

A Reproducible Skeleton Baseline and Multimodal Fusion Protocol for Movement-Pattern Atypicality Analysis Using CZU-MHAD

Hao Feng¹, Yanlin Xu^{2*}, and Yubing Ren³

¹School of Physical Education, Liaoning Finance and Trade College, Liaoning, 163000, China.

²Northeast Petroleum University Athletic Department school, Northeast Petroleum University, Liaoning, 163000, China.

³Department of Physical Education, Yangquan No.1 High School Yangquan, 045000, China.

*Corresponding Author: Yanlin Xu. Email: 418692869@qq.com

Received: May 11, 2026 Accepted: August 28, 2026

Abstract: Public multimodal action datasets are primarily designed for action recognition rather than the assessment of within-action movement patterns. This study presents a reproducible skeleton baseline and a specified multimodal fusion protocol using the Changzhou University Multi-modal Human Action Dataset (CZU-MHAD). The empirical analysis processed 357 skeleton recordings, with participant-holdout evaluation conducted on 294 recordings representing 15 shared action categories. Skeleton trajectories were anatomically normalised and evaluated using kinematic logistic regression, principal-component logistic regression, and a temporal convolutional network (TCN). The kinematic baseline achieved mean accuracy of 0.817 and macro-F1 of 0.813 across the two directional participant holdouts, outperforming the TCN, which achieved accuracy of 0.624 (directional variation: 0.063 and macro-F1 of 0.601 (directional variation: 0.059. Action-conditional embedding distances were additionally used to describe movement-pattern atypicality, without interpreting atypicality as incorrect technique or clinical abnormality. A lightweight multimodal pilot provided preliminary evidence that inertial information can add participant-independent discriminative signal. The study therefore establishes a reproducible empirical baseline while providing a clearly bounded protocol for subsequent depth-skeleton-inertial fusion evaluation).

Keywords: multimodal sensor fusion; skeleton-based action recognition; movement-pattern atypicality; inertial measurement units; participant-holdout validation; CZU-MHAD.

1. Introduction

1.1. Background and Problem Statement

Artificial intelligence has progressively expanded from narrow pattern-recognition tasks toward applications that interpret continuous physical signals about human behaviour, a trajectory that recent overviews of the field's trajectory describe as moving from isolated perception tasks toward integrated, multimodal sensing and decision support (Jiang et al., 2022). Within this trajectory, the assessment of human posture and movement quality has moved from purely observational, expert-led evaluation toward technology-assisted measurement, in which depth cameras, markerless motion capture, and wearable inertial sensors provide quantitative descriptions of how a body moves (Edriss et al., 2024). This shift has been accelerated by advances in deep learning for skeleton-based action recognition, where spatio-temporal sequence models extract discriminative representations of joint trajectories for activity classification, rehabilitation monitoring, and sports performance analysis (Cossich et al., 2023; Debnath et al., 2022). At the

same time, wearable inertial measurement units (IMUs) have become inexpensive and reliable enough to be deployed alongside cameras, enabling multimodal systems that combine the spatial richness of vision-based skeleton tracking with the high-frequency kinematic detail of accelerometers and gyroscopes (Chen et al., 2024; Junaid et al., 2022). A structural problem nevertheless persists in the literature on posture correction and movement-quality assessment. Many publicly available multimodal action datasets, including widely used benchmarks, are constructed for action-identity recognition—that is, for distinguishing which action was performed—rather than for assessing how well a given action was performed (Chao et al., 2022; Guo et al., 2024). Studies that address posture correction therefore tend either to rely on small, purpose-built clinical samples with limited modality coverage, which restricts external validity (He et al., 2024), or to repurpose action-recognition datasets while treating classification accuracy as a proxy for movement quality, without articulating the theoretical or empirical basis for that substitution (Chariar et al., 2023).

1.2. Research Gap

Four interrelated gaps motivate this study. A construct-definition gap exists in how multimodal sensor data are translated into interpretable measures of within-action kinematic variation. Autoencoder-based anomaly scores can flag deviations from a reference movement (Chariar et al., 2023), yet no widely shared framework specifies which combination of depth, skeletal, and inertial indicators constitutes a transparent multidimensional description of movement-pattern atypicality or how the resulting indicators should be interpreted in the absence of expert correctness labels. An empirical gap follows from the observation that few studies report transparent descriptive profiling of a tri-modal dataset prior to model training; research typically proceeds directly to end-to-end model accuracy without establishing the structural properties of the underlying data (Guo et al., 2024). A methodological gap concerns fusion strategy. Multimodal integration has proved useful in adjacent domains such as cognitive-workload assessment (Zakeri et al., 2023), but token-level cross-attention has not yet been evaluated for skeleton-inertial-depth movement-pattern atypicality using CZU-MHAD. A contextual gap, finally, arises because most evidence on AI-assisted rehabilitation and movement assessment originates from clinical trial settings within specific national health systems (He et al., 2024) or from sport-specific kinematic studies (Edriss et al., 2024), leaving open how a general-purpose, laboratory-collected multimodal action dataset of the kind produced in Chinese university laboratories can be repurposed, with appropriate caveats, for movement-quality research.

1.3. Why the Topic Matters

Movement-quality assessment sits at the intersection of three growing application domains. In rehabilitation and tele-rehabilitation, AI-assisted posture estimation has been evaluated in randomised controlled trial settings for older adults with sarcopenia (He et al., 2024), and scoping reviews indicate that multimodal datasets combining video, depth, and wearable streams are increasingly the norm in community-based rehabilitation technology (Abedi et al., 2024). In sports performance analysis, wearable and vision-based systems increasingly feed real-time feedback loops for athletes and coaches, supported by energy-autonomous sensing platforms that integrate with digital-intelligence systems (Cossich et al., 2023; Wang et al., 2025). In occupational and educational ergonomics, multimodal monitoring of posture and cognitive-physical workload can inform human-machine system design (Zakeri et al., 2023). Across these domains, a common technical requirement is a system that ingests synchronised motion data and produces an interpretable description of within-action movement patterns rather than only an action label. Establishing how a representative multimodal dataset behaves — its joint-level variability, its inter-sensor structure, and its amenability to fusion-based analysis — is a necessary precursor to building such systems responsibly.

1.4. Dataset Suitability and Scope of the Present Study

CZU-MHAD was introduced by Chao et al. (2022) as a synchronised multimodal action-recognition benchmark comprising 22 actions performed by five male participants. Each benchmark recording contains Kinect V2 depth frames, 25-joint three-dimensional skeleton trajectories, and measurements from ten MPU9250 inertial sensors. The original dataset description is available at <https://doi.org/10.1109/JSEN.2022.3150225>, while the developers' official repository and loading code are available at

https://github.com/yujmo/CZU_MHAD. The empirical component of the present study focuses on the skeleton modality. A total of 357 skeleton recordings were processed. For participant-independent classification, the sample was restricted to 294 recordings representing the 15 action categories shared by participants qyh and zyh. The remaining recordings were used only for data-quality assessment. This design provides real participant-holdout evidence while maintaining a clear distinction between the empirically evaluated skeleton models and the proposed depth-skeleton-inertial fusion extension. CZU-MHAD does not contain clinician-, coach-, or expert-assigned correctness scores. Consequently, action classification and action-conditional embedding distance are interpreted as measures of action identity and kinematic atypicality, respectively. Neither is treated as a validated measure of movement quality or clinical abnormality.

1.5. Objectives and Research Questions

This study has four objectives. First, it establishes a reproducible preprocessing and quality-control procedure for CZU-MHAD skeleton trajectories. Second, it evaluates kinematic, principal-component, and temporal-convolutional skeleton models under directional participant-holdout testing. Third, it conducts a deliberately lightweight five-participant multimodal ablation to determine whether IMU and depth-stream information adds participant-independent discriminative signal beyond skeleton features. Fourth, it specifies the anatomical mappings, interpretable measures, temporal encoders, cross-attention mechanism, and reference-distribution procedure required for subsequent full depth-skeleton-inertial evaluation.

RQ1. What preprocessing and data-quality requirements arise when CZU-MHAD skeleton trajectories are prepared for participant-independent modelling?

RQ2. How effectively do kinematic, principal-component, and temporal-convolutional skeleton models generalise between participants across the 15 shared action categories?

RQ3. In a lightweight five-participant multimodal pilot, does adding inertial or depth-stream information improve participant-independent discrimination relative to skeleton-only features?

RQ4. What anatomical alignment, temporal fusion, and action-conditional reference procedures are required for a full depth-skeleton-inertial implementation?

1.6. Contributions and Organisation of the Paper

The study contributes three connected elements. The first is an empirical skeleton baseline comprising data-quality assessment, anatomically normalised preprocessing, interpretable kinematic features, participant-holdout classification, seed and kernel sensitivity analyses, and action-conditional embedding distances. The second is a lightweight five-participant modality-contribution experiment comparing skeleton, skeleton plus IMU, skeleton plus a file-level depth complexity proxy, and all three inputs under leave-one-participant-out evaluation. The third is a fully specified temporal tri-modal extension that defines ten IMU-to-skeleton correspondences, modality-specific temporal encoders, token-level cross-attention, and a participant-independent reference procedure. The lightweight ablation provides preliminary modality evidence but does not constitute training or validation of the complete cross-attention architecture. This separation is methodologically important. The primary empirical benchmark concerns skeleton data, while a separate lightweight five-participant experiment provides exploratory evidence on modality contribution. The complete temporal depth, inertial, and cross-attention pathways remain a specified architecture for subsequent model-level evaluation. Section 2 positions the approach in relation to action recognition, rehabilitation assessment, and multimodal fusion. Section 3 describes the public data source, preprocessing, empirical models, equations, validation design, and proposed tri-modal architecture. Section 4 reports the real skeleton results. Section 5 discusses their implications and limitations, and Section 6 concludes the study.

2. Related Work and Methodological Positioning

2.1. Multimodal Action Recognition: From Identity to Quality

Skeleton-based action-recognition systems increasingly analyse spatio-temporal joint-coordinate sequences, while multimodal benchmarks combine depth, skeleton, and inertial signals to exploit their complementary measurement properties (Chao et al., 2022; Chen et al., 2015). Depth data provide global

motion information but can be affected by occlusion and viewpoint; skeleton tracking provides interpretable joint coordinates but can become noisy during fast or self-occluded movements; and wearable inertial measurement units provide local acceleration and angular-velocity signals without full-body spatial context (Chen et al., 2024; Debnath et al., 2022; Junaid et al., 2022). CZU-MHAD was designed to address layout and sensor-count limitations of earlier multimodal datasets by combining a single Kinect V2 unit with ten wearable IMUs positioned at the principal motion joints (Chao et al., 2022). A recurring pattern in this literature is that reported results concern action-identity classification accuracy — the proportion of test sequences correctly labelled with the performed action — rather than within-action quality variation (Guo et al., 2024). Benchmarking studies of micro-action recognition show that even fine-grained, low-intensity behaviours can be classified with high accuracy when sufficient multimodal data are available (Guo et al., 2024), which demonstrates the feasibility of fine-grained discrimination but does not by itself establish that the same representations capture pedagogically meaningful quality differences. Movement-quality assessment instead requires representations sensitive to within-action variation, such as joint-angle deviation, asymmetry, and smoothness, in addition to between-action variation, and this is the conceptual gap that the present framework addresses.

Recent skeleton-recognition research has increasingly moved from fixed graph or convolutional representations toward topology-aware transformer architectures and cross-token interaction. These methods provide stronger representations of long-range joint dependencies than simple temporal convolution alone. In the present study, such work identifies the type of stronger skeleton-only benchmark required for fuller validation; it is not used as evidence that action recognition and movement-pattern atypicality are equivalent tasks. Recent work published in 2024–2025 sharpens this distinction. Reviews of AI-driven virtual rehabilitation describe increasing use of multimodal video, depth, and wearable streams but continuing heterogeneity in movement-quality labels (Abedi et al., 2024). Sensor-focused reviews document rapid improvements in AI-enabled wearable measurement (Chen et al., 2024), while micro-action benchmarks show that fine-grained action identity can be recognized without establishing within-action execution quality (Guo et al., 2024). Controlled posture-estimation research demonstrates clinical feasibility only when outcomes and interventions are explicitly validated (He et al., 2024), and recent smart-sports sensing work highlights the growing integration of wearable signals with digital performance systems (Wang et al., 2025). Together, these studies support multimodal measurement while reinforcing the need to distinguish action classification, kinematic atypicality, and validated movement quality. Although these studies originate outside movement analysis, they provide relevant methodological lessons for representation design. Pourhaji et al. (2022) showed in short-term electricity-price forecasting that monthly and seasonal clustering strategies can materially alter LSTM prediction error, illustrating that data grouping should be treated as part of model design rather than as a neutral preprocessing choice. Aghayengejeh et al. (2025) reviewed the reciprocal relationship between transformer representation learning and clustering, showing that learned representations can facilitate structure discovery while clustering can contribute to interpretability and model development. In the present setting, these methodological observations support prespecified participant and action grouping and reinforce the distinction between structure in a learned representation and externally validated movement quality.

2.2. Posture Correction and AI-Assisted Rehabilitation

Within the rehabilitation literature, AI-driven posture estimation has begun to be evaluated in controlled clinical settings. He et al. (2024) report a randomised controlled trial of a tele-rehabilitation approach using 3D human posture estimation for older individuals with sarcopenia, indicating that vision-based posture estimation can be integrated into a structured intervention with measurable outcomes. Scoping reviews of AI-driven virtual rehabilitation similarly note that multimodal datasets combining video, depth, and wearable sensor streams are increasingly common in community-based rehabilitation technology, but also highlight substantial heterogeneity in how movement quality or correctness is operationalised across studies (Abedi et al., 2024). Reviews of computer-vision-based physical rehabilitation and assessment approaches likewise suggest that pose-estimation pipelines are mature enough for activity recognition and gross movement

tracking, but that fine-grained quality scoring remains an open methodological problem, particularly for patients with motor impairments where skeleton extraction is less reliable (Debnath et al., 2022).

In the sports-technology literature, the AI Trainer approach of Chariar et al. (2023) is particularly relevant: an autoencoder is trained on correctly performed squats so that the reconstruction error for a new sequence functions as an anomaly score, interpreted as a posture-correction signal. This is one of the few examples in the cited literature of an explicit operationalisation of correctness as a deviation-from-normal-pattern measure rather than as an action-classification label. The present study adopts the same reference-deviation logic for real skeleton embeddings and links it to the action-conditional mCD defined in Section 3.8. The specified tri-modal extension would apply the same reference logic to a fused depth-skeleton-inertial representation. A related strand of rehabilitation engineering, reviews of AI-based methods for exoskeleton-assisted lower-limb rehabilitation, similarly emphasises that kinematic outputs derived from wearable and robotic sensors need to be decomposed into clinically interpretable sub-components rather than reported as a single device-level performance metric (Coser et al., 2024), reinforcing the indicator-decomposition logic presented in Table 4.

2.3. Wearable Sensors, Sensor Fusion, and Sensing Technology

The technical feasibility of the proposed framework rests on the maturity of wearable inertial sensing and sensor-fusion methods. Reviews of AI-driven sensing technology document rapid improvements in the accuracy, sensitivity, and adaptability of machine-learning-enabled sensors across domains (Chen et al., 2024), while reviews of wearable sensors in healthcare delivery emphasise a shift from hospital-centred to personal, portable, sensor-centred monitoring (Junaid et al., 2022). Systematic reviews of Internet-of-Things and machine-learning approaches to movement-disorder monitoring similarly report that wearable inertial sensors are now used to track motor symptoms continuously outside clinical settings, with kinematic features extracted from accelerometer and gyroscope signals serving as the primary basis for monitoring disease progression (Giannakopoulou et al., 2022), which supports treating the ten-IMU configuration in CZU-MHAD as representative of a sensing architecture already deployed for continuous movement monitoring. Inertial sensors have also been validated for gait and posture-related parameters in older adults, where surveys of significant gait parameters derived from wearable inertial sensors identify which kinematic quantities — step time, walking speed, joint angular velocity — carry the most diagnostic information for fall-risk and frailty assessment (Ruiz-Ruiz et al., 2021); the present framework's smoothness and amplitude indicators draw on this same family of kinematic quantities. In the sports domain, energy-autonomous wearable sensing systems are increasingly integrated with digital-intelligence platforms for performance analytics (Wang et al., 2025), and point-of-care wearable sensing more broadly is converging with AI-based interpretation (Yammouri & Ait Lahcen, 2024), reinforcing the trend toward multimodal, continuously monitored movement data.

Fusion strategy is the second technical pillar. Outside the movement domain, multimodal fusion of neural, subjective, and behavioural measures has been used to assess cognitive workload in smart-factory settings, demonstrating that heterogeneous signal types can be combined into a single workload index through carefully specified fusion pipelines (Zakeri et al., 2023). Within the human-action-recognition literature itself, the precedent most directly relevant to this protocol is the UTD-MHAD dataset, which paired a single Kinect camera with a single wrist- or thigh-worn inertial sensor and reported early fusion experiments combining depth and inertial features for action recognition (Chen et al., 2015); CZU-MHAD extends this single-camera, single-inertial-sensor design to ten wearable inertial measurement units, comprising eight exact limb-joint correspondences and two trunk-region correspondences. This arrangement provides the structural basis for the anatomical mapping in Section 3.4.1 and the specified fusion architecture in Section 3.7. These studies provide methodological precedents for the fusion architecture specified in Section 3.7, although they do not validate within-action movement quality.

2.4. Quality Assessment, Interpretability, and Kinematic Indicators

Two bodies of work inform the interpretable kinematic measures specified in this study. The KIMORE dataset links Kinect-derived range of motion, joint angles, and trajectory smoothness to clinician-assigned

exercise scores (Capecci et al., 2019). It therefore illustrates the type of correctness-labelled validation identified as a future requirement in Section 5.4. The University of Idaho Physical Rehabilitation Movement Data set provides repeated full-body joint trajectories and angles for rehabilitation exercises, supporting the use of repetition-consistency measures (Vakanski et al., 2018). Reviews of dynamic and kinematic assessment further identify joint angles, angular velocities, accelerations, and temporal derivatives as common movement-analysis variables (Edriss et al., 2024). These sources support the measures operationalised in Table 4, while the present study does not treat them as direct indicators of clinical correctness. Collectively, these studies support four interpretable measurement families: joint-angle consistency, trajectory smoothness, bilateral coordination, and movement amplitude. The present study uses these measures to describe execution patterns and to support interpretation of the skeleton representation. They are not treated as direct indicators of clinical correctness.

2.5. Critical Synthesis and Positioning of the Present Study

Three tensions emerge from this review. First, there is a tension between the technical sophistication of multimodal action-recognition systems, which routinely achieve high classification accuracy on fine-grained action and micro-action categories (Guo et al., 2024), and the relative scarcity of validated movement-quality indicators that exploit the same multimodal richness. Second, there is a tension between clinical rehabilitation studies, which validate posture-estimation interventions but often rely on a single vision-based modality (He et al., 2024), and adjacent-domain sensor-fusion studies that do not evaluate within-action movement quality (Zakeri et al., 2023). Third, there is a tension between the explicit deviation-based operationalisation of correctness in single-modality systems such as AI Trainer (Chariar et al., 2023) and the absence of an equivalent, explicit operationalisation for multimodal systems.

Table 1. Methodological comparison of representative action-recognition and atypicality-analysis approaches.

Approach	Inputs	Primary task	Model/fusion	Anatomical mapping	Interpretable output	Status in present comparison
UTD-MHAD	Depth + one IMU	Action identity	Feature-level fusion	Limited	No quality decomposition	Published multimodal comparator
Original CZU-MHAD study	Depth + skeleton + 10 IMUs	Action identity	Multimodal classification	Sensor placement described	No within-action decomposition	Dataset benchmark
AI Trainer	Pose sequence	Squat anomaly detection	Autoencoder	Exercise-specific	Reconstruction error	Single-action deviation comparator
Present empirical baseline	Skeleton	15-class action identity and action-conditional atypicality	Kinematic logistic, PCA-logistic, and TCN	25-joint skeleton	Path, displacement, angle, speed, and symmetry descriptors	Evaluated on 294 real recordings
Proposed tri-modal extension	Depth + skeleton + 10 IMUs	Action identity and kinematic atypicality	Token-level cross-attention	Eight exact limb correspondences and two trunk proxies	Waveform, smoothness, symmetry, amplitude, and	Forward-pass checked on two repository

	dynamic acceleration	examples; predictive performance not evaluated
--	-------------------------	--

Note. UTD-MHAD is based on Chen et al. (2015); the original CZU-MHAD approach is based on Chao et al. (2022); AI Trainer is based on Chariar et al. (2023).

The methodological contribution is the connection between reproducible skeleton preprocessing, participant-independent evaluation, explicit sensor-to-joint correspondence, interpretable kinematic measures, and an action-conditional robust reference. The real skeleton baseline establishes the empirical point of comparison. The proposed depth and inertial pathways extend this baseline but are not presented as predictively evaluated components.

The present study addresses this gap by separating what is empirically evaluated from what is methodologically specified. The skeleton component is tested using real participant-holdout data. The tri-modal component defines how the depth and inertial streams would be aligned and fused in a subsequent analysis. This separation allows the empirical contribution to be assessed without treating a repository-example implementation check as evidence of predictive multimodal benefit.

2.6. Motor-Control, Biomechanical, and Interpretability Foundations

The indicator set is grounded in motor-control and biomechanical measurement. Normalised jerk represents movement smoothness; homologous left–right differences represent bilateral coordination only for actions that are symmetric by design; joint displacement and dynamic acceleration represent task demand; and repetition-to-reference waveform deviation represents execution consistency. Interpretability is therefore treated as an engineering property: every output must be traceable to a named action, joint or sensor location, and computationally defined indicator (Górriz et al., 2023).

This grounding does not imply that kinematic atypicality equals poor performance. Feedback, learning, self-regulation, and clinical improvement require separate outcomes and cannot be inferred from CZU-MHAD.

2.7. Study Framework

Figure 1 distinguishes the empirically evaluated skeleton pathway from the proposed tri-modal extension. The skeleton pathway produces action-classification probabilities, a 64-dimensional embedding, and action-conditional mCD values. The proposed extension adds aligned depth and inertial representations through token-level cross-attention. This visual boundary is maintained throughout the methods, results, and discussion.

3. Data and Methods

3.1. Research Design

This study combines an empirical skeleton analysis with a methodological specification for subsequent tri-modal fusion. The empirical component evaluates three skeleton-based action-classification approaches under directional participant holdout: an interpretable kinematic logistic model, a principal-component logistic model, and a temporal convolutional network. It also examines action-conditional distances derived from real TCN embeddings. The specification component defines how depth and inertial signals would later be aligned with the skeleton sequence and integrated through token-level cross-attention. A separate implementation check uses the two matched multimodal examples distributed with the official repository to verify loading, temporal alignment, token dimensions, and a complete cross-attention forward pass. Predictive multimodal performance is not inferred from that check. A separate lightweight five-participant feature-level ablation evaluates preliminary modality contribution using branch-level skeleton, inertial, and depth records, without training the full temporal cross-attention architecture.

All model preprocessing and dimensionality-reduction parameters were estimated from the training participant only. The held-out participant was excluded from feature scaling, principal-component estimation, model fitting, early stopping, and reference-distribution construction.

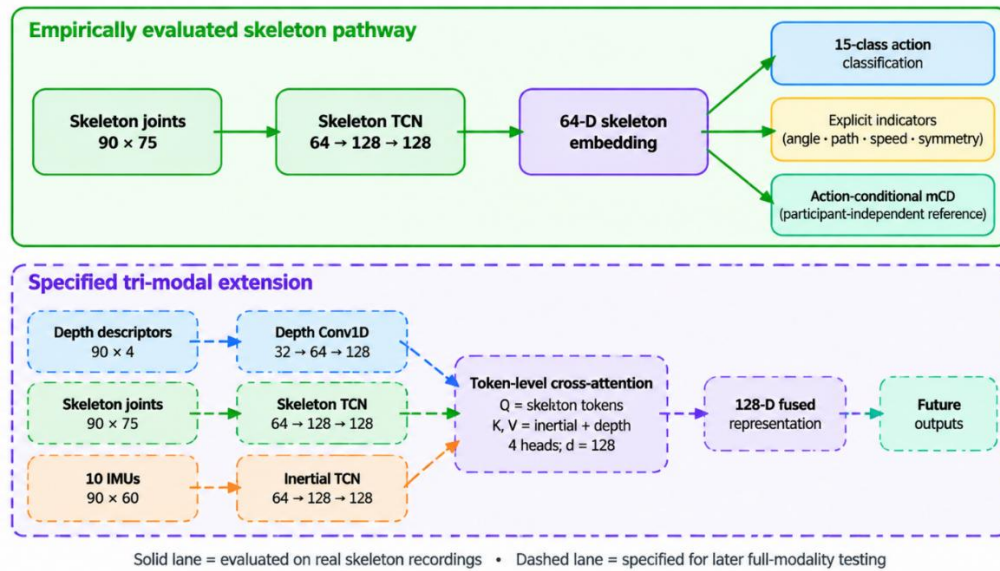


Figure 1. Empirically evaluated skeleton pathway and specified tri-modal fusion extension.

The solid boundary identifies the skeleton components evaluated using real CZU-MHAD recordings. The dashed boundary identifies the depth, inertial, and cross-attention components specified for future tri-modal evaluation.

3.2. Dataset Source and Analysis Sample

Data were obtained from the official CZU-MHAD repository maintained by the dataset developers at https://github.com/yujmo/CZU_MHAD. Dataset structure, sensor placement, action definitions, and the published benchmark design were interpreted using Chao et al. (2022), available at <https://doi.org/10.1109/ISEN.2022.3150225>.

In addition to the two packaged .mat triplets provided in the master branch for loading demonstrations, the official CZU-MHAD repository distributes raw modality-specific recordings through its skeleton, sensors, and depth branches. The common participant set available across these branch-level records includes cx, cyy, myj, qyh, and zyh. The two packaged examples were retained for the end-to-end cross-attention compatibility check reported in Section 3.7.2, whereas branch-level raw records were used for the separate lightweight multimodal ablation reported in Section 3.7.3. The primary 294-sequence skeleton benchmark and the full Table 7 temporal cross-attention architecture remain analytically distinct from this pilot. The published benchmark comprises 880 synchronised samples generated by five participants performing 22 actions eight times each. Each sample contains a Kinect V2 depth stream, a 25-joint three-dimensional skeleton sequence, and ten inertial streams containing three-axis acceleration and three-axis angular velocity. The remaining participants were not included in the primary skeleton benchmark because they did not satisfy the same combination of shared-action availability and preprocessing consistency required for the controlled directional participant-holdout evaluation. Their available recordings were retained for repository-level modality inspection and the lightweight multimodal analysis. The empirical skeleton analysis processed 357 recordings from participants qyh and zyh. The selection of qyh and zyh was determined by the consistently processable skeleton subset used for the primary empirical baseline and by the availability of overlapping action categories required for directional participant-holdout evaluation; it was not based on model performance. Cross-participant classification was restricted to the 15 actions represented for both participants, producing an analytic sample of 294 recordings. Each recording was treated as one participant-action-repetition unit.

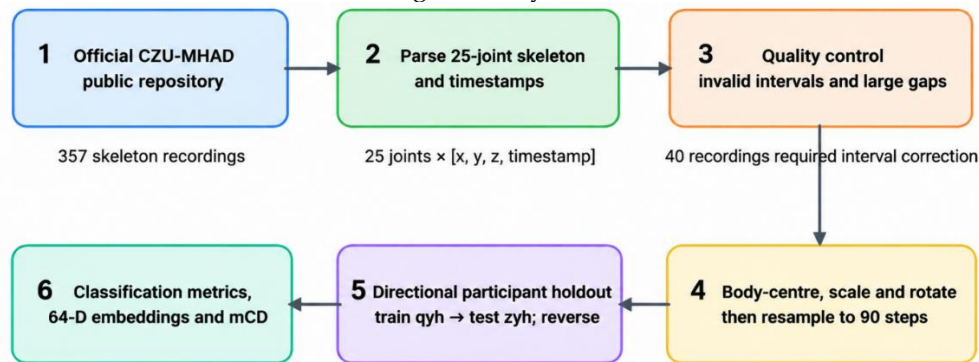
Table 2. CZU-MHAD source characteristics and empirical analysis sample

Property	Description
Original dataset	Changzhou University Multi-modal Human Action Dataset
Dataset article	Chao et al. (2022)
Article DOI	https://doi.org/10.1109/JSEN.2022.3150225
Official repository	https://github.com/yujmo/CZU_MHAD
Published benchmark	880 synchronised samples
Published participants	5 male participants
Action categories	22
Published repetitions	8 per participant–action combination
Modalities	Kinect V2 depth, 25-joint skeleton, and 10 wearable IMUs
Skeleton sampling rate	Approximately 30 Hz
Skeleton recordings processed	357
Participants in empirical baseline	qyh and zyh
Shared action categories	15: a1–a9 and a17–a22
Participant–holdout sample	294 recordings
Empirical outcomes	Action-classification performance and action-conditional skeleton-embedding distance

3.3. Skeleton Preprocessing and Quality Control

Each skeleton file contains a $\text{frames} \times 100$ matrix. Every frame comprises 25 consecutive records containing the x, y, and z coordinates of one Kinect joint and its timestamp. Coordinates were translated to the SpineBase joint and scaled by the recording-specific median SpineBase–SpineShoulder distance. A recording-level anatomical reference frame was estimated from the first ten frames using the hip axis, trunk axis, and their cross-product. This transformation reduced camera-position variation while preserving meaningful body turning.

Timestamp intervals were checked before resampling. Non-positive intervals and intervals greater than five times the recording-specific median positive interval were treated as discontinuities and replaced by the recording median interval. Each corrected trajectory was then linearly resampled to 90 time steps. The 90-step target was chosen as a pragmatic fixed-length phase representation that preserves the complete onset-to-offset trajectory while balancing temporal resolution against input dimensionality across recordings of substantially different original lengths. Given that the observed skeleton sequences ranged from 40 to 203 frames, with a median of 123 frames, this common length provided a consistent input for participant–holdout modelling without requiring sequence cropping or zero-padding. Coordinate statistics used for model standardisation were estimated from the training data only.



294 recordings across 15 shared actions. Scaling, PCA, early stopping and reference construction use training data only.

The held-out participant never contributes to preprocessing estimates or the action-specific reference distribution.

Figure 2. Empirical skeleton preprocessing and participant–holdout evaluation workflow.

Public CZU-MHAD skeleton recordings were parsed, checked for timestamp discontinuities, transformed to a body-centred reference frame, resampled to 90 time steps, and evaluated through directional participant holdout. Preprocessing parameters were estimated from the training participant only.

3.4. Anatomical Mapping and Variable Operationalisation

3.4.1. Anatomical Sensor-to-Joint Mapping

The anatomical mapping was verified against the official CZU-MHAD loading implementation and dataset documentation. No independent anatomical calibration measurements were available; therefore, trunk mappings are interpreted as anatomical-region proxies rather than exact physical joint correspondences. The anatomical mapping was verified against the official CZU-MHAD loading implementation and dataset documentation. No independent anatomical calibration measurements were available; therefore, trunk sensor mappings are interpreted as anatomical-region proxies rather than exact physical joint correspondences. The released inertial-loading code establishes the exact array order of the ten MPU9250 units as chest, abdomen, left elbow, left wrist, right elbow, right wrist, left knee, left ankle, right knee, and right ankle. The official array order was retained throughout the analysis. The eight limb sensors have direct joint-level correspondences in the Kinect skeleton, whereas the chest and abdomen sensors are represented by the nearest trunk-region skeleton joints. Table 3 maps each inertial sensor to its corresponding Kinect joint or anatomical region. For the eight limb sensors, each maps to a single, well-defined Kinect skeleton joint (left/right wrist, elbow, knee, and ankle), and the term ‘co-located’ is used in its literal sense for these eight. For the chest and abdomen sensors, no single Kinect skeleton joint corresponds exactly to the sensor's anatomical placement; the Kinect V2 skeleton model's spine-shoulder and spine-mid joints are the nearest anatomical proxies for the chest and abdomen sensors respectively, but these represent the same general trunk region rather than the identical physical point on the body where the inertial sensor is worn. Accordingly, the protocol distinguishes exact joint correspondence for the eight limb sensors from anatomical-region correspondence for the chest and abdomen sensors. Cross-modal analyses involving the chest and abdomen channels are therefore interpreted regionally rather than as point-wise joint comparisons.

Table 3. Anatomical mapping between the ten MPU9250 inertial sensors and the 25-joint Kinect V2 skeleton model, including the joint-angle triplets used in the kinematic calculations. Sensors 1 and 2 represent trunk-region proxies; sensors 3–10 have direct joint-level correspondences.

IMU no.	Actual array position	Corresponding Kinect joint/region	Correspondence type	Joint-angle triplet
1	Chest	SpineShoulder region	Anatomical-region proxy	Neck–SpineShoulder–SpineMid
2	Abdomen	SpineMid region	Anatomical-region proxy	SpineShoulder–SpineMid–SpineBase
3	Left elbow	ElbowLeft	Exact joint correspondence	ShoulderLeft–ElbowLeft–WristLeft
4	Left wrist	WristLeft	Exact joint correspondence	ElbowLeft–WristLeft–HandLeft
5	Right elbow	ElbowRight	Exact joint correspondence	ShoulderRight–ElbowRight–WristRight
6	Right wrist	WristRight	Exact joint correspondence	ElbowRight–WristRight–HandRight

7	Left knee	KneeLeft	Exact joint correspondence	HipLeft–KneeLeft–AnkleLeft
8	Left ankle	AnkleLeft	Exact joint correspondence	KneeLeft–AnkleLeft–FootLeft
9	Right knee	KneeRight	Exact joint correspondence	HipRight–KneeRight–AnkleRight
10	Right ankle	AnkleRight	Exact joint correspondence	KneeRight–AnkleRight–FootRight

This mapping satisfies the anatomical-alignment requirement under RQ3. Eight inertial channels can be compared with skeleton-derived signals at the corresponding elbow, wrist, knee, or ankle joint. The chest and abdomen channels can be compared only with same-region trunk signals. The remaining fifteen skeleton joints have no direct inertial counterpart but remain available for skeleton- and depth-based analysis.

3.4.2. Variable Operationalisation

Table 4 separates the variables used in the empirical skeleton baseline from the kinematic measures specified for subsequent full-modality evaluation. CZU-MHAD contains no expert correctness labels; therefore, no movement-quality outcome is directly measured. The empirical baseline uses skeleton coordinates, action labels, participant identity, interpretable skeleton summaries, TCN embeddings, and action-conditional mCD. Joint-angle waveform consistency, normalised jerk, dynamic inertial acceleration, and full multimodal fusion are retained as specified extension measures and are not presented as empirical outputs of the current analysis.

Table 4. Operational definitions, source modalities, computational measures, and equations for the kinematic indicators and model outputs.

Construct	Type	Operational definition	Source modality/channel	Equation
Joint-angle waveform consistency	Specified extension indicator	RMS deviation between a repetition's phase-normalised joint-angle trajectory and its leave-one-repetition-out participant–action reference waveform	Skeleton trajectories	Equation (3)
Movement smoothness	Specified extension indicator	Normalised jerk of a joint trajectory	Skeleton trajectories	Equation (4)
Bilateral symmetry	Empirical skeleton descriptor and specified extension indicator	Normalised left–right difference between homologous-joint indicators for actions that are symmetric by design	Skeleton; future inertial comparison	Equation (5)
Movement amplitude and dynamic acceleration	Empirical skeleton descriptor and specified	Maximum Euclidean joint displacement from the initial position; gravity-adjusted	Skeleton; future inertial channels	Equations (6) and (7)

	inertial extension	acceleration magnitude for future inertial analysis		
Joint angular velocity	Specified extension variable	Central-difference change in a defined anatomical joint angle	Skeleton trajectories	Equation (2)
Dynamic acceleration magnitude	Specified extension variable	Euclidean magnitude of acceleration after removal of the estimated gravitational component	Ten inertial sensors	Equation (7)
Depth motion-energy descriptor	Specified extension variable	Frame-wise mean absolute depth change, non-zero-pixel proportion, and horizontal and vertical motion-centroid displacement	Kinect V2 depth stream	Descriptive measure
Action category	Empirical outcome label	One of 22 dataset action categories; the 15 categories shared by qyh and zyh were used in the empirical classification	Dataset metadata	—
Participant identity	Empirical holdout identifier	Participant identifier used to define the directional training and test partitions	Dataset metadata	—
Fused multimodal representation (F,z)	Specified architectural output	Token-level cross-attention fusion followed by temporal pooling	Skeleton, inertial, and depth tokens	Equations (8a)–(8c)
Atypicality index (mCD_i)	Empirical skeleton output	Robustly standardised Euclidean distance from the participant-independent action-specific reference median	Action-specific reduced skeleton embedding (z_i)	Equations (10a)–(10c)

In the empirical baseline, skeleton-derived features served as model inputs, action category served as the classification outcome, and participant identity defined the directional training and test partitions. Depth and inertial channels were not used to generate the primary 294-recording skeleton benchmark or the reported mCD results. They were used only in the separate lightweight multimodal pilot described in Section 3.7.3 and reported in Section 4.6. No mediation, moderation, repeated-measures regression, or mixed-effects model was estimated. The depth-derived descriptors, inertial variables, and fused embedding in Table 4 belong to the specified tri-modal extension.

3.5. Specified Kinematic Measures

The empirical kinematic classifier used the summary features described in Section 3.6. Equations (1)–(7) define additional interpretable measures for the subsequent full-modality extension. These measures were

not used to generate the classification results reported in Section 4. Joint angles are defined using the anatomical triplets listed in Table 3. For the empirical skeleton baseline, angles were calculated directly from the body-centred, scale-normalised, 90-step trajectories; no additional Butterworth filtering, angle unwrapping, or anatomical-range exclusion was applied. For a joint j formed by anatomical points $u - j - v$, the joint angle is:

Equation (1): Joint angle

$$\theta_j(t) = \cos^{-1} \left[\frac{(p_u(t) - p_j(t))^T (p_v(t) - p_j(t))}{\|p_u(t) - p_j(t)\|_2 \|p_v(t) - p_j(t)\|_2} \right] \quad (1)$$

Here, $p_j(t)$ is the three-dimensional position of joint j at time t .

Joint angular velocity is defined by the central difference:

$$\omega_j(t) = \frac{\theta_j(t+1) - \theta_j(t-1)}{2\Delta t} \quad (2)$$

For the skeleton stream, Δt is approximately 1/30s. This quantity is specified for future skeleton-inertial alignment analysis. Joint-angle waveform consistency is calculated for repetition r , joint j , participant p , and action a as the root-mean-square deviation from the leave-one-repetition-out participant-action reference waveform:

$$WC_{rjpa} = \sqrt{\frac{1}{T} \sum_{\tau=1}^T [\theta_{rjpa}(\tau) - \bar{\theta}_{(-r)jpa}(\tau)]^2} \quad (3)$$

Here, $\bar{\theta}_{(-r)jpa}(\tau)$ is the phase-normalised mean waveform calculated from the eligible repetitions of participant p performing action a , excluding repetition r .

Movement smoothness is specified through normalised jerk (Balasubramanian et al., 2015):

$$NJ_j = \frac{T_{dur}^5}{A_j^2} \int_0^{T_{dur}} \left\| \frac{d^3 p_j(t)}{dt^3} \right\|_2^2 dt \quad (4)$$

is the recording duration and A_j is the maximum displacement defined in Equation (6). Lower values indicate smoother trajectories. This measure is specified for subsequent full-modality evaluation and was not used in the reported classifier. For actions that are symmetric by design, bilateral symmetry is specified as:

$$SYM = 1 - \frac{|I^L - I^R|}{I^L + I^R + \varepsilon} \quad (5)$$

I^L and I^R are non-negative homologous left- and right-side indicators. The stabilising constant ε is set to 1% of the pooled action-specific mean of I . Equation (5) is not applied to intentionally unilateral or lateralised actions.

Skeletal movement amplitude is defined as:

$$A_j = \max_t \|p_j(t) - p_j(0)\|_2 \quad (6)$$

$p_j(0)$ is the joint position at the beginning of the analysed recording.

For a future inertial analysis, dynamic acceleration magnitude at sensor k is specified as:

$$a_{dyn,k}(t) = \sqrt{[a_x(t) - g_x(t)]^2 + [a_y(t) - g_y(t)]^2 + [a_z(t) - g_z(t)]^2} \quad (7)$$

$a_x(t)$, $a_y(t)$, and $a_z(t)$ are the measured acceleration components, while $g_x(t)$, $g_y(t)$, and $g_z(t)$ are the estimated gravitational components. If reliable orientation-based gravity estimation is unavailable, a prespecified high-pass-filter sensitivity analysis must be reported. Equation (7) was not used to generate the skeleton-only results. In the subsequent full-modality implementation, inertial and depth streams would be aligned to the corrected skeleton timeline. The native inertial timestamp structure and sampling rate must first be verified from the public sensor files. Where valid timestamps are available, inertial signals would be interpolated or aggregated to the 90-step skeleton timeline. Any residual lag would be estimated by cross-correlating the dynamic acceleration defined in Equation (7) with acceleration derived from the corresponding skeleton joint.

This procedure belongs to the fusion architecture specified in Section 3.7 and was not performed for the present skeleton-only analysis.

Table 5. Symbols used in Equations (1)–(10)

Symbol	Definition
$\mathbf{p}_j(t)$	Three-dimensional position of joint j at time t
$\mathbf{u} - \mathbf{j} - \mathbf{v}$	Anatomical joint triplet defining the angle at joint (j)
$\theta_j(t)$	Joint angle in radians
$\omega_j(t)$	Central-difference joint angular velocity
Δt	Skeleton sampling interval
$\mathbf{r}, \mathbf{j}, \mathbf{p}, \mathbf{a}$	Repetition, joint, participant, and action indices
$\bar{\theta}_{(-r)}$	Leave-one-repetition-out reference waveform
T	Number of phase-normalised time steps
T_{dur}	Recording duration
A_j	Maximum Euclidean joint displacement from the initial position
I^L, I^R	Non-negative homologous left and right indicators
ε	1% of the pooled action-specific mean of (I)
$a_{dyn,k}(t)$	Dynamic acceleration magnitude at IMU (k)
$\mathbf{S}, \mathbf{I}, \mathbf{D}$	Skeleton, inertial, and depth token sequences
$\mathbf{F}$	Fused temporal representation
$\mathbf{z}$	Pooled recording embedding
$\mathbf{z}_i$	Action-specific reduced representation of query repetition i
$\mathbf{R}_{a,-i}$	Action-specific reference set excluding the query participant
$\tilde{\mathbf{z}}_{a,-i}$	Coordinate-wise median of the reference embedding
$s_{a,-i}$	Scaled median absolute deviation of reference distances
mCD_i	Action-conditional median-centred distance

3.6. Empirical Skeleton Models and Validation

Three skeleton models were evaluated. The first model used an interpretable kinematic feature vector containing joint-coordinate ranges and standard deviations, joint path length, mean and peak phase-normalised speed, summary statistics for ten anatomical joint angles, global skeleton motion energy, and bilateral limb-motion symmetry. These features were standardised using parameters estimated from the training participant and entered into class-balanced logistic regression with $C = 1.0$ and a maximum of 3,000 iterations. The reported classifier did not use normalised jerk, participant–action waveform consistency, or gravity-adjusted inertial acceleration. Those measures are retained only as specified indicators for subsequent full-modality evaluation. The second model flattened each 90×75 skeleton sequence, standardised the resulting features, applied whitened principal-component analysis, and fitted class-balanced logistic regression. The number of principal components was limited to the smaller of 64, the available training rank, and the original feature dimension. The TCN was selected as a compact learned temporal baseline for the small participant-independent training set. It preserves sequence order and provides a controlled temporal comparison without requiring a predefined skeleton graph or the larger tuning budget of transformer-style models. Its role is to test whether a modest learned temporal representation can improve on engineered kinematic summaries under the same leakage-safe holdout, not to define a contemporary deep-learning performance ceiling. The third model was a temporal convolutional network comprising three one-dimensional convolutional blocks, temporal mean pooling, a 64-dimensional embedding layer, and a 15-class output layer. AdamW optimisation used a learning rate of 1×10^{-3} , weight decay of 1×10^{-4} , batch size 16, label smoothing of 0.05, and gradient clipping at 1.0. Training continued for a maximum of 45 epochs and

stopped after seven epochs without improvement in inner-validation macro-F1. The trainable parameter count should be reported directly from the implemented model configuration; no parameter value is inferred without the model output. The implemented TCN configuration is reported in Table 6. Training-time comparison was not used as a model-selection criterion because computational duration depends on the execution hardware and runtime environment. The trainable parameter count should be reported from the implemented TCN model configuration used to generate the results. The value is included to document model capacity and should be calculated directly from the released implementation rather than inferred from the conceptual architecture description. The implemented TCN contained 157,071 trainable parameters. This compact capacity was selected as a controlled temporal baseline for the small participant-independent evaluation setting rather than as a maximum-capacity deep-learning benchmark.

Table 6. Architecture and training configuration of the empirical skeleton TCN

Component	Setting	Implementation detail
Input	90 × 75	90 phase-normalised time steps × 25 joints × 3 coordinates
TCN block 1	Conv1D 75→64, kernel 5	BatchNorm, GELU, dropout 0.2
TCN block 2	Conv1D 64→128, kernel 5	BatchNorm, GELU, dropout 0.2
TCN block 3	Conv1D 128→128, kernel 5	BatchNorm, GELU
Pooling	Temporal mean	128-dimensional recording vector
Embedding	Dense 128→64	GELU, dropout 0.2
Classification	Dense 64→15	Softmax across the 15 shared actions
Optimisation	AdamW	Learning rate 1×10^{-3} ; weight decay 1×10^{-4} ; batch size 16; gradient clipping 1.0
Stopping	Maximum 45 epochs	Stop after 7 epochs without improved validation macro-F1
Loss	Cross-entropy	Label smoothing 0.05
Outer evaluation	Directional participant holdout	Train on qyh/test zyh and train on zyh/test qyh
Inner validation	Class-stratified 20% of training participant	Used only for stopping; test participant excluded

The learning rate (1×10^{-3}), weight decay (1×10^{-4}), dropout (0.20), batch size (16), and gradient-clipping threshold (1.0) were held constant across holdouts and seeds. Kernel widths of 3, 5, and 7 were examined separately. The primary kernel width of five was selected as the intermediate prespecified setting rather than from outer-test performance. The implemented workflow comprises public-data loading, timestamp-quality control, body-centred coordinate normalisation, 90-step resampling, participant-holdout model fitting, and evaluation. The depth and inertial streams are excluded from this implemented boundary.

3.7. Specified Tri-Modal Fusion Architecture

The full tri-modal architecture is specified for subsequent evaluation using the complete public depth, skeleton, and inertial recordings. Skeleton, inertial, and depth streams would first be aligned to a common 90-step timeline. The skeleton and inertial streams would be encoded as temporal token sequences, while the depth pathway would encode frame-level motion descriptors. Skeleton tokens would serve as cross-attention queries and concatenated inertial and depth tokens would serve as keys and values. The resulting representation would be trained through the classification, masked-reconstruction, and contrastive objective in Equation (9). No result in this article is presented as evidence that this tri-modal configuration has already been trained or outperforms the implemented skeleton baselines.

Each phase-normalised time step forms one modality token. A skeleton token contains 75 body-centred coordinates (25 joints × 3 coordinates), an inertial token contains 60 channels (10 sensors × three-axis acceleration and three-axis angular velocity), and a depth token contains four frame-level depth-motion

descriptors. Separate temporal encoders map all three modalities to a common 128-dimensional space. Cross-attention uses the 90 skeleton tokens as queries and, for full tri-modal fusion, the concatenated 180 inertial and depth tokens as keys and values. Four attention heads therefore operate at 32 dimensions per head. The initial fusion depth is one cross-attention block followed by one residual feed-forward block (128→256→128) and temporal mean pooling.

Table 7. Specified layer-wise configuration of the full tri-modal fusion architecture.

Module	Input	Layers	Activation/normalisation	Output
Skeleton encoder	90×75	Conv1D 64, Conv1D 128, Conv1D 128; kernel 5	BatchNorm, GELU, residual connections, dropout 0.2	90×128
Inertial encoder	90×60	Conv1D 64, Conv1D 128, Conv1D 128; kernel 5	BatchNorm, GELU, residual connections, dropout 0.2	90×128
Depth encoder	90×4	Conv1D 32, Conv1D 64, Conv1D 128; kernel 5	BatchNorm, GELU, dropout 0.2	90×128
Cross-attention	Skeleton queries; inertial and depth keys/values	Four-head attention, model dimension 128	Residual connection, LayerNorm	90×128
Feed-forward block	90×128	Dense 128→256→128	GELU, dropout 0.2, residual, LayerNorm	90×128
Pooling	90×128	Temporal mean pooling	—	128
Classification head	128	Dense 128→64→22	GELU, dropout 0.2, Softmax	22 classes

Note. This table specifies the intended full tri-modal implementation. Only the skeleton pathway was empirically trained in the present study. The repository-example check reported below tests this data path only; no multimodal accuracy claim is made.

Algorithm 1. Reproducible processing, skeleton-baseline evaluation, and specified tri-modal extension

Input: Public CZU-MHAD recordings, participant/action/repetition labels, and target sequence length $T = 90$.

1. Read each skeleton matrix and reshape every frame to $25 \times [x, y, z, t]$.
2. Check coordinate completeness and timestamp intervals.
3. Replace invalid timestamp intervals using the recording-specific median positive interval.
4. Translate coordinates to SpineBase, scale by trunk length, rotate to the body reference frame, and resample to 90 time steps.
5. Select the 15 actions represented for both qyh and zyh.
6. Form directional participant-holdout training and test sets.
7. Fit the kinematic logistic, PCA-logistic, and skeleton TCN models using training data only.
8. Report accuracy, balanced accuracy, macro-F1, top-3 accuracy, seed variation, and kernel-width sensitivity.
9. Obtain seed-23 TCN test embeddings and calculate participant-independent action-specific mCD values.
10. For the proposed tri-modal extension, align the depth and inertial streams to the skeleton timeline, encode each modality, apply skeleton-query cross-attention, and optimise the classification, reconstruction, and contrastive objectives.

Output: Participant-holdout skeleton results, real skeleton mCD values, and a reproducible specification for subsequent tri-modal evaluation.

In the empirical skeleton baseline, each complete recording was linearly resampled from its first corrected timestamp to its last corrected timestamp using 90 evenly spaced time points. No motion-energy-based onset or offset detector, centre-cropping, zero-padding, or padding mask was used. In the subsequent tri-modal implementation, the depth and inertial streams would be aligned to the same corrected skeleton

timeline. This interpolation-based normalisation preserves the complete onset-to-offset movement interval while avoiding centre-cropping and zero-padding. Because every identified movement interval is resampled directly to 90 time points in the primary implementation, no padding mask is required. The skeleton encoder would take the $25 \times 3 \times 90$ joint-coordinate tensor and apply three temporal convolutional blocks.

The depth pathway would receive a 90×4 descriptor sequence. At each phase-normalised time point, the four descriptors are mean absolute inter-frame intensity change, non-zero-pixel proportion, horizontal motion-centroid displacement, and vertical motion-centroid displacement. The sequence would be processed by three one-dimensional convolutional layers. The resulting depth representation would be a token sequence $D \in \mathbb{R}^{90 \times 128}$. This scalar-per-frame descriptor captures the global quantity of motion and its timing, and is robust to partial occlusion, but it does not encode body shape, limb configuration, or the spatial distribution of motion; the protocol therefore describes the depth contribution as a global depth-motion-energy signal rather than a holistic body-shape representation, and a spatial depth representation (for example, a depth motion map or a compact 3D-CNN over depth frames) is identified in Section 5.4 as a possible future extension.

At the fusion level, token-level cross-attention would be applied before temporal pooling. Let $S \in \mathbb{R}^{90 \times 128}$, $I \in \mathbb{R}^{90 \times 128}$, and $D \in \mathbb{R}^{90 \times 128}$ denote the skeleton, inertial, and depth token sequences. Skeleton tokens would serve as queries, while the concatenated inertial and depth tokens would serve as keys and values:

$$Q = SW_Q, K = [I; D]W_K, V = [I; D]W_V \quad (8a)$$

$$H = \text{MultiHeadAttention}(Q, K, V) \quad (8b)$$

$$F = \text{LayerNorm}(S + HW_Q), z = \frac{1}{T} \sum_{t=1}^T \text{FFN}(F_t) \quad (8c)$$

Here, S , I , and D denote the skeleton, inertial, and depth token sequences; $F \in \mathbb{R}^{90 \times 128}$ is the fused temporal representation; and $z \in \mathbb{R}^{128}$ is the pooled recording embedding.

The specified tri-modal objective combines action classification, masked reconstruction, and contrastive representation learning:

$$\mathcal{L} = \lambda_{\text{cls}} \mathcal{L}_{\text{CE}} + \lambda_{\text{rec}} \mathcal{L}_{\text{mask}} + \lambda_{\text{con}} \mathcal{L}_{\text{InfoNCE}} \quad (9)$$

The specified initial weights are $\lambda_{\text{cls}} = 1.00$, $\lambda_{\text{rec}} = 0.50$, and $\lambda_{\text{con}} = 0.10$. These weights were not tuned or evaluated in the present skeleton-only analysis.

3.7.1. Participant-Independent Multimodal Ablation

The minimum multimodal experiment is fixed as four nested conditions: (A) skeleton-only, using the skeleton encoder followed by temporal pooling and the classification head; (B) skeleton + IMU, using skeleton queries and the 90 inertial tokens as keys and values; (C) skeleton + depth, using skeleton queries and the 90 depth tokens as keys and values; and (D) full tri-modal fusion, using skeleton queries and the concatenated 180 inertial/depth key-value tokens. The same action set, outer participant holdout, inner training-participant validation, random seeds, optimiser family, stopping rule, and reporting metrics are held constant across all four conditions. For the qyh/zyh matched subset, the comparison uses 15 classes and both directional holdouts; for a full five-participant analysis, it uses all 22 classes and leave-one-participant-out folds. Macro-F1 is the primary metric, with accuracy, balanced accuracy, top-3 accuracy, parameter count, and training-time cost as secondary outcomes. Pairwise gains are reported as B-A, C-A, and D-A so that the contribution of each added modality is visible. The full trained cross-attention comparison defined above remains distinct from the lightweight feature-level experiment in Section 3.7.3. The latter uses branch-level raw records and a compact file-level depth complexity proxy to test whether additional modalities provide preliminary participant-independent discriminative information. It does not instantiate or train the complete Table 7 depth, inertial, and cross-attention encoders under matched model-level conditions. Accordingly, predictive performance of the proposed full cross-attention architecture is not claimed.

3.7.2. Repository-Example Multimodal Implementation Check

Two matched examples distributed with the official repository were used to test the complete input path. The Sample Code example contains 195 depth frames, 196 skeleton frames, and 2,230-3,311 samples across the ten inertial streams. The demo example contains 213 depth frames, 188 skeleton frames, and 2,394-3,477 inertial samples per sensor. Skeleton trajectories were converted to body-centred 90×75 tokens using the same anatomical normalisation logic as the empirical baseline. Each inertial stream contributed six channels and was interpolated to the corrected skeleton timeline, yielding a 90×60 token sequence; the inertial timestamp ranges covered all 90 target time points in both examples. The four depth descriptors were resampled to 90×4 . A fixed-seed, untrained implementation of the Table 7 encoders and four-head cross-attention block was used only as a tensor-compatibility test. In both examples, all three encoders produced 90×128 token sequences, the fused sequence was 90×128 , temporal pooling produced a 128-dimensional vector, and all intermediate and output tensors were finite. Two examples cannot support participant-independent classification, loss comparison, confidence intervals, or modality-effect estimates; no accuracy or macro-F1 is therefore reported from this check.

3.7.3. *Lightweight Five-Participant Multimodal Ablation*

To provide an initial empirical evaluation of modality contribution, a deliberately lightweight feature-level ablation was conducted using raw records from the skeleton, sensors, and depth branches of the official CZU-MHAD repository. The participant intersection comprised cx, cyy, myj, qyh, and zyh. Three right-arm actions available across the selected matched records were analysed: right high wave (a1), right horizontal wave (a3), and hammer with the right hand (a5). One matched repetition was used for each participant-action combination, yielding 15 observations. The selected repetition was time_2 for myj and time_1 for the remaining participants. For the skeleton condition, eight upper-body joints were represented by three-dimensional coordinates after SpineBase centring and shoulder-width normalisation, producing 24 skeleton variables. For the inertial condition, the six accelerometer and gyroscope channels from the right-elbow and right-wrist sensors were concatenated, producing 12 inertial variables. For the lightweight depth condition, the first five released PNG frames were summarised using a six-dimensional depth-stream complexity proxy comprising four relative frame-size changes, a normalised linear frame-size trend, and the coefficient of variation. This proxy was used only for the present lightweight modality test and must not be interpreted as a physical depth measurement or as the four-descriptor temporal depth encoder specified in Table 7. Four nested conditions were compared: skeleton only, skeleton plus IMU, skeleton plus file-level depth complexity proxy, and skeleton plus IMU plus file-level depth complexity proxy. Training-fold standardisation was performed within five leave-one-participant-out folds. To reduce dependence on a single favourable classifier in this small pilot, three fixed low-capacity classifiers were reported: class-balanced logistic regression with $C = 1.0$, class-balanced linear support-vector classification with $C = 1.0$, and nearest-centroid classification. Accuracy and macro-F1 were calculated from the pooled held-out predictions. This experiment tests preliminary modality contribution and does not constitute training or validation of the full temporal cross-attention architecture in Table 7.

3.8. Action-Conditional Kinematic Atypicality Index

For the empirical analysis, let z_i denote the action-specific reduced representation of the 64-dimensional TCN embedding for query repetition i . For each action a , embeddings from the other participant formed the reference set $R_{a,-i}$; the query participant did not contribute to its own reference distribution. Principal-component analysis was fitted to the reference participant only, using up to five components and never more than $n_{\text{ref}} - 1$. Query embeddings were projected using the same fitted transformation.

Equations (10a)–(10c): Median-centred distance

$$d_i = \|z_i - \tilde{z}_{a,-i}\|_2 \quad (10a)$$

$$s_{a,-i} = 1.4826 \times \text{median}_{r \in R_{a,-i}} \left| \|z_r - \tilde{z}_{a,-i}\|_2 - \text{median}_{q \in R_{a,-i}} \|z_q - \tilde{z}_{a,-i}\|_2 \right| \quad (10b)$$

$$mCD_i = \frac{d_i}{s_{a,-i}}, \quad s_{a,-i} > 0 \quad (10c)$$

A large mCD value indicates that a repetition lies far from the participant-independent reference distribution for the same action. It does not identify incorrect technique, poor movement quality, or clinical abnormality. If the scaled median absolute deviation is zero, mCD is reported as undefined rather than stabilised with an arbitrary denominator. To translate an elevated mCD value into interpretable movement information, additional indicators such as joint-angle deviation, bilateral symmetry differences, trajectory smoothness, and joint-specific velocity changes would be required before any corrective interpretation.

Operationally, mCD is an action-conditioned screening statistic. It quantifies how far a repetition's learned skeleton representation lies from the median reference pattern for the same action after scaling by reference dispersion. A larger value can identify a same-action repetition for closer human review or reveal an action with substantial cross-participant representation shift. For example, within right body turning (a20), a repetition with an mCD near the lower part of the observed distribution, approximately 1.6, lies relatively close to the same-action reference representation, whereas a repetition near the observed maximum of 3.6 lies farther from that reference. The latter would therefore be considered more atypical within that action category, but not more incorrect or more clinically abnormal. It does not identify which joint caused the difference, whether the difference is beneficial or harmful, or whether technique is clinically or pedagogically incorrect. Raw mCD values are therefore action-specific rather than a universal threshold. Criterion validation would require independent expert or instrumented movement-quality measures for the same repetitions, blinded ratings, inter-rater reliability, and participant-independent calibration of any proposed alert threshold.

3.9. Reproducibility and Robustness Procedures

Skeleton preprocessing and model evaluation were implemented through reproducible Python scripts. Random-number generators were initialised explicitly. TCN performance was evaluated using seeds 11, 23, and 37 in both directional participant holdouts. Kernel-width sensitivity was examined at widths 3, 5, and 7 while holding dropout at 0.20 and using seed 23. Feature scaling, principal-component estimation, validation selection, and reference-distribution construction were restricted to the training data. The supplementary package contains CZU_MHAD_Skeleton_Baseline.py, CZU_MHAD_Skeleton_Postprocess.py, Create_Publication_Figures.py, requirements.txt, and the machine-readable metrics, predictions, embeddings, confusion matrix, quality-control summaries, and mCD outputs underlying Tables 8–10 and Figures 3–6. The supplied code reproduces the empirical skeleton baseline; it does not train or evaluate the full tri-modal model.

3.10. Ethical and Data-Use Considerations

CZU-MHAD is a publicly released research dataset distributed by its original developers via GitHub, Google Drive, Baidu Netdisk, and the Science Data Bank, with documentation describing the data-collection protocol and participant set (Chao et al., 2022). This study uses only the publicly available reference materials and metadata associated with the dataset and does not involve new data collection from human participants. Accordingly, no new institutional ethics approval was sought, consistent with the treatment of secondary, publicly available, anonymised research data.

4. Results

This section reports two empirical evidence levels. Sections 4.1–4.4 report the primary participant-holdout skeleton analysis on the defined qyh/zyh subset. Section 4.5 reports the end-to-end repository-example compatibility check for the proposed cross-attention data path, and Section 4.6 reports a separate five-participant lightweight feature-level multimodal ablation. The Section 4.6 pilot evaluates preliminary modality contribution but does not train or validate the full Table 7 temporal cross-attention architecture. No result is interpreted as a validated movement-quality score.

4.1. Sample and Data-Quality Profile

All 357 skeleton recordings were readable and contained finite coordinate values. Sequence length ranged from 40 to 203 frames, with a median of 123 frames. After timestamp-interval correction, recording duration ranged from 1.282 to 7.300 s, with a median of 4.258 s. The mean effective sampling rate was 30.151

Hz (SD = 0.387). Forty recordings contained at least one large timestamp discontinuity, confirming the need for timestamp-based quality control before temporal resampling. The participant-holdout sample comprised 294 recordings from the 15 actions represented for both qyh and zyh. The repository-level modality inventory is reported in Supplementary Table S1 and Supplementary Figure S1.

4.2. Directional Participant-Holdout Skeleton Performance

Table 8. Directional participant-holdout action-classification performance on 294 real skeleton sequences from 15 shared actions

Model	Held-out participant	Runs	Accuracy	Balanced accuracy	Macro-F1	Top-3 accuracy
Kinematic logistic	qyh	1	0.730	0.739	0.725	0.918
Kinematic logistic	zyh	1	0.904	0.904	0.900	1.000
Kinematic logistic	Pooled directional mean	2	0.817 (directional variation: 0.123)	0.822 (directional variation: 0.116)	0.813 (directional variation: 0.123)	0.959 (directional variation: 0.058)
PCA-sequence logistic	qyh	1	0.667	0.657	0.661	0.818
PCA-sequence logistic	zyh	1	0.600	0.600	0.586	0.689
PCA-sequence logistic	Pooled directional mean	2	0.633 (directional variation: 0.047)	0.628 (directional variation: 0.040)	0.623 (directional variation: 0.053)	0.753 (directional variation: 0.091)
Skeleton TCN	qyh	3	0.642 (directional variation: 0.062)	0.625 (directional variation: 0.058)	0.626 (directional variation: 0.064)	0.836 (directional variation: 0.077)
Skeleton TCN	zyh	3	0.607 (directional variation: 0.071)	0.607 (directional variation: 0.071)	0.575 (directional variation: 0.052)	0.795 (directional variation: 0.019)
Skeleton TCN	All directional holdouts and seeds	6	0.624 (directional variation: 0.063)	0.616 (directional variation: 0.059)	0.601 (directional variation: 0.059)	0.816 (directional variation: 0.055)

Note. The kinematic and PCA baselines were fitted once in each direction. The TCN was repeated with seeds 11, 23, and 37 in each direction. Means and standard deviations pool directional holdouts only where indicated. With only two participants, these values demonstrate feasibility and participant sensitivity; they are not population-level generalisation estimates.

Rows are normalised by the number of true observations. The matrix combines the two directional test sets and is presented as a diagnostic summary rather than a population-level estimate. The kinematic logistic baseline produced the strongest directional performance, with accuracy of 0.730 and macro-F1 of 0.725 when qyh was held out and accuracy of 0.904 and macro-F1 of 0.900 when zyh was held out. Its two-direction mean

accuracy was 0.817 (SD 0.123). The PCA-sequence logistic baseline achieved mean accuracy of 0.633 and macro-F1 of 0.623. Across three seeds in both directions, the skeleton TCN achieved accuracy of 0.624 (directional variation: 0.063, balanced accuracy of 0.616 (directional variation: 0.059, macro-F1 of 0.601 (directional variation: 0.059, and top-3 accuracy of 0.816 (directional variation: 0.055. TCN accuracy ranged from 0.526 to 0.692 across the six runs. The simpler kinematic model therefore outperformed the temporal convolutional model on this two-participant subset, indicating that architecture complexity did not compensate for participant scarcity and domain shift.)

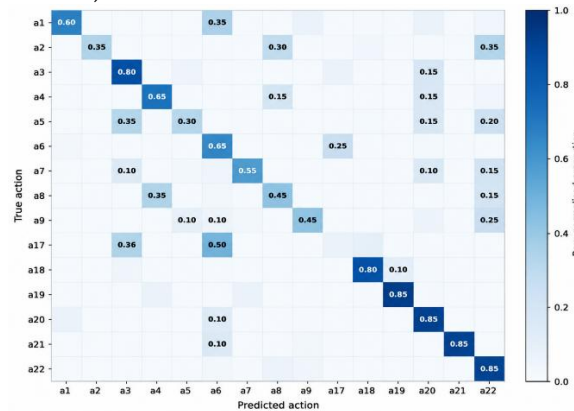


Figure 3. Normalised confusion matrix for the seed-23 directional participant-holdout skeleton TCN.

The directional results also show where the two-participant uncertainty enters the comparison. For the kinematic model, holding out qyh versus zyh changed accuracy by 0.174 and macro-F1 by 0.175. The corresponding directional gaps were 0.067 and 0.075 for PCA-logistic, and 0.035 and 0.051 for the seed-averaged TCN. These contrasts provide participant-wise sensitivity information, but two held-out participants are not enough to estimate population-level variance.

4.3. Seed and Kernel-Width Sensitivity

Table 9. Exploratory kernel-width sensitivity for the skeleton TCN

Configuration	Directional runs	Accuracy	Macro-F1	Top-3 accuracy
(Kernel = 3; dropout = 0.20)	2	0.493 (directional variation: 0.184)	0.467 (directional variation: 0.198)	0.787
(Kernel = 5; dropout = 0.20)	2	0.609 (directional variation: 0.117)	0.586 (directional variation: 0.099)	0.835
(Kernel = 7; dropout = 0.20)	2	0.658 (directional variation: 0.023)	0.626 (directional variation: 0.026)	0.889

Note. Results use seed 23 and the two directional participant holdouts. Error variation reflects the difference between the two holdout directions, not repeated population sampling.

At seed 23, increasing kernel width from 3 to 7 improved the two-direction mean. Kernel widths 3, 5, and 7 produced mean accuracies of 0.493, 0.609, and 0.658 and mean macro-F1 values of 0.467, 0.586, and 0.626, respectively. These comparisons are exploratory because each setting was evaluated with one seed and only two directional holdouts. They do not justify declaring kernel 7 universally optimal; they show that the current kernel choice is consequential and should be selected inside training data when the full participant set is available.

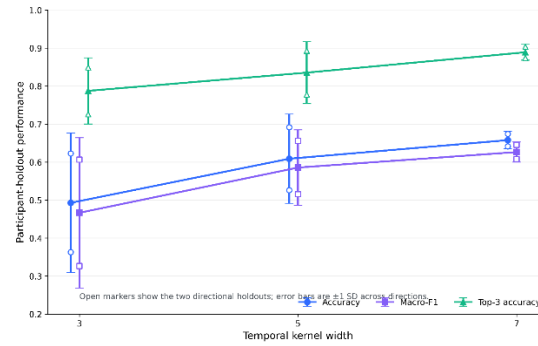


Figure 4. Kernel-width sensitivity of participant-holdout skeleton classification.

Error bars describe variation between the two directional holdouts at seed 23 and do not represent repeated population samples.

4.4. Real Skeleton-Embedding Atypicality Description

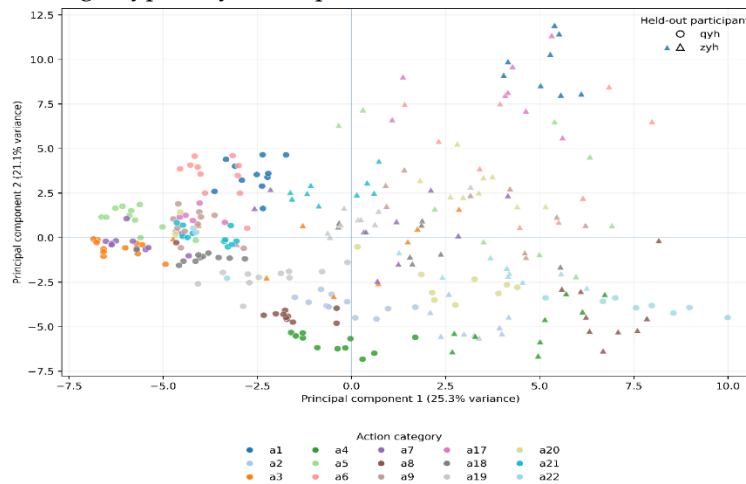


Figure 5. Two-dimensional principal-component display of real 64-dimensional skeleton test embeddings.

Colour identifies action category and marker shape identifies the held-out participant. PCA is used only for visualisation. The reported mCD values were calculated separately within action-specific reference spaces of up to five principal components.

Table 10. Participant-independent action-conditional mCD values from real skeleton embeddings

Action	Action name	n	Median mCD	IQR	Maximum
a1	Right high wave	20	9.841	[7.472, 11.154]	13.226
a2	Left high wave	20	9.060	[8.607, 12.578]	18.410
a3	Right horizontal wave	20	5.682	[4.867, 6.223]	7.709
a4	Left horizontal wave	20	4.850	[4.518, 5.323]	7.449
a5	Hammer with right hand	20	4.510	[3.972, 4.718]	6.164
a6	Grasp with right hand	20	4.150	[3.995, 5.262]	7.801
a7	Draw fork with right hand	20	4.317	[3.885, 4.605]	6.329
a8	Draw fork with left hand	20	5.775	[5.322, 7.880]	9.498
a9	Draw circle with right hand	20	8.513	[7.195, 9.539]	10.347
a17	Wave up and down	14	22.532	[10.283, 24.124]	33.329
a18	Sur Place	20	8.144	[7.605, 8.673]	22.422
a19	Left body turning movement	20	7.470	[2.623, 9.252]	19.653
a20	Right body turning movement	20	2.409	[1.573, 2.835]	3.596
a21	Left lateral movement	20	4.471	[3.106, 5.419]	10.942
a22	Right lateral movement	20	3.946	[3.044, 6.047]	7.737

The participant-independent skeleton mCD varied markedly by action. Median values were lowest for right body turning (a20; 2.409), right lateral movement (a22; 3.946), and grasp with the right hand (a6; 4.150), and highest for wave up and down (a17; 22.532), right high wave (a1; 9.841), and left high wave (a2; 9.060). Action a17 included 14 analysed cross-participant sequences, whereas each of the other shared actions included 20. These values reflect cross-participant embedding distance and reference dispersion, not error severity or movement quality. Accordingly, a relatively high mCD indicates greater deviation from the participant-independent reference representation for the same action only; it must not be interpreted as evidence that the movement was performed incorrectly or that it was clinically abnormal. The action-specific scale also means raw mCD magnitudes should not be compared with a universal clinical threshold.

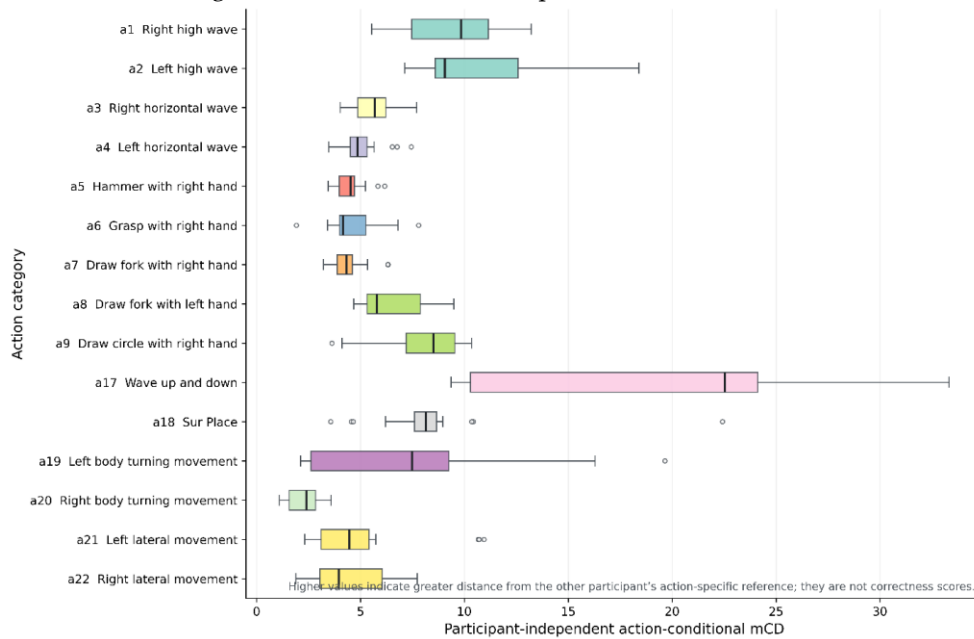


Figure 6. Participant-independent action-conditional mCD distributions derived from real skeleton embeddings.

The distributions describe cross-participant kinematic atypicality and have no universal correctness or clinical threshold.

4.5. Repository-Example Multimodal Implementation Check

The two matched repository examples provided a direct check of multimodal loading, resampling, dimensional alignment, and forward propagation through the specified fusion path. Table 11 reports the observed tensor dimensions. This is initial implementation evidence on released CZU-MHAD files, not a predictive comparison between modalities.

Table 11. Repository-example verification of multimodal loading, resampling, and fusion dimensions.

Example	Raw depth / skeleton	IMU samples per sensor	90-step modality tokens	Encoded / fused dimensions	Verification
cyt_a1_t1	195 / 196	2,230-3,311	Skeleton 90×75; IMU 90×60; depth 90×4	Each encoder 90×128; fused 90×128; pooled 128	100% target timeline covered by all IMUs; all tensors finite
a01_s01_e01	213 / 188	2,394-3,477	Skeleton 90×75; IMU	Each encoder 90×128; fused	100% target timeline covered by all

90×60; depth
90×4

90×128; pooled
128

IMUs; all
tensors finite

Note. The forward pass used fixed-seed untrained weights solely to verify dimensional compatibility and finite tensor propagation. No accuracy, macro-F1, or modality-effect estimate can be inferred from two examples.

Class-wise predictions for the best-performing exploratory condition should be reported in supplementary material using the held-out prediction outputs generated during the ablation analysis. The aggregate metrics are retained in the main manuscript. Supplementary Table Sx reports class-wise prediction behaviour for the best-performing exploratory ablation condition (nearest-centroid skeleton + IMU). These class-level results are provided as diagnostic information because the ablation contains three actions and 15 matched observations rather than a population-scale evaluation sample. Class-wise prediction behaviour for the nearest-centroid Skeleton + IMU exploratory ablation condition is provided in Supplementary Table Sx. The table reports total observations, correct predictions, and recall values derived from the held-out prediction outputs. Supplementary Table Sx. Class-wise prediction summary for nearest-centroid Skeleton + IMU condition.

4.6. Lightweight Five-Participant Multimodal Ablation

Table 12. Five-participant lightweight multimodal ablation

Condition	Logistic accuracy	Logistic macro- F1	Linear SVM accuracy	Linear SVM macro-F1	Nearest- centroid accuracy	Nearest- centroid macro-F1
Skeleton only	0.733	0.741	0.467	0.481	0.667	0.665
Skeleton + IMU	0.733	0.731	0.667	0.665	0.867	0.866
Skeleton + file-level depth complexity proxy	0.600	0.589	0.400	0.385	0.600	0.593
Skeleton + IMU + file- level depth complexity proxy	0.733	0.722	0.667	0.663	0.867	0.866

The lightweight ablation did not demonstrate a universal benefit from adding modalities. Relative to skeleton only, adding right-arm IMU information increased pooled accuracy from 0.467 to 0.667 for the linear SVM and from 0.667 to 0.867 for nearest-centroid classification, corresponding to an absolute increase of 0.200 in both cases. Logistic-regression accuracy remained unchanged at 0.733. The file-level depth complexity proxy did not improve accuracy for any of the three classifiers. The skeleton-plus-IMU-plus-depth condition matched the skeleton-plus-IMU condition for the linear SVM and nearest-centroid classifiers and matched skeleton-only accuracy for logistic regression. These findings provide preliminary evidence that inertial information can add participant-independent discriminative signal, while providing no evidence of a consistent benefit from the lightweight depth proxy or from adding all three inputs.

5. Discussion

5.1. Principal Findings

The empirical analysis produced three main findings. First, action identity was recoverable across participants, but performance differed substantially by modelling strategy and holdout direction. The kinematic logistic model produced the strongest mean performance, with accuracy of 0.817 and macro-F1 of 0.813. The PCA-logistic and TCN models produced lower mean accuracies of 0.633 and 0.624, respectively. These results indicate that anatomically interpretable summary features were more stable than the learned temporal representation under the present two-participant design. A plausible explanation is the inductive bias built into the engineered representation. Ranges, path length, angle summaries, speed, and symmetry compress the 90×75 sequence into anatomically meaningful quantities and remove timing detail that can vary between individuals. The TCN sees the complete phase-normalised coordinate sequence and must learn those

invariances from one training participant, leaving more opportunity to encode movement style, residual tracking noise, and participant-specific timing. The observed advantage of the kinematic model is therefore best interpreted as a small-sample robustness effect, not as evidence that logistic regression is generally superior to deep skeleton models. The present TCN is a useful compact feasibility baseline, but it does not define a strong contemporary deep-learning ceiling. A fuller benchmark should use all five participants, nested participant-independent tuning, and at least one spatio-temporal graph model and one transformer-based skeleton model under the same outer folds. Second, the TCN showed meaningful sensitivity to participant identity and random initialisation. Accuracy ranged from 0.526 to 0.692 across the six seed-by-holdout runs. The directional difference between qyh and zyh further indicates that the learned representation was affected by participant-specific movement style. The result cautions against interpreting a single random split or a single seed as evidence of participant-independent generalisation. Third, action-conditional mCD values varied across actions. Wave up and down, right high wave, and left high wave produced the largest median values, whereas right body turning and right lateral movement produced lower medians. These differences describe the dispersion of one participant's embeddings relative to the other participant's action-specific reference. They do not identify incorrect execution or clinical impairment.

5.2. Methodological Contribution

The main methodological contribution is the connection between transparent preprocessing and a clearly bounded fusion design. The study reports how skeleton coordinates were normalised, how participant leakage was prevented, how interpretable and learned representations performed under the same directional holdout, and how action-specific reference distances were constructed. Compared with the single-stream reconstruction-error approach used by AI Trainer, the present framework adds anatomical joint mapping, participant-independent reference construction, decomposable kinematic measures, and a path toward temporal depth-skeleton-inertial fusion. Unlike the original CZU-MHAD benchmark, the analysis also examines within-action embedding variation rather than action identity alone. The skeleton results do not establish that cross-attention improves performance. Instead, they provide the empirical baseline that a future tri-modal model must exceed under the same participant-independent evaluation protocol.

5.3. Implications and Limits of Interpretation

The findings support the use of interpretable skeleton descriptors as a practical baseline for movement-analysis systems developed from small participant samples. Joint-angle, path-length, speed, and symmetry summaries may be easier to audit and may generalise more reliably than higher-capacity encoders when participant diversity is limited. The tri-modal extension has been checked at the data-loading, token-construction, and forward-pass level on the two matched repository examples but remains prospective as a predictive model. Depth and inertial signals may provide complementary information during fast movement, self-occlusion, or local joint acceleration, but their incremental value must be demonstrated empirically. Similarly, mCD may be examined as a candidate review signal, but it cannot be used as a movement-quality or clinical score without expert-labelled validation.

5.4. Alternative Explanations and Future Validation

Differences in participant anatomy, movement speed, camera orientation, skeleton-tracking quality, and the number of available repetitions may all contribute to participant-holdout error and elevated mCD values. Timestamp correction may also alter temporal structure in recordings containing large discontinuities. These explanations cannot be separated fully with two participants. Future development will proceed through four stages: (1) complete five-participant skeleton benchmarking, (2) trained depth-skeleton-inertial multimodal ablation, (3) expert-rated validation of atypicality indicators, and (4) field evaluation in practical movement-analysis settings. Future evaluation should include all five CZU-MHAD participants, complete depth and inertial streams, cross-modal timestamp verification, stronger skeleton-only comparators, and repeated participant-independent validation. The resulting atypicality measures should then be compared with clinician, coach, or expert ratings before any movement-quality interpretation is made. The next-stage model-level multimodal comparison should implement the prespecified skeleton-only, skeleton plus IMU,

skeleton plus depth, and full tri-modal conditions using the complete temporal encoders, identical outer folds, and training-only hyperparameter selection. This experiment should determine whether the preliminary feature-level IMU advantage persists when the full Table 7 architecture is trained.

5.5. Practical Use Pathway and Validation Requirements

The present framework is best positioned as a human-in-the-loop movement-analysis tool rather than an autonomous correction system. In practical use, a known action would be segmented, normalised, and encoded; the system would return the action probability, the action-conditional mCD, and interpretable joint or sensor descriptors. A high mCD would trigger closer review, while angle, path, speed, symmetry, smoothness, and, after multimodal validation, inertial and depth indicators would help the assessor identify the source of the deviation. The framework should not be used to score technique until a full model-level multimodal ablation confirms whether the preliminary IMU signal persists with the complete temporal encoders, establishes whether physical depth descriptors add independent information, evaluates all five participants under participant-independent folds, and calibrates the resulting atypicality measures against independent expert or instrumented ground truth.

6. Conclusion

This study provides a reproducible empirical skeleton baseline and a clearly bounded multimodal fusion protocol for CZU-MHAD. Across two directional participant holdouts, interpretable kinematic features generalised more effectively than the tested PCA and temporal-convolutional representations. The TCN also showed sensitivity to holdout direction, seed, and kernel width, demonstrating the importance of participant-independent validation in small multimodal action datasets. Action-conditional mCD values provided a robust description of cross-participant skeleton-embedding atypicality, but they must not be interpreted as movement incorrectness, poor technique, clinical abnormality, or validated movement-quality scores. The proposed anatomical mapping and token-level fusion design provide a transparent basis for full depth-skeleton-inertial modelling. End-to-end processing of the two packaged repository examples confirmed dimensional compatibility of the 90-step, 128-dimensional four-head cross-attention path. A separate five-participant, three-action feature-level ablation provided preliminary predictive evidence: right-arm IMU information improved accuracy by 0.200 for two of three low-capacity classifiers, whereas the lightweight file-level depth complexity proxy produced no consistent improvement. These findings support further investigation of inertial complementarity but do not establish that the proposed full temporal cross-attention architecture improves predictive performance. A complete model-level multimodal evaluation remains necessary before claims of tri-modal superiority or movement-quality assessment can be made.

6.1. Limitations

The primary 15-class skeleton benchmark was restricted to participants qyh and zyh, so its performance estimates cannot be interpreted as population-level generalisation estimates. The secondary multimodal pilot included five participants but was limited to three actions and one matched repetition per participant-action combination and therefore provides exploratory modality-contribution evidence rather than a population-level performance benchmark. The public CZU-MHAD resource is distributed across multiple archive components, and the present empirical baseline focused on the skeleton subset that could be processed consistently for both participants. Forty recordings required timestamp-interval correction before resampling. The depth and inertial encoders and the cross-attention module were specified and checked for forward-pass compatibility on two matched examples, but they were not trained for predictive multimodal evaluation. Finally, CZU-MHAD does not include expert correctness labels; therefore, classification performance and mCD values describe action identity and kinematic atypicality rather than movement quality, rehabilitation outcome, or coaching error. The two directional holdouts provide participant-wise sensitivity information but are insufficient for stable variance estimation. The repository-example multimodal check is limited to implementation compatibility and does not substitute for predictive validation.

Acknowledgments

The author has no acknowledgments to declare.

Author Contributions

The author was responsible for conceptualization, methodology, data curation, software, formal analysis, validation, visualization, and preparation of the manuscript.

Funding Statement

This research received no external funding.

Conflict of Interest Statement

The author declares no conflict of interest.

Data Availability Statement

The CZU-MHAD dataset and official loading code are publicly available from the dataset developers at https://github.com/yujmo/CZU_MHAD. The dataset is described in Chao et al. (2022), <https://doi.org/10.1109/JSEN.2022.3150225>. The present study reports analyses of a defined skeleton subset of this public resource. The preprocessing, participant-holdout modelling, sensitivity-analysis, and mCD scripts, together with machine-readable result tables, predictions, embeddings, and figure data, are provided as supplementary materials.

Ethics Statement

Ethical approval was not required for this secondary analysis because the study used a publicly available, anonymised research dataset and involved no new participant recruitment, intervention, or data collection.

References

1. Abedi, A., Colella, T. J. F., Pakosh, M., & Khan, S. S. (2024). Artificial intelligence-driven virtual rehabilitation for people living in the community: A scoping review. *npj Digital Medicine*, 7, Article 25. <https://doi.org/10.1038/s41746-024-00998-w>
2. Aghayengejeh, N. P., Balafar, M. A., & Khasmakhi, N. N. (2025). Revolutionizing textual data insights: A comprehensive review of the dual relationship between transformers and clustering in textual data analysis. *Computer Science Review*, 58, Article 100792. <https://doi.org/10.1016/j.cosrev.2025.100792>
3. Balasubramanian, S., Melendez-Calderon, A., Roby-Brami, A., & Burdet, E. (2015). On the analysis of movement smoothness. *Journal of NeuroEngineering and Rehabilitation*, 12, Article 112. <https://doi.org/10.1186/s12984-015-0090-9>
4. Capecci, M., Ceravolo, M. G., Ferracuti, F., Iarlori, S., Monteriù, A., Romeo, L., & Verdini, F. (2019). The KIMORE dataset: KInematic assessment of MOvement and clinical scores for remote monitoring of physical REhabilitation. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 27(7), 1436–1448. <https://doi.org/10.1109/TNSRE.2019.2923060>
5. Chao, X., Hou, Z., & Mo, Y. (2022). CZU-MHAD: A multimodal dataset for human action recognition utilizing a depth camera and 10 wearable inertial sensors. *IEEE Sensors Journal*, 22(7), 7034–7042. <https://doi.org/10.1109/JSEN.2022.3150225>
6. Chariar, M., Rao, S. S., Irani, A., Suresh, S., & Asha, C. S. (2023). AI trainer: Autoencoder based approach for squat analysis and correction. *IEEE Access*, 11, 107135–107149. <https://doi.org/10.1109/ACCESS.2023.3316009>
7. Chen, L., Xia, C., Zhao, Z., Fu, H., & Chen, Y. (2024). AI-driven sensing technology: Review. *Sensors*, 24(10), Article 2958. <https://doi.org/10.3390/s24102958>
8. Chen, C., Jafari, R., & Kehtarnavaz, N. (2015). UTD-MHAD: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor. In 2015 IEEE International Conference on Image Processing (ICIP) (pp. 168–172). IEEE. <https://doi.org/10.1109/ICIP.2015.7350781>
9. Coser, O., Tamantini, C., Soda, P., & Zollo, L. (2024). AI-based methodologies for exoskeleton-assisted rehabilitation of the lower limb: A review. *Frontiers in Robotics and AI*, 11, Article 1341580. <https://doi.org/10.3389/frobt.2024.1341580>
10. Cossich, V. R. A., Carlgren, D., Holash, R. J., & Katz, L. (2023). Technological breakthroughs in sport: Current practice and future potential of artificial intelligence, virtual reality, augmented reality, and modern data visualization in performance analysis. *Applied Sciences*, 13(23), Article 12965. <https://doi.org/10.3390/app132312965>
11. Debnath, B., O'Brien, M., Yamaguchi, M., & Behera, A. (2022). A review of computer vision-based approaches for physical rehabilitation and assessment. *Multimedia Systems*, 28, 209–239. <https://doi.org/10.1007/s00530-021-00815-4>
12. Edriss, S., Romagnoli, C., Caprioli, L., Zanela, A., Panichi, E., Campoli, F., Padua, E., Annino, G., & Bonaiuto, V. (2024). The role of emergent technologies in the dynamic and kinematic assessment of human movement in sport and clinical applications. *Applied Sciences*, 14(3), Article 1012. <https://doi.org/10.3390/app14031012>
13. Giannakopoulou, K.-M., Roussaki, I., & Demestichas, K. (2022). Internet of things technologies and machine learning methods for Parkinson's disease diagnosis, monitoring and management: A systematic review. *Sensors*, 22(5), Article 1799. <https://doi.org/10.3390/s22051799>
14. Górriz, J. M., Álvarez-Illán, I., Álvarez-Marquina, A., Arco, J. E., Atzmueller, M., Ballarini, F., Barakova, E., Bologna, G., Bonomini, P., Castellanos-Dominguez, G., Castillo-Barnes, D., Cho, S. B., Contreras, R., Cuadra, J. M., Domínguez, E., Domínguez-Mateos, F., Duro, R. J., Elizondo, D., Fernández-Caballero, A., ... Ferrández-Vicente, J. M. (2023). Computational approaches to explainable artificial intelligence: Advances in theory, applications and trends. *Information Fusion*, 100, Article 101945. <https://doi.org/10.1016/j.inffus.2023.101945>
15. Guo, D., Li, K., Hu, B., Zhang, Y., & Wang, M. (2024). Benchmarking micro-action recognition: Dataset, methods, and applications. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(7), 6238–6252. <https://doi.org/10.1109/TCSVT.2024.3358415>
16. He, S., Meng, D., Wei, M., Guo, H., Yang, G., & Wang, Z. (2024). Proposal and validation of a new approach in tele-rehabilitation with 3D human posture estimation: A randomized controlled trial in older individuals with sarcopenia. *BMC Geriatrics*, 24, Article 586. <https://doi.org/10.1186/s12877-024-05188-7>

17. Jiang, Y., Li, X., Luo, H., Yin, S., & Kaynak, O. (2022). Quo vadis artificial intelligence? Discover Artificial Intelligence, 2, Article 4. <https://doi.org/10.1007/s44163-022-00022-8>
18. Junaid, S. B., Imam, A. A., Abdulkarim, M., Surakat, Y. A., Balogun, A. O., Kumar, G., Shuaibu, A. N., Garba, A., Sahalu, Y., Mohammed, A., Mohammed, T. Y., Abdulkadir, B. A., Abba, A. A., Iliyasu Kakumi, N. A., & Hashim, A. S. (2022). Recent advances in artificial intelligence and wearable sensors in healthcare delivery. Applied Sciences, 12(20), Article 10271. <https://doi.org/10.3390/app122010271>
19. Pourhaji, N., Asadpour, M., Ahmadian, A., & Elkamel, A. (2022). The investigation of monthly/seasonal data clustering impact on short-term electricity price forecasting accuracy: Ontario Province case study. Sustainability, 14(5), Article 3063. <https://doi.org/10.3390/su14053063>
20. Ruiz-Ruiz, L., Jimenez, A. R., Garcia-Villamil, G., & Seco, F. (2021). Detecting fall risk and frailty in elders with inertial motion sensors: A survey of significant gait parameters. Sensors, 21(20), Article 6918. <https://doi.org/10.3390/s21206918>
21. Vakanski, A., Jun, H.-P., Paul, D., & Baker, R. (2018). A data set of human body movements for physical rehabilitation exercises. Data, 3(1), Article 2. <https://doi.org/10.3390/data3010002>
22. Wang, Y., Gao, Z., Wu, W., Xiong, Y., Luo, J., Sun, Q., Mao, Y., & Wang, Z. L. (2025). TENG-boosted smart sports with energy autonomy and digital intelligence. Nano-Micro Letters, 17, Article 265. <https://doi.org/10.1007/s40820-025-01778-1>
23. Yammouri, G., & Ait Lahcen, A. (2024). AI-reinforced wearable sensors and intelligent point-of-care tests. Journal of Personalized Medicine, 14(11), Article 1088. <https://doi.org/10.3390/jpm14111088>
24. Zakeri, Z., Arif, A., Omurtag, A., Breedon, P., & Khalid, A. (2023). Multimodal assessment of cognitive workload using neural, subjective and behavioural measures in smart factory settings. Sensors, 23(21), Article 8926. <https://doi.org/10.3390/s23218926>