AI Virtual Coach with Multi-Stage Pipelines for Gender, Exercise Recognition, and Pose-Based Assessment

Anonymous Author(s)
Paper under double-blind review

***Abstract*—We present an AI Virtual Coach composed of three models: (i) Gender Recognition, (ii) Exercise Recognition, and (iii) Exercise Assessment. The assessment module performs aspect-level scoring for 15 resistance-training exercises using temporal pose sequences extracted from RGB video via MediaPipe/BlazePose, repetition segmentation, and a compact temporal CNN regressor. We evaluate on a small dataset of 51 volunteers (10 female, 41 male), captured simultaneously in front and side views using mobile phones. Each exercise includes 6–13 volunteers (after alignment), and each volunteer is rated on five aspects (0–10) by two coaches. We use a reliability-weighted label $y = 0.25y_{C1} + 0.75y_{C2}$, a subject-disjoint protocol, and 10 randomized runs per exercise per view with best-checkpoint selection.**

***Index Terms*—Virtual coach, pose estimation, action quality assessment, exercise assessment, temporal CNN, rep segmentation, MediaPipe**

I. Introduction

Contributions. We contribute:

- A practical modular virtual-coach pipeline (gender, exercise recognition, assessment).
- A pose-based assessment method that segments repetitions, aggregates multiple reps per volunteer, and predicts five aspect scores on a 0–10 scale.
- A subject-disjoint multi-run evaluation protocol for small datasets, saving the best model per exercise and view.

II. Related Work

A. Action Quality Assessment

FLEX introduces a large-scale multimodal, multiview dataset and benchmarks for action quality assessment. CoRe proposes group-aware contrastive regression for learning AQA from comparisons.

B. Pose-Based Motion Modeling

MediaPipe/BlazePose enables real-time pose landmark extraction from RGB video. Gait-ViT demonstrates the effectiveness of Transformer-based temporal modeling on gait sequences.

C. What Makes Our Work Different

Most AQA work targets broad action categories and relies on large-scale training data, while our setting focuses on 15 gym exercises with a constrained dataset (51 volunteers) and two consumer-camera views. Unlike single-score AQA, we predict five *aspect-level* scores per exercise (0–10), and explicitly model repetitions via segmentation and aggregation so multi-rep videos can be graded fairly.

III. Data Collection

A. Participants

The dataset was collected from 51 recreationally active volunteers, including 41 males (80.4%) and 10 females (19.6%). Participants ranged in age from 15 to 29 years (mean: 21.3 years), with the majority belonging to the 21–22 age group.

Participant heights ranged from 158 cm to 190 cm (mean: 174.2 cm), while body weights ranged from 52 kg to 110 kg (mean: 77.5 kg). Corresponding Body Mass Index (BMI) values ranged approximately from 18 to 37, with an average BMI of 24.7.

All participants reported no current musculoskeletal injuries at the time of recording. Participation was voluntary, and a verbal informed agreement was obtained from all volunteers, as well as approval from gym staff prior to data acquisition. All collected data were anonymized using numerical identifiers to preserve participant privacy.

B. Exercise Set

The dataset includes recordings of 15 resistance training exercises covering both upper- and lower-body movements:

- Dumbbell Shoulder Press
- Hammer Curls
- Standing Dumbbell Front Raises
- Lateral Raises
- Bulgarian Split Squat
- EZ-Bar Curls
- Incline Dumbbell Bench Press
- Overhead Triceps Extension
- Shrugs
- Weighted Squats
- Seated Biceps Curls
- Triceps Kickbacks
- Rows
- Deadlift
- Calf Raises

Overall, the dataset contains a larger variety of upper-body exercise types than lower-body exercise types. However, the number of volunteers contributing to each exercise was approximately comparable across upper- and lower-body movements, reflecting natural participation during the recording sessions.

C. Recording Setup and Protocol

All recordings were conducted indoors at multiple local gym facilities in Egypt. Videos were captured using a smartphone camera at a resolution of 1080 × 1920 pixels and a frame rate of 30 frames per second, in landscape orientation. Lighting conditions varied across sessions, resulting in mixed indoor illumination.

Camera height and distance were not strictly fixed and varied depending on the exercise, though recordings generally captured either the full body or were positioned at approximately waist level. Each video corresponds to a single exercise set lasting approximately 30 to 60 seconds and containing 8 to 12 repetitions, recorded as a continuous sequence.

Participants performed exercises using self-selected weights to emulate realistic training conditions. Exercise execution was performed naturally, allowing for intra- and inter-subject variability in movement quality, which is essential for form assessment modeling.

D. Viewpoint Acquisition

To capture complementary kinematic information, each exercise was recorded from two viewpoints:

- Front view
- Side view

A total of 154 videos per viewpoint were recorded, resulting in 308 videos overall. Each video was segmented into repetition-level clips, where each clip corresponds to a single exercise repetition. Viewpoints were manually labeled during preprocessing.

E. Expert Annotation

Each recorded exercise set was annotated by two certified fitness coaches based on a joint review of the corresponding front-view and side-view videos. Annotations were performed at the video (set) level, rather than at the individual repetition level.

For each exercise, coaches evaluated five predefined biomechanical aspects specific to that movement, such as joint alignment, range of motion, stability, and movement control. These criteria were defined in advance for each exercise to ensure consistency across annotations.

The resulting annotation scores represent an overall quality assessment of the exercise set and were used as supervisory signals for training and evaluating the proposed exercise assessment model. Repetition-level clips extracted from each video inherit the corresponding set-level annotation.

F. Dataset Limitations

The dataset exhibits a gender imbalance, with a higher proportion of male participants, as well as a limited number of lower-body exercise samples. These limitations are primarily due to cultural and logistical constraints during data collection. Nevertheless, the dataset captures substantial inter-subject variability and realistic execution patterns, making it suitable for evaluating virtual coaching and exercise assessment systems.

G. Pose Data Generation

1) Pose Landmark Extraction: For each video frame, 3D human pose landmarks were extracted using the MediaPipe Pose Landmarker (lite, float16) [**?**]. The model produces 33 body landmarks per frame, each with (x, y, z) coordinates, where x and y represent normalized image-plane coordinates and z represents depth from the camera plane. Detection and tracking confidence thresholds were set to 0.3. Frames were converted from BGR to RGB color space for MediaPipe compatibility, and a fresh `PoseLandmarker` instance was created per video to ensure proper timestamp handling in VIDEO mode.

2) 3D Landmark Normalization: To achieve scale and translation invariance with respect to camera distance and subject body size, all landmarks were normalized in 3D space using a torso-length-based transformation. First, the pelvis center $\mathbf{p}$ was computed as the midpoint of the left hip (landmark 23) and right hip (landmark 24):

$$\mathbf{p} = \frac{1}{2}(\mathbf{h}_{23} + \mathbf{h}_{24}), \tag{1}$$

where $\mathbf{h}_i = (x_i, y_i, z_i)$. Similarly, the mid-shoulder point $\mathbf{s}$ was computed as the midpoint of the left shoulder (landmark 11) and right shoulder (landmark 12):

$$\mathbf{s} = \frac{1}{2}(\mathbf{h}_{11} + \mathbf{h}_{12}). \tag{2}$$

The torso length L in 3D space was then calculated as:

$$L = \|\mathbf{s} - \mathbf{p}\| = \sqrt{(s_x - p_x)^2 + (s_y - p_y)^2 + (s_z - p_z)^2}. \tag{3}$$

Finally, each landmark $\mathbf{h}_i$ was normalized as:

$$\mathbf{h}_i^{\text{norm}} = \frac{\mathbf{h}_i - \mathbf{p}}{L}. \tag{4}$$

Frames with invalid torso length ($L < 10^{-6}$) were discarded to ensure numerical stability.

3) Joint Angle Computation: From the normalized 3D landmarks, nine biomechanical joint angles were computed per frame using the 3D dot product formula. For each angle defined by a triplet of landmarks (a, b, c) where b is the joint vertex, the angle θ was calculated as:

$$\theta = \arccos\left(\frac{\mathbf{v}_1 \cdot \mathbf{v}_2}{\|\mathbf{v}_1\|\|\mathbf{v}_2\|}\right), \tag{5}$$

where $\mathbf{v}_1 = \mathbf{h}_a - \mathbf{h}_b$ and $\mathbf{v}_2 = \mathbf{h}_c - \mathbf{h}_b$ are vectors in 3D space. The nine angles extracted were: left and right elbow (landmarks 11-13-15 and 12-14-16), left and right shoulder (13-11-23 and 14-12-24), left and right hip (11-23-25 and 12-24-26), left and right knee (23-25-27 and 24-26-28), and torso lean (computed from mid-shoulder and pelvis alignment).

Pipeline diagram placeholder

Video → Pose → Rep Segmentation → CNN Assessment (+ Gender Recognition, + Exercise Recognition)

Fig. 1. End-to-end system overview. Replace with your pipeline figure.

4) Temporal Feature Representation: For each exercise repetition, the nine joint angle time series were resampled to a fixed temporal length of $T_{\text{fixed}} = 50$ frames using linear interpolation to enable batch processing and model training. Let $\theta_j^{(t)}$ denote the j-th angle at original time index $t \in [0, T_{\text{orig}} - 1]$. We defined a normalized time axis $\tau \in [0, 1]$ and applied linear interpolation to obtain angle values at target time indices $\tau' \in [0, 1]$ uniformly spaced with T_{fixed} points. This resampling process produced temporal feature tensors of shape $(N_{\text{reps}}, 50, 9)$, where N_{reps} is the total number of repetitions across all volunteers and exercises.

Edge cases were handled as follows: sequences with $T_{\text{orig}} = 0$ were filled with zeros, single-frame sequences ($T_{\text{orig}} = 1$) were replicated across all 50 timesteps, and sequences with $T_{\text{orig}} > 1$ underwent standard linear interpolation.

5) Tempo Preservation: Since resampling to a fixed length inherently discards information about exercise execution speed, we explicitly preserved tempo as separate feature arrays. For each repetition, we computed: (i) **tempo_duration_sec**, the FPS-normalized duration in seconds calculated as $D = T_{\text{orig}}/\text{FPS}$, where FPS is the video frame rate; (ii) **tempo_frame_count**, the total number of frames in the original video; and (iii) **tempo_fps**, the original video FPS. These tempo features allow the model to distinguish between fast and slow executions of the same exercise, which is critical for movement quality assessment.

All processed pose data, including temporal angle sequences and tempo metadata, were stored in compressed NumPy archive (NPZ) format with data type `float32` for use in exercise recognition and assessment tasks.

IV. Dataset and Annotations

A. Data Collection

We collected an in-house dataset with 51 volunteers (10 female, 41 male) performing 15 resistance-training exercises. Videos were captured with mobile phones in *front* and *side* views simultaneously at ∼30 FPS.

B. Annotation Protocol

Each volunteer is scored by two coaches (C1 and C2) for five exercise-specific aspects on a 0–10 scale. We compute a reliability-weighted label: $y = 0.25\, y_{\text{C1}} + 0.75\, y_{\text{C2}}$.

V. System Overview

[conference]IEEEtran

cite amsmath,amssymb graphicx booktabs multirow xcolor url array

***Abstract*—We present an AI Virtual Coach composed of three models: (i) Gender Recognition, (ii) Exercise Recognition, and (iii) Exercise Assessment. The assessment module performs aspect-level scoring for 15 resistance-training exercises using temporal pose sequences extracted from RGB video via MediaPipe/BlazePose, repetition segmentation, and a compact temporal CNN regressor. We evaluate on a small dataset of 51 volunteers (10 female, 41 male), captured simultaneously in front and side views using mobile phones. Each exercise includes 6–13 volunteers (after alignment), and each volunteer is rated on five aspects (0–10) by two coaches. We use a reliability-weighted label $y = 0.25y_{C1} + 0.75y_{C2}$, a subject-disjoint protocol, and 10 randomized runs per exercise per view with best-checkpoint selection.**

***Index Terms*—Virtual coach, pose estimation, action quality assessment, exercise assessment, temporal CNN, rep segmentation, MediaPipe**

VI. Introduction

Contributions. We contribute:

- A practical modular virtual-coach pipeline (gender, exercise recognition, assessment).
- A pose-based assessment method that segments repetitions, aggregates multiple reps per volunteer, and predicts five aspect scores on a 0–10 scale.
- A subject-disjoint multi-run evaluation protocol for small datasets, saving the best model per exercise and view.

VII. Related Work

A. Action Quality Assessment

FLEX introduces a large-scale multimodal, multiview dataset and benchmarks for action quality assessment. CoRe proposes group-aware contrastive regression for learning AQA from comparisons.

B. Pose-Based Motion Modeling

MediaPipe/BlazePose enables real-time pose landmark extraction from RGB video. Gait-ViT demonstrates the effectiveness of Transformer-based temporal modeling on gait sequences.

C. What Makes Our Work Different

Most AQA work targets broad action categories and relies on large-scale training data, while our setting focuses on 15 gym exercises with a constrained dataset (51 volunteers) and two consumer-camera views. Unlike single-score AQA, we predict five *aspect-level* scores per exercise (0–10), and explicitly model repetitions via segmentation and aggregation so multi-rep videos can be graded fairly.

VIII. Data Collection

A. Participants

The dataset was collected from 51 recreationally active volunteers, including 41 males (80.4%) and 10 females (19.6%). Participants ranged in age from 15 to 29 years (mean: 21.3 years), with the majority belonging to the 21–22 age group.

Participant heights ranged from 158 cm to 190 cm (mean: 174.2 cm), while body weights ranged from 52 kg to 110 kg (mean: 77.5 kg). Corresponding Body Mass Index (BMI) values ranged approximately from 18 to 37, with an average BMI of 24.7.

All participants reported no current musculoskeletal injuries at the time of recording. Participation was voluntary, and a verbal informed agreement was obtained from all volunteers, as well as approval from gym staff prior to data acquisition. All collected data were anonymized using numerical identifiers to preserve participant privacy.

B. Exercise Set

The dataset includes recordings of 15 resistance training exercises covering both upper- and lower-body movements:

- Dumbbell Shoulder Press
- Hammer Curls
- Standing Dumbbell Front Raises
- Lateral Raises
- Bulgarian Split Squat
- EZ-Bar Curls
- Incline Dumbbell Bench Press
- Overhead Triceps Extension
- Shrugs
- Weighted Squats
- Seated Biceps Curls
- Triceps Kickbacks
- Rows
- Deadlift
- Calf Raises

Overall, the dataset contains a larger variety of upper-body exercise types than lower-body exercise types. However, the number of volunteers contributing to each exercise was approximately comparable across upper- and lower-body movements, reflecting natural participation during the recording sessions.

C. Recording Setup and Protocol

All recordings were conducted indoors at multiple local gym facilities in Egypt. Videos were captured using a smartphone camera at a resolution of 1080×1920 pixels and a frame rate of 30 frames per second, in landscape orientation. Lighting conditions varied across sessions, resulting in mixed indoor illumination.

Camera height and distance were not strictly fixed and varied depending on the exercise, though recordings generally captured either the full body or were positioned at approximately waist level. Each video corresponds to a single exercise set lasting approximately 30 to 60 seconds and containing 8 to 12 repetitions, recorded as a continuous sequence.

Participants performed exercises using self-selected weights to emulate realistic training conditions. Exercise execution was performed naturally, allowing for intra- and inter-subject variability in movement quality, which is essential for form assessment modeling.

D. Viewpoint Acquisition

To capture complementary kinematic information, each exercise was recorded from two viewpoints:

- Front view
- Side view

A total of 154 videos per viewpoint were recorded, resulting in 308 videos overall. Each video was segmented into repetition-level clips, where each clip corresponds to a single exercise repetition. Viewpoints were manually labeled during preprocessing.

E. Expert Annotation

Each recorded exercise set was annotated by two certified fitness coaches based on a joint review of the corresponding front-view and side-view videos. Annotations were performed at the video (set) level, rather than at the individual repetition level.

For each exercise, coaches evaluated five predefined biomechanical aspects specific to that movement, such as joint alignment, range of motion, stability, and movement control. These criteria were defined in advance for each exercise to ensure consistency across annotations.

The resulting annotation scores represent an overall quality assessment of the exercise set and were used as supervisory signals for training and evaluating the proposed exercise assessment model. Repetition-level clips extracted from each video inherit the corresponding set-level annotation.

F. Dataset Limitations

The dataset exhibits a gender imbalance, with a higher proportion of male participants, as well as a limited number of lower-body exercise samples. These limitations are primarily due to cultural and logistical constraints during data collection. Nevertheless, the dataset captures substantial inter-subject variability and realistic execution patterns, making it suitable for evaluating virtual coaching and exercise assessment systems.

G. Pose Data Generation

1) Pose Landmark Extraction: For each video frame, 3D human pose landmarks were extracted using the MediaPipe Pose Landmarker (lite, float16) [**?**]. The model produces 33 body landmarks per frame, each with (x, y, z) coordinates, where x and y represent normalized image-plane coordinates and z represents depth from the camera plane. Detection and tracking confidence thresholds were set to 0.3. Frames were converted from BGR to RGB color space for MediaPipe compatibility, and a fresh `PoseLandmarker` instance was created per video to ensure proper timestamp handling in VIDEO mode.

2) 3D Landmark Normalization: To achieve scale and translation invariance with respect to camera distance and subject body size, all landmarks were normalized in 3D space using a torso-length-based transformation. First, the pelvis center $\mathbf{p}$ was computed as the midpoint of the left hip (landmark 23) and right hip (landmark 24):

$$\mathbf{p} = \frac{1}{2}(\mathbf{h}_{23} + \mathbf{h}_{24}), \tag{6}$$

where $\mathbf{h}_i = (x_i, y_i, z_i)$. Similarly, the mid-shoulder point $\mathbf{s}$ was computed as the midpoint of the left shoulder (landmark 11) and right shoulder (landmark 12):

$$\mathbf{s} = \frac{1}{2}(\mathbf{h}_{11} + \mathbf{h}_{12}). \tag{7}$$

The torso length L in 3D space was then calculated as:

$$L = \|\mathbf{s} - \mathbf{p}\| = \sqrt{(s_x - p_x)^2 + (s_y - p_y)^2 + (s_z - p_z)^2}. \tag{8}$$

Finally, each landmark $\mathbf{h}_i$ was normalized as:

$$\mathbf{h}_i^{\text{norm}} = \frac{\mathbf{h}_i - \mathbf{p}}{L}. \tag{9}$$

Frames with invalid torso length ($L < 10^{-6}$) were discarded to ensure numerical stability.

3) Joint Angle Computation: From the normalized 3D landmarks, nine biomechanical joint angles were computed per frame using the 3D dot product formula. For each angle defined by a triplet of landmarks (a, b, c) where b is the joint vertex, the angle θ was calculated as:

$$\theta = \arccos\left(\frac{\mathbf{v}_1 \cdot \mathbf{v}_2}{\|\mathbf{v}_1\|\|\mathbf{v}_2\|}\right), \tag{10}$$

where $\mathbf{v}_1 = \mathbf{h}_a - \mathbf{h}_b$ and $\mathbf{v}_2 = \mathbf{h}_c - \mathbf{h}_b$ are vectors in 3D space. The nine angles extracted were: left and right elbow (landmarks 11-13-15 and 12-14-16), left and right shoulder (13-11-23 and 14-12-24), left and right hip (11-23-25 and 12-24-26), left and right knee (23-25-27 and 24-26-28), and torso lean (computed from mid-shoulder and pelvis alignment).

4) Temporal Feature Representation: For each exercise repetition, the nine joint angle time series were resampled to a fixed temporal length of $T_{\text{fixed}} = 50$ frames using linear interpolation to enable batch processing and model training. Let $\theta_j^{(t)}$ denote the j-th angle at original time index $t \in [0, T_{\text{orig}} - 1]$. We defined a normalized time axis $\tau \in [0, 1]$ and applied linear interpolation to obtain angle values at target time indices $\tau' \in [0, 1]$ uniformly spaced with T_{fixed} points. This resampling process produced temporal feature tensors of shape $(N_{\text{reps}}, 50, 9)$, where N_{reps} is the total number of repetitions across all volunteers and exercises.

Edge cases were handled as follows: sequences with $T_{\text{orig}} = 0$ were filled with zeros, single-frame sequences ($T_{\text{orig}} = 1$) were replicated across all 50 timesteps, and sequences with $T_{\text{orig}} > 1$ underwent standard linear interpolation.

5) Tempo Preservation: Since resampling to a fixed length inherently discards information about exercise execution speed, we explicitly preserved tempo as separate feature arrays. For each repetition, we computed: (i) **tempo_duration_sec**, the FPS-normalized duration in seconds calculated as $D = T_{\text{orig}}/\text{FPS}$, where FPS is the video frame rate; (ii) **tempo_frame_count**, the total number of frames in the original video; and (iii) **tempo_fps**, the original video FPS. These tempo features allow the model to distinguish between fast and slow executions of the same exercise, which is critical for movement quality assessment.

Pipeline diagram placeholder
Video → Pose → Rep Segmentation → CNN Assessment
(+ Gender Recognition, + Exercise Recognition)

Fig. 2. End-to-end system overview. Replace with your pipeline figure.

All processed pose data, including temporal angle sequences and tempo metadata, were stored in compressed NumPy archive (NPZ) format with data type `float32` for use in exercise recognition and assessment tasks.

IX. Dataset and Annotations

A. Data Collection

We collected an in-house dataset with 51 volunteers (10 female, 41 male) performing 15 resistance-training exercises. Videos were captured with mobile phones in *front* and *side* views simultaneously at ~30 FPS.

B. Annotation Protocol

Each volunteer is scored by two coaches (C1 and C2) for five exercise-specific aspects on a 0–10 scale. We compute a reliability-weighted label: $y = 0.25\, y_{\text{C1}} + 0.75\, y_{\text{C2}}$.

X. System Overview

XI. Methodology

A. Gender Recognition (Zeyad's Module)

B. Exercise Recognition (Ahmed's Module)

C. Assessment Module (Pose-Based)

1) Pose Extraction: Pose landmarks were extracted and processed as described in Section 3.5 (Pose Data Generation), yielding temporal feature sequences of shape $(N_{\text{reps}}, 50, 9)$ representing nine joint angles over 50 time steps, along with tempo metadata for each repetition.

2) Repetition Segmentation: We detect repetitions using exercise/view-specific biomechanical signals (e.g., arm elevation for raises; hip/knee depth for squats) and a robust threshold-based up–down–up (or down–up–down) cycle logic with hysteresis and minimum-duration constraints.

3) Rep Aggregation: Since labels are per-volunteer overall performance (not per-rep), we aggregate multiple reps from the same volunteer into a single subject-level sample using masked pooling/attention over rep embeddings.

4) Temporal CNN Regressor: We use a compact temporal CNN to encode per-rep sequences. The model predicts five aspect scores; training uses normalized labels in $[0, 1]$ with MSE loss, and outputs are rescaled to 0–10 at inference.

XII. Experimental Setup

A. Protocol

For each exercise and view, we use **subject-disjoint** splits (a volunteer appears only in train or test). We repeat training for 10 randomized runs and select the best checkpoint (lowest MAE) per exercise/view.

B. Metric

We report Mean Absolute Error (MAE) on the 0–10 scale. In our current implementation, the reported MAE is a macro average computed across all aspects and all test subjects for that run.

XIII. Results

A. Assessment Performance (Front vs Side)

XIV. Discussion

XV. Limitations and Future Work

XVI. Conclusion