

Automatic dance movement quality assessment via spatiotemporal skeletal representation learning

Zhen Ji

azal tejp2316624@outlook.com

Hangzhou Polytechnic University

Panpan Wang

Nanchang Institute of Technology

Article

Keywords: Dance Movement Quality Assessment, Spatiotemporal Skeletal Representation, Reference-Conditioned Alignment, Action Quality Assessment, Order-Consistent Learning

Posted Date: October 16th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10981960/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Automatic dance movement quality assessment via spatiotemporal skeletal representation learning

Zhen Ji^{1,*} and Panpan Wang²

¹Art teaching and Research Department, Hangzhou Polytechnic University, HangZhou, 310018, China

²School of Computer Information Engineering, Nanchang Institute of Technology, Nanchang, 330044, China

*azaltejp2316624@outlook.com

ABSTRACT

Automatic dance movement quality assessment requires models to distinguish execution quality from movement identity while remaining robust to differences in choreography, timing, and data acquisition. This study proposes a Reference-Aware Spatiotemporal Deviation Quality Network (RSDQ-Net) for skeletal dance assessment. Heterogeneous motion sources are unified into a 19-node canonical skeleton and encoded through relation-aware spatiotemporal representation learning. A learnable reference-conditioned phase alignment establishes soft target-reference correspondence under a temporal phase prior, enabling spatial execution deviation and temporal phase deviation to be modeled separately. An order-consistent latent quality formulation further integrates ordinal expertise labels and pairwise human preferences without imposing an artificial absolute score. Experiments on AIST++, ImperialDance, and CDRG-UNIMAS-TIKTOK evaluate dance representation, ordinal quality assessment, and human-ranked quality prediction under complementary protocols. RSDQ-Net achieves 93.6% accuracy on AIST++, a QWK of 0.814 with an ordinal MAE of 0.207 on ImperialDance, and 81.4% pairwise accuracy with a Kendall τ_b of 0.638 on CDRG. Ablation, transfer, robustness, and qualitative analyses further support the effectiveness of reference-aware alignment and explicit spatiotemporal deviation modeling, highlighting its potential for reference-guided dance training and assessment.

Key Words: Dance Movement Quality Assessment, Spatiotemporal Skeletal Representation, Reference-Conditioned Alignment, Action Quality Assessment, Order-Consistent Learning

1 Introduction

Dance training requires performers to reproduce complex movement patterns with technical consistency, temporal control, and whole-body coordination. Dance quality depends on the interaction of spatial posture, temporal dynamics, and inter-joint coordination, meaning that visually similar performances may still differ substantially in joint configuration, timing, movement amplitude, or coordination¹. In practical training, these characteristics are commonly evaluated through instructor observation and comparison with technical demonstrations. Although expert judgment remains important, repeated manual assessment is time-consuming and can be affected by subjective variation². Automatic movement assessment therefore provides a promising means of supporting reproducible and scalable analysis of dance execution.

Advances in computer vision and human motion understanding have established a technical basis for automated dance analysis³. Human action recognition has progressed from appearance-based RGB modeling to pose estimation and skeleton-based representations, which describe movement through structured body keypoints and reduce sensitivity to clothing, illumination, and background variation⁴. Graph-based models further represent joints as nodes and anatomical connections as edges, allowing spatial body relationships to be learned jointly with temporal motion patterns⁵. Such representations are particularly suitable for dance because coordinated body configurations and movement trajectories carry more direct information about execution than visual appearance. Most skeleton-based approaches, however, remain focused on recognizing which action is performed rather than evaluating how well it is executed⁶.

This distinction is critical for dance movement assessment. Correct recognition of a dance movement does not imply correct execution of that movement. Two performers may complete the same choreography while differing in body alignment, joint trajectories, timing, balance, or coordination. Conventional recognition models generally suppress within-class variation to obtain stable semantic categories, whereas action quality assessment must preserve and interpret these variations⁷. A useful assessment model must therefore combine movement semantics with sensitivity to execution-dependent differences, making direct adaptation of classification-oriented action recognition insufficient for reliable dance quality evaluation.

Reference-based assessment offers a natural solution because dance training commonly involves comparison with a professional or standardized demonstration⁸. Existing approaches often measure target-reference similarity through joint-angle differences, trajectory distances, pose similarity, or dynamic time warping⁹. Such techniques can compensate for differences

in sequence duration and execution speed, but fixed alignment is independent of downstream quality supervision. Temporal warping may also remove meaningful timing errors when a performer reaches approximately correct poses but executes movement phases too early or too late¹⁰. Reference-based assessment should therefore establish correspondence without eliminating quality-relevant temporal differences. This requirement motivates a learnable phase alignment mechanism that can identify semantically corresponding motion phases while preserving temporal deviation as part of the quality evidence¹¹.

Another challenge is the heterogeneity of quality supervision in public dance datasets. ImperialDance represents proficiency through ordered expertise levels,

$$\text{Beginner} < \text{Intermediate} < \text{Expert}, \quad (1)$$

whereas CDRG-UNIMAS-TIKTOK provides human pairwise judgments that define relative relations such as

$$Q_i > Q_j. \quad (2)$$

These annotations cannot be assumed to share a common continuous scoring scale, and assigning arbitrary numerical scores would introduce unsupported assumptions. Both forms of supervision nevertheless encode movement-quality ordering¹². Expertise labels provide coarse ordinal constraints, whereas pairwise judgments describe relative quality between individual performances. Their shared ordering structure provides a basis for learning a common latent quality representation across heterogeneous supervision schemes¹³.

To address these challenges, this study proposes the Reference-Aware Spatiotemporal Deviation Quality Network (RSDQ-Net) for automatic dance movement quality assessment from skeletal motion¹⁴. Heterogeneous motion sources are converted into a unified canonical skeletal representation and encoded using anatomical and coordination relationships. A reference-conditioned phase alignment mechanism establishes soft correspondences between target and reference motions under a temporal phase prior. The resulting correspondence is used to separate spatial movement deviation from temporal phase deviation rather than reducing both to a single similarity measure. These features are mapped to a shared latent quality variable that supports different ordered supervision forms¹⁵. AIST++ is used for dance-specific skeletal representation learning, ImperialDance evaluates ordinal quality assessment under choreography-disjoint settings, and CDRG-UNIMAS-TIKTOK evaluates human-ranked quality assessment under subject-independent real-world conditions.

The main contributions of this study are threefold. A reference-conditioned learnable phase alignment mechanism is developed to establish quality-aware target-reference correspondence while preserving temporal execution differences, enabling spatial and temporal deviations to be modeled separately. An order-consistent quality learning strategy is introduced to integrate ordinal expertise labels and human pairwise rankings within a shared latent quality space without constructing an artificial continuous score. A cross-source skeletal evaluation framework is established across AIST++, ImperialDance, and CDRG-UNIMAS-TIKTOK to evaluate dance representation learning, unseen-choreography ordinal quality assessment, and real-world subject-independent quality ranking within a unified setting.

2 Problem Formulation

Dance movement quality assessment is formulated as a skeletal motion analysis problem in which heterogeneous dance recordings are represented in a common anatomical and temporal space. AIST++ provides three-dimensional dance motion, ImperialDance contains motion-capture sequences with a different skeletal definition, and CDRG-UNIMAS-TIKTOK consists of real-world RGB videos from which three-dimensional body motion is reconstructed. All sequences are transformed into a shared canonical representation before dataset-specific quality supervision is introduced¹⁶.

2.1 Canonical Dance Skeletal Sequence Representation

A dance performance is represented as a temporal sequence of three-dimensional skeletal configurations,

$$X = \{x_{t,v}\} \in \mathbb{R}^{T \times 19 \times 3}, \quad (3)$$

where T denotes the sequence length and v indexes the 19 canonical skeletal nodes covering the trunk, upper limbs, and lower limbs. AIST++, ImperialDance, and reconstructed CDRG motion are mapped to the same node ordering and anatomical connectivity. To retain complementary movement information, each canonical node is represented by its normalized position $P_{t,v}$, temporal displacement $V_{t,v}$, anatomical bone vector $B_{t,v}$, and observation confidence $C_{t,v}$. The velocity is computed from consecutive normalized positions, the bone vector is defined relative to the anatomical parent node, and confidence reflects observation reliability when available¹⁷. The resulting node feature is

$$F_{t,v} = [P_{t,v}, V_{t,v}, B_{t,v}, C_{t,v}] \in \mathbb{R}^{10}, \quad F \in \mathbb{R}^{T \times 19 \times 10}. \quad (4)$$

This representation provides a common skeletal feature space while retaining spatial configuration, local motion, anatomical structure, and observation reliability. In the following problem formulation, X_s and X_r denote the target and reference motions conceptually, while their implementation-level encoder inputs are specified in Section 3.

2.2 Reference-Conditioned Dance Movement Quality Assessment

Dance movement quality assessment aims to determine how well a target performance is executed relative to an appropriate technical reference. Let X_s and X_r denote the target and reference motions, respectively¹⁸. They may have different sequence lengths because the same choreography can be performed with different durations and local timing patterns.

The assessment problem is formulated as

$$q = \mathcal{Q}(X_s, X_r), \quad q \in \mathbb{R}, \quad (5)$$

where $\mathcal{Q}(\cdot)$ denotes a reference-conditioned quality assessment function¹⁹. The latent variable q represents relative movement quality rather than an absolute score because the datasets used in this study do not share a common continuous rating scale. For two performances evaluated under compatible reference conditions, $q_i > q_j$ indicates that performance i is estimated to exhibit higher execution quality than performance j . The magnitude of the difference represents separation within the learned quality space rather than a directly interpretable difference in professional dance scores²⁰.

2.3 Ordered Quality Supervision Across Heterogeneous Datasets

ImperialDance and CDRG-UNIMAS-TIKTOK provide different forms of movement-quality supervision²¹. ImperialDance describes proficiency through ordered expertise levels, whereas CDRG represents quality through human pairwise preferences. Although their annotation formats differ, both constrain relative quality ordering rather than define a common absolute score²².

For ImperialDance, the ordinal labels are encoded as

$$y^{ord} \in \{0, 1, 2\}, \quad 0 \equiv \text{Beginner} < 1 \equiv \text{Intermediate} < 2 \equiv \text{Expert}. \quad (6)$$

This structure preserves the ordered distance between expertise levels rather than treating them as independent categories²³.

For CDRG, two performances X_i and X_j from the same dance challenge are compared through

$$y_{ij}^{rank} = \mathbb{I}(Q_i > Q_j), \quad (7)$$

where $y_{ij}^{rank} = 1$ indicates that performance i is preferred over j . This supervision specifies relative quality without requiring absolute scores²⁴.

ImperialDance encourages $q_{\text{Beginner}} < q_{\text{Intermediate}} < q_{\text{Expert}}$, while a positive CDRG preference requires $q_i > q_j$. The learning objective is consequently to preserve observed quality ordering without constructing an artificial continuous scoring scale.

3 Reference-Aware Spatiotemporal Deviation Quality Network

3.1 Framework Overview

The proposed Reference-Aware Spatiotemporal Deviation Quality Network (RSDQ-Net) formulates dance movement quality assessment as a reference-conditioned comparison problem. As illustrated in Figure 1, the target and reference motions are converted to the canonical node features and global trajectories defined in Section 3.2 and processed by a shared spatiotemporal skeletal encoder.

The target and reference representations are obtained using shared encoder parameters,

$$H_s = E_\theta(F_s^{can}, G_s), \quad H_r = E_\theta(F_r^{can}, G_r). \quad (8)$$

Because corresponding movement phases may occur at different temporal positions, RSDQ-Net learns a soft target-reference correspondence rather than relying on direct frame-wise comparison. The aligned representations are used to derive spatial execution differences and temporal phase deviations, which are subsequently aggregated into the quality-sensitive representation Z_q . The final latent movement quality is obtained as

$$q = g_\phi(Z_q), \quad (9)$$

where q represents relative execution quality rather than an absolute professional score. Training follows two stages. AIST++ is used to initialize the shared encoder with dance-specific motion representations, after which ImperialDance and CDRG-UNIMAS-TIKTOK are used to optimize reference alignment, deviation modeling, and order-consistent quality learning²⁵. The detailed encoder, alignment mechanism, deviation representation, quality formulation, and optimization objectives are described in the following subsections²⁶.

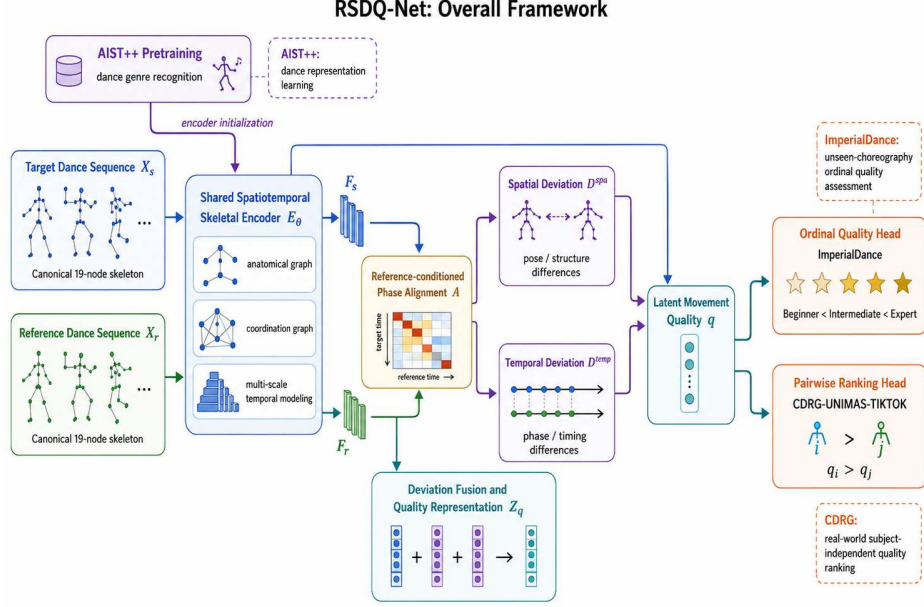


Figure 1. Overall architecture of RSDQ-Net. Target and reference motions are encoded in a shared skeletal representation space, followed by reference-conditioned phase alignment, spatiotemporal deviation modeling, and order-consistent quality learning. AIST++ is used for dance-specific representation pretraining, while ImperialDance and CDRG-UNIMAS-TIKTOK provide ordinal and pairwise quality supervision.

3.2 Cross-Dataset Canonical Skeletal Representation

The three datasets differ in their original motion representations. AIST++ provides SMPL-based three-dimensional motion, ImperialDance provides 21-node motion-capture sequences, and CDRG-UNIMAS-TIKTOK contains RGB videos from which three-dimensional body motion is reconstructed. To enable shared modeling, all sequences are converted to the same 19-node canonical skeleton, as illustrated in Figure 2.

The canonical topology contains 19 nodes: Hip, Ab, Chest, Neck, Head, bilateral Shoulder, Upper Arm, Forearm, Hand, Thigh, Shin, and Foot nodes. It preserves the trunk and bilateral limb structures while excluding source-specific landmarks that cannot be matched consistently across datasets²⁷. AIST++ motion is mapped from SMPL to this topology using direct anatomical correspondences where available. Segment-centered landmarks for the upper arms, forearms, thighs, and shins are obtained as the midpoint of the corresponding adjacent joints²⁸. ImperialDance is converted by removing its left and right toe nodes. For CDRG, each RGB video is first converted to three-dimensional body motion using a frozen monocular reconstruction model and is subsequently mapped through the same SMPL-to-canonical transformation. The reconstruction model is used only during preprocessing and is not optimized using dance-quality supervision²⁹. Spatial normalization reduces differences caused by global position and performer scale while preserving meaningful whole-body displacement. Let r_t denote the Hip position at frame t . Root-centered local coordinates, the relative global trajectory, and the sequence-level body scale are defined jointly as

$$\hat{x}_{t,v} = \frac{x_{t,v} - r_t}{s}, \quad \hat{g}_t = \frac{r_t - r_1}{s}, \quad s = \frac{1}{T} \sum_{t=1}^T \|x_{t,\text{Neck}} - x_{t,\text{Hip}}\|_2. \quad (10)$$

Initial-orientation normalization is applied once per sequence rather than independently at every frame so that subsequent turns remain part of the motion. An orthonormal body coordinate frame is constructed from the mean initial Hip-to-Neck direction and the left-to-right shoulder direction, with the shoulder axis orthogonalized against the vertical body axis and the third axis obtained by their cross product. Let R_0 denote the resulting orientation matrix. The canonical local coordinates and preserved global trajectory are then

$$P_{t,v} = R_0^\top \hat{x}_{t,v}, \quad G_t = R_0^\top \hat{g}_t. \quad (11)$$

All sequences are temporally resampled to 30 Hz while preserving their physical duration. Sequence lengths therefore remain variable rather than being forced to a common duration. Within each mini-batch, shorter sequences are padded and

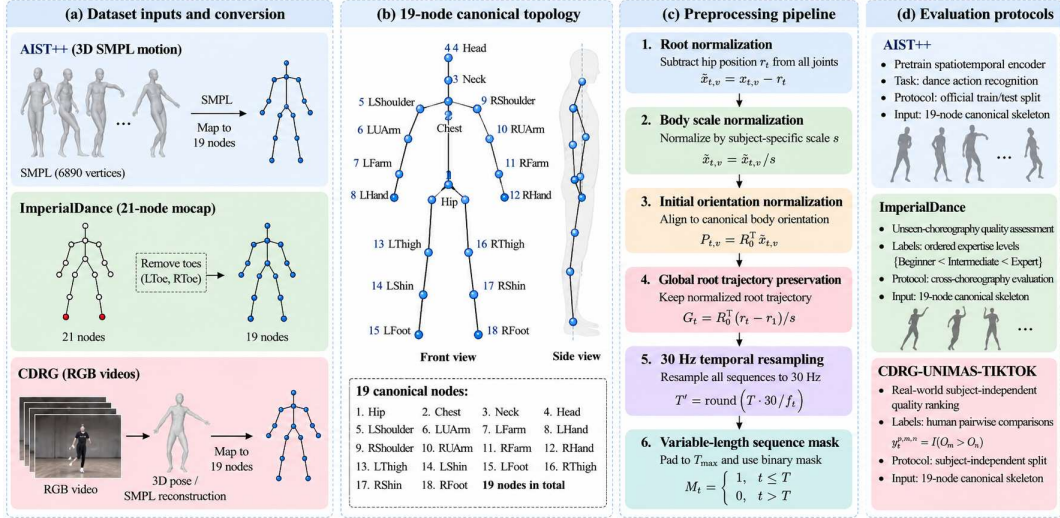


Figure 2. Cross-dataset skeletal unification and evaluation protocols. AIST++ and reconstructed CDRG motion are mapped from SMPL to a common 19-node topology, while ImperialDance is converted from its original 21-node skeleton. All sequences undergo spatial normalization, 30-Hz temporal resampling, and variable-length masking before dataset-specific evaluation.

accompanied by a binary validity mask M_t , where $M_t = 1$ denotes a valid frame and $M_t = 0$ denotes padding. Masked positions are excluded from temporal aggregation, attention, alignment, and loss computation³⁰.

After topology conversion and normalization, the three datasets share the same canonical representation,

$$X^{can} \in \mathbb{R}^{T' \times 19 \times 3}, \quad F^{can} \in \mathbb{R}^{T' \times 19 \times 10}, \quad G \in \mathbb{R}^{T' \times 3}. \quad (12)$$

3.3 Relation-Aware Spatiotemporal Skeletal Encoder

Dance execution depends not only on local anatomical connectivity but also on coordinated interactions between spatially separated body regions. RSDQ-Net therefore represents the 19-node canonical skeleton using two complementary relations³¹. The anatomical relation follows the physical skeletal structure, while the coordination relation connects non-local but functionally related regions, including bilateral shoulders, hands, thighs, and feet. Their adjacency matrices are denoted by A^{anat} and A^{coord} , respectively. The encoder operates on the canonical node features F^{can} and preserved global trajectory G defined in Section 3.2. Self-connections are added to each graph and the corresponding adjacency matrices are symmetrically normalized. Starting from $H^{(0)} = F^{can}$, relation-aware spatial propagation is performed as

$$H^{(l+1)} = \sum_{r \in \{anat, coord\}} \alpha_r^{(l)} \tilde{A}_r H^{(l)} W_r^{(l)}, \quad (13)$$

where $\tilde{A}_r$ is the normalized adjacency matrix for relation r , $W_r^{(l)}$ is its learnable transformation, and $\alpha_r^{(l)}$ is a learnable weight normalized across the two relations. A nonlinear activation is applied after relation aggregation³². The output of the final graph layer is denoted by $H^{spa} = H^{(L)}$. Temporal dynamics are subsequently modeled using parallel temporal convolutions with dilation factors $\{1, 2, 4\}$. Their outputs are concatenated and projected to obtain a multi-scale node representation, which captures both local transitions and longer-range movement patterns³³. The resulting multi-scale node representation is retained as $H_{t,v}$ for subsequent body-region deviation modeling. To incorporate whole-body displacement, the node features are aggregated into frame-wise motion tokens and combined with the normalized global trajectory,

$$h_t = \frac{1}{19} \sum_{v=1}^{19} H_{t,v} + W_G G_t, \quad (14)$$

where W_G projects the three-dimensional global trajectory into the frame-level feature space.

A lightweight scaled dot-product temporal self-attention layer with padding masking and a residual connection is then applied to the frame-wise tokens to capture dependencies between distant movement phases³⁴. Its output constitutes the frame-level skeletal representation,

$$H = E_\theta(F^{can}, G) \in \mathbb{R}^{T' \times d}. \quad (15)$$

3.4 Reference-Conditioned Phase Alignment

Target and reference performances of the same choreography may differ in sequence length, local execution speed, and phase progression. Direct frame-wise comparison can therefore confuse harmless temporal displacement with genuine execution differences³⁵. RSDQ-Net instead estimates soft correspondence between the encoded target and reference sequences in the learned motion space³⁶.

For target position t and reference position j , the alignment weight is defined as

$$a_{tj} = \text{softmax}_j \left[\frac{(W_Q H_t^s) (W_K H_j^r)^\top}{\sqrt{d_a}} - \lambda \left| \frac{t}{T_s} - \frac{j}{T_r} \right| \right], \quad (16)$$

where W_Q and W_K project the frame representations to an alignment space of dimension d_a . The phase-prior strength is constrained to be non-negative by setting $\lambda = \text{softplus}(\lambda_0)$, with λ_0 learned jointly with the network.

The alignment weights form a soft correspondence matrix $A = [a_{tj}]$ between target and reference phases. The reference representation aligned to target position t is

$$\bar{H}_t^r = \sum_{j=1}^{T_r} a_{tj} H_j^r. \quad (17)$$

Soft correspondence allows neighboring reference states to contribute when phase boundaries are uncertain, rather than imposing a hard one-to-one match.

3.5 Spatiotemporal Movement Deviation Representation

The alignment learned in Section 3.4 establishes correspondence between target and reference phases but does not determine whether their remaining differences arise from spatial execution or temporal progression. RSDQ-Net therefore preserves these two sources of deviation separately³⁷. The alignment weights are first applied to the retained node-level reference features. The aligned reference feature and the corresponding spatial deviation are defined jointly as

$$\bar{H}_{t,v}^r = \sum_{j=1}^{T_r} a_{tj} H_{j,v}^r, \quad D_{t,v}^{spa} = |H_{t,v}^s - \bar{H}_{t,v}^r|. \quad (18)$$

This node-level formulation retains body-region information and allows spatial differences in posture and inter-joint configuration to remain distinguishable after phase correspondence has been established.

Temporal deviation is derived directly from the alignment distribution. The expected reference phase, its displacement from the normalized target phase, and the local change in this displacement are

$$\begin{aligned} \mu_t &= \sum_{j=1}^{T_r} a_{tj} \frac{j}{T_r}, \\ d_t^{temp} &= \mu_t - \frac{t}{T_s}, \\ \Delta d_t^{temp} &= d_t^{temp} - d_{t-1}^{temp}, \quad \Delta d_1^{temp} = 0. \end{aligned} \quad (19)$$

Here, d_t^{temp} measures phase displacement relative to the reference, while Δd_t^{temp} captures local changes in phase progression associated with acceleration or delay.

Spatial and temporal information is integrated with the target representation at each node,

$$D_{t,v} = [H_{t,v}^s, D_{t,v}^{spa}, d_t^{temp}, \Delta d_t^{temp}]. \quad (20)$$

The temporal terms are shared across nodes at the same target phase, whereas the spatial deviation remains node specific. Retaining $H_{t,v}^s$ provides movement context so that similar deviation magnitudes can be interpreted differently across different motion states. The canonical skeleton is divided into upper-body, lower-body, and trunk regions. Masked mean pooling over the valid temporal positions and corresponding node subsets produces regional representations z_U , z_L , and z_T , while pooling over all nodes produces the global representation z_G . These representations are concatenated and passed through a nonlinear projection to obtain the quality-sensitive representation z_q . This preserves both regional execution differences and whole-body spatiotemporal information for subsequent quality learning.

3.6 Order-Consistent Movement Quality Learning

ImperialDance and CDRG-UNIMAS-TIKTOK provide different annotation forms but share the quality-ordering structure established in Section 2.3. RSDQ-Net therefore uses the scalar latent quality $q = g_\phi(Z_q)$ defined in Section 3.1 as their common representation, where larger values indicate higher relative execution quality rather than an absolute professional score.

Differences between dataset-specific supervision scales are handled through a monotonic calibration,

$$q_d = a_d q + b_d, \quad a_d = \text{softplus}(\alpha_d) > 0, \quad (21)$$

where $d \in \{\text{Imp}, \text{CDRG}\}$. The positive scale factor allows the location and dispersion of the latent quality representation to adapt to each dataset without reversing its ordering.

For ImperialDance, the three expertise levels are modeled using two ordered thresholds and cumulative ordinal probabilities,

$$\begin{aligned} \tau_1 &= \gamma_1, & \tau_2 &= \tau_1 + \text{softplus}(\gamma_2), \\ P(y^{\text{ord}} > k \mid q_{\text{Imp}}) &= \sigma(q_{\text{Imp}} - \tau_{k+1}), & k &\in \{0, 1\}. \end{aligned} \quad (22)$$

The construction guarantees ordered thresholds and preserves the ordinal relationship among Beginner, Intermediate, and Expert performances.

For CDRG, pairwise preference is predicted from the difference between the calibrated latent qualities,

$$P(i \succ j) = \sigma(q_{\text{CDRG},i} - q_{\text{CDRG},j}) = \sigma[a_{\text{CDRG}}(q_i - q_j)]. \quad (23)$$

The dataset-specific offset cancels in pairwise comparison, while the positive calibration scale preserves the ordering of q_i and q_j . Ordinal expertise labels and pairwise human preferences can therefore constrain the same latent quality variable without requiring a shared absolute scoring scale.

3.7 Training Objectives and Optimization

RSDQ-Net is trained in two stages. Stage I initializes the skeletal encoder using dance-genre recognition on AIST++, whereas Stage II adapts the complete framework to movement-quality assessment using alternating mini-batches from ImperialDance and CDRG-UNIMAS-TIKTOK. During Stage I, the temporally pooled encoder representation is passed to a dance-genre classification head and optimized using standard multi-class cross-entropy. The resulting encoder parameters provide the initialization for subsequent quality learning. This pretraining stage supplies dance-specific motion representations but does not use movement-quality annotations. For ImperialDance, the cumulative ordinal targets are defined as $z_{n,k} = \mathbb{I}(y_n^{\text{ord}} > k)$ for $k \in \{0, 1\}$. Using the cumulative probabilities defined in Section 3.6, the ordinal objective is

$$\mathcal{L}_{\text{ord}} = -\frac{1}{2N_o} \sum_{n=1}^{N_o} \sum_{k=0}^1 [z_{n,k} \log \hat{z}_{n,k} + (1 - z_{n,k}) \log(1 - \hat{z}_{n,k})]. \quad (24)$$

For CDRG-UNIMAS-TIKTOK, the pairwise probability defined in Section 3.6 is optimized using binary cross-entropy,

$$\mathcal{L}_{\text{rank}} = -\frac{1}{N_r} \sum_{(i,j)} [y_{ij}^{\text{rank}} \log \hat{y}_{ij}^{\text{rank}} + (1 - y_{ij}^{\text{rank}}) \log(1 - \hat{y}_{ij}^{\text{rank}})], \quad (25)$$

where N_r denotes the number of training pairs in the mini-batch.

The learned phase correspondence is regularized using the expected reference phase μ_t defined in Section 3.5. Backward phase transitions are penalized by a monotonicity term, while abrupt changes in local phase progression are controlled by a smoothness term,

$$\begin{aligned} \mathcal{L}_{\text{mon}} &= \frac{1}{T_s - 1} \sum_{t=2}^{T_s} [\max(0, \mu_{t-1} - \mu_t)]^2, \\ \mathcal{L}_{\text{sm}} &= \frac{1}{T_s - 2} \sum_{t=3}^{T_s} [(\mu_t - \mu_{t-1}) - (\mu_{t-1} - \mu_{t-2})]^2, \\ \mathcal{L}_{\text{align}} &= \mathcal{L}_{\text{mon}} + \eta \mathcal{L}_{\text{sm}}. \end{aligned} \quad (26)$$

Only valid temporal positions contribute to these terms.

During Stage II, ImperialDance and CDRG mini-batches alternately update the shared encoder, alignment module, deviation representation, and latent quality mapping. The objective applied to each mini-batch is

$$\mathcal{L} = \begin{cases} \lambda_o \mathcal{L}_{ord} + \lambda_a \mathcal{L}_{align}, & \text{ImperialDance batch,} \\ \lambda_r \mathcal{L}_{rank} + \lambda_a \mathcal{L}_{align}, & \text{CDRG batch.} \end{cases} \quad (27)$$

The loss weights are selected using the validation partitions and remain fixed during final evaluation. This alternating optimization allows ordinal expertise labels and pairwise human preferences to constrain the same latent quality representation without requiring the two datasets to share an absolute scoring scale.

4 Experimental Setup

Experiments were designed to evaluate dance-specific skeletal representation learning, ordinal movement-quality assessment, and human-ranked quality assessment under complementary evaluation settings. AIST++, ImperialDance, and CDRG-UNIMAS-TIKTOK were used according to their available annotations and acquisition conditions. All datasets were converted to the canonical representation described in Section 3.2 before being processed by the proposed model. Dataset characteristics and evaluation protocols are summarized in Table 1.

4.1 Datasets and Evaluation Protocols

AIST++³⁸ contains 1,408 three-dimensional dance sequences covering 10 dance genres and provides SMPL-based body motion. The dataset is used for dance-specific representation pretraining rather than quality supervision because it does not provide movement-quality annotations. We follow the official subject-based partition consisting of 868 training, 70 validation, and 470 test sequences. The test set contains performers unseen during training, allowing the representation ability of the skeletal encoder to be evaluated without subject overlap. The encoder is first optimized for dance-genre recognition on this dataset and then transferred to the quality-assessment tasks. ImperialDance³⁹ provides three-dimensional motion-capture recordings covering 20 choreographies performed at Beginner, Intermediate, and Expert levels. The original motion is recorded using a 21-node skeletal representation and is converted to the 19-node canonical topology described in Section 3.2. To prevent the model from exploiting choreography-specific patterns, evaluation is conducted using five-fold choreography-disjoint cross-validation. In each fold, 12 choreographies are assigned to training, four to validation, and four to testing. All expertise levels and repeated performances belonging to the same choreography remain within the same partition. For reference-conditioned assessment, an Expert demonstration of the corresponding choreography is used as the reference motion. The selected reference is excluded from the target samples evaluated under that reference. Consequently, test choreographies are not observed during parameter learning, although their Expert demonstrations are available as references at assessment time. This setting is therefore regarded as unseen-choreography reference-guided assessment rather than fully reference-free choreography generalization. Performance is aggregated across the five folds. CDRG-UNIMAS-TIKTOK⁴⁰ contains 240 real-world dance videos from 20 subjects performing 12 TikTok dance challenges. Unlike AIST++ and ImperialDance, it does not provide native three-dimensional skeletal motion; the RGB videos are therefore converted to three-dimensional body motion using the frozen reconstruction procedure described in Section 3.2. Human comparisons provide pairwise quality relations between performances of the same challenge, making the dataset suitable for evaluating agreement with human-ranked movement quality. A five-fold subject-disjoint protocol is adopted for CDRG. In each fold, 12 subjects are used for training, four for validation, and four for testing, with all 12 challenge videos of the same subject retained in the same partition. Pair generation is performed only after subject splitting. This procedure prevents performances from the same subject from appearing across training and evaluation pairs. For each challenge, the reference performance is selected only from the training subjects and remains fixed for the corresponding fold, ensuring that reference construction does not introduce information from validation or test subjects.

Table 1. Datasets and evaluation protocols used in the experiments.

Dataset	Data	Supervision	Split protocol	Experimental role
AIST++	1,408 3D sequences, 10 genres	Dance genre	Official subject-based split	Encoder pretraining
ImperialDance	20 choreographies, 5 subjects, 3 levels	Beginner < Intermediate < Expert	5-fold choreography-disjoint	Ordinal quality assessment
CDRG-UNIMAS-TIKTOK	240 videos, 20 subjects, 12 challenges	Human pairwise preference	5-fold subject-disjoint	Pairwise quality ranking

4.2 Baseline Methods

Baseline methods are selected separately for the three experimental tasks because AIST++, ImperialDance, and CDRG-UNIMAS-TIKTOK provide different supervision forms and evaluation objectives. All baselines are evaluated under the

same data partitions described in Section 4.1, and skeleton-based models use the unified canonical representation whenever applicable. For dance representation learning on AIST++, four representative skeleton-based action recognition models are compared. ST-GCN serves as a classical spatial-temporal graph convolutional baseline that models skeletal joints using a predefined anatomical graph. 2s-AGCN extends graph convolution with adaptive graph learning and complementary joint and bone streams. CTR-GCN improves spatial modeling through channel-wise topology refinement, while HD-GCN introduces hierarchical decomposition to capture skeletal relations at multiple structural levels. These methods provide increasingly strong reference points for evaluating whether the proposed relation-aware encoder can learn competitive dance-specific skeletal representations. For ordinal quality assessment on ImperialDance, both reference-based and learning-based methods are included. DTW + Joint Distance measures frame-level skeletal discrepancy after dynamic time warping and provides a non-learned reference-comparison baseline. ST-GCN + CE and CTR-GCN + CE use cross-entropy classification over the three expertise levels and therefore treat Beginner, Intermediate, and Expert as nominal categories. ST-GCN + Ordinal replaces the conventional classification head with an ordinal prediction objective, enabling the contribution of ordered supervision to be isolated from the proposed architecture. 2T-GCN is also included as a stronger graph-based movement assessment baseline. All methods are evaluated using the same five-fold choreography-disjoint protocol. For human-ranked quality assessment on CDRG, the original codebook-based matching approach is evaluated using Codebook + Edit Distance and its variant without codebook compression. Both variants are rerun under the subject-disjoint protocol defined in Section 4.1. DTW-3D directly compares reconstructed three-dimensional skeletal trajectories after temporal alignment and serves as a geometry-based reference baseline. ST-GCN + RankNet and CTR-GCN + RankNet combine representative skeleton encoders with pairwise ranking supervision, allowing comparison between the proposed reference-aware framework and conventional graph representations trained with the same ranking objective. Pair construction for all CDRG baselines is performed only after subject splitting to maintain identical leakage-free evaluation conditions.

4.3 Evaluation Metrics

Evaluation metrics are selected according to the supervision form of each dataset. For AIST++, dance-category recognition is evaluated using Accuracy and Macro-F1. Accuracy measures the overall proportion of correctly classified sequences, while Macro-F1 gives equal importance to each dance genre and is therefore less affected by class imbalance. For ImperialDance, ordinal quality assessment is evaluated using Accuracy, Macro-F1, Ordinal MAE, quadratic weighted kappa (QWK), and Spearman’s rank correlation coefficient ρ . Ordinal MAE is computed from the absolute difference between predicted and ground-truth expertise levels and therefore penalizes larger ordinal errors more strongly. QWK measures agreement while accounting for the ordered distance between categories, whereas Spearman’s ρ evaluates consistency between predicted latent quality and the ordinal expertise ordering. For CDRG-UNIMAS-TIKTOK, pairwise quality assessment is evaluated using Pairwise Accuracy, Kendall’s τ_b , Spearman’s ρ , and NDCG@5. Pairwise Accuracy measures agreement with individual human preferences, whereas Kendall’s τ_b and Spearman’s ρ evaluate the consistency between predicted and human-derived performance rankings within each dance challenge. NDCG@5 evaluates the quality of the highest-ranked performances and complements the global rank-correlation metrics. For the ImperialDance and CDRG-UNIMAS-TIKTOK quality-assessment experiments, results are reported as mean $\pm$ standard deviation over 25 fold-seed evaluations obtained from five cross-validation folds and five random seeds. RSDQ-Net and the corresponding learning-based baselines are evaluated using identical data partitions and random seeds, allowing matched comparisons. Statistical significance is assessed using two-sided paired Wilcoxon signed-rank tests over the matched fold-seed evaluations. QWK and Ordinal MAE are treated as the primary metrics for ImperialDance, whereas Pairwise Accuracy and Kendall’s τ_b are treated as the primary metrics for CDRG-UNIMAS-TIKTOK. Holm correction is applied separately within each task across the two prespecified primary metrics, with adjusted $p < 0.05$ considered statistically significant.

4.4 Implementation Details

All experiments are implemented in PyTorch 2.4.1 and conducted on an NVIDIA RTX 4090 GPU with 24 GB of memory. The models are optimized using AdamW with an initial learning rate of 1×10^{-3} and a weight decay of 1×10^{-4} . The batch sizes are 64 for AIST++ pretraining and 16 for quality learning. Stage I is trained for 100 epochs on AIST++, while Stage II is trained for 150 epochs using alternating mini-batches from ImperialDance and CDRG-UNIMAS-TIKTOK. During Stage II, the learning rate is reduced to 3×10^{-4} and adjusted using a cosine annealing scheduler with a 5-epoch linear warm-up and a minimum learning rate of 1×10^{-6} . Model selection is based exclusively on the corresponding validation partition. The skeletal encoder uses a hidden dimension of 128 and three relation-aware graph layers. Multi-scale temporal convolutions employ a kernel size of 5 with dilation factors $\{1, 2, 4\}$ as defined in Section 3.3. A dropout rate of 0.3 is applied to the encoder and quality-learning layers. Unless otherwise stated, all compared learning-based methods are trained under consistent optimization settings where applicable. For reference-conditioned alignment, the query and key projections have dimension 64.

The phase-prior parameter is constrained by

$$\lambda = \text{softplus}(\lambda_0), \quad (28)$$

with λ_0 initialized to 0 and learned jointly with the network. The alignment smoothness coefficient η in $\mathcal{L}_{align}$ is set to 0.1. The Stage II loss weights are

$$\lambda_o = 1.0, \quad \lambda_r = 1.0, \quad \lambda_a = 0.1, \quad (29)$$

and are selected using the validation partitions before final testing. Equal weights are assigned to the ordinal and pairwise quality objectives so that neither supervision source dominates optimization, while the smaller alignment-regularization weight constrains phase correspondence without overwhelming the task-specific quality losses. For CDRG-UNIMAS-TIKTOK, three-dimensional body motion is reconstructed using WHAM with the publicly released pretrained weights. Experiments are repeated using random seeds $\{0, 1, 2, 3, 4\}$ for weight initialization, mini-batch sampling, and other stochastic operations. Each of the five cross-validation folds is evaluated under all five random seeds, resulting in 25 fold-seed evaluations for each method. Reported means and standard deviations for the fold-based quality-assessment experiments are computed over these 25 evaluations. Identical folds and random seeds are used for RSDQ-Net and the corresponding learning-based baselines to support matched statistical comparison.

5 Experimental Results and Analysis

5.1 Dance Skeletal Representation Performance

As shown in Table 2, the proposed RSDQ Encoder achieves the best performance on AIST++, obtaining an Accuracy of 93.6% and a Macro-F1 of 93.3%. Compared with the classical ST-GCN, these results correspond to improvements of 5.2 and 5.4 percentage points, respectively, indicating that conventional propagation restricted primarily to predefined anatomical connectivity is less effective for representing the complex coordination patterns present in dance motion. The performance increases progressively from ST-GCN to 2s-AGCN, CTR-GCN, and HD-GCN, reflecting the benefit of increasingly sophisticated skeletal topology modeling. Among the compared baselines, HD-GCN provides the strongest results, reaching 93.0% Accuracy and 92.6% Macro-F1. RSDQ Encoder further improves these metrics by 0.6 and 0.7 percentage points, respectively. Although the margin over HD-GCN is moderate, this result is meaningful because dance-genre recognition is used primarily to verify the representation capability of the shared encoder rather than as the final quality-assessment objective. The improvement can be attributed to the relation-aware skeletal representation used in the proposed encoder. In contrast to models that mainly emphasize anatomical connectivity, RSDQ Encoder jointly models anatomical and coordination relations, enabling interactions between spatially separated body regions, such as bilateral limbs, to contribute directly to motion representation. This property is particularly relevant to dance, where genre-specific movement patterns frequently depend on coordinated whole-body configurations rather than isolated joint trajectories. The multi-scale temporal convolutions with dilation factors $\{1, 2, 4\}$ further capture motion characteristics at different temporal ranges, while the temporal attention layer complements local convolutional modeling by representing dependencies between distant movement phases. The small gap between Accuracy and Macro-F1 also indicates that the improvement is not concentrated in a limited subset of dance genres. These results support the use of the AIST++-pretrained RSDQ Encoder as the shared skeletal representation backbone for the subsequent ordinal and pairwise movement-quality assessment tasks.

Table 2. Dance-genre recognition performance on AIST++.

Method	Accuracy (%)	Macro-F1 (%)
ST-GCN ⁴¹ .	88.4	87.9
2s-AGCN ⁴² .	90.7	90.2
CTR-GCN ⁴³ .	92.2	91.7
HD-GCN ⁴⁴ .	93.0	92.6
RSDQ Encoder	93.6	93.3

5.2 Ordinal Dance Movement Quality Assessment

The first main quality-assessment experiment evaluates whether RSDQ-Net can distinguish ordered differences in dance execution on ImperialDance. Table 3 reports the results under the five-fold choreography-disjoint protocol. Because Beginner, Intermediate, and Expert form an ordered sequence rather than independent categories, QWK and Ordinal MAE are emphasized together with conventional classification metrics.

Table 3. Ordinal dance movement quality assessment results on ImperialDance. Results are reported as mean $\pm$ standard deviation over 25 fold–seed evaluations from five choreography-disjoint folds and five random seeds.

Method	Accuracy (%) $\uparrow$	Macro-F1 (%) $\uparrow$	Ordinal MAE $\downarrow$	QWK $\uparrow$	Spearman ρ $\uparrow$
DTW ⁴⁵ + Joint Distance	61.8 $\pm$ 4.3	61.1 $\pm$ 4.5	0.468 $\pm$ 0.052	0.548 $\pm$ 0.049	0.563 $\pm$ 0.047
ST-GCN ⁴¹ + CE	69.7 $\pm$ 3.7	69.1 $\pm$ 3.8	0.351 $\pm$ 0.041	0.657 $\pm$ 0.042	0.673 $\pm$ 0.040
CTR-GCN ⁴³ + CE	72.5 $\pm$ 3.4	71.9 $\pm$ 3.5	0.316 $\pm$ 0.038	0.694 $\pm$ 0.039	0.711 $\pm$ 0.037
ST-GCN ⁴¹ + Ordinal	73.8 $\pm$ 3.2	73.3 $\pm$ 3.3	0.292 $\pm$ 0.036	0.722 $\pm$ 0.036	0.739 $\pm$ 0.034
2T-GCN ⁴⁶	76.7 $\pm$ 2.9	76.2 $\pm$ 3.0	0.261 $\pm$ 0.033	0.755 $\pm$ 0.032	0.771 $\pm$ 0.031
RSDQ-Net	81.2 $\pm$ 2.4	80.8 $\pm$ 2.5	0.207 $\pm$ 0.027	0.814 $\pm$ 0.027	0.828 $\pm$ 0.026

RSDQ-Net achieves the strongest ordinal quality-assessment performance on ImperialDance, with an Accuracy of 81.2%, a Macro-F1 of 80.8%, an Ordinal MAE of 0.207, a QWK of 0.814, and a Spearman correlation of 0.828. Compared with the strongest baseline, 2T-GCN, RSDQ-Net improves QWK from 0.755 to 0.814 and reduces Ordinal MAE from 0.261 to 0.207. The latter corresponds to a relative reduction of approximately 20.7% in ordinal error. The improvements are also reflected in Accuracy and Macro-F1, which increase by 4.5 and 4.6 percentage points, respectively. These results indicate that explicitly modeling target–reference correspondence and ordered quality structure provides additional information beyond conventional graph-based movement representations. Paired Wilcoxon signed-rank tests over the 25 matched fold–seed evaluations, followed by Holm correction across the two primary ImperialDance metrics, yielded an adjusted $p = 0.004$ for QWK and $p = 0.006$ for Ordinal MAE, indicating statistically significant differences from 2T-GCN.

5.3 Human-Ranked Dance Quality Assessment

The second main quality-assessment experiment evaluates whether the latent quality produced by RSDQ-Net is consistent with human pairwise judgments on CDRG-UNIMAS-TIKTOK. Table 4 reports the results under the five-fold subject-disjoint protocol, with pair generation performed only after subject splitting.

Table 4. Human-ranked dance quality assessment results on CDRG-UNIMAS-TIKTOK. Results are reported as mean $\pm$ standard deviation over 25 fold–seed evaluations from five subject-disjoint folds and five random seeds.

Method	Pairwise Accuracy (%) $\uparrow$	Kendall τ_b $\uparrow$	Spearman ρ $\uparrow$	NDCG@5 $\uparrow$
Codebook + Edit Distance	65.8 $\pm$ 4.1	0.386 $\pm$ 0.052	0.521 $\pm$ 0.055	0.801 $\pm$ 0.032
Codebook + Edit Distance w/o Compression	63.7 $\pm$ 4.4	0.354 $\pm$ 0.056	0.487 $\pm$ 0.059	0.784 $\pm$ 0.035
DTW-3D	68.9 $\pm$ 3.8	0.438 $\pm$ 0.048	0.579 $\pm$ 0.051	0.826 $\pm$ 0.030
ST-GCN + RankNet	73.6 $\pm$ 3.3	0.512 $\pm$ 0.043	0.663 $\pm$ 0.046	0.861 $\pm$ 0.027
CTR-GCN + RankNet	76.8 $\pm$ 3.0	0.566 $\pm$ 0.039	0.713 $\pm$ 0.041	0.887 $\pm$ 0.024
RSDQ-Net	81.4 $\pm$ 2.5	0.638 $\pm$ 0.034	0.782 $\pm$ 0.036	0.918 $\pm$ 0.021

RSDQ-Net achieves the strongest human-ranked quality-assessment performance on CDRG-UNIMAS-TIKTOK, obtaining a Pairwise Accuracy of 81.4% and a Kendall τ_b of 0.638. The strongest baseline, CTR-GCN + RankNet, achieves 76.8% and 0.566 on the same metrics, corresponding to an improvement of 4.6 percentage points in Pairwise Accuracy and 0.072 in Kendall τ_b . The higher Pairwise Accuracy indicates that RSDQ-Net more frequently reproduces the direction of individual human preferences, whereas the improvement in Kendall τ_b demonstrates stronger agreement between predicted and human-derived performance rankings across each dance challenge. The non-learning reference-comparison methods exhibit substantially lower performance. Codebook + Edit Distance obtains a Pairwise Accuracy of 65.8% and a Kendall τ_b of 0.386, while DTW-3D improves these values to 68.9% and 0.438. These results suggest that fixed symbolic matching and geometric trajectory alignment capture coarse motion similarity but cannot fully represent perceptual execution-quality differences. ST-GCN + RankNet and CTR-GCN + RankNet achieve stronger results because direct pairwise supervision allows their skeletal representations to adapt to relative quality ordering. Nevertheless, their remaining gap to RSDQ-Net shows that ranking supervision alone does not provide an explicit mechanism for determining where and how a target performance deviates from a technically appropriate reference. The improvement of RSDQ-Net can be attributed to the proposed reference-conditioned phase alignment and spatiotemporal deviation representation. The learnable alignment establishes soft correspondence between semantically related movement phases while retaining temporal displacement, allowing spatial body-configuration errors and phase-progression differences to contribute separately to quality estimation. RSDQ-Net also achieves a Spearman correlation of 0.782 and an NDCG@5 of 0.918, exceeding CTR-GCN + RankNet by 0.069 and 0.031, respectively. The lower standard deviations across all four metrics further indicate more stable subject-independent performance. Paired Wilcoxon signed-rank tests over the 25 matched fold–seed evaluations, followed by Holm correction across the two primary CDRG metrics, yielded

an adjusted $p = 0.005$ for Pairwise Accuracy and $p = 0.008$ for Kendall’s τ_b , indicating statistically significant differences from the strongest baseline.

5.4 Ablation Analysis

Ablation experiments are conducted to identify the contributions of the main components of RSDQ-Net. Table 5 reports the results on ImperialDance and CDRG-UNIMAS-TIKTOK using the primary metrics of the two quality-assessment tasks. All variants follow the same evaluation protocols and optimization settings as the complete model.

Table 5. Ablation analysis of the proposed RSDQ-Net. Results are reported as mean $\pm$ standard deviation over 25 fold–seed evaluations.

Variant	QWK $\uparrow$	Ordinal MAE $\downarrow$	Pairwise Accuracy (%) $\uparrow$	Kendall τ_b $\uparrow$
RSDQ-Net	0.814 $\pm$ 0.027	0.207 $\pm$ 0.027	81.4 $\pm$ 2.5	0.638 $\pm$ 0.034
w/o AIST++ Pretraining	0.788 $\pm$ 0.031	0.229 $\pm$ 0.030	79.6 $\pm$ 2.8	0.612 $\pm$ 0.038
w/o Coordination Graph	0.782 $\pm$ 0.033	0.234 $\pm$ 0.032	79.2 $\pm$ 2.9	0.603 $\pm$ 0.040
w/o Reference Alignment	0.724 $\pm$ 0.041	0.286 $\pm$ 0.039	74.9 $\pm$ 3.5	0.529 $\pm$ 0.049
DTW Alignment	0.758 $\pm$ 0.036	0.254 $\pm$ 0.035	77.2 $\pm$ 3.1	0.566 $\pm$ 0.044
w/o Spatial Deviation	0.765 $\pm$ 0.035	0.247 $\pm$ 0.034	77.8 $\pm$ 3.0	0.575 $\pm$ 0.043
w/o Temporal Deviation	0.748 $\pm$ 0.038	0.261 $\pm$ 0.036	76.5 $\pm$ 3.2	0.553 $\pm$ 0.046
w/o Order-Consistent Learning	0.776 $\pm$ 0.034	0.239 $\pm$ 0.033	78.4 $\pm$ 2.9	0.590 $\pm$ 0.041

The ablation results demonstrate that each component of RSDQ-Net contributes to movement-quality assessment across both supervision settings. Removing AIST++ pretraining decreases QWK from 0.814 to 0.788 and Pairwise Accuracy from 81.4% to 79.6%, showing that dance-specific representation initialization provides a transferable prior for subsequent quality learning. Removing the coordination graph produces a comparable degradation, reducing QWK to 0.782 and Kendall τ_b to 0.603. This result indicates that anatomical connectivity alone cannot fully represent the bilateral and cross-body coordination patterns that characterize dance execution. The most substantial degradation occurs when reference-conditioned alignment is removed. QWK decreases to 0.724, Ordinal MAE increases from 0.207 to 0.286, Pairwise Accuracy falls to 74.9%, and Kendall τ_b declines to 0.529. Replacing the proposed alignment with DTW partially recovers performance, achieving a QWK of 0.758 and Kendall τ_b of 0.566, but remains clearly below the complete model. This difference supports the proposed learnable alignment mechanism because DTW establishes temporal correspondence without jointly optimizing correspondence according to downstream quality supervision. Removing spatial deviation reduces QWK to 0.765 and Pairwise Accuracy to 77.8%, whereas removing temporal deviation results in larger decreases to 0.748 and 76.5%, respectively. The stronger effect of temporal deviation suggests that phase-progression errors provide particularly important evidence for distinguishing performances that contain similar poses but differ in execution timing. Replacing order-consistent learning with dataset-specific prediction heads also reduces performance, yielding a QWK of 0.776 and Kendall τ_b of 0.590. The degradation across both datasets indicates that the shared latent quality formulation provides a more coherent representation of ordered expertise levels and human pairwise preferences. Taken together, these results show that the gains of RSDQ-Net arise from the complementary interaction of dance-specific pretraining, coordination-aware skeletal encoding, reference-conditioned alignment, explicit spatiotemporal deviation modeling, and order-consistent quality learning.

5.5 Reference Alignment and Quality Learning Analysis

The reference-alignment mechanism and latent quality space are further examined to determine whether the improvements of RSDQ-Net arise from meaningful phase correspondence and ordered quality representation. Table 6 compares different alignment mechanisms and reference strategies using the primary metrics on ImperialDance and CDRG-UNIMAS-TIKTOK.

The alignment comparison shows a consistent improvement as increasingly quality-aware correspondence mechanisms are introduced. Without alignment, RSDQ-Net obtains a QWK of 0.724 and a Kendall τ_b of 0.529. Linear alignment provides modest gains, while DTW further improves QWK to 0.758 and Kendall τ_b to 0.566, indicating that compensating for sequence-length and execution-speed differences is beneficial. Nevertheless, both fixed alignment strategies remain clearly below the learnable alternatives because their correspondence is determined independently of the downstream quality objective. Learnable alignment without the phase prior increases QWK to 0.789 and Pairwise Accuracy to 79.3%, demonstrating the advantage of estimating target-reference correspondence directly in the learned motion space. Incorporating the temporal phase prior further raises QWK to 0.814 and Kendall τ_b to 0.638 while reducing Ordinal MAE from 0.229 to 0.207. The gains of 0.025 in QWK and 0.031 in Kendall τ_b indicate that semantic similarity alone can produce ambiguous correspondence between repeated or visually similar poses, whereas the proposed phase constraint suppresses implausible long-range matches while retaining meaningful local timing variation. Reference selection produces a similarly consistent pattern. Random eligible

Table 6. Analysis of alignment and reference strategies. Results are reported as mean $\pm$ standard deviation over 25 fold-seed evaluations.

Strategy	QWK $\uparrow$	Ordinal MAE $\downarrow$	Pairwise Accuracy (%) $\uparrow$	Kendall τ_b $\uparrow$
<i>Alignment strategy</i>				
No Alignment	0.724 $\pm$ 0.041	0.286 $\pm$ 0.039	74.9 $\pm$ 3.5	0.529 $\pm$ 0.049
Linear Alignment	0.741 $\pm$ 0.039	0.268 $\pm$ 0.037	76.0 $\pm$ 3.3	0.548 $\pm$ 0.047
DTW	0.758 $\pm$ 0.036	0.254 $\pm$ 0.035	77.2 $\pm$ 3.1	0.566 $\pm$ 0.044
Learnable Alignment w/o Phase Prior	0.789 $\pm$ 0.031	0.229 $\pm$ 0.030	79.3 $\pm$ 2.8	0.607 $\pm$ 0.039
Full RSDQ Alignment	0.814 $\pm$ 0.027	0.207 $\pm$ 0.027	81.4 $\pm$ 2.5	0.638 $\pm$ 0.034
<i>Reference strategy</i>				
Random Eligible Reference	0.768 $\pm$ 0.037	0.245 $\pm$ 0.035	77.9 $\pm$ 3.2	0.578 $\pm$ 0.044
Training-Set Representative Reference	0.797 $\pm$ 0.030	0.222 $\pm$ 0.029	80.0 $\pm$ 2.7	0.615 $\pm$ 0.038
Protocol Reference	0.814 $\pm$ 0.027	0.207 $\pm$ 0.027	81.4 $\pm$ 2.5	0.638 $\pm$ 0.034

references achieve a QWK of 0.768 and Kendall τ_b of 0.578, whereas selecting a representative reference from the training set improves these values to 0.797 and 0.615. The protocol reference provides the strongest performance, reaching 0.814 QWK, 0.207 Ordinal MAE, 81.4% Pairwise Accuracy, and 0.638 Kendall τ_b . Relative to the training-set representative strategy, the protocol reference improves QWK by 0.017 and Pairwise Accuracy by 1.4 percentage points. These results indicate that reference-conditioned deviation modeling benefits not only from learnable phase correspondence but also from a stable and technically appropriate reference. In ImperialDance, the designated Expert demonstration provides choreography-specific technical guidance, while the CDRG protocol restricts reference selection to training subjects, avoiding information leakage from evaluation subjects.

Figure 3 qualitatively illustrates the reference-conditioned alignment patterns for representative Expert, Intermediate, and Beginner performances. In the selected Expert example, the alignment distribution is concentrated near a coherent diagonal trajectory and the corresponding d_t^{temp} remains relatively close to zero. The Intermediate example exhibits a broader alignment band and larger local phase variation, whereas the selected Beginner example shows a more dispersed correspondence pattern and larger temporal fluctuations. These examples illustrate how the learned alignment can represent different target-reference phase relationships without forcing strict frame-wise synchronization. The visualization is intended to show the behavior of the learned correspondence mechanism in representative samples rather than to establish a deterministic relationship between expertise level and alignment structure.

Figure 4 examines whether the learned latent quality is consistent with the ordering information provided by the two quality datasets. Figure 4(a) shows the test-set distributions of q for Beginner, Intermediate, and Expert performances in ImperialDance. The distributions are presented together with their within-group variability, allowing the ordinal tendency of the learned representation to be examined without assuming that the three expertise levels are perfectly separable. Figure 4(b) relates the predicted latent quality to the human-derived performance ranking within each CDRG dance challenge. Because rankings are defined within challenges rather than across the 12 challenge categories, the human ranks are normalized before aggregation across challenges. The resulting visualization therefore evaluates whether higher human-ranked performances tend to receive higher latent-quality values. The strength of this association is quantified by the rank-correlation metrics reported in Table 4, rather than inferred solely from the visual trend.

5.6 Cross-Dataset Transfer and Generalization

As shown in Table 7 training the complete quality-learning framework from scratch produces a QWK of 0.788 ± 0.031 on ImperialDance and a Kendall τ_b of 0.612 ± 0.038 on CDRG-UNIMAS-TIKTOK. Initializing the skeletal encoder using AIST++ increases these values to 0.799 ± 0.029 and 0.621 ± 0.036 , respectively, while reducing ImperialDance Ordinal MAE from 0.229 ± 0.030 to 0.220 ± 0.029 . These gains indicate that the spatial and temporal motion representations acquired through dance-genre recognition transfer beyond semantic classification and provide a useful prior for execution-quality modeling. The reduced standard deviations further suggest that dance-specific pretraining improves the stability of the learned representation across different folds and random initializations. End-to-end adaptation to the target quality dataset provides further improvements. Fine-tuning on ImperialDance reaches a QWK of 0.807 ± 0.028 and an Ordinal MAE of 0.212 ± 0.028 , whereas fine-tuning on CDRG achieves a Pairwise Accuracy of $80.9 \pm 2.6\%$ and a Kendall τ_b of 0.629 ± 0.035 . Scratch denotes joint quality learning with a randomly initialized encoder. AIST++ Pretrained uses the AIST++-pretrained encoder as a frozen representation backbone, whereas the dataset-specific transfer settings fine-tune the pretrained model on ImperialDance or CDRG. Joint Quality Training denotes the complete end-to-end Stage II optimization using alternating mini-batches from both quality datasets. A more stringent evaluation transfers the model trained through AIST++ and ImperialDance directly to CDRG

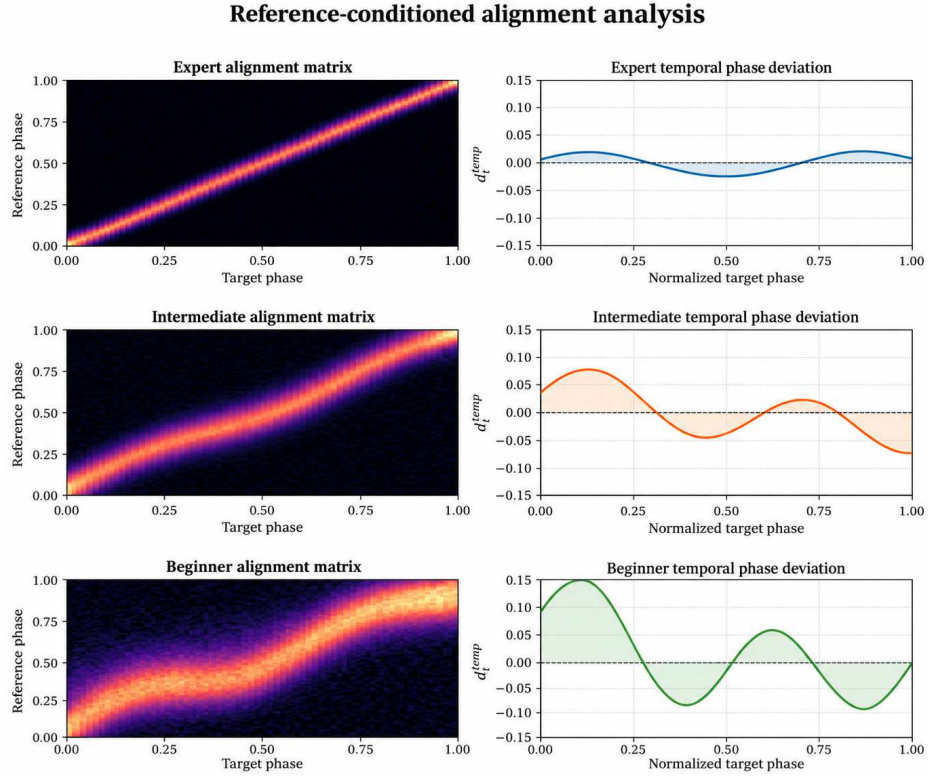


Figure 3. Representative model outputs of the reference-conditioned phase alignment for Expert, Intermediate, and Beginner target performances. Alignment matrices are shown together with the corresponding temporal phase-deviation trajectories d_t^{temp} . The examples illustrate learned target–reference correspondence patterns and are not intended as deterministic templates for each expertise level.

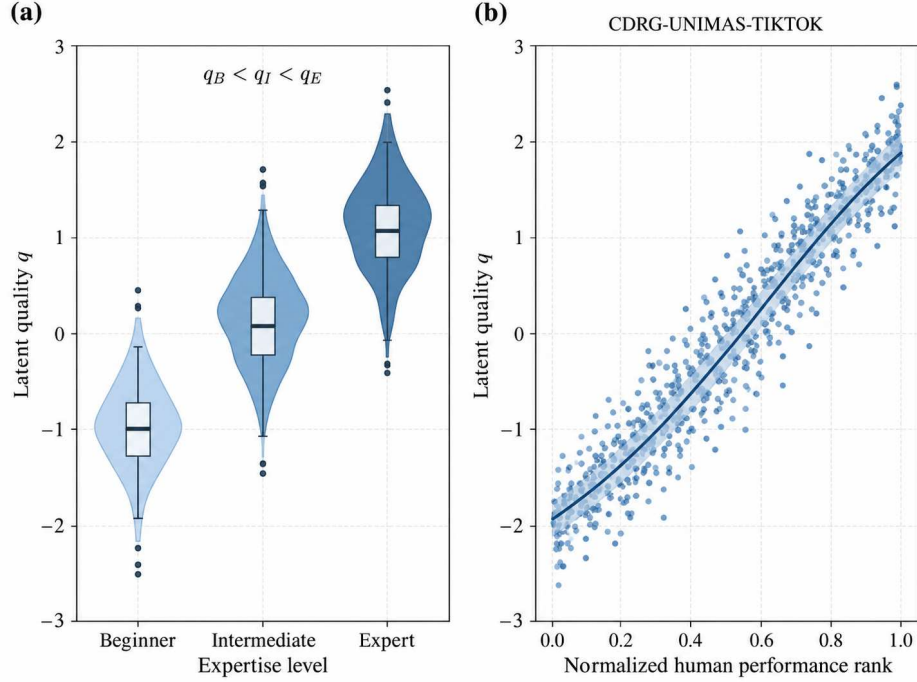


Figure 4. Learned latent quality on the two quality-assessment datasets. (a) Test-set distributions of latent quality q for Beginner, Intermediate, and Expert performances in ImperialDance. (b) Relationship between latent quality and normalized human-derived performance rank within CDRG dance challenges. Human rankings are defined within each challenge and normalized before cross-challenge visualization.

without using CDRG quality annotations. The zero-shot model obtains a Pairwise Accuracy of $72.7 \pm 3.6\%$ and a Kendall τ_b of 0.491 ± 0.047 . Although these results remain below supervised CDRG adaptation, the 22.7-percentage-point margin above random pairwise ordering indicates that the ordered quality representation learned from Beginner, Intermediate, and Expert supervision retains transferable information about relative execution quality. The larger variability under this setting is also reasonable because the model must generalize across differences in performers, motion acquisition, and annotation format without target-domain quality supervision. Joint quality training provides the strongest and most stable results, achieving a QWK of 0.814 ± 0.027 and an Ordinal MAE of 0.207 ± 0.027 on ImperialDance together with a Pairwise Accuracy of $81.4 \pm 2.5\%$ and a Kendall τ_b of 0.638 ± 0.034 on CDRG. Compared with dataset-specific fine-tuning, joint optimization improves QWK by 0.007 and CDRG Pairwise Accuracy by 0.5 percentage points. The moderate but consistent gains can be attributed to the proposed order-consistent quality learning strategy, through which ordinal expertise labels and pairwise human preferences jointly constrain the shared latent quality variable rather than forming independent dataset-specific decision spaces. Together with the AIST++-initialized relation-aware skeletal encoder, this formulation enables RSDQ-Net to retain dance-specific motion structure while learning quality cues that generalize across heterogeneous supervision conditions.

Table 7. Cross-dataset transfer and generalization results. Results are reported as mean $\pm$ standard deviation over 25 fold-seed evaluations where applicable.

Training Strategy	Imperial QWK $\uparrow$	Imperial Ordinal MAE $\downarrow$	CDRG Pairwise Acc. (%) $\uparrow$	CDRG Kendall τ_b $\uparrow$
Scratch	0.788 ± 0.031	0.229 ± 0.030	79.6 ± 2.8	0.612 ± 0.038
AIST++ Pretrained	0.799 ± 0.029	0.220 ± 0.029	80.2 ± 2.7	0.621 ± 0.036
AIST++ $\rightarrow$ ImperialDance	0.807 ± 0.028	0.212 ± 0.028	—	—
AIST++ $\rightarrow$ CDRG	—	—	80.9 ± 2.6	0.629 ± 0.035
AIST++ $\rightarrow$ ImperialDance $\rightarrow$ CDRG Zero-Shot	—	—	72.7 ± 3.6	0.491 ± 0.047
Joint Quality Training	0.814 ± 0.027	0.207 ± 0.027	81.4 ± 2.5	0.638 ± 0.034

5.7 Robustness Analysis

Figure 5 evaluates the stability of RSDQ-Net under coordinate corruption, missing skeletal observations, and temporal execution perturbations. Under the unperturbed setting, RSDQ-Net retains the performance reported in the main experiments, reaching a

QWK of 0.814 on ImperialDance and a Pairwise Accuracy of 81.4% on CDRG-UNIMAS-TIKTOK. As the coordinate-noise level increases, both RSDQ-Net and the strongest baselines exhibit progressive degradation, but the proposed model maintains a consistently larger performance margin. At $\sigma = 0.05$, RSDQ-Net retains a QWK of approximately 0.74 on ImperialDance and a Pairwise Accuracy of 76.2% on CDRG, whereas the corresponding baseline performance decreases more sharply. This behavior suggests that the relation-aware skeletal encoder is less dependent on individual joint coordinates because anatomical and non-local coordination relations provide complementary structural context when local positions become unreliable. Coordinate noise is added to the normalized target joint positions before derived motion features are recomputed. Missing-joint perturbation masks the selected target joints throughout the complete sequence. Temporal speed perturbation is applied only to the target motion, while the corresponding reference remains unchanged. A similar pattern is observed under missing-joint corruption. Performance remains relatively stable at low masking rates and decreases more substantially as r_m approaches 20–30%. Nevertheless, RSDQ-Net preserves a clear advantage throughout the perturbation range. On CDRG, Pairwise Accuracy decreases from 81.4% without masking to 75.3% at a 30% missing-joint rate, corresponding to a reduction of 6.1 percentage points. In comparison, CTR-GCN + RankNet decreases from 76.8% to 66.8%, a substantially larger reduction of 10.0 percentage points. The smaller degradation of RSDQ-Net is consistent with the proposed combination of anatomical and coordination graphs, through which information from spatially related and bilaterally coordinated body regions can partially compensate for unavailable joints. The reference-conditioned comparison further reduces reliance on isolated observations by evaluating deviations within the context of the complete target-reference motion correspondence. Temporal speed perturbation produces a different robustness pattern. Performance is highest around the original speed factor $r_s = 1.0$ and decreases toward both $0.8\times$ and $1.2\times$, but the degradation of RSDQ-Net is comparatively moderate. On CDRG, Pairwise Accuracy changes from 81.4% at the nominal speed to 78.2% and 78.6% at $0.8\times$ and $1.2\times$, respectively, whereas CTR-GCN + RankNet falls from 76.8% to 71.9% and 71.7%. This robustness is consistent with the behavior expected from the reference-conditioned phase-alignment mechanism, which establishes soft correspondence in the learned motion space rather than assuming fixed frame-to-frame synchronization. Moderate differences in execution speed can therefore be absorbed through local phase correspondence, while the remaining phase displacement is retained through d_t^{temp} instead of being eliminated completely. The three perturbation experiments collectively indicate that the robustness of RSDQ-Net does not originate from a single form of invariance. Structural robustness is supported by relation-aware anatomical and coordination modeling, observation robustness benefits from contextual target-reference comparison, and temporal robustness is primarily associated with the learnable phase-alignment mechanism. Importantly, performance still decreases under severe perturbations, showing that the proposed framework does not artificially suppress quality-relevant variation. Instead, it maintains greater stability than the strongest baselines while remaining sensitive to substantial spatial and temporal deviations, which is desirable for movement-quality assessment.

5.8 Qualitative and Computational Analysis

Figure 6 visualizes representative deviation patterns produced by RSDQ-Net for Expert, Intermediate, and Beginner performances. In the selected Expert example, the learned spatial deviation responses are relatively limited and the temporal phase-deviation trajectory remains close to zero across most of the sequence. The Intermediate example shows more pronounced responses in several body regions together with moderate temporal variation. The selected Beginner example exhibits broader spatial deviation responses and larger fluctuations in temporal phase deviation. These visualizations illustrate how RSDQ-Net represents target–reference differences through complementary spatial and temporal signals. They should not be interpreted as localized expert-annotated movement errors, because the evaluated datasets do not provide frame-level or joint-level error annotations. The spatial maps therefore represent learned deviation responses generated by the model, while d_t^{temp} reflects the estimated phase displacement derived from the learned alignment. The qualitative patterns are used only to illustrate model behavior; the effectiveness of spatial and temporal deviation modeling is evaluated quantitatively through the ablation experiments.

As shown in Table 8, RSDQ-Net introduces moderate computational overhead relative to conventional graph-based baselines while remaining within a practical complexity range. The proposed model contains 3.84 million parameters, which is 0.72 million more than ST-GCN but 0.52 million fewer than 2T-GCN. This result indicates that the performance gains of RSDQ-Net are not obtained simply by substantially increasing model capacity. The relatively limited parameter growth is mainly due to parameter sharing between the target and reference skeletal encoders, while the additional computation arises primarily from evaluating both sequences and constructing their reference-conditioned correspondence. RSDQ-Net requires 21.2 GFLOPs per sample, compared with 15.4 GFLOPs for ST-GCN and 16.8 GFLOPs for CTR-GCN. The corresponding inference time is 8.1 ms per sample, representing an increase of 2.5 ms over ST-GCN and 1.8 ms over CTR-GCN. This additional cost is associated with the learnable phase-alignment matrix, node-level spatial deviation computation, and temporal phase-deviation extraction. Nevertheless, RSDQ-Net remains slightly more efficient than 2T-GCN, which requires 23.7 GFLOPs and 8.9 ms per sample. Peak GPU memory usage follows a similar pattern. RSDQ-Net requires 618 MB, compared with 468 MB for

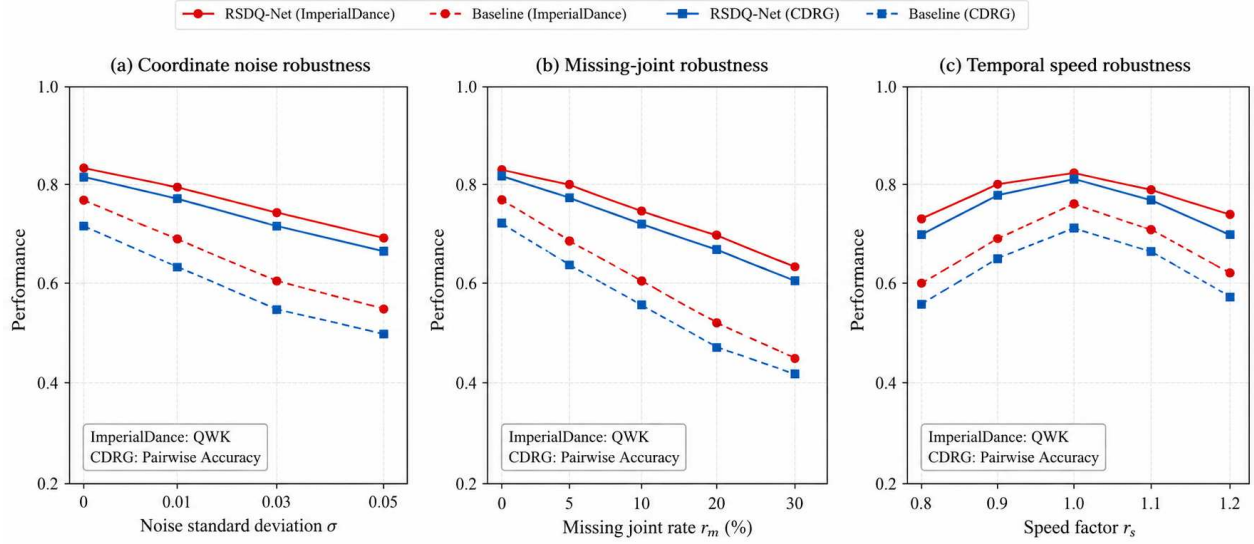


Figure 5: Robustness comparison between RSDQ-Net and the strongest baseline under three perturbation conditions. (a) Coordinate noise with $\sigma \in \{0, 0.01, 0.03, 0.05\}$. (b) Randomly missing joints at rates of 0–30%. (c) Temporal speed perturbation with factors from 0.8 to 1.2.

Figure 5. Robustness comparison between RSDQ-Net and the strongest task-specific baselines under controlled perturbations. ImperialDance is evaluated using QWK, whereas CDRG-UNIMAS-TIKTOK is evaluated using Pairwise Accuracy. (a) Coordinate noise with $\sigma \in \{0, 0.01, 0.03, 0.05\}$. (b) Random missing-joint corruption with rates from 0 to 30%. (c) Temporal speed perturbation with factors from 0.8 to 1.2.

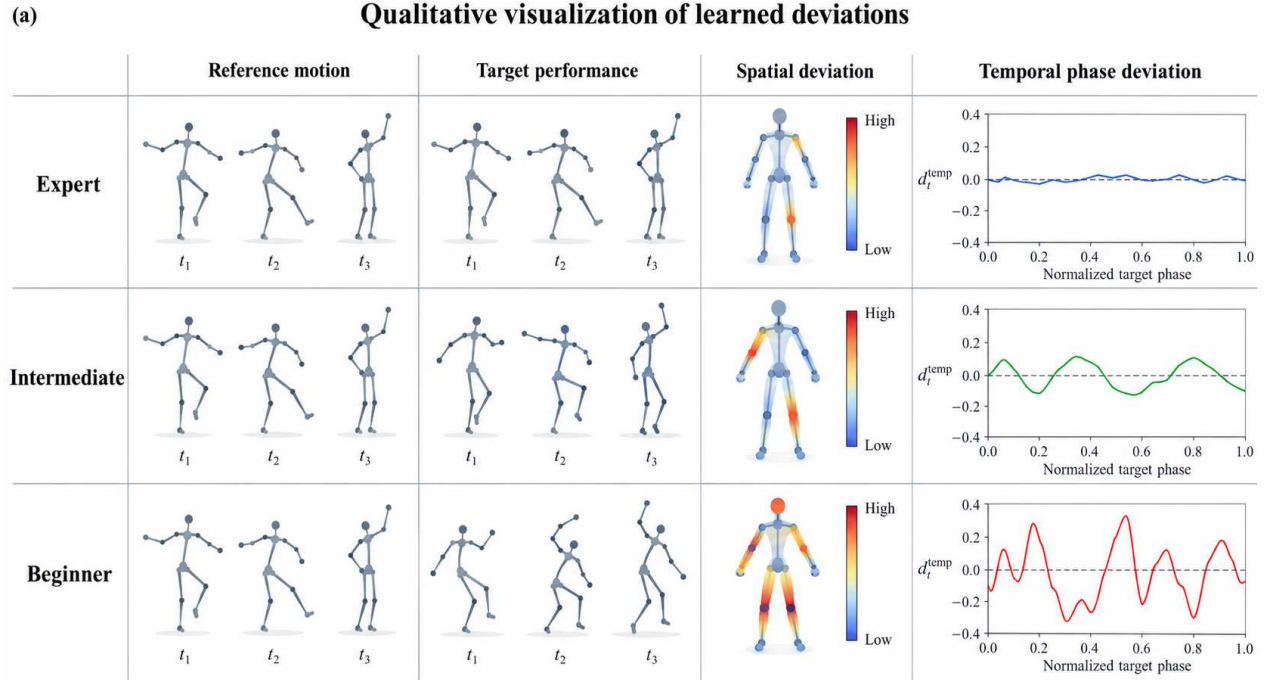


Figure 6. Representative learned deviation responses produced by RSDQ-Net for Expert, Intermediate, and Beginner performances. Model-derived spatial deviation and temporal phase-deviation patterns are visualized relative to the corresponding reference motion. These responses are learned model representations and should not be interpreted as ground-truth joint-level error localization.

ST-GCN, 506 MB for CTR-GCN, and 642 MB for 2T-GCN. The increased memory requirement is expected because both target and reference representations must be retained during reference-conditioned comparison. Importantly, this additional computational burden accompanies substantial improvements in the main quality-assessment tasks, including a QWK increase from 0.755 to 0.814 over the strongest ImperialDance baseline and a Pairwise Accuracy increase from 76.8% to 81.4% over the strongest CDRG baseline. The results therefore indicate that the proposed reference-aware modeling achieves a favorable trade-off between movement-quality assessment performance and computational complexity, making it suitable for offline evaluation and interactive assisted-training scenarios.

Table 8. Computational cost comparison of RSDQ-Net and learning-based baselines.

Method	Parameters (M) ↓	GFLOPs ↓	Inference Time (ms/sample) ↓	Peak Memory (MB) ↓
ST-GCN + CE / RankNet	3.12	15.4	5.6	468
CTR-GCN + CE / RankNet	1.49	16.8	6.3	506
2T-GCN	4.36	23.7	8.9	642
RSDQ-Net	3.84	21.2	8.1	618

6 Discussion

6.1 Main Findings and Methodological Implications

The results support the central premise of this study that dance movement quality is better represented as reference-conditioned execution deviation than as a conventional action-recognition problem. Across ordinal and pairwise assessment settings, RSDQ-Net benefits from combining dance-specific skeletal representation learning, reference-aware comparison, and ordered quality supervision. The methodological implications can be understood from three aspects corresponding to the main design choices of the framework. Reference-conditioned alignment is effective because target and reference performances rarely progress through a choreography at identical temporal rates. Direct frame-wise comparison confounds execution differences with harmless duration variation, while fixed temporal warping may over-correct timing discrepancies that are themselves relevant to movement quality. The proposed alignment instead learns soft correspondence in the motion representation space while retaining a temporal phase prior. This design allows semantically similar movement phases to be associated despite moderate acceleration or delay, but discourages implausible matches between repeated poses occurring at distant phases. More importantly, the learned alignment matrix is not used only to warp the reference sequence. Its phase displacement is retained as quality evidence, allowing correspondence estimation and quality assessment to remain coupled rather than treating alignment as an independent preprocessing operation. Separating spatial and temporal deviations is necessary because visually similar execution errors can arise from different underlying mechanisms. Spatial deviation describes discrepancies in body configuration after corresponding movement phases have been identified, including differences in posture, limb arrangement, and coordinated skeletal structure. Temporal deviation instead describes how the target progresses through those phases relative to the reference. Combining both into a single distance would make a performer with accurate poses but inappropriate timing difficult to distinguish from one with correct timing but inaccurate body configuration. RSDQ-Net therefore preserves spatial deviation, phase displacement, and local phase variation as complementary signals before quality aggregation. The ablation and alignment analyses indicate that neither component alone fully represents execution quality, supporting the view that dance assessment requires both configurational accuracy and temporal control. The unified treatment of ordinal and pairwise supervision follows from their common relational structure. ImperialDance does not provide a continuous quality score but defines the ordered relation $\text{Beginner} < \text{Intermediate} < \text{Expert}$, whereas CDRG provides instance-level preferences indicating whether one performance is judged better than another. Although these annotations differ in granularity, both constrain the relative ordering of performances. Representing quality through a scalar latent variable q therefore avoids assigning arbitrary numerical distances between expertise levels or assuming that the two datasets share an absolute rating scale. Dataset-specific monotonic calibration adapts the latent variable to different supervision distributions without changing its ordering, while ordinal thresholds and pairwise ranking losses impose complementary constraints on the same quality space. This formulation has a broader methodological implication for action quality assessment with heterogeneous annotations. Quality labels collected from different environments often differ in scale, granularity, and annotation procedure, making direct score-level fusion difficult to justify. Learning a shared order-consistent representation provides an alternative when the supervision sources agree primarily on relative quality rather than absolute magnitude. In the present setting, the resulting framework separates dance semantics, reference correspondence, execution deviation, and quality ordering into distinct but jointly optimized components, providing a more interpretable formulation of dance movement quality than direct classification or similarity-based assessment.

6.2 Implications for Automatic Dance Training

The proposed framework has practical value in training scenarios where an appropriate reference demonstration is available. Under this setting, assessment can be formulated as

$$\text{Target Motion} + \text{Reference Motion} \rightarrow \text{Relative Quality}, \quad (30)$$

allowing the system to estimate how a learner’s execution compares with a technically appropriate demonstration without requiring an absolute scoring standard. Such a formulation is particularly suitable for self-training and automatic feedback. Learners can repeatedly perform the same choreography and obtain relative quality estimates together with learned spatial and temporal deviation patterns. These outputs may help identify whether performance differences are associated primarily with body configuration, temporal progression, or broader coordination patterns. Because the framework operates on skeletal motion, it can also support movement screening by identifying performances that deviate substantially from a reference and may therefore require closer review. The reference-conditioned design is also relevant to remote dance instruction, where continuous in-person observation may not be available. A standardized or instructor-provided demonstration can serve as the reference, while a learner’s recorded movement is analyzed using the same comparison framework. This may provide instructors with an additional source of structured information for reviewing repeated practice sessions or prioritizing performances that require further attention. The intended role of RSDQ-Net is therefore to support, rather than replace, professional dance instruction. The learned quality variable represents relative execution quality under a particular reference and does not capture all dimensions considered by expert teachers, such as artistic interpretation, expressiveness, stylistic intention, or pedagogical context. Automatic assessment is more appropriately viewed as a complementary tool for scalable practice monitoring and movement-focused feedback within instructor-guided or self-directed training environments.

6.3 Limitations and Future Work

Several limitations should be considered when interpreting the present findings. The three datasets differ substantially in dance style, acquisition environment, skeletal source, and annotation form. Although the canonical representation and shared quality formulation reduce part of this heterogeneity, the current evaluation does not fully eliminate dataset-specific characteristics. Broader validation on additional dance genres, recording conditions, and performer populations will therefore be necessary to further establish generalizability. CDRG-UNIMAS-TIKTOK introduces an additional source of uncertainty because its three-dimensional skeletal motion is reconstructed from monocular RGB video rather than obtained directly from motion capture. Reconstruction errors, occlusion, and viewpoint ambiguity may affect the estimated skeletal trajectories and consequently the downstream deviation representation. More robust pose recovery, including multi-view reconstruction or uncertainty-aware motion estimation, could improve assessment reliability in unconstrained environments. The current datasets also do not provide frame-level or joint-level expert annotations describing where and why execution errors occur. The qualitative visualizations in this study therefore represent learned spatial and temporal deviations rather than validated expert error localization. Future datasets containing fine-grained expert feedback could enable direct evaluation of whether model-derived deviation patterns correspond to pedagogically meaningful technical errors. The latent quality variable learned by RSDQ-Net represents relative quality ordering rather than an absolute professional score. This design is appropriate for heterogeneous ordinal and pairwise supervision, but it limits direct interpretation in settings that require standardized numerical grading. Future work may investigate calibrated scoring models when sufficiently reliable expert ratings become available. The present framework also focuses primarily on skeletal execution quality. Dance performance is influenced by additional factors that are not explicitly modeled here, including synchronization with music, rhythmic structure, stylistic expression, and higher-level artistic interpretation. Integrating music–dance synchronization, multi-view motion reconstruction, and fine-grained expert feedback may provide a more comprehensive assessment framework while preserving the reference-aware and order-consistent principles introduced in this study.

7 Conclusions

This study addressed automatic dance movement quality assessment from skeletal motion by distinguishing movement recognition from execution-quality evaluation. A Reference-Aware Spatiotemporal Deviation Quality Network (RSDQ-Net) was proposed to compare a target performance with an appropriate reference through a shared skeletal encoder, learnable phase alignment, explicit spatial and temporal deviation modeling, and an order-consistent latent quality representation. The framework was designed to preserve quality-relevant differences rather than suppress them as within-class variation. Experiments across AIST++, ImperialDance, and CDRG-UNIMAS-TIKTOK evaluated the framework at complementary levels. AIST++ verified the dance-specific representation capability of the encoder, ImperialDance examined ordinal quality assessment under choreography-disjoint evaluation, and CDRG evaluated agreement with human pairwise rankings under subject-disjoint conditions. The combined results indicate that reference-conditioned alignment, separated spatial and temporal deviations, and shared quality ordering provide a consistent basis for modeling heterogeneous forms of dance-quality supervision.

These findings highlight the potential of reference-aware skeletal representation learning for scalable dance training and movement assessment when suitable demonstrations are available. Future work will extend the framework toward richer expert supervision, more robust multi-view motion reconstruction, and multimodal modeling of factors such as music–dance synchronization and stylistic expression.

Conflict of Interest Statement

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Author Contributions

Z.J. conceived and designed the study, developed the methodology, supervised the research, and drafted the manuscript. P.W. conducted data processing, model implementation, experiments, validation, and visualization. Both authors contributed to data analysis, interpretation of the results, manuscript revision, and approved the final version of the manuscript.

Funding

No Funding.

Data Availability Statement

The datasets analysed in this study are publicly available from their respective original sources. AIST++, ImperialDance, and CDRG-UNIMAS-TIKTOK were used in accordance with their corresponding dataset access and licensing conditions. The processed data splits and implementation code used in this study will be made available at <https://github.com/ZhenJi0407/RSDQ-Net-Dance-Quality-Assessment.git>.

References

1. Chen, L. Visual language transformer framework for multimodal dance performance evaluation and progression monitoring. *Sci. Reports* **15**, 30649 (2025).
2. Xu, H. *et al.* Dancefix: An exploration in group dance neatness assessment through fixing abnormal challenges of human pose. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, 8869–8877 (2025).
3. Xu, J. *et al.* Finediving: A fine-grained dataset for procedure-aware action quality assessment. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2939–2948 (IEEE, 2022).
4. Bai, Y. *et al.* Action quality assessment with temporal parsing transformer. In *European conference on computer vision*, 422–438 (Springer, 2022).
5. Xu, J., Yin, S., Zhao, G., Wang, Z. & Peng, Y. Fineparser: A fine-grained spatio-temporal action parser for human-centric action quality assessment. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 14628–14637 (IEEE, 2024).
6. Xu, J., Rao, Y., Zhou, J. & Lu, J. Procedure-aware action quality assessment: Datasets and performance evaluation. *Int. journal computer vision* **132**, 6069–6090 (2024).
7. Yu, X., Rao, Y., Zhao, W., Lu, J. & Zhou, J. Group-aware contrastive regression for action quality assessment. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 7899–7908 (IEEE, 2021).
8. Zhou, C. & Huang, Y. Uncertainty-driven action quality assessment. *arXiv preprint arXiv:2207.14513* (2022).
9. Parmar, P., Gharat, A. & Rhodin, H. Domain knowledge-informed self-supervised representations for workout form assessment. In *European conference on computer vision*, 105–123 (Springer, 2022).
10. Ji, Y. *et al.* Localization-assisted uncertainty score disentanglement network for action quality assessment. In *Proceedings of the 31st ACM International Conference on Multimedia*, 8590–8597 (2023).
11. Ke, X., Xu, H., Lin, X. & Guo, W. Two-path target-aware contrastive regression for action quality assessment. *Inf. Sci.* **664**, 120347 (2024).
12. Gedamu, K., Ji, Y., Yang, Y., Shao, J. & Shen, H. T. Self-supervised sub-action parsing network for semi-supervised action quality assessment. *IEEE transactions on image processing* **33**, 6057–6070 (2024).

13. Zeng, L.-A. & Zheng, W.-S. Multimodal action quality assessment. *IEEE Transactions on Image Process.* **33**, 1600–1613 (2024).
14. Zhou, K. *et al.* Cofinal: Enhancing action quality assessment with coarse-to-fine instruction alignment. *arXiv preprint arXiv:2404.13999* (2024).
15. Zhang, S. *et al.* Narrative action evaluation with prompt-guided multimodal interaction. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 18430–18439 (IEEE, 2024).
16. Ding, Y. *et al.* 2m-af: A strong multi-modality framework for human action quality assessment with self-supervised representation learning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 1564–1572 (2024).
17. Han, R., Zhou, K., Atapour-Abarghouei, A., Liang, X. & Shum, H. P. Finecausal: A causal-based framework for interpretable fine-grained action quality assessment. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 6008–6017 (IEEE, 2025).
18. Zhou, K., Shum, H. P., Li, F. W., Zhang, X. & Liang, X. Phi: Bridging domain shift in long-term action quality assessment via progressive hierarchical instruction. *IEEE transactions on image processing* **34**, 3718–3732 (2025).
19. Zhou, Y. *et al.* Blockgcgn: Redefine topology awareness for skeleton-based action recognition. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2049–2058 (IEEE, 2024).
20. Wu, C. *et al.* Scd-net: Spatiotemporal clues disentanglement network for self-supervised skeleton-based action recognition. In *Proceedings of the AAAI conference on artificial intelligence*, vol. 38, 5949–5957 (2024).
21. Qi, M., Ye, H., Peng, J. & Ma, H. Action quality assessment via hierarchical pose-guided multi-stage contrastive regression. *IEEE Transactions on Image Process.* (2025).
22. Xu, H. *et al.* Language-guided audio-visual learning for long-term sports assessment. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 23967–23977 (IEEE, 2025).
23. Xu, H. *et al.* Quality-guided vision-language learning for long-term action quality assessment. *IEEE Transactions on Multimed.* (2025).
24. Dong, X., Liu, X., Li, W., Adeyemi-Ejeye, A. & Gilbert, A. Uil-aqa: Uncertainty-aware clip-level interpretable action quality assessment: X. dong *et al.* *Int. J. Comput. Vis.* **134**, 24 (2026).
25. Liu, J. *et al.* Vision-based human action quality assessment: A systematic review. *Expert. Syst. with applications* **263**, 125642 (2025).
26. Zhu, W. *et al.* Motionbert: A unified perspective on learning human motion representations. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 15039–15053 (IEEE, 2023).
27. Mehraban, S., Adeli, V. & Taati, B. Motionagformer: Enhancing 3d human pose estimation with a transformer-gcnformer network. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 6905–6915 (IEEE, 2024).
28. Shi, L., Zhang, Y., Cheng, J. & Lu, H. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 12018–12027 (IEEE, 2019).
29. Chi, H.-g. *et al.* Infogcn: Representation learning for human skeleton-based action recognition. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 20154–20164 (IEEE, 2022).
30. Lee, J., Lee, M., Lee, D. & Lee, S. Hierarchically decomposed graph convolutional networks for skeleton-based action recognition. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 10410–10419 (IEEE, 2023).
31. Lee, J. *et al.* Leveraging spatio-temporal dependency for skeleton-based action recognition. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 10221–10230 (IEEE, 2023).
32. Zhou, H., Liu, Q. & Wang, Y. Learning discriminative representations for skeleton based action recognition. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10608–10617 (IEEE, 2023).
33. Wang, X., Xu, X. & Mu, Y. Neural koopman pooling: Control-inspired temporal dynamics encoding for skeleton-based action recognition. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10597–10607 (IEEE, 2023).
34. Lin, L., Zhang, J. & Liu, J. Actionlet-dependent contrastive learning for unsupervised skeleton-based action recognition. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2363–2372 (IEEE, 2023).
35. Xiang, W., Li, C., Zhou, Y., Wang, B. & Zhang, L. Generative action description prompts for skeleton-based action recognition. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 10242–10251 (IEEE, 2023).

36. Liu, X., Zhou, S., Wang, L. & Hua, G. Parallel attention interaction network for few-shot skeleton-based action recognition. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 1379–1388 (IEEE, 2023).
37. Shin, S., Kim, J., Halilaj, E. & Black, M. J. Wham: Reconstructing world-grounded humans with accurate 3d motion. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2070–2080 (IEEE, 2024).
38. Li, R., Yang, S., Ross, D. A. & Kanazawa, A. Ai choreographer: Music conditioned 3d dance generation with aist++. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 13381–13392 (Ieee, 2021).
39. Zhong, Y. & Demiris, Y. Dancemvp: Self-supervised learning for multi-task primitive-based dance performance assessment via transformer text prompting. In *Proceedings of the AAAI conference on artificial intelligence*, vol. 38, 10270–10278 (2024).
40. Hipiny, I., Ujir, H., Alias, A. A., Shanat, M. & Ishak, M. K. Who danced better? ranked tiktok dance video dataset and pairwise action quality assessment method. *Int. J. Adv. Intell. Informatics* **9**, 96–107 (2023).
41. Yan, S., Xiong, Y. & Lin, D. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the AAAI conference on artificial intelligence*, vol. 32 (2018).
42. Song, Y.-F., Zhang, Z., Shan, C. & Wang, L. Constructing stronger and faster baselines for skeleton-based action recognition. *IEEE transactions on pattern analysis machine intelligence* **45**, 1474–1488 (2022).
43. Chen, Y. *et al.* Channel-wise topology refinement graph convolution for skeleton-based action recognition. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 13339–13348 (IEEE, 2021).
44. Duan, H. *et al.* Skeletr: Towards skeleton-based action recognition in the wild. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 13588–13598 (IEEE, 2023).
45. Jiang, Y., Deng, T. & Nasr, K. M. Baseline tracking in bipolar charge signal process by dynamic time warping algorithm. *IEEE Sensors J.* **23**, 4540–4547 (2022).
46. Bruce, X., Liu, Y., Chan, K. C., Yang, Q. & Wang, X. Skeleton-based human action evaluation using graph convolutional network for monitoring alzheimer’s progression. *Pattern Recognit.* **119**, 108095 (2021).