPJPE measures the mean 3D Euclidean distance between predicted and reference joint positions. Procrustes-aligned MPJPE (P-MPJPE) is computed after independently aligning each frame for translation, rotation, and scale. Together, these metrics quantify the reduction associated with alignment but do not identify a unique physical cause of error. Position error alone cannot describe temporal behavior: two sequences may have similar MPJPE but different frame-to-frame fluctuations. We therefore also evaluated frame-to-frame displacement error using MPJVE.

Previous robustness studies have linked 2D keypoint quality to 3D reconstruction accuracy. Hoang et al. [32] examined the effects of occlusion, motion blur, and pixel noise on 2D-to-3D lifting. Other approaches address robustness through occlusion-aware modeling [33], diffusion models [34], and graph-attention spatiotemporal networks [35]. Position accuracy is often emphasized, while the effect of small 2D perturbations on frame-to-frame displacement error receives less attention. We examine this effect using controlled synthetic keypoint perturbations.

III. Dataset, Evaluation Pipeline, and Metrics

A. Synchronized Paired Dataset

Six standardized movements were included: quiet standing, stepping in place, squatting, front straight-leg raise, side straight-leg raise, and seated knee extension. Together, they represent static weight bearing, hip flexion and abduction, knee flexion–extension, and dynamic weight shifting. Participants stood quietly for 10 s and performed 10 alternating steps during stepping in place. Each squatting set comprised five repetitions with a brief pause at a comfortable depth. Each of the three unilateral movements was performed separately on the left and right sides, with five repetitions per side per set.

The monocular RGB and optical motion-capture systems operated during the same recording sessions. The RGB camera was positioned approximately 3 m in front of each participant and recorded 1920 × 1080 video at 30 Hz. The optical motion-capture system used 8 infrared cameras operating at 100 Hz. The monocular camera was calibrated using a planar checkerboard [36], yielding a mean reprojection error of 0.1437 pixels. Ten healthy participants (4 female, 6 male; age 20–25 years) with no history of lower-limb injury were recruited. Ethics approval and written informed consent were obtained. The reprojection error characterizes intrinsic camera calibration and is distinct from cross-system extrinsic registration accuracy.

Motion-capture trajectories were extracted from FBX files and converted to consistent units. Stored camera orientation and translation parameters were used to transform the 3D reference poses into camera coordinates, followed by pelvis-root centering. Video and motion-capture frames were paired using their nominal sampling-rate ratio:

$$t_m = min\{max(round[(100/30) \cdot t_v], 0), T_m - 1\} \quad (1)$$

where, t_v is the video frame index, t_m is the corresponding motion capture frame index, and T_m is the number of motion capture frames. No independent measurement of the inter-system timing offset was available.

BlazePose's 33 keypoints were mapped to the 17-joint Human3.6M representation. Joint indices 1 and 4 represent the right and left hips; 2 and 5 the knees; and 3 and 6 the ankles. After interpolation of anomalous frames, trajectory smoothing, and coordinate normalization, the 2D keypoints were paired with the 3D reference positions. Quiet standing, stepping, and squatting contributed one sample per set, whereas each side of the three unilateral movements contributed a separate sample. This yielded 450 samples and approximately 202,500 frames. Each sample averaged approximately 15 s, equivalent to about 450 frames at 30 Hz. The common 17-joint representation enables paired evaluation, although anatomical landmark correspondence may influence joint-level errors.

Fig. 2 illustrates the paired acquisition and frame-matching workflow of monocular RGB video and optical motion capture.

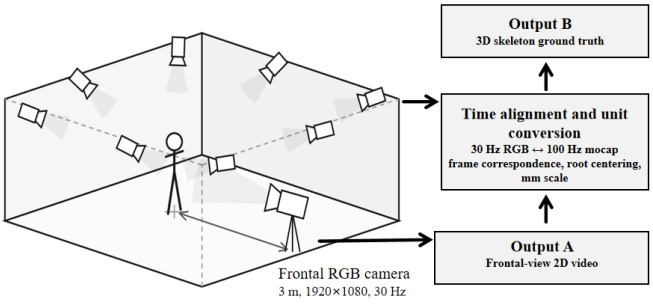

Fig. 2. Synchronized acquisition and pairing of monocular RGB video and optical motion capture.

Each participant completed 5 recording sets; the first 4 were used for training (360 samples) and the fifth for testing (90 samples). All participants and movement types were represented in the test set, but the same individuals also appeared in the training set. The results therefore describe within-participant, across-recording-set performance rather than generalization to unseen participants (Section V-C).

B. Evaluation Pipeline

The 3D pose estimation pipeline comprises two stages (Fig. 3). First, BlazePose detects 2D keypoints in each frame, which are then mapped to the 17-joint skeleton. Second, a dilated temporal convolutional network uses the 2D keypoint sequence to estimate the center frame's 3D joint coordinates. Dilated convolution expands the temporal receptive field to incorporate neighboring frames:

$$y(t) = \sum_{k=0}^{K-1} w(k) \cdot x(t - dk) \quad (2)$$

where, x denotes the input sequence, w the convolution kernel of length K, and d the dilation factor.

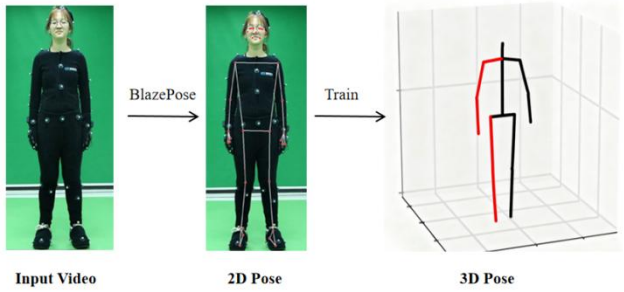

Fig. 3. Illustration of the two-stage 2D-to-3D pose estimation pipeline.

The network receives 243 frames of 17 two-dimensional joint coordinates (243 × 34). It comprises an initial convolutional layer and four residual blocks, with kernel size $W = 3$, $C = 1024$ channels, and dropout $p = 0.25$. The residual blocks use dilation factors of 3, 9, 27, and 81; the output contains the center frame's 17 three-dimensional joint positions (1 × 51). The model was pretrained on Human3.6M [37] and then fine-tuned on our dataset. Fine-tuning used BlazePose 2D coordinates as inputs and optical motion-capture 3D coordinates as labels. Training used the Adam optimizer for 120 epochs, with an initial learning rate $\eta = 0.001$ multiplied by $\alpha = 0.95$ after each epoch. Sequence boundaries were padded to provide the required 243-frame input window. Test-time augmentation averaged original and horizontally flipped predictions after restoring left-right joint correspondence.

C. Evaluation Metrics

Spatial error was measured with MPJPE and P-MPJPE. Let Ji denote the predicted position of the i-th joint, Ji* the ground truth, and P(·) the Procrustes similarity transform independently fitted to all joints in each frame. Then

$$e_{MPJPE} = (1/N) \sum_{i=1}^{N} \|J_i - J_i^*\|_2 \quad (3)$$

$$e_{P-MPJPE} = (1/N) \sum_{i=1}^{N} \|P(J_i) - J_i^*\|_2 \quad (4)$$

To describe the proportion of error that can be removed by Procrustes alignment, we define the alignment gain g:

$$g = (e_{MPJPE} - e_{P-MPJPE})/e_{MPJPE} \times 100\% \quad (5)$$

A larger g indicates a greater relative error reduction after Procrustes alignment; it does not uniquely identify the physical

sources of error. A separate Procrustes similarity transformation was fitted to each frame before the joint-position errors were averaged.

Frame-to-frame displacement error was quantified using mean per-joint velocity error (MPJVE), defined as:

$$e_{MPJVE} = [1/N(T-1)] \sum_{i=1}^{N} \sum_{t=1}^{T-1} \|\Delta J_{i,t} - \Delta J^{*}_{i,t}\|_2 \quad (6)$$

In Eq. (6), $\Delta J_{i,t} = J_{i,t+1} - J_{i,t}$ is the frame-to-frame displacement of joint i, and the asterisk denotes the corresponding ground truth. The first-order frame differences in Eq. (6) are not divided by the time interval Δt; therefore, MPJVE measures frame-to-frame displacement error (mm/frame), not velocity in mm/s or test-retest reliability.

Independent zero-mean Gaussian noise (seed 42) was added to 2D pixel coordinates before normalization, with model weights fixed. At 1080-pixel image height, normalized $\sigma \in \{0, 0.005, 0.010, 0.020, 0.040\}$ corresponds to pixel noise standard deviations of 0, 5.4, 10.8, 21.6, and 43.2 pixels. MPJPE, P-MPJPE, and MPJVE were computed at each noise level. The existing single-frame model was compared with the 243-frame temporal model. Table I averages six movement classes equally; Tables III and IV instead average 11 logged action labels. Table IV uses the second recorded no-noise evaluation for $\sigma = 0$.

IV. Results

A. Overall Performance

The fine-tuned pipeline achieved lower MPJPE than the direct 3D output of BlazePose under the reported evaluation protocol. Table I shows a reduction in mean MPJPE from 121.70 mm to 32.03 mm (73.7%). The comparison reflects the evaluated locally fine-tuned pipeline and direct BlazePose 3D output, which differ in their use of local 3D supervision.

TABLE I. Movement-Wise Position Errors and Alignment Gain

Movement	Base	MPJPE	P-MPJPE	Drop (%)	Align. (%)
Quiet standing	118.53	11.21	9.96	90.5	11.2
Side straight-leg raise	121.84	19.78	15.06	83.8	23.9
Front straight-leg raise	130.08	24.82	21.77	80.9	12.3
Squatting	137.15	35.50	26.96	74.1	24.1
Stepping in place	120.68	38.03	31.15	68.5	18.1
Seated knee extension	101.90	62.83	53.24	38.3	15.3
Mean	121.70	32.03	26.36	73.7	17.7

Note: Base denotes direct BlazePose 3D MPJPE. Position errors are in mm; Drop and Align. are percentages. Align. denotes frame-wise Procrustes alignment gain as defined in Eq. (5). The Mean row assigns equal weight to each of the six movements.

B. Error Distribution Across Movements and Joints

Errors varied substantially across movements, as summarized in Table I. Quiet standing had the lowest MPJPE (11.21 mm). Side and front straight-leg raises yielded 19.78 mm and 24.82 mm, respectively. Squatting and stepping yielded 35.50 mm and 38.03 mm, while seated knee extension had the highest error (62.83 mm). The mean MPJPE and P-MPJPE across the six movements were 32.03 mm and 26.36 mm. The more than 50-mm range between movements highlights the importance of reporting movement-specific results alongside the overall mean.

Quiet standing and straight-leg raises typically preserve a frontal orientation and may involve less depth variation, which could partly explain their lower errors. Squatting and stepping involve greater flexion–extension, weight shifting, and depth changes; both showed higher errors. Seated knee extension changes the support configuration and may increase occlusion around the knee. These movement characteristics provide plausible context for the observed error differences.

The difference between MPJPE and P-MPJPE quantifies the error reduction from frame-wise similarity alignment; it is not a measure of physical camera or body drift. Squatting and side straight-leg raise had alignment gains of 24.1% and 23.9%, respectively, indicating a larger alignment-related reduction without establishing its physical cause. Quiet standing and front straight-leg raise showed smaller alignment gains (11.2% and 12.3%, respectively), indicating less relative error reduction after alignment.

Table II reports hip, knee, and ankle MPJPE for each movement. For several movements, knee and ankle errors were larger than hip errors: 9.62, 13.16, and 19.06 mm for quiet standing; 16.76, 31.54, and 38.59 mm for stepping; and 15.95, 28.51, and 39.16 mm for front straight-leg raise. The hip is closer to the pelvis root, whereas knee and ankle estimates may be more affected by upstream localization error or 2D keypoint jitter. Both are plausible contributors to the observed joint-level pattern.

TABLE II. MPJPE of Key Lower-Limb Joints by Movement (mm)

Movement	Hip	Knee	Ankle	Lower-limb mean
Quiet standing	9.62	13.16	19.06	13.95
Side straight-leg raise	13.20	26.07	23.68	20.98
Front straight-leg raise	15.95	28.51	39.16	27.87
Squatting	25.41	25.04	31.41	27.29
Stepping in place	16.76	31.54	38.59	28.96
Seated knee extension	49.97	79.65	69.39	66.34
Mean	21.82	34.00	36.88	30.90

Note: Values average the left and right joints of each type. Lower-limb mean is the unweighted average of hip, knee, and ankle MPJPE.

Seated knee extension was an exception: knee error (79.65 mm) exceeded ankle error (69.39 mm). This pattern may be associated with knee flexion–extension and changes in joint visibility. During squatting, hip, knee, and ankle errors were 25.41, 25.04, and 31.41 mm. Their comparable magnitudes are consistent with the movement's relatively large alignment-sensitive error component.

C. Temporal Stability and Sensitivity to 2D Noise

Table III reports MPJVE, expressed as frame-to-frame displacement error (mm/frame), by movement. Quiet standing yielded 0.40 mm/frame; front and side straight-leg raises each yielded approximately 2.7 mm/frame; and squatting and stepping each yielded 5.91 mm/frame. For seated knee extension, MPJVE was 3.40 mm/frame on the right and 19.04

mm/frame on the left. Given the limited number of samples per side, this asymmetry is descriptive and should not be interpreted as a consistent left–right effect.

TABLE III. MPJVE BY MOVEMENT (MM/FRAME)

Movement	MPJVE	Movement dynamics
Quiet standing	0.40	Static (no substantial motion)
Side straight-leg raise	2.69	Moderate (unilateral frontal-plane swing)
Front straight-leg raise	2.77	Moderate (unilateral sagittal-plane swing)
Squatting	5.91	High (whole-body cyclic flexion–extension)
Stepping in place	5.91	High (alternating bilateral stepping)
Seated knee extension (right)	3.40	Moderate (unilateral knee extension–flexion)
Seated knee extension (left)	19.04	Moderate (abnormal temporal fluctuation)
Mean	4.65	—

Note: MPJVE is defined in Eq. (6) and reported in mm/frame. The overall value (4.65) is the unweighted mean across 11 logged action labels in the first retained evaluation pass, not the average of the seven displayed movement groups. Left–right differences are descriptive.

Table IV presents the errors at each noise level; Fig. 4 illustrates their trends. Increasing σ from 0 to 0.005 (5.4 pixels) raised MPJPE from 33.5 to 33.8 mm (0.9%) and P-MPJPE from 29.0 to 29.3 mm (approximately 1.0%). Over the same range, MPJVE increased from 4.89 to 16.31 mm/frame (approximately 234%). Thus, under independent synthetic noise, frame-to-frame displacement error changed substantially more than position error. At σ = 0.040, MPJPE reached 44.8 mm and MPJVE reached 20.29 mm/frame. These responses characterize the fixed model under controlled synthetic perturbations. Adjacent-frame differencing inherently amplifies independent frame-wise noise, contributing to the larger MPJVE increase. The σ = 0 baseline reflects the second retained evaluation averaged across 11 action labels, rather than the six movement classes used in Table I.

TABLE IV. CHANGES IN 3D ERROR UNDER DIFFERENT LEVELS OF 2D NOISE

σ/H	σ (px)	MPJPE	P-MPJPE	MPJVE	Pos. Δ%	Disp. Δ%
0.000	0	33.5	29.0	4.89	0.0	0.0
0.005	5.4	33.8	29.3	16.31	+0.9	+234
0.010	10.8	37.9	32.2	18.64	+13.1	+281
0.020	21.6	39.2	34.1	19.15	+17.0	+292
0.040	43.2	44.8	39.4	20.29	+33.7	+315

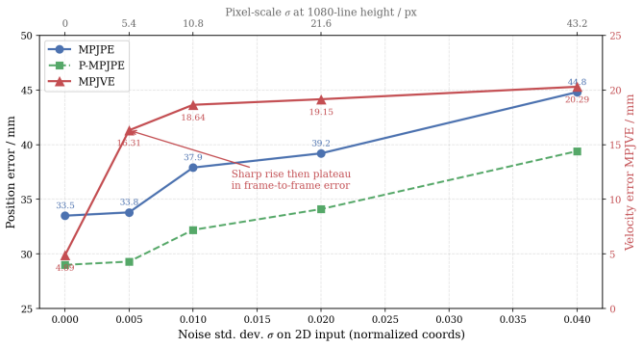

Fig. 4. Changes in 3D error with 2D keypoint noise level.

Table V compares the single-frame baseline with the 243-frame temporal model. The reported MPJPE was 38.87 mm for the single-frame baseline and 32.03 mm for the temporal model (17.6% lower). Corresponding P-MPJPE values were 29.53 mm and 26.36 mm (10.7% lower). The reported difference characterizes the evaluated model configurations rather than a temporal-window-only ablation.

TABLE V. SINGLE-FRAME BASELINE VERSUS THE 243-FRAME TEMPORAL MODEL (MM)

Method	Temporal window	MPJPE	P-MPJPE
Single-frame baseline	1 frame	38.87	29.53
Temporal model	243 frames	32.03	26.36

Note: Both models followed the reported training procedure. The comparison summarizes the evaluated model configurations.

D. Kinematic Measurement Agreement

To complement 3D position errors, we report existing kinematic agreement results for the same 90 test samples. Selected joint angles were compared with optical motion capture using mean bias and Bland–Altman 95% limits of agreement (LOA). Together, mean bias and limits of agreement describe complementary aspects of joint-angle measurement agreement.

TABLE VI. SELECTED KINEMATIC AGREEMENT RESULTS (DEGREES)

Movement / selected angle	Bias (deg)	95% LOA (deg)
Standing - left-knee frontal	+1.17	0.47 to 1.87
Stepping - trunk vertical	-0.01	-5.12 to 5.11
Squatting - knee ROM	-1.30	-15.48 to 12.89
Front raise - hip angle	+3.57	-4.07 to 11.21
Side raise - hip angle	-2.92	-7.96 to 2.11
Seated knee ext. - hip angle	-11.01	-28.99 to 6.97

Note: Previously computed values for the same 90 test samples. LOA = mean bias ± 1.96 SD. The standing knee measure approximates frontal hip–knee–ankle alignment rather than the clinical Q-angle. The side-raise hip angle is defined in the sagittal plane.

Squatting showed a knee-angle bias of −1.30° (95% LOA: −15.48° to 12.89°), whereas the hip-angle bias during seated knee extension was −11.01° (95% LOA: −28.99° to 6.97°). Reporting both bias and limits of agreement captures average differences as well as their spread.

V. DISCUSSION

A. Sources of Error

The movement-dependent errors are consistent with the challenges posed by monocular depth ambiguity. Squatting includes substantial sagittal-plane flexion–extension, while side straight-leg raises may change support and trunk posture. Both movements exhibited alignment gains close to 24%. Since an independent similarity transformation was fitted to each frame, these gains indicate sensitivity to pose alignment rather than a measured amount of physical drift. The distinction between alignment-related and joint-specific error is also relevant to the evaluation framework discussed in OpenCapBench [4].

Joint visibility and movement configuration provide further context for the observed error distribution. Hip position errors

were generally smaller than knee or ankle errors, consistent with the distal-joint patterns discussed by Martinez et al. [20]. Seated knee extension differed from this trend, with knee error exceeding ankle error. Changes in joint visibility and movement amplitude may contribute to this movement-specific pattern.

The synthetic-noise experiment reveals a different sensitivity in temporal measures. MPJVE increased much more sharply than MPJPE as independent perturbations were applied to the 2D keypoints. Because MPJVE is based on adjacent-frame differences, its response partly reflects the amplification of independent noise by temporal differencing. This observation complements the robustness analysis of Hoang et al. [32]. Movement-dependent performance has also been documented by Rode et al. [5] and in AthletePose3D [6].

B. Practical Implications

Quiet standing and both straight-leg raises showed relatively low position and frame-to-frame displacement errors, supporting their use in exploratory frontal-view assessments of posture and movement trends. Squatting and stepping exhibited larger alignment-sensitive and distal-joint errors, for which movement-specific interpretation of 3D coordinates is useful. The higher position error during seated knee extension points to an opportunity for task-specific refinement in detailed quantitative applications. The selected joint-angle comparisons in Table VI add biomechanical context to these position-based results.

The distinction between position accuracy and temporal error is important for downstream analysis. The 243-frame model achieved a lower MPJPE than the evaluated single-frame baseline, whereas the synthetic-noise analysis showed a substantially greater relative change in MPJVE than in MPJPE. Under the laboratory protocol, the lower-error movements are promising for exploratory assessments of posture and slow movement trends. Analyses of velocity, acceleration, or impact-related features may benefit from smoothing or confidence-based filtering of the 2D keypoints or reconstructed 3D trajectories.

C. Scope and Limitations

This study provides a controlled, movement-specific evaluation involving 10 healthy adults and a frontal camera under laboratory conditions. Training and testing used separate recording sets from the same participants, so the reported results characterize within-participant performance. Frame pairing relied on nominal sampling rates, while noise sensitivity was evaluated with independent Gaussian keypoint perturbations. Building on this protocol, future research can assess unseen participants and clinical populations, additional camera views, measured timing offsets and extrinsic registration accuracy, realistic detector noise, and independent test–retest reliability.

VI. Conclusion

Paired monocular RGB video and optical motion capture were used to characterize a BlazePose-based temporal 3D lifting pipeline across six lower-limb movements. The fine-tuned pipeline achieved a mean MPJPE of 32.03 mm under the within-participant recording-set split. Error profiles differed by movement and joint: quiet standing and straight-leg raises had lower position errors, seated knee extension had the highest, and squatting and side straight-leg raise showed comparatively large gains after frame-wise similarity alignment. With independent 2D keypoint noise at a standard deviation of 5.4 pixels, MPJPE increased by approximately 0.9% while MPJVE increased by about 234%. These findings show why position, alignment-sensitive, and frame-to-frame displacement errors should be considered together when interpreting monocular lower-limb pose estimates. The framework also provides a basis for extending evaluation to additional participant groups, camera views, detector conditions, and repeated measurements.

Acknowledgment

The authors thank the participants for their time and cooperation and the laboratory staff for assistance with data collection and equipment setup.

Declarations

Conflicts of interest: The authors declare that they have no conflicts of interest or personal relationships that could have appeared to influence the work reported in this paper.

Ethics approval: The study was approved by the Ethics Committee of Anqing Normal University (Approval No. SWYX-Y-2021-13). The study procedures adhered to the principles of the Declaration of Helsinki.

Consent for publication: Written informed consent was obtained from the participants for the publication of their identifiable images and video recordings in this article.

Data availability: The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Use of AI tools: Generative AI tools assisted with English-language editing and manuscript formatting. The authors reviewed the text, verified the data and references, and accept full responsibility for the final manuscript.

References

[1] K. B. Diallo and M. A. Akhloufi, “Deep human pose estimation: A conceptual review of paradigms, progress, and frontiers,” Computers, vol. 15, no. 6, Art. no. 366, 2026.

[2] Y. Guo, T. H. Gao, A. S. Dong, et al., “A survey of the state of the art in monocular 3D human pose estimation: Methods, benchmarks, and challenges,” Sensors, vol. 25, no. 8, Art. no. 2409, 2025.

[3] S. D. Uhlrich, A. Falisse, Ł. Kidziński, et al., “OpenCap: Human movement dynamics from smartphone videos,” PLoS Comput. Biol., vol. 19, no. 10, Art. no. e1011462, 2023.

[4] Y. Gozlan, A. Falisse, S. Uhlrich, et al., “OpenCapBench: A benchmark to bridge pose estimation and biomechanics,” in Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV), 2025, pp. 4056–4065.

[5] D. Rode, A. Dunkel, R. Willi, P. Wolf, M. Xiloyannis, and R. Riener, “Assessment of monocular human pose estimation models for clinical movement analysis,” Sci. Rep., vol. 15, Art. no. 38767, 2025.

[6] C. Yeung, T. Suzuki, R. Tanaka, Z. Yin, and K. Fujii, “AthletePose3D: A benchmark dataset for 3D human pose estimation and kinematic validation in athletic movements,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), 2025, pp. 5991–6002.

[7] T. Fukushima, P. Blauberger, T. G. Russomanno, and M. Lames, "The potential of human pose estimation for motion capture in sports: A validation study," Sports Eng., vol. 27, no. 1, Art. no. 19, 2024.

[8] D. Pavllo, C. Feichtenhofer, D. Grangier, and M. Auli, "3D human pose estimation in video with temporal convolutions and semi-supervised training," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2019, pp. 7753–7762.

[9] C. Zheng, S. Zhu, M. Mendieta, T. Yang, C. Chen, and Z. Ding, "3D human pose estimation with spatial and temporal transformers," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), 2021, pp. 11656–11665.

[10] B. X. B. Yu, Z. Zhang, Y. Liu, S.-H. Zhong, Y. Liu, and C. W. Chen, "GLA-GCN: Global-local adaptive graph convolutional network for 3D human pose estimation from monocular video," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), 2023, pp. 8818–8829.

[11] W. Zhu, X. Ma, Z. Liu, L. Liu, W. Wu, and Y. Wang, "MotionBERT: A unified perspective on learning human motion representations," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), 2023, pp. 15085–15099.

[12] L. E. Molteni and G. Andreoni, "Comparing the accuracy of markerless motion analysis and optoelectronic system for measuring gait kinematics of lower limb," Bioengineering, vol. 12, no. 4, Art. no. 424, 2025.

[13] R. Hu, Y. Diao, Y. Wang, et al., "Effective evaluation of HGcnMLP method for markerless 3D pose estimation of musculoskeletal diseases patients based on smartphone monocular video," Front. Bioeng. Biotechnol., vol. 11, Art. no. 1335251, 2024.

[14] S. Natraj, T. Messmer, Y. Fujii, et al., "3D pose estimation for scalable remote gait kinematics assessment," npj Digit. Med., vol. 9, no. 1, Art. no. 37, 2026.

[15] G. Pavlakos, X. Zhou, K. G. Derpanis, and K. Daniilidis, "Coarse-to-fine volumetric prediction for single-image 3D human pose," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2017, pp. 7025–7034.

[16] B. Tekin, I. Katircioglu, M. Salzmann, V. Lepetit, and P. Fua, "Structured prediction of 3D human pose with deep neural networks," in Proc. Brit. Mach. Vis. Conf. (BMVC), 2016, pp. 130.1–130.11.

[17] H.-S. Fang, Y. Xu, W. Wang, X. Liu, and S.-C. Zhu, "Learning pose grammar to encode human body configuration for 3D pose estimation," in Proc. AAAI Conf. Artif. Intell., 2018, pp. 6821–6828.

[18] X. Sun, J. Shang, S. Liang, and Y. Wei, "Compositional human pose regression," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 2017, pp. 2602–2611.

[19] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2017, pp. 7291–7299.

[20] J. Martinez, R. Hossain, J. Romero, and J. J. Little, "A simple yet effective baseline for 3D human pose estimation," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 2017, pp. 2640–2649.

[21] V. Bazarevsky, I. Grishchenko, K. Raveendran, T. Zhu, F. Zhang, and M. Grundmann, "BlazePose: On-device real-time body pose tracking," arXiv preprint arXiv:2006.10204, 2020.

[22] J. Zhang, Z. Tu, J. Yang, Y. Chen, and J. Yuan, "MixSTE: Seq2seq mixed spatio-temporal encoder for 3D human pose estimation in video," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2022, pp. 13232–13242.

[23] Q. Zhao, C. Zheng, M. Liu, P. Wang, and C. Chen, "PoseFormerV2: Exploring frequency domain for efficient and robust 3D human pose estimation," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2023, pp. 8877–8886.

[24] S. Mehraban, V. Adeli, and B. Taati, "MotionAGFormer: Enhancing 3D human pose estimation with a Transformer-GCNFormer network," in Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV), 2024, pp. 6920–6930.

[25] Y. Huang, J. Liu, K. Xian, and R. C. Qiu, "PoseMamba: Monocular 3D human pose estimation with bidirectional global-local spatio-temporal state space model," in Proc. AAAI Conf. Artif. Intell., vol. 39, no. 4, 2025, pp. 3842–3850.

[26] W. Bao, Z. Ma, D. Liang, X. Yang, and T. Niu, "Pose ResNet: 3D human pose estimation based on self-supervision," Sensors, vol. 23, no. 6, Art. no. 3057, 2023.

[27] X.-G. Wu, H.-J. Xie, X.-C. Niu, et al., "Transformer-based weakly supervised 3D human pose estimation," J. Vis. Commun. Image Represent., vol. 109, Art. no. 104432, 2025.

[28] Y. L. Lima, T. Collings, M. Hall, et al., "Validity and reliability of trunk and lower-limb kinematics during squatting, hopping, jumping and side-stepping using OpenCap markerless motion capture application," J. Sports Sci., vol. 42, no. 19, pp. 1847–1858, 2024.

[29] S. Dill, A. Ahmadi, M. Grimmer, et al., "Accuracy evaluation of 3D pose reconstruction algorithms through stereo camera information fusion for physical exercises with MediaPipe Pose," Sensors, vol. 24, no. 23, Art. no. 7772, 2024.

[30] R. Maskeliūnas, A. Kulikajevas, R. Damaševičius, et al., "Biomac3D: 2D-to-3D human pose analysis model for tele-rehabilitation based on Pareto optimized deep-learning architecture," Appl. Sci., vol. 13, no. 2, Art. no. 1116, 2023.

[31] M. Medrano-Paredes, C. Fernández-González, H. Saoudi, et al., "Paving the way towards kinematic assessment using monocular video: A preclinical benchmark of state-of-the-art deep-learning-based 3D human pose estimators against inertial sensors in daily living activities," Artif. Intell. Rev., vol. 59, no. 8, Art. no. 181, 2026.

[32] T. H. Hoang, M. Zehni, H. Phan, et al., "Improving the robustness of 3D human pose estimation: A benchmark dataset and learning from noisy input," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), 2024, pp. 113–123.

[33] H. Zheng, H. Li, W. Dai, et al., "HiPART: Hierarchical pose autoregressive transformer for occluded 3D human pose estimation," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2025, pp. 16807–16817.

[34] J. Li, Z. Bai, D. Kong, et al., "3D human pose estimation based on conditional dual-branch diffusion," Multimedia Syst., vol. 31, no. 1, Art. no. 5, 2025.

[35] J. Liu, J. Rojas, Y. Li, et al., "A graph attention spatio-temporal convolutional network for 3D human pose estimation in video," in Proc. IEEE Int. Conf. Robot. Autom. (ICRA), 2021, pp. 3374–3380.

[36] Z. Zhang, "A flexible new technique for camera calibration," IEEE Trans. Pattern Anal. Mach. Intell., vol. 22, no. 11, pp. 1330–1334, 2000.

[37] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments," IEEE Trans. Pattern Anal. Mach. Intell., vol. 36, no. 7, pp. 1325–1339, 2014.