

# EgoBalanceFormer: Predicting Human Balance from Egocentric Pose and VR-Native Motion

Md Mushfiqur Azam , Nipa Anjum , John Quarles , and Kevin Desai 

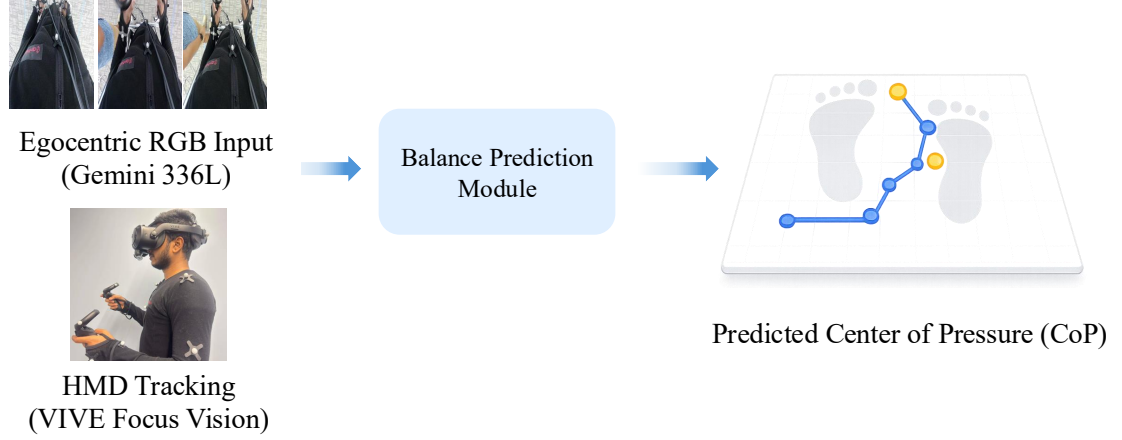


Figure 1: Our method predicts the Center of Pressure (CoP) from egocentric RGB observations and HMD tracking signals acquired using a Gemini 336L camera mounted to a VIVE Focus Vision HMD.

**Abstract**—Maintaining balance is important for safe interaction in immersive virtual reality (VR), where visual–vestibular conflicts can affect postural stability. Most prior works on VR-based balance analysis rely primarily on motion signals from the head-mounted display (HMD), which capture head movement but provide only limited information about full-body biomechanics. As a result, directly predicting biomechanical balance indicators, such as the Center of Pressure (CoP), from VR sensing data remains challenging. In this paper, we present EgoBalanceFormer, a multimodal framework for predicting human balance during VR interaction by leveraging egocentric pose reconstruction as a body-aware sensing modality together with VR-native HMD motion signals. Unlike prior VR balance approaches that rely primarily on sparse device motion such as HMD trajectories, our approach uses egocentric vision to reconstruct full-body pose and derive biomechanical cues—including lower-body kinematics and center-of-mass (CoM) dynamics that are critical for modeling balance control. These pose-derived biomechanical cues are fused with VR-native HMD motion signals and processed through temporal encoders to capture motion dynamics over time. We further apply bidirectional cross-attention between lower-body joint kinematics, CoM dynamics, rotational features, and HMD motion features to capture complementary biomechanical relationships. The refined multimodal representations are fused and decoded using a transformer with a learnable CoP query to regress  $\text{CoP}_X$  and  $\text{CoP}_Y$ . We evaluate EgoBalanceFormer against several machine learning and deep learning baselines for CoP prediction. Experimental results show improved accuracy and stronger robustness across participants, demonstrating that egocentric pose-derived biomechanical cues significantly enhance balance prediction compared with using VR-device motion signals alone.

**Index Terms**—Virtual Reality, Balance Prediction, Egocentric Pose, Center of Pressure.

## 1 INTRODUCTION

Immersive virtual reality (VR) enables users to interact with virtual environments through natural body movements using head-mounted displays (HMDs) and handheld controllers (Figure 1). At the same time, immersive VR can affect postural stability, since visual immersion may introduce conflicts among visual, vestibular, and proprioceptive cues, leading to increased sway or instability during standing interaction [32, 34, 40]. This makes balance monitoring an important component of safe VR interaction.

A widely used biomechanical measure of balance is the Center of Pressure (CoP), which represents the point of application of the ground reaction force on the support surface [10, 28]. CoP trajectories are commonly used to quantify postural sway and stability in biomechanics and clinical assessment [10, 28]. However, direct CoP measurement typically requires specialized hardware such as force plates or balance boards [10, 44], which limits its practicality in consumer-grade VR settings. This motivates computational approaches that can infer balance-related signals from sensing modalities available in VR.

Recent progress in egocentric pose estimation and VR motion tracking provides a promising direction for this problem. Egocentric vision can recover body motion despite strong occlusion and viewpoint challenges [38, 48], while sparse HMD and controller signals have been shown to contain useful cues for full-body motion reconstruction in VR [2, 5, 21]. However, most existing VR balance studies focus on rehabilitation protocols, fall detection, or stable/unstable classification rather than time-indexed regression of biomechanical variables such as CoP [7, 32]. Moreover, prior approaches largely rely on sparse VR-device motion and rarely exploit egocentric visual input to recover

• The authors are with The University of Texas at San Antonio.

E-mail: mdmushfiqur.azam@utsa.edu

E-mail: nipa.anjum@utsa.edu

E-mail: john.quarles@utsa.edu

E-mail: kevin.desai@utsa.edu

Manuscript received xx xxx. 202x; accepted xx xxx. 202x. Date of Publication xx xxx. 202x; date of current version xx xxx. 202x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxx/TVCG.202x.xxxxxx

body-aware biomechanical cues for balance prediction.

To address this gap, we propose EgoBalanceFormer, a multimodal spatio-temporal framework for time-indexed real-valued CoP prediction during VR interaction. Our central idea is to use egocentric pose reconstruction as a body-aware sensing pathway for balance prediction. Specifically, we estimate egocentric 3D body pose from RGB observations and derive lower-body kinematics and center-of-mass (CoM) dynamics that are not directly observable from HMD motion alone. These pose-derived biomechanical cues are combined with VR-native HMD motion and processed through spatio-temporal transformer encoders. The resulting modality-specific representations are refined through bidirectional cross-attention and decoded with a learnable CoP query to regress  $\text{CoP}_X$  and  $\text{CoP}_Y$ .

The primary novelty of this work is the formulation and evaluation of egocentric pose as a body-aware sensing modality for CoP prediction during VR interaction. Prior VR balance work relies on sparse device motion; prior egocentric pose work treats reconstruction as the end goal. Neither estimates a continuous biomechanical balance variable from egocentric observation. Our contribution lies in what is sensed rather than in a new general-purpose Transformer component: egocentric pose recovers lower-body motion that HMD tracking cannot observe, and from it we derive biomechanically motivated lower-body kinematics, rotational dynamics, and pelvis-relative CoM cues. EgoBalanceFormer is the task-specific architecture built around this formulation, temporally encoding these cues and fusing them with HMD motion through pose-centric bidirectional cross-attention.

We evaluate EgoBalanceFormer against several machine learning and deep learning baselines for CoP prediction. The results show that combining egocentric pose-derived biomechanical cues with VR-device motion improves prediction accuracy and robustness across participants. More broadly, this work takes a step toward balance-aware VR systems that can support safer immersive interaction and future rehabilitation-oriented applications.

The main contributions of this work are summarized as follows:

- We formulate and evaluate egocentric pose as a body-aware sensing modality for CoP prediction during VR interaction, complementing the limited full-body information available from HMD motion alone.
- We derive lower-body kinematics, pelvis-relative CoM dynamics, and segment rotational kinematics from estimated egocentric pose, and quantify the contribution of these biomechanical cues.
- We propose a pose-anchored multimodal architecture that refines lower-body representations against CoM, rotational, and HMD cues through bidirectional cross-attention and a learnable CoP query.
- Through baseline comparisons, modality ablations, pose-normalization analysis, and forecast-horizon experiments, we demonstrate the value of egocentric pose-derived cues for CoP prediction.

## 2 RELATED WORK

Recent work has studied balance in virtual reality from several related perspectives, including postural stability in immersive environments, rehabilitation and fall-risk reduction, and learning-based analysis of user instability. Existing studies suggest that VR can both challenge balance and provide a controlled setting for observing how users respond to postural demands [30, 32, 46]. At the same time, researchers have started to view safety in VR more broadly as a prediction and adaptation problem, although most prior work still emphasizes assessment, intervention, or classification rather than continuous estimation of biomechanical balance signals such as Center of Pressure (CoP) [23, 24, 29, 36]. Since our work predicts CoP from egocentric pose, body dynamics, pelvis-relative center of mass (CoM), and VR-device motion, the most relevant prior studies relate to VR balance analysis, motion-based balance cues, and learning-based modeling. The following subsections review these directions and position our approach in relation to them.

### 2.1 Balance, Instability, and Fall-Related Research in VR

A substantial body of VR research has examined how immersive environments affect postural stability and balance control. Prior studies have shown that VR can serve both as a source of balance challenge and as a practical platform for balance assessment. For example, Rosiak *et al.* [32] studied the feasibility of HMD-based balance evaluation, Wittstein *et al.* [46] used VR for dynamic posturography and sensory organization assessment, and Ribeiro *et al.* [30] designed an immersive environment to induce balance disturbances through visual perturbations. Schedler *et al.* [34] further reviewed balance performance in real and virtual environments and highlighted the promise of VR-based assessment despite heterogeneous protocols and measures. Overall, these works primarily focus on assessment, perturbation analysis, or intervention rather than continuous estimation of biomechanical balance variables from multimodal motion cues.

VR has also been studied as a tool for fall-risk reduction and rehabilitation. Systematic reviews and meta-analyses have reported improvements in balance-related outcomes in older adults, including standing balance, gait, lower-limb strength, and fear of falling [6, 15, 23]. In rehabilitation-oriented populations, Kalron *et al.* [22] and Cortés-Pérez *et al.* [12] reported improved postural control and reduced fear of falling in people with Multiple Sclerosis (MS), while later reviews emphasized both the promise of immersive therapy and the variability of existing protocols [4, 7]. Taken together, these studies establish balance and fall-related risk as important themes in VR research, but most remain centered on intervention outcomes or clinical assessment rather than learning a continuous mapping from VR-observable motion to biomechanical balance signals.

### 2.2 Improving Safety in VR

Another important research direction considers safety more broadly in immersive environments, often through predicting or mitigating adverse user states such as cybersickness, disorientation, or instability [14, 24, 29, 36]. Prior work has explored this from several angles, including demographic-based machine learning models for cybersickness prediction [29], personalized prediction frameworks that account for individual differences [36], and multimodal physiological models using signals such as EDA and ECG [24]. Although these studies do not target balance directly, they are relevant because they frame VR safety as a sensing-and-prediction problem rather than only retrospective evaluation.

Safety has also been studied through adaptive intervention and feedback [14, 25, 26]. Mahmud *et al.* [25] showed that auditory, vibrotactile, and visual feedback can improve standing balance in immersive VR for users with balance impairments, and later reported that multimodal feedback can further support balance in people with MS, albeit with possible increases in mental load and fatigue [26]. Dissanayake *et al.* [14] introduced VRsense, an explainable system for analyzing design factors related to cybersickness and user comfort. Collectively, these works suggest that safer VR systems will likely depend on sensing, prediction, and adaptive response working together. In this context, time-indexed CoP estimates could provide a complementary sensing input for future balance-aware VR systems. Rather than predicting cybersickness or providing only reactive feedback, a model can estimate ongoing biomechanical stability directly from egocentric and VR-native motion inputs.

### 2.3 Learning-Based Modeling of Balance Cues

Learning-based balance modeling outside VR provides additional context for this problem. Chen *et al.* [10] reviewed force-plate-based upright balance assessment and reinforced the importance of Center of Pressure trajectories as a core quantitative measure of postural control. Quijoux *et al.* [28] summarized commonly used CoP-derived variables for characterizing standing balance. To move beyond force-plate-only setups, Wantanajittikul *et al.* [44] used machine learning with triaxial inertial sensing to predict CoP trajectories from wearable data. These works support the idea that balance-related variables can be inferred from motion signals rather than measured only by dedicated hardware.

At the same time, recent progress in egocentric and sparse-sensor motion reconstruction is highly relevant to continuous balance prediction in VR. *xR-EgoPose* [39] demonstrated that body kinematics can be estimated from headset-mounted cameras despite strong self-occlusion and limited body visibility. *EgoBody3M* [48] extended this direction with egocentric body tracking on a VR headset. *EgoPoser* [21] showed that full-body motion can also be inferred from sparse head-and-hand VR signals. These methods are important for our setting because they provide access to lower-body kinematics, body dynamics, and center-of-mass-related structure from egocentric or VR-native sensing. However, most prior work uses these motion representations for pose reconstruction, avatar control, or classification of stable versus unstable states rather than continuous CoP regression. In contrast, our work uses estimated egocentric pose as an intermediate representation, derives pelvis-relative CoM and motion dynamics from it, and fuses these cues with HMD motion to predict time-indexed real-valued CoP values in VR.

### 3 DATASET

In this study, we use the VRBalance dataset [1]. The analyzed subset comprised 18 adults (11 female and 7 male), aged  $42.6 \pm 12.7$  years (range: 25–75), with a mean height of  $167.5 \pm 13.4$  cm and body mass of  $78.7 \pm 25.1$  kg. Participants completed seven randomized, counter-balanced two-minute *Song Beater* trials, striking color-matched targets with the controllers while standing under spatial or CoP-dependent auditory feedback at low, medium, and high levels.

Participants stood barefoot on a BTrackS balance board while secured by a safety harness. CoP, full-body motion, egocentric RGB, and HMD/controller motion were recorded at approximately 2, 100, 22, and 60 Hz using the balance board, a 14-camera Vicon system, an HMD-mounted Gemini 336L, and the VR devices, respectively. Before recording, Vicon was calibrated and a participant-specific skeleton was initialized from a T-pose; all streams were timestamped.

Vicon pose and balance-board CoP provide supervision, while inference uses estimated egocentric pose and HMD motion. Controller motion was recorded but excluded because it is task-specific. Preprocessing comprises CoP outlier removal, timestamp alignment, temporal-window construction, and fold-wise normalization. The protocol captures interactive standing VR but represents only one rhythm-game scenario on a flat, stationary surface, excluding seated interaction, locomotion, and other VR genres.

#### 3.1 Gemini-Vicon Synchronization

We perform temporal alignment using timestamp-based nearest-neighbor synchronization. The Gemini data was captured at 22 Hz while the Vicon system recorded at 100 Hz, so direct frame-to-frame alignment was not possible. For each Gemini frame  $k$ , we select the Vicon sample whose timestamp is closest in time:

$$J_k^* = \arg \min_j \left| t_j^{\text{vicon}} - t_k^{\text{gemini}} \right|, \quad \left| t_{J_k^*}^{\text{vicon}} - t_k^{\text{gemini}} \right| \leq \Delta_{\max}. \quad (1)$$

Here  $t_j^{\text{vicon}}$  and  $t_k^{\text{gemini}}$  denote the timestamps of the  $j$ -th Vicon sample and the  $k$ -th Gemini frame, respectively. To account for system jitter, we set the maximum allowable time difference to  $\Delta_{\max} = 23$  ms. Gemini frames without a matching Vicon sample within this tolerance window were excluded from the dataset.

#### 3.2 Data Preprocessing

We preprocess the multimodal VR and egocentric streams before training to improve robustness, reduce supervision noise, and ensure consistent alignment across modalities. Our preprocessing pipeline consists of three main steps: robust outlier removal for the CoP labels, timestamp-based synchronization across streams, and z-score normalization of both inputs and targets. These steps follow standard practices in robust statistics and machine learning for noisy sensor data [9, 16, 17, 33].

##### 3.2.1 Robust Outlier Removal for CoP

The raw balance-board measurements occasionally contain abrupt spikes in  $\text{CoP}_x$  or  $\text{CoP}_y$ , typically due to transient sensor artifacts or unstable contact. To reduce the influence of such samples, we apply a Median Absolute Deviation (MAD) [33] based filter to the CoP trajectories before synchronization.

For each CoP axis, we compute the median and MAD, and then measure the robust standardized deviation of each sample as

$$z_i = \frac{|c_i - \text{median}(c)|}{1.4826 \cdot \text{MAD}(c) + \epsilon}. \quad (2)$$

A CoP sample is removed if either axis exceeds a predefined threshold. In our experiments, we use a MAD threshold of 4.0. Compared with mean/std-based filtering, MAD is more robust to extreme contamination and is therefore well suited to sensor streams with occasional spikes [33]. Prior to synchronization, MAD-based outlier removal discarded 266 out of 22,151 CoP samples (1.20%).

##### 3.2.2 Timestamp-Based Cross-Modal Synchronization

The multimodal streams are recorded at different frequencies. The pose and VR-device streams are sampled at approximately 22 Hz, whereas the CoP signal is sampled at around 2 Hz. Therefore, direct index-wise alignment is not valid. Instead, we use timestamp-based nearest-neighbor synchronization.

For each CoP timestamp  $t_k^{\text{cop}}$ , we find the closest pose timestamp  $t_j^{\text{pose}}$  by

$$j_k^* = \arg \min_j \left| t_j^{\text{pose}} - t_k^{\text{cop}} \right|. \quad (3)$$

The match is accepted only if the time difference is below a maximum allowable offset  $\Delta_{\max} = 50$  ms. Once the synchronized index is identified, we construct one training example using a fixed temporal context of  $T = 11$  consecutive frames ending at  $j_k^*$ , and assign the corresponding CoP value at time  $t_k^{\text{cop}}$  as the target label. Samples without enough preceding frames to form the full 11-frame window are discarded.

This synchronization strategy is important because the CoP at a given instant depends not only on the current posture but also on recent motion history. By aligning each label to a short temporal sequence, the model can learn how body dynamics, VR-device motion, and pose changes jointly influence balance.

##### 3.2.3 Feature-Wise Z-Score Normalization

The input streams contain heterogeneous variables with different numeric ranges, including joint positions, center of mass, joint rotations, and HMD motion. Without normalization, features with larger magnitudes can dominate optimization and make training unstable. We therefore compute normalization statistics from the training split only and reuse them for validation and test sets to avoid information leakage [9, 16].

**Input normalization.** All input channels are normalized using feature-wise z-score scaling:

$$\hat{u} = \frac{u - \mu_u}{\sigma_u + \epsilon}, \quad (4)$$

where  $\mu_u$  and  $\sigma_u$  are the mean and standard deviation of feature channel  $u$ , computed over the training set only.

**Target normalization.** The CoP targets are also normalized using z-score scaling:

$$\hat{y} = \frac{y - \mu_y}{\sigma_y + \epsilon}, \quad (5)$$

where  $\mu_y$  and  $\sigma_y$  are computed from the training targets only. During evaluation, predictions are inverse-transformed back to centimeters before reporting RMSE, MAE, and Correlation, ensuring that all final metrics remain physically interpretable.

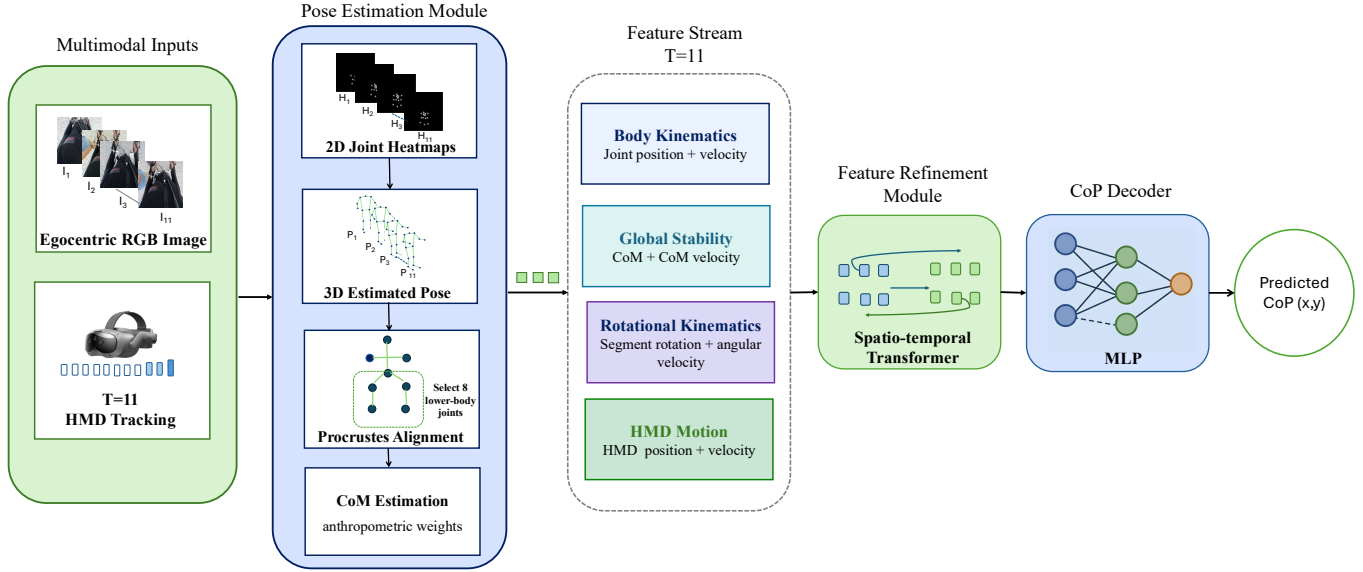


Figure 2: Overview of the proposed EgoBalanceFormer for time-indexed real-valued balance prediction in VR. For each sample, a synchronized window of  $T = 11$  frames is used. Egocentric observations are first converted to 3D pose and associated motion cues, then fused with HMD motion features. The fused multimodal representation is processed by a query-based transformer decoder, followed by a lightweight MLP regression head that predicts the CoP target ( $\text{CoP}_X, \text{CoP}_Y$ ).

## 4 METHODOLOGY

Our goal is to predict the Center of Pressure (CoP) in VR from short egocentric motion sequences. For each synchronized sample, we use a temporal window of  $T = 11$  consecutive frames and regress one aligned CoP target ( $\text{CoP}_X, \text{CoP}_Y$ ). Our pipeline first estimates 3D body pose from egocentric RGB input and derives motion-based biomechanical cues, then combines these pose-derived signals with HMD motion in a multimodal spatio-temporal predictor, as illustrated in Figure 2.

### 4.1 Egocentric Pose Estimation

Egocentric pose estimation provides body-aware motion cues required for CoP prediction that cannot be inferred from HMD motion alone. While HMD tracking captures head movement, it does not encode lower-body configuration or joint kinematics related to balance. Estimating 3D body pose from egocentric RGB therefore provides an intermediate representation for deriving lower-body motion and center-of-mass cues used in balance prediction.

To select a suitable egocentric pose estimation backbone, we evaluated several recent egocentric 3D pose estimation models on VRBalance dataset [1] using standard pose accuracy and efficiency metrics, including Mean Per Joint Position Error (MPJPE) [19], Procrustes Aligned MPJPE (PA-MPJPE) [27], latency, FLOPs, and parameter count, as summarized in Table 1. Although EgoPoseFormer [47] achieves slightly lower MPJPE and PA-MPJPE, it requires substantially higher inference latency and computational cost. In contrast, UnrealEgo [3] provides competitive pose accuracy while being significantly more efficient, reducing latency substantially and requiring far fewer FLOPs than most alternatives. This efficiency is important in our setting because egocentric pose estimation operates as an intermediate

module within a balance-prediction pipeline processing continuous VR interaction sequences. We estimate body kinematics from egocentric RGB observations using a monocular adaptation of the *UnrealEgo* pose estimation framework [3]. For each frame, the model predicts 2D joint heatmaps from the egocentric image and regresses a pelvis-relative 3D pose. The resulting pose sequence is then used to derive lower-body kinematics, rotational dynamics, and pelvis-relative center-of-mass features for CoP prediction.

The pose estimation module is trained directly on the VRBalance dataset [1] using Vicon motion capture data as ground-truth supervision. Training and evaluation follow the same subject-wise 10-fold cross-validation protocol used for the CoP prediction model.

#### 4.1.1 2D Heatmap Generation

For each egocentric RGB frame  $I_t$ , we first predict 2D joint heatmaps using a pretrained heatmap network trained on the EgoPW and SceneEgo datasets [42, 43]. Similar to prior heatmap-based egocentric pose pipelines, this module provides explicit spatial supervision that bridges the RGB input and downstream 3D pose regression [3, 49].

The heatmap network uses a ResNet-18 encoder followed by an FPN-style decoder with unified skip connections [45] to preserve multi-scale spatial details while producing localized joint confidence maps. For each frame  $t$ , the output is a 15-channel heatmap tensor

$$H_t \in \mathbb{R}^{15 \times 64 \times 64}, \quad (6)$$

where each channel corresponds to one body joint. The heatmap prediction module is trained with pixel-wise binary cross-entropy loss,

$$\mathcal{L}_{2D} = \text{BCEWithLogitsLoss}(\hat{H}, H), \quad (7)$$

Table 1: Comparison of egocentric 3D pose estimation models evaluated on VRBalance [1] dataset. Best results are shown in **bold** and second-best results are underlined.

| Model              | MPJPE ↓      | PA-MPJPE ↓   | Latency (ms) ↓ | FLOPs ↓      | Params ↓      |
|--------------------|--------------|--------------|----------------|--------------|---------------|
| UnrealEgo [3]      | <u>86.55</u> | <u>65.87</u> | <b>1.57</b>    | <b>0.47G</b> | 70.67M        |
| EgoGlass [49]      | 94.41        | 68.76        | <u>3.73</u>    | 15.95G       | 36.64M        |
| EgoPoseFormer [47] | <b>86.27</b> | <b>65.21</b> | 5.28           | <u>7.41G</u> | <b>14.18M</b> |
| EgoVisionSpan [20] | 282.15       | 207.56       | 5.88           | 44.27G       | <u>26.87M</u> |
| xR-EgoPose [39]    | 102.99       | 79.28        | 3.83           | 25.42G       | 47.81M        |

where  $H$  and  $\hat{H}$  denote the ground-truth and predicted heatmaps, respectively. These heatmaps provide a structured intermediate representation of joint locations and serve as reliable spatial cues for the subsequent 3D pose regression stage.

#### 4.1.2 3D Pose Regression

The predicted 2D heatmaps are next converted into 3D joint coordinates using a monocular autoencoder-based pose regressor. Following the idea used in UnrealEgo [3], the heatmaps are first compressed into a latent representation and then decoded into a 3D pose. These intermediate representations help the network retain body-structure information and improve robustness to the strong self-occlusions and perspective distortion common in egocentric views. Rather than regressing 3D pose directly from RGB input, using heatmaps provides a more stable structural cue for recovering plausible body geometry. For each frame  $t$ , the regressor  $f_{\text{AE}}$  maps the heatmap tensor  $H_t$  to a root-relative 3D pose:

$$\hat{\mathbf{P}}_t = f_{\text{AE}}(H_t) \in \mathbb{R}^{15 \times 3},$$

where  $\hat{\mathbf{P}}_t$  denotes the estimated 3D coordinates of a 15-joint skeleton centered at the pelvis. Before passing the pose sequence to the temporal model, we apply rigid Procrustes alignment [41] to reduce frame-to-frame geometric variation caused by global translation and orientation noise.

Although the full 3D pose is estimated, CoP is influenced most directly by the lower body. We therefore keep only the lower-limb joints providing the lower-body pose representation:

$$\hat{\mathbf{P}}_t^{lb} \in \mathbb{R}^{8 \times 3},$$

Restricting the representation to these joints allows the downstream model to focus on the kinematics most relevant to balance while also reducing unnecessary feature dimensions.

#### 4.1.3 Lower-Body Joint Velocity

To capture short-term body dynamics, we compute lower-body joint velocity from the estimated pose sequence using finite differences. For interior frames, we use a central-difference approximation:

$$\mathbf{V}_t^{lb} = \frac{\hat{\mathbf{P}}_{t+1}^{lb} - \hat{\mathbf{P}}_{t-1}^{lb}}{2} f_s, \quad (8)$$

where  $f_s = 22$  Hz is the input sampling rate. Forward and backward differences are used at sequence boundaries.

#### 4.1.4 Pose-Derived Rotational Kinematics

From the estimated 3D pose sequence, we construct lower-body rotational kinematic features using limb segment geometry. For each frame  $t$ , segment direction vectors are computed from adjacent joints (hip→knee, knee→ankle, ankle→foot) for both legs. We then form a 3-channel rotational descriptor per lower-body joint:

$$\mathbf{r}_t = [\psi_t, \theta_t, \phi_t],$$

where  $\psi_t$  and  $\theta_t$  denote yaw and pitch from segment direction, and  $\phi_t$  is a signed flexion proxy angle derived from neighboring segment pairs. Yaw and pitch are computed from each normalized segment direction in spherical coordinates with respect to the pelvis-centered frame. The flexion proxy is defined as the signed angle between adjacent limb segments (e.g., hip→knee→ankle). All angles are represented in radians and temporally unwrapped before differencing to prevent  $2\pi$  discontinuities. Aggregating all lower-body joints gives

$$\mathbf{R}_t^{lb} \in \mathbb{R}^{8 \times 3}.$$

To capture angular dynamics, rotational velocities are computed by temporal finite differences:

$$\dot{\mathbf{R}}_t^{lb} = \frac{\mathbf{R}_{t+1}^{lb} - \mathbf{R}_{t-1}^{lb}}{2} f_s,$$

where  $f_s$  is the sampling rate. These pose-derived rotational states and velocities are used as auxiliary inputs to the spatio-temporal balance predictor.

#### 4.1.5 CoM and CoM Velocity

We explicitly use both the center of mass (CoM) position and its temporal derivative as model inputs. CoM position is estimated from the full 15-joint egocentric pose using fixed anthropometric weights  $w_j$  ( $\sum_j w_j = 1$ ) derived from standard body segment mass distributions [13]:

$$\mathbf{C}_t = \sum_{j=1}^{15} w_j \hat{\mathbf{p}}_{t,j}, \quad (9)$$

where  $\hat{\mathbf{p}}_{t,j} \in \mathbb{R}^3$  is the pelvis-relative 3D coordinate of joint  $j$  at time  $t$ . To capture dynamic stabilization behavior, we further compute CoM velocity by central difference:

$$\dot{\mathbf{C}}_t = \frac{\mathbf{C}_{t+1} - \mathbf{C}_{t-1}}{2} f_s, \quad (10)$$

with  $f_s$  the sampling frequency. Thus,  $\mathbf{C}_t$  encodes instantaneous mass distribution, while  $\dot{\mathbf{C}}_t$  captures how that distribution is changing over time.

#### 4.2 HMD Motion Features

In addition to pose-derived position cues, we incorporate head-mounted display (HMD) motion as complementary signals for balance prediction. We include HMD motion as a compact VR-native cue. Let  $\mathbf{h}_t \in \mathbb{R}^3$  denote the HMD position at frame  $t$ , expressed in centimeters. Its velocity is computed using a central-difference approximation:

$$\dot{\mathbf{h}}_t = \frac{\mathbf{h}_{t+1} - \mathbf{h}_{t-1}}{2} f_s, \quad (11)$$

where  $f_s$  is the sampling frequency. The HMD motion provides additional information about the head movement that are useful for modeling balance.

#### 4.3 Spatio-Temporal CoP Predictor

For each synchronized sample, we use a temporal window of  $T = 11$  frames and predict a single CoP target at the final time step. The input sequence is

$$\mathcal{X}_k = \{\hat{\mathbf{P}}_t^{lb}, \mathbf{V}_t^{lb}, \mathbf{C}_t, \dot{\mathbf{C}}_t, \mathbf{R}_t^{lb}, \dot{\mathbf{R}}_t^{lb}, \mathbf{h}_t, \dot{\mathbf{h}}_t\}_{t=t_k-10}^{t_k}, \quad (12)$$

where the modalities correspond to lower-body pose, pose velocity, center of mass (CoM), CoM velocity, lower-body rotations, rotation velocity, and HMD motion cues.

As illustrated in Figure 3, the predictor is organized into two stages: a *spatio-temporal refinement module* that encodes and fuses modality-specific temporal features, and a *CoP decoding stage* that maps the fused representation to the final ( $\text{CoP}_X, \text{CoP}_Y$ ) prediction.

The lower-body joint stream acts as the anchor branch, since CoP is most directly influenced by leg and foot motion. For each frame, lower-body joint positions and velocities are embedded jointly with joint identity, followed by spatial attention across joints to capture within-frame structural dependencies. The resulting frame-level representation is then processed by temporal convolution and temporal self-attention to model both short-range motion changes and temporal relationships.

The remaining modalities are encoded in parallel using dedicated temporal branches. The CoM branch captures global body stabilization through  $\mathbf{C}_t$  and  $\dot{\mathbf{C}}_t$ , the rotation branch models lower-limb orientation dynamics through  $\mathbf{R}_t^{lb}$  and  $\dot{\mathbf{R}}_t^{lb}$ , and the HMD branch encodes head-set motion from  $\mathbf{h}_t$  and  $\dot{\mathbf{h}}_t$ . Keeping these branches separate before fusion allows each modality to preserve its own temporal structure and semantic role in balance prediction.

To couple local lower-body motion with these complementary signals, we apply joint-centric bidirectional cross-attention in a stage-wise manner. Let  $\mathbf{J} \in \mathbb{R}^{T \times D}$  denote the temporally encoded joint features and let  $\mathbf{X} \in \mathbb{R}^{T \times D}$  denote an auxiliary modality branch, such as CoM, rotation, or HMD motion. We first refine the joint representation using  $\mathbf{X}$  as memory:

$$\mathbf{J}' = \text{CA}(\mathbf{J}, \mathbf{X}), \quad (13)$$

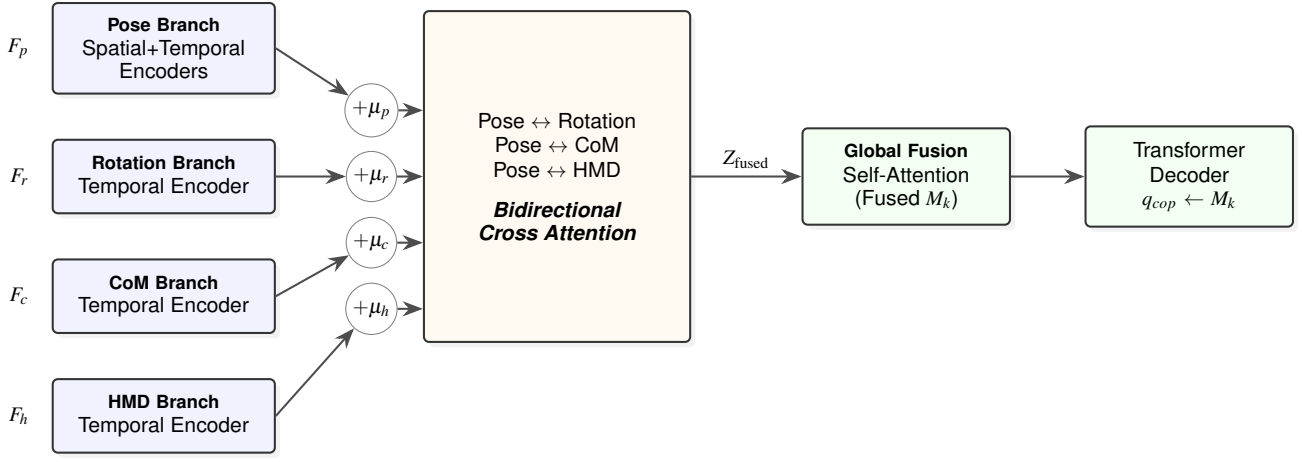


Figure 3: Our spatio-temporal feature refinement module. Modality-specific features from the Pose (\$F\_p\$), Rotation (\$F\_r\$), center-of-mass (\$F\_c\$), and HMD (\$F\_h\$) branches are first temporally encoded and augmented with modality embeddings (\$\mu\_p, \mu\_r, \mu\_c, \mu\_h\$). A joint-pose centric bidirectional cross-attention stage models interactions between joint-pose features and auxiliary modalities (Rotation, CoM and HMD), yielding the refined representation \$Z\_{fused}\$. The fused tokens are then aggregated by global fusion self-attention to form multimodal memory \$M\_k\$, which is decoded by a Transformer using a learnable CoP query \$q\_{cop}\$ to regress (CoP<sub>x</sub>, CoP<sub>y</sub>).

and then update the modality branch using the refined joint representation:

$$\mathbf{X}' = \text{CA}(\mathbf{X}, \mathbf{J}'), \quad (14)$$

so that both streams are mutually informed. This refinement is applied sequentially for the active modality branches.

After pairwise refinement, the branch features are concatenated and augmented with modality embeddings, then passed through a global fusion attention layer to form the shared multimodal memory \$\mathbf{M}\_k\$. This fused memory summarizes the joint, CoM, rotational, and HMD cues across the full temporal window. The final CoP prediction is obtained through a transformer decoder with a learnable CoP query \$\mathbf{q}\_{cop}\$. The query attends to the fused memory \$\mathbf{M}\_k\$, and the resulting decoder output is mapped to the 2D CoP prediction by an MLP head:

$$\hat{\mathbf{y}}_k = g_{\text{MLP}}(\text{Decoder}(\mathbf{q}_{cop}, \mathbf{M}_k)), \quad (15)$$

where \$\hat{\mathbf{y}}\_k \in \mathbb{R}^2\$ corresponds to the predicted coordinates (CoP<sub>x</sub>, CoP<sub>y</sub>).

This design keeps lower-body motion as the primary reference stream while allowing CoM, rotational, and HMD cues to iteratively refine the representation before decoding. The spatio-temporal refinement module therefore captures complementary biomechanical and VR-native dynamics, while the decoder and MLP head translate the fused representation into the final CoP estimate.

#### 4.4 Loss Function

Balance prediction is formulated as a 2D regression problem, where the model predicts the Center of Pressure (CoP) coordinates \$\hat{\mathbf{y}}\_k = [\widehat{\text{CoP}}\_{X,k}, \widehat{\text{CoP}}\_{Y,k}]\$ for each sample \$k\$, with ground truth \$\mathbf{y}\_k = [\text{CoP}\_{X,k}, \text{CoP}\_{Y,k}]\$. The model is trained using the Huber loss with parameter \$\delta = 0.5\$:

$$\mathcal{L} = \frac{1}{N} \sum_{k=1}^N \text{Huber}_{\delta}(\hat{\mathbf{y}}_k - \mathbf{y}_k), \quad (16)$$

where \$N\$ denotes the number of training samples. The Huber loss is used because it behaves quadratically for small errors and linearly for large errors, providing stable optimization while reducing sensitivity to occasional noisy CoP measurements.

### 5 EXPERIMENTS

#### 5.1 Experimental Setup

**Implementation details.** All experiments were conducted on a Lambda Vector workstation running Ubuntu Linux, equipped with 256 GB RAM

and an NVIDIA RTX 5000 Ada GPU with 32 GB VRAM. The system uses NVIDIA driver version 570.211.01 with CUDA 12.8. All models were implemented in PyTorch and trained using GPU acceleration. We evaluate the proposed model using subject-wise 10-fold cross-validation [8] on 18 subjects. In subject-wise 10-fold CV with 18 subjects, folds are necessarily imbalanced. We partition subjects as 8 folds with 2 subjects and 2 folds with 1 subject (disjoint test subjects per fold), with the remaining subjects used for training and validation. This protocol prevents identity leakage and reports performance on unseen subjects in every fold. Feature normalization is applied fold-wise using statistics computed only from the training partition of each fold, and the same transform is applied for validation and test data.

Each sample uses an 11-frame window (\$T=11\$, approximately 0.5 s at 22 Hz) containing lower-body joint position and velocity (\$T \times 8 \times 6\$), pelvis-relative CoM position and velocity (\$T \times 6\$), segment rotation and angular velocity (\$T \times 8 \times 6\$), and HMD position and velocity (\$T \times 6\$). The pose branch maps joint position and velocity to 64-D embeddings, adds learned joint-identity embeddings, and applies a two-layer spatial Transformer (2 heads, FFN 256) followed by attention pooling and a \$64 \rightarrow 128\$ projection. Depthwise and pointwise temporal convolutions (kernels 3 and 1) capture short-range motion. The CoM, rotation, and HMD streams are independently projected to 128-D. All branches use sinusoidal temporal encoding and a two-layer temporal Transformer (8 heads, FFN 512, dropout 0.2).

Using pose as the anchor stream, bidirectional cross-attention sequentially refines it with CoM, rotation, and HMD, using one Transformer decoder layer per direction (\$d=128\$, 8 heads, FFN 512). After refinement, learned modality embeddings are added and the branch tokens are concatenated and processed by a one-layer global-fusion block. A learnable \$1 \times 128\$ CoP query attends to the fused memory through a one-layer Transformer decoder, followed by a \$128 \rightarrow 64 \rightarrow 2\$ MLP. Controller motion was recorded but excluded because it is strongly task-specific in *Song Beater*.

**Training details.** We train all models using AdamW with mini-batch optimization (batch\_size=64), initial learning rate \$1 \times 10^{-4}\$, and weight decay \$1 \times 10^{-5}\$, for up to 50 epochs per fold. The learning rate is updated with cosine annealing (CosineAnnealingLR, \$\eta\_{\min} = 10^{-6}\$). Gradient clipping with max norm 1.0 is applied for stable optimization. The main objective is Huber loss with \$\delta = 0.5\$, and early stopping with patience 10 is used based on validation RMSE to mitigate overfitting.

#### 5.2 Evaluation Metrics

We evaluate prediction performance using standard regression metrics, including root mean squared error (RMSE), mean absolute error (MAE),

Table 2: Comparison of CoP prediction performance across methods. All baseline methods use VR-HMD motion signals only. EgoBalanceFormer (HMD only) uses the proposed architecture with matched HMD input to isolate the architecture contribution. EgoBalanceFormer (EgoPose+HMD) is the full proposed pipeline incorporating egocentric pose-derived biomechanical features.

| Method                                | CoP <sub>X</sub> |             |              | CoP <sub>Y</sub> |             |              | Threshold Accuracy ↑(%) |              |              |
|---------------------------------------|------------------|-------------|--------------|------------------|-------------|--------------|-------------------------|--------------|--------------|
|                                       | RMSE↓            | MAE↓        | Correlation↑ | RMSE↓            | MAE↓        | Correlation↑ | ≤1 cm                   | ≤2 cm        | ≤3 cm        |
| XGBoost                               | 2.13             | 1.73        | 0.47         | 1.65             | 1.30        | 0.62         | 20.41                   | 52.57        | 70.75        |
| RandomForest                          | 2.15             | 1.82        | 0.45         | 1.75             | 1.41        | 0.61         | 19.81                   | 51.08        | 68.40        |
| ExtraTrees                            | 2.04             | 1.64        | 0.45         | 1.79             | 1.43        | 0.64         | 18.88                   | 49.72        | 69.80        |
| Lasso Regression                      | 1.85             | 1.48        | 0.46         | 1.52             | 1.25        | 0.64         | 19.80                   | 54.77        | 78.33        |
| LSTM                                  | 1.82             | 1.42        | 0.41         | 1.37             | 1.12        | 0.65         | 25.04                   | 61.07        | 80.77        |
| GRU                                   | 1.79             | 1.39        | 0.43         | 1.39             | 1.12        | 0.65         | 25.03                   | 61.31        | 81.53        |
| DeepTCN                               | 1.98             | 1.57        | 0.42         | 1.39             | 1.21        | 0.63         | 24.55                   | 60.32        | 80.56        |
| CNN-LSTM                              | 1.73             | 1.35        | 0.46         | 1.43             | 1.14        | 0.64         | 25.47                   | 61.23        | 81.44        |
| EgoBalanceFormer (HMD only)           | 1.67             | 1.31        | 0.51         | 1.37             | 1.09        | 0.65         | 27.07                   | 64.37        | 83.52        |
| <b>EgoBalanceFormer (EgoPose+HMD)</b> | <b>1.54</b>      | <b>1.18</b> | <b>0.56</b>  | <b>1.20</b>      | <b>0.95</b> | <b>0.73</b>  | <b>31.57</b>            | <b>68.37</b> | <b>87.05</b> |

and Pearson’s correlation between predicted and ground-truth CoP values for both CoP<sub>X</sub> and CoP<sub>Y</sub>.

## 6 RESULTS

This section presents quantitative comparisons with ML and DL baselines and analyzes the effect of input modalities, pose normalization, and motion features on CoP prediction performance.

### 6.1 Comparison with ML and DL Baselines

Table 2 reports RMSE, MAE, and Pearson’s correlation for CoP<sub>X</sub> and CoP<sub>Y</sub> under subject-wise 10-fold cross-validation, together with fold-aggregated threshold accuracy at 1, 2, and 3 cm. Following prior VR balance-related work, all baseline models are trained using only VR-HMD motion signals, which represent the standard and commonly available device-motion cue in VR systems. EgoBalanceFormer is evaluated in two settings: *HMD-only*, to isolate the contribution of the proposed architecture under the same input condition as the baselines, and *Egocentric Pose + HMD*, which corresponds to the full proposed framework.

Under the HMD-only setting, EgoBalanceFormer achieves the strongest overall performance among HMD-based models, indicating that the proposed spatio-temporal design is effective even when restricted to sparse HMD motion cues. Adding egocentric pose-derived biomechanical features further improves performance across all metrics, yielding the lowest error on both axes (CoP<sub>X</sub>: RMSE = 1.54 cm, MAE = 1.18 cm; CoP<sub>Y</sub>: RMSE = 1.20 cm, MAE = 0.95 cm), the highest correlation ( $r = 0.56$  for CoP<sub>X</sub> and  $r = 0.73$  for CoP<sub>Y</sub>), and the best threshold accuracy (31.57%, 68.37%, and 87.05% within 1, 2, and 3 cm, respectively). This improvement over the HMD-only variant indicates that the gain comes not only from the architecture, but also from the added body-aware information provided by egocentric pose.

Among the deep learning baselines, CNN-LSTM [35] and GRU [11] are the strongest competitors in terms of absolute error. CNN-LSTM achieves RMSE values of 1.73 cm and 1.43 cm on CoP<sub>X</sub> and CoP<sub>Y</sub>, respectively, while GRU obtains 1.79 cm and 1.39 cm. These results suggest that standard temporal models can capture coarse trends from HMD motion, but remain limited without explicit biomechanical cues. Among the classical regression baselines, Lasso Regression [37] achieves the lowest overall error (RMSE<sub>X</sub>=1.85 cm, RMSE<sub>Y</sub>=1.52 cm), suggesting that part of the relationship between HMD motion and CoP can be captured through sparse linear dependencies. The tree-based ensemble models (XGBoost, RandomForest, and ExtraTrees) show higher errors overall, indicating that purely device-motion-based features remain insufficient for accurate balance prediction. Among these methods, ExtraTrees slightly improves RMSE on CoP<sub>X</sub>, while XGBoost produces stronger correlation and accuracy on CoP<sub>Y</sub>, but both remain substantially below the proposed EgoBalanceFormer. The threshold accuracy results further reinforce this advantage. Compared with the strongest non-EgoBalanceFormer baselines, the full

model improves accuracy within 1 cm from 25.47% (CNN-LSTM) to 31.57%, within 2 cm from 61.31% (GRU) to 68.37%, and within 3 cm from 81.53% (GRU) to 87.05%. These results show that EgoBalanceFormer is stronger than existing HMD-only baselines under matched inputs, and that incorporating egocentric pose-derived biomechanical information provides an additional and meaningful gain for continuous CoP prediction.

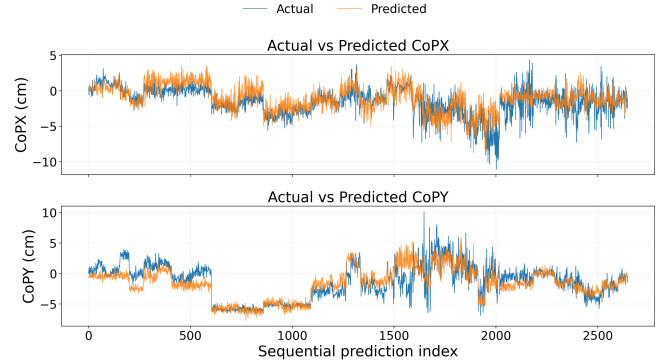


Figure 4: Actual vs Predicted Center of Pressure (CoP) along both axes.

### 6.2 Qualitative Analysis of Predicted CoP Trajectories

Figure 4 illustrates representative ground-truth and predicted CoP trajectories for a held-out test subject. Along the CoP<sub>X</sub> axis, the model tracks the overall drift and oscillatory pattern of the signal, although some rapid short-duration variations are slightly attenuated. In contrast, predictions along CoP<sub>Y</sub> show closer alignment with the ground truth, consistent with the stronger quantitative results reported in Table 2. Across both axes, the predicted trajectories remain temporally smooth and largely free of spurious spikes. This behavior reflects the effect of the Huber loss and the short temporal window (11 frames), which encourage stable regression while reducing sensitivity to transient noise.

### 6.3 Ablation on Input Modalities

Table 3 shows that the usefulness of each modality differs across axes. For CoP<sub>X</sub>, *Estimated Pose* alone achieves the lowest RMSE (1.51) and shares the lowest MAE (1.16) with *Estimated Pose + Rotational Features*. This indicates that pose-derived body kinematics already capture the main cues required for lateral balance prediction. Adding other modalities does not reduce these errors further, although multi-modal fusion improves correlation: *Estimated Pose + HMD Features* increases correlation to 0.53, and the full multimodal setting achieves the highest value (0.56), suggesting better temporal agreement with the target trajectory.

For CoP<sub>Y</sub>, both *Estimated Pose + CoM Features* and *Estimated Pose +*

Table 3: Ablation study on input modality combinations for CoP prediction.

| Input Modalities                             | CoP <sub>X</sub> |             |              | CoP <sub>Y</sub> |             |              |
|----------------------------------------------|------------------|-------------|--------------|------------------|-------------|--------------|
|                                              | RMSE↓            | MAE↓        | Correlation↑ | RMSE↓            | MAE↓        | Correlation↑ |
| <i>Estimated Pose</i>                        | <b>1.51</b>      | <b>1.16</b> | 0.51         | 1.24             | 0.99        | 0.71         |
| <i>Estimated Pose + CoM Features</i>         | 1.53             | 1.19        | 0.49         | <b>1.17</b>      | <b>0.92</b> | 0.72         |
| <i>Estimated Pose + HMD Features</i>         | 1.54             | 1.18        | 0.53         | <b>1.17</b>      | 0.93        | 0.71         |
| <i>Estimated Pose + Rotational Features</i>  | 1.53             | <b>1.16</b> | 0.52         | 1.24             | 0.98        | 0.71         |
| <i>CoM Features + Rotational Features</i>    | 1.76             | 1.36        | 0.40         | 1.34             | 1.06        | 0.66         |
| <i>CoM Features + HMD Features</i>           | 1.70             | 1.35        | 0.51         | 1.34             | 1.09        | 0.66         |
| <i>Estimated Pose + Rotation + CoM + HMD</i> | 1.54             | 1.18        | <b>0.56</b>  | 1.20             | 0.95        | <b>0.73</b>  |

Table 4: Effect of Procrustes alignment on CoP prediction performance. Both settings use egocentric pose estimates; the second applies Procrustes alignment to normalize global pose variations before prediction.

| Pose Representation | CoP <sub>X</sub> |             |              | CoP <sub>Y</sub> |             |              | Threshold Accuracy ↑(%) |              |              |
|---------------------|------------------|-------------|--------------|------------------|-------------|--------------|-------------------------|--------------|--------------|
|                     | RMSE↓            | MAE↓        | Correlation↑ | RMSE↓            | MAE↓        | Correlation↑ | ≤1 cm                   | ≤2 cm        | ≤3 cm        |
| Raw                 | 1.68             | 1.31        | 0.49         | 1.34             | 1.06        | 0.67         | 27.50                   | 64.69        | 82.65        |
| Procrustes Aligned  | <b>1.54</b>      | <b>1.18</b> | <b>0.56</b>  | <b>1.20</b>      | <b>0.95</b> | <b>0.73</b>  | <b>31.57</b>            | <b>68.37</b> | <b>87.05</b> |

*HMD Features* achieve the lowest RMSE (1.17), improving over pose-only (1.24). Among them, *Estimated Pose + CoM Features* yields the best MAE (0.92) and slightly higher correlation (0.72). This suggests that anterior–posterior CoP prediction benefits from complementary global motion information.

In contrast, configurations excluding pose entirely (*CoM + Rotational* and *CoM + HMD*) produce the highest errors on both axes, indicating that pose is the strongest standalone modality. The full multimodal setting (*Pose + Rotation + CoM + HMD*) does not minimize RMSE but achieves the highest correlation for both CoP<sub>X</sub> (0.56) and CoP<sub>Y</sub> (0.73), suggesting that combining modalities mainly improves trajectory consistency.

#### 6.4 Effect of Procrustes Alignment on Pose-Based CoP Prediction

We evaluate the effect of removing global pose variations by comparing CoP prediction using raw egocentric poses and poses aligned through Procrustes analysis [41]. As shown in Table 4, applying Procrustes alignment consistently improves performance across all metrics. For CoP<sub>X</sub>, RMSE decreases from 1.68 cm to 1.54 cm and MAE from 1.31 cm to 1.18 cm, while correlation increases from 0.49 to 0.56. Similar improvements are observed for CoP<sub>Y</sub>, where RMSE decreases from 1.34 cm to 1.20 cm and MAE from 1.06 cm to 0.95 cm, with correlation increasing from 0.67 to 0.73. Threshold accuracy also improves, with predictions within 3 cm increasing from 82.65% to 87.05%.

#### 6.5 Forecast Horizon on CoP Prediction

We analyze the effect of forecast horizon on CoP prediction accuracy. Table 5 shows that performance degrades gradually as the horizon increases from  $\Delta = 0.5$  s to 2.0 s. RMSE increases slightly from 1.65 cm to 1.73 cm for CoP<sub>X</sub> and from 1.24 cm to 1.27 cm for CoP<sub>Y</sub>, while correlation decreases from 0.45 to 0.33 and from 0.71 to 0.64, respectively. Threshold accuracy follows a similar trend, with predictions within 1 cm decreasing from 29.69% to 28.13% and predictions within 3 cm decreasing from 85.01% to 83.19%. Figure 5 qualitatively illustrates that the predicted trajectories continue to follow the overall temporal trend of the ground-truth CoP despite the increasing horizon. These results indicate that the proposed multimodal biomechanical representation captures balance dynamics that remain informative for near-future CoP prediction.

## 7 DISCUSSION

The results indicate that time-indexed CoP prediction in VR benefits from incorporating body-aware biomechanical cues rather than relying only on sparse device motion. This is consistent with prior biomechanical understanding that CoP and CoM capture complementary aspects of postural control, with CoP reflecting corrective action at the support surface and CoM reflecting whole-body mass redistribution [31]. In our setting, pelvis-relative CoM and lower-body kinematic features derived from egocentric pose provide information that HMD motion alone cannot encode, which likely explains the consistent improvement over HMD-only baselines. More broadly, whereas prior VR balance work has largely focused on head-motion sway, intervention, or force-plate-based assessment [30, 32, 46], our findings suggest that egocentric pose estimation provides a body-aware sensing pathway for recovering balance cues during VR interaction.

The ablation results further suggest that the relevance of biomechanical cues is axis-dependent. This is consistent with postural control theory, which shows that mediolateral and anterior–posterior balance are regulated differently and depend on the CoM relative to the base of support [18]. In our experiments, lateral CoP prediction benefited most from lower-body pose and rotational cues, suggesting stronger dependence on local ankle and hip adjustment dynamics. In contrast, anterior–posterior prediction improved more clearly when CoM features were added, consistent with greater involvement of whole-body mass redistribution and trunk dynamics [31]. This axis-dependent pattern was not assumed a priori and provides useful insight into how different motion representations contribute to balance estimation.

One notable result is that estimated pose alone achieves slightly lower CoP<sub>X</sub> RMSE than the full multimodal setting. A plausible explanation is that, under the current dataset size and sparse 2 Hz CoP supervision, adding CoM, rotational, and HMD cues introduces variability that is not always resolved into lower point-wise error. At the same time, the full model yields higher correlation, suggesting a trade-off between minimizing instantaneous error and preserving trajectory consistency.

The improvement from Procrustes alignment suggests that pose consistency is important when estimated skeletons are used as predictive signals. Removing frame-to-frame global orientation and translation reduces geometric noise and allows the model to focus on relative joint configurations that are more directly related to balance state. Likewise, the modest degradation with increasing forecast horizons indicates that the learned representation captures short-term stabilization dynamics rather than instantaneous posture.

Table 5: Effect of forecast horizon on CoP prediction performance. Results are reported for increasing prediction horizons ( $\Delta = 0.5\text{--}2.0\text{ s}$ ).

| Forecast Horizons    | CoP <sub>X</sub>  |                  |                        | CoP <sub>Y</sub>  |                  |                        | Threshold Accuracy $\uparrow$ (%) |                |                |
|----------------------|-------------------|------------------|------------------------|-------------------|------------------|------------------------|-----------------------------------|----------------|----------------|
|                      | RMSE $\downarrow$ | MAE $\downarrow$ | Correlation $\uparrow$ | RMSE $\downarrow$ | MAE $\downarrow$ | Correlation $\uparrow$ | $\leq 1$<br>cm                    | $\leq 2$<br>cm | $\leq 3$<br>cm |
| $\Delta = 1$ (0.5 s) | <b>1.65</b>       | <b>1.27</b>      | <b>0.45</b>            | <b>1.24</b>       | <b>0.98</b>      | <b>0.71</b>            | <b>29.69</b>                      | 65.40          | <b>85.01</b>   |
| $\Delta = 2$ (1.0 s) | 1.69              | 1.29             | 0.38                   | 1.25              | 0.98             | 0.68                   | 29.22                             | 65.44          | 84.10          |
| $\Delta = 3$ (1.5 s) | 1.69              | 1.30             | 0.36                   | 1.25              | 0.98             | 0.67                   | 28.61                             | <b>65.57</b>   | 83.38          |
| $\Delta = 4$ (2.0 s) | 1.73              | 1.36             | 0.33                   | 1.27              | 0.99             | 0.64                   | 28.13                             | 63.53          | 83.19          |

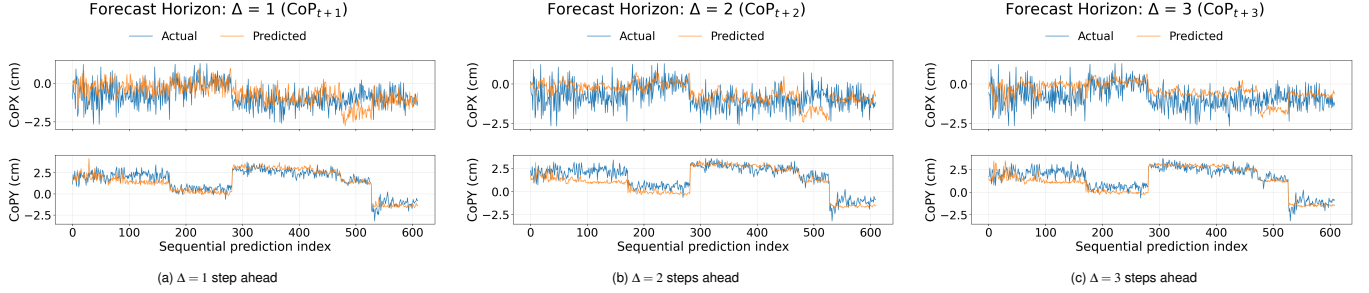


Figure 5: Actual and predicted CoP trajectories under different forecast horizons. Prediction accuracy gradually decreases as the horizon increases, while the model continues to follow the overall balance dynamics.

Our metrics capture complementary aspects of prediction quality: RMSE and MAE quantify CoP localization error, correlation measures preservation of sway trends, and the 1/2/3 cm accuracies report the fraction of estimates within fixed tolerance bands. Compared with the HMD-only model, adding pose-derived cues increases within-1 cm accuracy from 27.07% to 31.57% and CoP<sub>Y</sub> correlation from 0.65 to 0.73. The smaller gains in point-wise error suggest that the primary benefit lies in preserving sway dynamics and improving accuracy within tight spatial tolerances. These gains could support future instability detection or adaptive feedback, particularly when decisions depend on sway trends rather than centimetre-exact CoP. However, we demonstrate only time-indexed CoP regression, not fall-risk detection, instability classification, adaptive feedback, or safety improvement. These capabilities require an instability criterion, a decision policy, causal deployment, and closed-loop user validation. Thus, our contribution is an enabling sensing layer rather than a complete balance-aware VR system.

## 8 LIMITATIONS AND FUTURE WORK

Although EgoBalanceFormer demonstrates consistent improvements in time-indexed CoP prediction, some limitations remain. First, the current study includes 18 participants performing a single VR balancing task, which limits demographic diversity and ecological coverage. While subject-wise cross-validation provides a meaningful generalization test, future work should evaluate the approach on larger and diverse cohorts and across a wider range of VR interaction scenarios, including populations with balance impairments such as individuals with MS or vestibular dysfunction.

Second, the framework relies on an intermediate egocentric pose estimation stage whose errors propagate into the downstream balance predictor. Estimating lower-body pose from a head-mounted fisheye camera is challenging due to self-occlusion and limited foot visibility, and current models still produce MPJPE values above 86 mm on the VRBalance dataset [1]. Improving lower-body observability through end-to-end learning or enhanced sensing setups, such as stereo fisheye cameras, could reduce this source of noise.

Third, the current prototype adds a Gemini 336L camera, rigid mount, and power/data connection to the HMD. Consumer deployment would require comparable hardware or access to suitable headset cameras, camera calibration, stable mounting, and synchronization with HMD tracking. Vicon and the balance board provide supervision and evaluation ground truth during model development and are not required at inference.

Fourth, the CoM estimate uses fixed de Leva mass fractions rather than subject-specific anthropometry, which may limit generalization across body proportions. We also assume a flat, stationary support surface without explicit foot-ground contact modeling; uneven or moving surfaces remain untested and may require contact-aware sensing.

Finally, the balance board provides CoP ground truth at approximately 2 Hz after synchronization, which limits supervision of higher frequency postural adjustments and results in a single CoP target per temporal window. Future data collection with higher-frequency force platforms and models that predict short CoP trajectories rather than single endpoints may allow finer modeling.

## 9 CONCLUSION

We present EgoBalanceFormer, a multimodal framework for time-indexed real-valued CoP prediction in immersive VR that combines egocentric pose reconstruction with VR-native motion cues. By deriving lower-body kinematics, rotational dynamics, and pelvis-relative CoM features from egocentric observations, the approach captures biomechanical information not available from HMD motion alone. Experimental results show consistent improvements over both machine learning and deep learning baselines, while ablation, alignment, and forecast-horizon analyses highlight the value of biomechanically structured multimodal representations for balance prediction. Overall, this work demonstrates the feasibility of egocentric pose as a sensing pathway for CoP estimation and represents an enabling step toward future balance-aware VR systems.

## 10 ACKNOWLEDGEMENT

This material is partially based upon work supported by the National Science Foundation under Grant No. 2316240 and 2403411. Any opinions, findings, and conclusions or recommendations expressed herein are those of the author(s) and do not reflect National Science Foundation views.

## REFERENCES

- [1] VRBalance: A multimodal dataset for postural stability analysis in immersive vr. <https://anonymous.4open.science/r/VRBalance>, 2026. [Online; accessed Mar. 2026]. 3, 4, 9
- [2] K. Ahuja, V. Shen, C. M. Fang, N. Riopelle, A. Kong, and C. Harrison. Controllerpose: inside-out body capture with vr controller cameras. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pp. 1–13, 2022. 1

- [3] H. Akada, J. Wang, S. Shimada, M. Takahashi, C. Theobalt, and V. Golyanik. Unrealego: A new dataset for robust egocentric 3d human motion capture. In *European Conference on Computer Vision (ECCV)*, vol. 13666 of *Lecture Notes in Computer Science*, pp. 1–17, 2022. doi: 10.1007/978-3-031-20068-7\_1 4, 5
- [4] B. Alayidi, E. Al-Yahya, D. McNally, and S. P. Morgan. Exploring balance control mechanisms in people with multiple sclerosis in virtual reality environment: a systematic review. *Journal of NeuroEngineering and Rehabilitation*, 22:75, 2025. doi: 10.1186/s12984-025-01612-0 2
- [5] M. M. Azam and K. Desai. A survey on 3d egocentric human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1643–1654, 2024. 1
- [6] B. Azimi Sanavi, M. Talebpour, V. Saatchian, F. Oveysi Sani, and F. Mirakhor. Effects of virtual reality-based exercise training on risk of falling, walking capacity, and quality of life in healthy older adults: A systematic review and meta-analysis research. *Journal of Applied Gerontology*, p. 7334648251362512, 2025. doi: 10.1177/07334648251362512 2
- [7] E. B. Basalic, N. Roman, V. I. Tuchel, and R. S. Miclăuş. Virtual reality applications for balance rehabilitation and efficacy in addressing other symptoms in multiple sclerosis—a review. *Applied Sciences*, 14(10):4244, 2024. 1, 2
- [8] Y. Bengio and Y. Grandvalet. No unbiased estimator of the variance of k-fold cross-validation. *Journal of machine learning research*, 5(Sep):1089–1105, 2004. 6
- [9] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, New York, 2006. 3
- [10] B. Chen, P. Liu, F. Xiao, Z. Liu, and Y. Wang. Review of the upright balance assessment based on the force plate. *International journal of environmental research and public health*, 18(5):2696, 2021. 1, 2
- [11] K. Cho, B. van Merriënboer, C. Gülçehre, D. Bahdanau, F. Bougares, H. Schwenk et al. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1724–1734, 2014. 7
- [12] I. Cortés-Pérez, M. C. Osuna-Pérez, D. Montoro-Cárdenas, R. Lomas-Vega, E. Obrero-Gaitán, and F. A. Nieto-Escámez. Virtual reality-based therapy improves balance and reduces fear of falling in patients with multiple sclerosis: a systematic review and meta-analysis of randomized controlled trials. *Journal of NeuroEngineering and Rehabilitation*, 20:42, 2023. doi: 10.1186/s12984-023-01174-z 2
- [13] P. De Leva. Adjustments to zatsiorsky-seluyanov’s segment inertia parameters. *Journal of Biomechanics*, 29(9):1223–1230, 1996. 5
- [14] D. Dissanayake, C. Gupta, P. Sasikumar, and S. C. Nanayakkara. VRSense: An explainable system to help mitigate cybersickness in VR games. In *Proceedings of the Extended Abstracts of the 2025 CHI Conference on Human Factors in Computing Systems*, pp. 1–7, 2025. doi: 10.1145/3706599.3719891 2
- [15] Y. Gao, N. Wang, Y. Liu, and N. Liu. Effectiveness of virtual reality in preventing falls in non-disabled older adults: A meta-analysis and systematic review. *Geriatric Nursing*, 56:106–113, 2024. doi: 10.1016/j.gerinurse.2024.04.010 2
- [16] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, Cambridge, MA, 2016. 3
- [17] F. R. Hampel. The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69(346):383–393, 1974. doi: 10.1080/01621459.1974.10482962 3
- [18] A. L. Hof, M. Gazendam, and W. Sinke. The condition for dynamic stability. *Journal of Biomechanics*, 38(1):1–8, 2005. 8
- [19] 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 TPAMI*, 2014. 4
- [20] H. Jiang and V. K. Ithapu. Egocentric pose estimation from human vision span, 2021. 4
- [21] J. Jiang, P. Strel, M. Meier, and C. Holz. Egoposer: Robust real-time egocentric pose estimation from sparse and intermittent observations everywhere. In *European Conference on Computer Vision*, pp. 277–294. Springer, 2024. 1, 3
- [22] A. Kalron, I. Fonkatz, L. Frid, H. Baransi, and A. Achiron. The effect of balance training on postural control in people with multiple sclerosis using the CAREN virtual reality system: a pilot randomized controlled trial. *Journal of NeuroEngineering and Rehabilitation*, 13:13, 2016. doi: 10.1186/s12984-016-0124-y 2
- [23] J. Lee, S. Phu, S. R. Lord, and Y. Okubo. Effects of immersive virtual reality training on balance, gait and mobility in older adults: A systematic review and meta-analysis. *Gait & Posture*, 110:129–137, 2024. doi: 10.1016/j.gaitpost.2024.03.009 2
- [24] Y. Long, T. Wang, X. Liu, Y. Li, and D. Tao. Toward accurate cybersickness prediction in virtual reality: A multimodal physiological modeling approach. *Sensors*, 25(18):5828, 2025. doi: 10.3390/s25185828 2
- [25] M. R. Mahmud, A. Cordova, and J. Quarles. Auditory, vibrotactile, or visual? investigating the effective feedback modalities to improve standing balance in immersive virtual reality for people with balance impairments due to type 2 diabetes. In *2023 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 573–582, 2023. doi: 10.1109/ISMAR59233.2023.00072 2
- [26] M. R. Mahmud, A. Cordova, and J. Quarles. Multimodal feedback methods for advancing the accessibility of immersive virtual reality for people with balance impairments due to multiple sclerosis. *IEEE Transactions on Visualization and Computer Graphics*, 2024. doi: 10.1109/TVCG.2024.3456189 2
- [27] J. Martinez, R. Hossain, J. Romero, and J. J. Little. A simple yet effective baseline for 3d human pose estimation. In *ICCV*, 2017. 4
- [28] F. Quijoux, A. Nicolai, I. Chair, I. Bargiotas, D. Ricard, A. Yelnik et al. A review of center of pressure (COP) variables to quantify standing balance in elderly people: Algorithms and open-access code. *Physiological Reports*, 9(22):e15067, 2021. doi: 10.14814/phy2.15067 1, 2
- [29] A. N. Ramaseri-Chandra and H. Reza. Predicting cybersickness using machine learning and demographic data in virtual reality. *Electronics*, 13(7):1313, 2024. doi: 10.3390/electronics13071313 2
- [30] N. F. Ribeiro, A. Veloso, H. Pires, and C. P. Santos. Immersive ecological virtual environment for inducing balance disturbances. *IEEE Transactions on Visualization and Computer Graphics*, 31(12):10372–10383, 2025. doi: 10.1109/TVCG.2025.3598464 2, 8
- [31] S. B. Richmond, B. W. Fling, H. Lee, and D. S. Peterson. The assessment of center of mass and center of pressure during quiet stance: Current applications and future directions. *Journal of Biomechanics*, 123:110485, 2021. 8
- [32] O. Rosiak, N. Pietrzak, A. Szczesna, I. Kulczak, G. Zwoliński, D. Kamińska et al. The effect of immersive virtual reality on balance: an exploratory study on the feasibility of head-mounted displays for balance evaluation. *Scientific Reports*, 14(1):3481, 2024. 1, 2, 8
- [33] P. J. Rousseeuw and C. Croux. Alternatives to the median absolute deviation. *Journal of the American Statistical Association*, 88(424):1273–1283, 1993. doi: 10.1080/01621459.1993.10476408 3
- [34] S. Schedler, K. Gramann, M. W. Hill, and T. Muehlbauer. Balance performance of healthy young individuals in real versus matched virtual environments: a systematic scoping review. *Frontiers in Human Neuroscience*, 18:1422581, 2024. 1, 2
- [35] X. e. a. Shi. Convolutional lstm network: A machine learning approach for precipitation nowcasting. In *NeurIPS*, 2015. 7
- [36] U. Tasnim, R. Islam, K. Desai, and J. Quarles. Investigating personalization techniques for improved cybersickness prediction in virtual reality environments. *IEEE Transactions on Visualization and Computer Graphics*, 30(5):2368–2378, 2024. doi: 10.1109/TVCG.2024.3372122 2
- [37] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996. 7
- [38] D. Tome, T. Alldieck, P. Peluse, G. Pons-Moll, L. Agapito, H. Badino et al. Selfpose: 3d egocentric pose estimation from a headset mounted camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):6794–6806, 2020. 1
- [39] D. Tome, P. Peluse, L. Agapito, and H. Badino. xr-egopose: Egocentric 3d human pose from an hmd camera. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7728–7738, 2019. 3, 4
- [40] S. Ş. Tongar and Ç. Yazici-Mutlu. How virtual reality is impacting balance: An examination of postural stability. *Journal of Bodywork and Movement Therapies*, 38:81–85, 2024. 1
- [41] S. Umeyama. Least-squares estimation of transformation parameters between two point patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(4):376–380, 1991. 5, 8
- [42] J. Wang, L. Liu, W. Xu, K. Sarkar, D. Luvizon, and C. Theobalt. Estimating egocentric 3d human pose in the wild with external weak supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15153–15163, 2022. 4
- [43] J. Wang, D. Luvizon, W. Xu, L. Liu, K. Sarkar, and C. Theobalt. Scene-aware egocentric 3d human pose estimation. In *Proceedings of the*

*IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13031–13040, 2023. [4](#)

- [44] K. Wantanajittikul, C. Wiboonsuntharakoon, B. Chuatrakoon, and S. Kongsawasdi. Application of machine learning to predict trajectory of the center of pressure (cop) path of postural sway using a triaxial inertial sensor. *The Scientific World Journal*, 2022(1):9483665, 2022. [1](#), [2](#)
- [45] W. Weng and X. Zhu. Inet: convolutional networks for biomedical image segmentation. *Ieee Access*, 9:16591–16603, 2021. [4](#)
- [46] M. W. Wittstein, A. Crider, S. Mastrocola, and M. Guerenza Gonzalez. Use of virtual reality to assess dynamic posturography and sensory organization: Instrument validation study. *JMIR Serious Games*, 8(4):e19580, 2020. doi: 10.2196/19580 [2](#), [8](#)
- [47] C. Yang, A. Tkach, S. Hampali, L. Zhang, E. J. Crowley, and C. Keskin. Egoposeformer: A simple baseline for stereo egocentric 3d human pose estimation. In *European conference on computer vision*, pp. 401–417. Springer, 2024. [4](#)
- [48] A. Zhao, C. Tang, L. Wang, Y. Li, M. Dave, L. Tao et al. Egobody3m: Egocentric body tracking on a vr headset using a diverse dataset. In *European Conference on Computer Vision*, pp. 375–392. Springer, 2024. [1](#), [3](#)
- [49] D. Zhao, Z. Wei, J. Mahmud, and J.-M. Frahm. Egoglass: Egocentric-view human pose estimation from an eyeglass frame. In *2021 International Conference on 3D Vision (3DV)*, pp. 32–41, 2021. doi: 10.1109/3DV53792.2021.00014 [4](#)