Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.081251

ARTICLE

A Spatial-Temporal Normalized Contrastive Embedding for Robust Motion Similarity Retrieval

Seung-su Lee[1], **Young-Been Noh**[1], **HwaYoung Jeong**[2] **and Kwang-il Hwang**[1,*]

[1]Department of Embedded Systems Engineering, Incheon National University, Incheon, Republic of Korea
[2]Humanitas College, Kyung Hee University, Seoul, Republic of Korea

*Corresponding Author: Kwang-il Hwang. Email: hkwangil@inu.ac.kr

Received: 26 February 2026; Accepted: 22 May 2026; Published: 23 July 2026

ABSTRACT: Robust motion similarity retrieval from monocular 2D pose sequences is challenged by body-scale variation, viewpoint inconsistency, translation drift, and temporal misalignment. Existing contrastive skeleton learning methods primarily address action recognition and rarely integrate explicit geometric canonicalization for retrieval-oriented metric learning. This paper proposes a spatial-temporal normalized contrastive embedding framework that unifies structured nuisance suppression with scalable similarity representation learning. A four-stage normalization pipeline—torso-scale normalization, pelvis-centered alignment, posture-axis alignment, and phase-synchronized temporal resampling—removes geometric and temporal distortions prior to embedding. The normalized sequences are encoded using an acausal dilated temporal convolutional network trained with a hybrid contrastive objective combining NT-Xent and semi-hard triplet loss, enabling both global separation and fine-grained stylistic discrimination. A prototype-based representation further supports interpretable amateur-to-professional style mapping. Experiments on a golf swing benchmark achieve a Top-1 accuracy of 91.3%, outperforming BiLSTM and Dynamic Time Warping baselines. The framework establishes an invariant and interpretable paradigm for motion similarity retrieval applicable to broader human movement analysis tasks.

KEYWORDS: Motion similarity retrieval; spatial-temporal normalization; contrastive representation learning; temporal convolutional networks (TCN); prototype-based embedding

1 Introduction

Human motion similarity retrieval plays an increasingly important role in sports analytics, rehabilitation monitoring, and skill assessment. In golf biomechanics, trunk rotation amplitude, pelvis-driven motion, and weight transfer patterns are well-established indicators of swing quality [1–4]. Traditional motion evaluation systems often rely on wearable inertial measurement units (IMUs) to measure kinematic consistency and phase transitions [1]. Although sensor-based systems provide precise rotational and temporal measurements, they require physical instrumentation, suffer from placement sensitivity, and lack scalability for large-scale deployment.

The emergence of markerless pose estimation frameworks such as OpenPose [5] and BlazePose [6] has enabled scalable motion modeling using monocular RGB cameras. Recent golf swing analysis systems incorporate pose refinement networks and embedding-based feedback mechanisms to facilitate self-training and correction [7]. However, most of these approaches focus on posture-level classification or phase-specific evaluation rather than full-sequence similarity retrieval. Fine-grained motion similarity retrieval

requires consistent metric representation across entire motion cycles, which is fundamentally different from categorical action recognition.

Dynamic Time Warping (DTW) has been widely adopted to address temporal misalignment in sequence comparison [8,9]. While DTW compensates for phase shifts by aligning temporal sequences, it performs pairwise frame-level matching and exhibits quadratic computational complexity. More importantly, DTW does not learn invariant representations and remains sensitive to geometric distortions such as scale, translation, and in-plane rotation. These limitations restrict its scalability and robustness in retrieval-based applications.

Recent advances in contrastive and self-supervised representation learning have significantly improved skeleton-based motion modeling. Efficient spatio-temporal contrastive learning frameworks [10] and contrast-reconstruction approaches [11] demonstrate strong performance in action recognition. Deep metric learning methods further extend similarity modeling to embedding spaces [12]. Temporal convolutional networks (TCNs) [13,14] and contrastive learning paradigms such as SimCLR [15] and FaceNet [16] provide scalable representation learning mechanisms. However, most existing skeleton-based contrastive frameworks are optimized for action classification rather than retrieval-based metric consistency. Furthermore, structured geometric canonicalization is rarely integrated as an explicit pre-embedding stage, leaving learned representations vulnerable to nuisance factors.

Our previous works addressed similarity-based swing evaluation [17] and reinforcement learning-based motion correction incorporating temporal rhythm and kinematic stability [18]. These studies demonstrated the effectiveness of embedding-based analysis but did not provide a unified framework that explicitly integrates spatial-temporal canonicalization with contrastive temporal modeling for scalable similarity retrieval.

To address these limitations, this paper proposes a spatial-temporal normalized contrastive embedding framework for robust motion similarity retrieval from monocular 2D pose sequences. The proposed method systematically removes geometric and temporal nuisance factors prior to embedding and learns a discriminative representation optimized for retrieval-based similarity measurement.

The main contributions of this work are summarized as follows:

1. **Structured Spatial-Temporal Normalization:**

We introduce a four-stage normalization pipeline consisting of torso-scale normalization, pelvis-centered alignment, posture-axis alignment, and phase-synchronized temporal resampling, explicitly modeling invariance to scale, translation, rotation, and tempo variations.

2. **Acausal Dilated Temporal Embedding with Hybrid Contrastive Learning:**

We design an acausal temporal convolutional network (A-TCN) optimized with a hybrid objective combining NT-Xent and semi-hard triplet loss, enabling both global embedding separation and fine-grained inter-style discrimination.

3. **Prototype-Based Retrieval for Interpretability:**

We propose a prototype representation mechanism that supports efficient similarity retrieval and interpretable amateur-to-professional style mapping in the embedding space.

4. **Comprehensive Experimental Validation:**

Extensive experiments on a golf swing benchmark demonstrate significant performance improvements over BiLSTM and DTW baselines, achieving 91.3% Top-1 accuracy while producing compact and semantically structured embedding manifolds.

Although the present study focuses on golf swing analysis, the proposed framework is designed as a modular motion representation architecture in which the temporal normalization strategy can be adapted according to the target motion domain. Previous rehabilitation-oriented vision studies have also highlighted the importance of robust and invariant motion representations for analyzing diverse human movements [19]. Nevertheless, additional validation on broader motion domains such as gait analysis, rehabilitation exercise, and daily living activities remains an important direction for future work.

The remainder of this paper is organized as follows. Section 2 reviews related work on sensor-based motion analysis, vision-based modeling, sequence alignment, and contrastive skeleton representation learning. Section 3 formulates the motion similarity retrieval problem and defines the invariance objective. Section 4 presents the proposed spatial-temporal normalization framework. Section 5 describes the acausal temporal convolutional embedding network and hybrid contrastive objective. Section 6 analyzes computational complexity, receptive field characteristics, and stability properties. Section 7 introduces the experimental setup and evaluation protocol. Section 8 reports quantitative and qualitative experimental results, including ablation studies, retrieval comparison, clustering analysis, and amateur-to-professional mapping. Finally, Sections 9 and 10 discuss implications and conclude the paper.

2 Related Work

2.1 Sensor-Based Motion Analysis

IMU-based systems are widely used in golf biomechanics to measure rotational velocity and phase segmentation [1]. Biomechanical analyses reveal that trunk rotation amplitude and pelvis angular velocity significantly influence swing performance [2–4]. Despite their precision, wearable systems suffer from placement sensitivity and reduced accessibility.

2.2 Vision-Based Golf Motion Modeling

Markerless pose estimation techniques such as OpenPose [5] and BlazePose [6] enable scalable motion analysis. Recent golf systems employ embedding-based posture correction and feedback mechanisms [7]. However, these approaches typically evaluate posture consistency rather than full-sequence similarity retrieval.

2.3 Sequence Alignment and Motion Similarity

DTW remains a classical technique for temporal alignment [8,9]. Although effective for phase compensation, DTW requires exhaustive pairwise comparison and does not learn discriminative representations.

Deep metric learning approaches for motion similarity have recently emerged [12]. While these methods learn embedding distances, they often assume pre-aligned sequences and do not explicitly address geometric normalization.

2.4 Contrastive Skeleton Representation Learning

Contrastive learning has significantly improved skeleton-based motion representation. Spatio-temporal contrastive frameworks [10] and reconstruction-consistent self-supervised learning approaches [11] enhance representation robustness. Hierarchical and augmentation-consistent learning schemes further improve discriminative embedding quality.

However, these studies focus primarily on action recognition rather than retrieval-based similarity modeling. Structured geometric normalization is rarely integrated as a pre-embedding stage.

Recent advances in skeleton-based motion modeling have further expanded beyond conventional temporal sequence encoders by incorporating graph-based and attention-based architectures. Spatio-temporal graph convolutional networks (ST-GCNs) explicitly model the spatial connectivity among human joints and the temporal evolution of skeleton sequences, thereby providing a strong representation paradigm for skeleton-based action analysis. Transformer-based and attention-driven models have also been actively explored to capture long-range temporal dependencies and joint-level interactions in human motion sequences.

Recent studies such as Multi-degree Tail-aware Attention Network (MTAN) [20] and Spatio-Temporal Articulation and Coordination Co-attention Graph Network [21] demonstrate that attention mechanisms and graph-based modeling can effectively capture fine-grained articulation patterns and long-range temporal dependencies in human motion. These architectures provide strong alternatives for skeleton-based representation learning by explicitly modeling joint topology and temporal coordination.

However, most existing graph-based and transformer-based approaches are primarily optimized for action recognition or motion prediction rather than retrieval-oriented metric representation learning. In contrast, the proposed framework focuses on invariant similarity retrieval by explicitly integrating deterministic spatial-temporal canonicalization with hybrid contrastive embedding learning.

2.5 Positioning of the Proposed Framework

Our previous similarity framework [17] and reinforcement learning-based correction model [18] demonstrate the effectiveness of embedding-based swing analysis. Building upon these works, the present study integrates structured spatial-temporal normalization with hybrid contrastive embedding for scalable motion similarity retrieval.

Additionally, computer vision-based rehabilitation studies emphasize the importance of robust motion embedding in healthcare applications [19], highlighting the broader applicability of the proposed framework.

3 Problem Formulation

We represent a motion sequence extracted from a monocular RGB video as a 2D skeleton time series:

$$\mathbf{X} = \{\mathbf{x}_t\}_{t=1}^{T}, \mathbf{x}_t \in \mathbb{R}^{J\times 2} \tag{1}$$

where T is the number of frames and J denotes the number of joints. Each joint coordinate at frame t is written as

$$\mathbf{x}_t(j) = (u_{t,j}, v_{t,j}), j \in \{1, \ldots, J\} \tag{2}$$

3.1 Objective: Motion Similarity Retrieval in an Embedding Space

Given a database of motion sequences $\mathcal{D} = \{\mathbf{X}^{(n)}\}_{n=1}^{N}$ and a query sequence $\mathbf{X}^{(q)}$, the goal of motion similarity retrieval is to rank database samples by their similarity to the query. To enable scalable retrieval, we learn an embedding function:

$$f_\theta: \mathbb{R}^{T\times J\times 2} \to \mathbb{R}^{d} \tag{3}$$

that maps a motion sequence to a d-dimensional unit vector $\mathbf{z} = f_\theta(\mathbf{X})$ with $\| \mathbf{z} \|_2 = 1$. Similarity between motions is computed using cosine similarity:

$$s(\mathbf{X}_a, \mathbf{X}_b) = \mathbf{z}_a^{\top}\mathbf{z}_b, \mathbf{z}_a = f_\theta(\mathbf{X}_a), \mathbf{z}_b = f_\theta(\mathbf{X}_b) \tag{4}$$

The retrieval problem is then formulated as finding top-K nearest neighbors:

$$\mathrm{NN}_K(\mathbf{X}^{(q)}) = \arg\max_{\mathrm{X}\in\mathcal{D}}^{K} s(\mathbf{X}^{(q)}, \mathbf{X}) \tag{5}$$

3.2 Spatial-Temporal Nuisance Factors and Invariance Requirement

For monocular 2D pose sequences, similarity comparison is heavily affected by nuisance factors that are not directly related to motion style, including:

1. body-scale differences and camera distance,
2. translation drift due to subject location and framing,
3. in-plane rotation changes caused by camera viewpoint, and
4. temporal phase misalignment due to different swing tempo.

We model these nuisances as a composition of spatial-temporal transformations $\mathcal{T}$ applied to $\mathbf{X}$:

$$\mathbf{X}' = \mathcal{T}(\mathbf{X}) = \mathcal{T}_{\text{temp}} \circ \mathcal{T}_{\text{rot}} \circ \mathcal{T}_{\text{trans}} \circ \mathcal{T}_{\text{scale}}(\mathbf{X}) \tag{6}$$

The embedding should satisfy an invariance property:

$$f_\theta(\mathbf{X}) \approx f_\theta(\mathcal{T}(\mathbf{X})) \tag{7}$$

so that the learned representation reflects motion style rather than viewpoint, scale, or tempo.

3.3 Normalized Sequence and Learning Target

To explicitly enforce invariance, we introduce a deterministic normalization operator $g(\cdot)$ that converts the raw pose sequence into a normalized sequence:

$$\tilde{\mathrm{X}} = g(\mathrm{X}) \tag{8}$$

where $g(\cdot)$ corresponds to the four-stage spatial-temporal normalization described in Section 4 (torso-scale normalization, pelvis-centered alignment, posture-axis alignment, and phase-synchronized temporal resampling). The final embedding becomes

$$\mathbf{z} = f_\theta(\tilde{\mathbf{X}}) = f_\theta(g(\mathbf{X})) \tag{9}$$

3.4 Training Pairs/Triplets for Style Consistency

Let $y(\mathbf{X}) \in \{1, \ldots, C\}$ be the style identity (e.g., professional player ID or expert style label). During training, we sample:

- positive pairs $(\mathbf{X}_i, \mathbf{X}_j)$ such that $y(\mathbf{X}_i) = y(\mathbf{X}_j)$
- negative pairs such that $y(\mathbf{X}_i) \neq y(\mathbf{X}_k)$

The learning objective is to minimize intra-style embedding distances while maximizing inter-style distances. The objective encourages the similarity $z_i^\top z_j$ to increase when $y_i = y_j$, and to decrease when $y_i \neq y_k$, which is implemented via the hybrid contrastive objective in Section 5.

3.5 Prototype-Based Retrieval for Interpretability

For interpretability, we represent each style/class $\boldsymbol{c}$ with a prototype vector:

$$\mathrm{p}_c = \frac{\mathbf{1}}{|\mathcal{D}_c|} \sum_{\mathrm{X} \in \mathcal{D}_c} \boldsymbol{f}_{\boldsymbol{\theta}}\left(g\left(\mathrm{X}\right)\right), \text{then } \mathrm{p}_c \leftarrow \frac{\mathbf{p}_c}{\| \mathbf{p}_c \|_2} \tag{10}$$

where $\mathcal{D}_c \subset \mathcal{D}$ is the subset of sequences belonging to class $\boldsymbol{c}$. A query is mapped to the closest prototype to provide an explainable style match:

$$\hat{c} = \arg\max_c \mathrm{z}^{(q)\top} \mathrm{p}_c \tag{11}$$

This formulation enables both fast retrieval (via nearest-neighbor search in the embedding space) and prototype-based style mapping for amateur-to-professional analysis (Section 8.5).

4 Spatial-Temporal Normalization Framework

To suppress non-style-related variations while preserving intrinsic motion characteristics, we introduce a structured four-stage spatial-temporal normalization model. The framework systematically removes geometric and temporal inconsistencies prior to embedding, thereby ensuring that the learned representation reflects motion style rather than nuisance factors such as body size, camera distance, translation drift, in-plane rotation, or execution tempo.

As illustrated in Fig. 1, raw monocular RGB video frames are first converted into 2D pose sequences through a pose estimation model. The extracted joint coordinates are then processed through the proposed normalization module before being fed into the attention-enhanced temporal convolutional network (A-TCN). The resulting embeddings are finally used for prototype-based similarity retrieval. The normalization module serves as a structural regularizer that stabilizes embedding learning.

4.1 Torso-Scale Normalization

Body-size differences and camera distance variation introduce scale inconsistency across pose sequences. Without normalization, embedding vectors may inadvertently encode anthropometric differences rather than motion characteristics.

Let the torso reference length at a selected reference frame be defined as $\boldsymbol{L}_{\text{torso}}$. This value is typically computed using the Euclidean distance between shoulder and hip midpoints, providing a stable anatomical reference. Each joint coordinate is then scaled as follows:

$$\hat{\boldsymbol{x}}_t = \frac{\boldsymbol{x}_t}{\boldsymbol{L}_{\text{torso}}} \tag{12}$$

Through this operation, all joint coordinates are expressed in a dimensionless coordinate system normalized by torso length. This ensures that relative body proportions are preserved while removing dependency on absolute body size and camera zoom level.

By applying torso-scale normalization, sequences recorded under different distances or from players with different body builds become directly comparable in the embedding space.

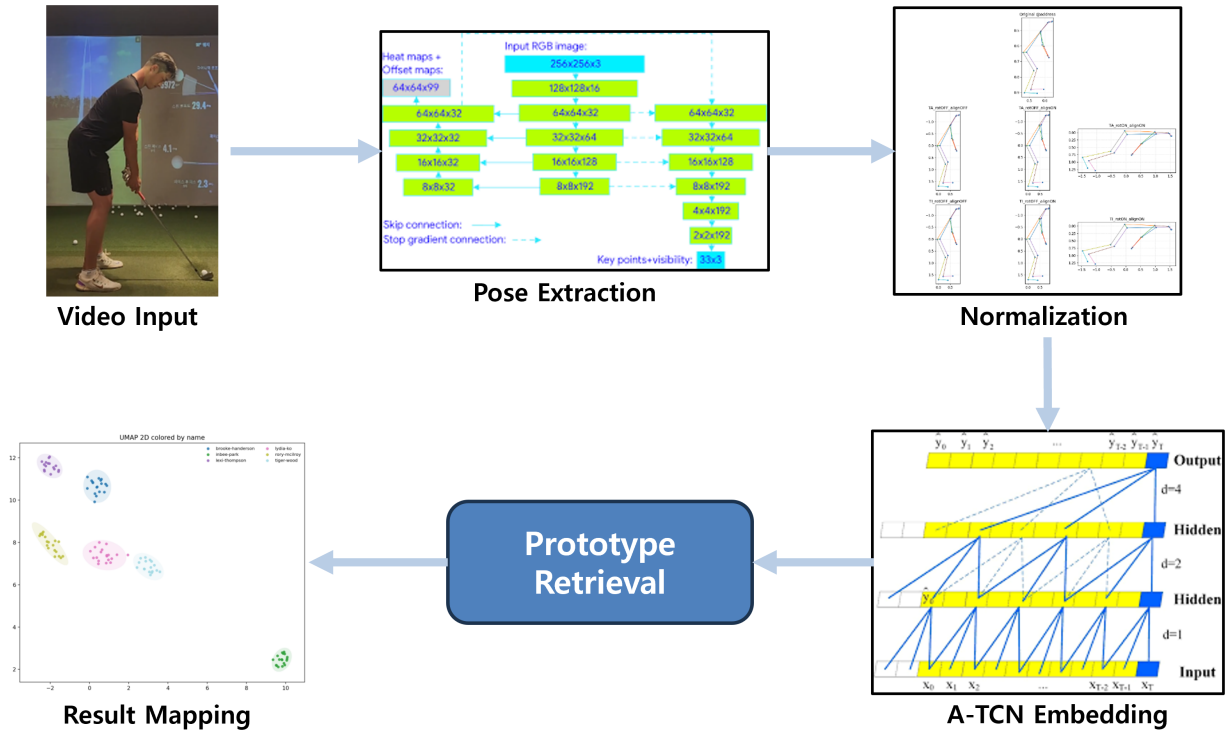

Figure 1: Overall pipeline of the proposed framework. Pose extraction → normalization → A-TCN embedding → prototype retrieval.

4.2 Pelvis-Centered Alignment

Even after scale normalization, translation variance remains due to subject positioning within the image frame. To address this issue, we adopt pelvis-centered alignment strategies.

Two variants are considered: static pelvis alignment recentering (SPAR) and dynamic pelvis alignment recentering (DPAR).

Static alignment is defined as:

$$\hat{\boldsymbol{x}}_t = \boldsymbol{x}_t - \boldsymbol{p}_{\text{ref}} \tag{13}$$

where $\boldsymbol{p}_{\text{ref}}$ denotes a fixed pelvis reference point, typically obtained from the initial address frame.

Dynamic alignment is defined as:

$$\hat{\boldsymbol{x}}_t = \boldsymbol{x}_t - \boldsymbol{p}_t \tag{14}$$

where $\boldsymbol{p}_t$ represents the pelvis center at each time step.

Static alignment preserves global displacement patterns over time because all frames are referenced to a single coordinate origin. This characteristic is particularly important in golf swing analysis, where weight transfer and lateral body shift are meaningful motion features. In contrast, dynamic alignment removes frame-wise translation completely, emphasizing relative limb articulation while suppressing global body movement.

Fig. 2 illustrates representative trajectory variations under different normalization configurations. SPAR preserves cumulative body displacement relative to a fixed reference frame, whereas DPAR suppresses

frame-wise translation more aggressively. Activating posture-axis alignment (PAA) substantially improves rotational consistency, and the additional application of PSR further stabilizes temporal phase alignment. Overall, the combination of SPAR, PAA, and PSR produces the most structurally coherent motion trajectories.

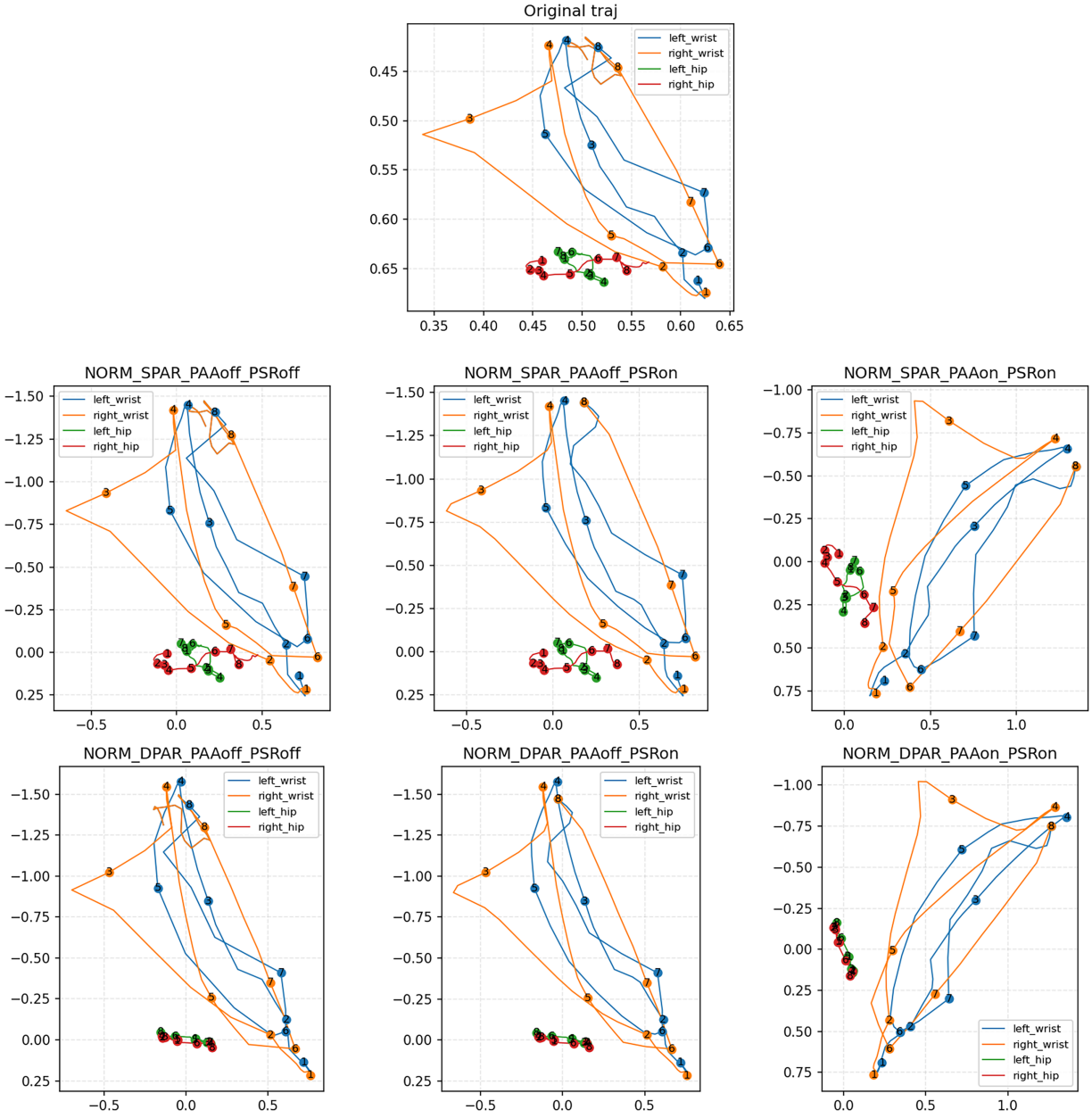

Figure 2: Trajectory comparison under different normalization configurations. SPAR and DPAR denote static and dynamic pelvis alignment recentering; PAA indicates posture axis alignment; PSR denotes phase-synchronized resampling. Static alignment preserves global displacement patterns, whereas dynamic alignment centers each frame independently. The combination of PAA and PSR further improves rotational and temporal consistency.

A key difference between SPAR and DPAR emerges in the preservation of global motion characteristics. SPAR maintains cumulative body displacement relative to a fixed reference frame, thereby retaining weight-shift dynamics that are stylistically meaningful in golf swing analysis. In contrast, DPAR centers each frame independently, suppressing global translation but potentially removing informative displacement cues. This observation supports the later ablation results showing that SPAR combined with PAA and PSR yields superior retrieval performance.

4.3 Posture Axis Alignment (PAA)

In-plane rotational variation arises from camera orientation and stance differences. To canonicalize skeleton orientation, we apply rotation correction based on the posture axis.

The rotation transformation is defined as:

$$x_t^{'} = R(\theta)x_t \tag{15}$$

where the rotation matrix is

$$R(\theta) = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \tag{16}$$

The angle $\boldsymbol{\theta}$ is computed from the shoulder-axis direction, ensuring that the upper-body orientation is aligned consistently across sequences.

This step reduces viewpoint-induced angular discrepancies while preserving relative joint configurations. As a result, inter-sequence comparison becomes less sensitive to camera tilt and stance orientation.

Fig. 3 demonstrates the effect of posture-axis alignment (PAA) under different normalization settings. Without PAA, noticeable orientation inconsistencies remain even after pelvis recentering. Once PAA is applied, upper-body orientation becomes geometrically aligned across frames, substantially improving rotational consistency. The effect becomes more stable when combined with PSR, indicating that spatial and temporal canonicalization complement each other.

4.4 Phase-Synchronized Temporal Resampling

Temporal misalignment is a major source of inconsistency in motion similarity analysis. Players perform swings at different speeds, leading to inconsistent frame-to-phase correspondence across sequences.

To align motion phases, we perform phase-synchronized temporal resampling. Given two adjacent frames $\boldsymbol{x_k}$ and $\boldsymbol{x_{k+1}}$, an interpolated coordinate at normalized time $\boldsymbol{\tau}$ is computed as:

$$\boldsymbol{x'(\tau)} = (\mathbf{1} - \boldsymbol{\alpha})\boldsymbol{x_k} + \boldsymbol{\alpha x_{k+1}} \tag{17}$$

where $\boldsymbol{\alpha} \in [\mathbf{0}, \mathbf{1}]$ controls the interpolation ratio.

By detecting key motion events (e.g., backswing top and impact) and resampling sequences to a fixed temporal resolution, we ensure that semantically equivalent swing phases are temporally aligned across samples. This operation removes tempo dependency while preserving dynamic structure.

Fig. 4 presents wrist coordinate trajectories under different combinations of spatial normalization components and phase-synchronized resampling (PSR). The original temporal signal exhibits inconsistent phase boundaries across key swing events (e.g., take-away, top, impact, and follow-through), resulting in irregular alignment of semantically corresponding motion segments.

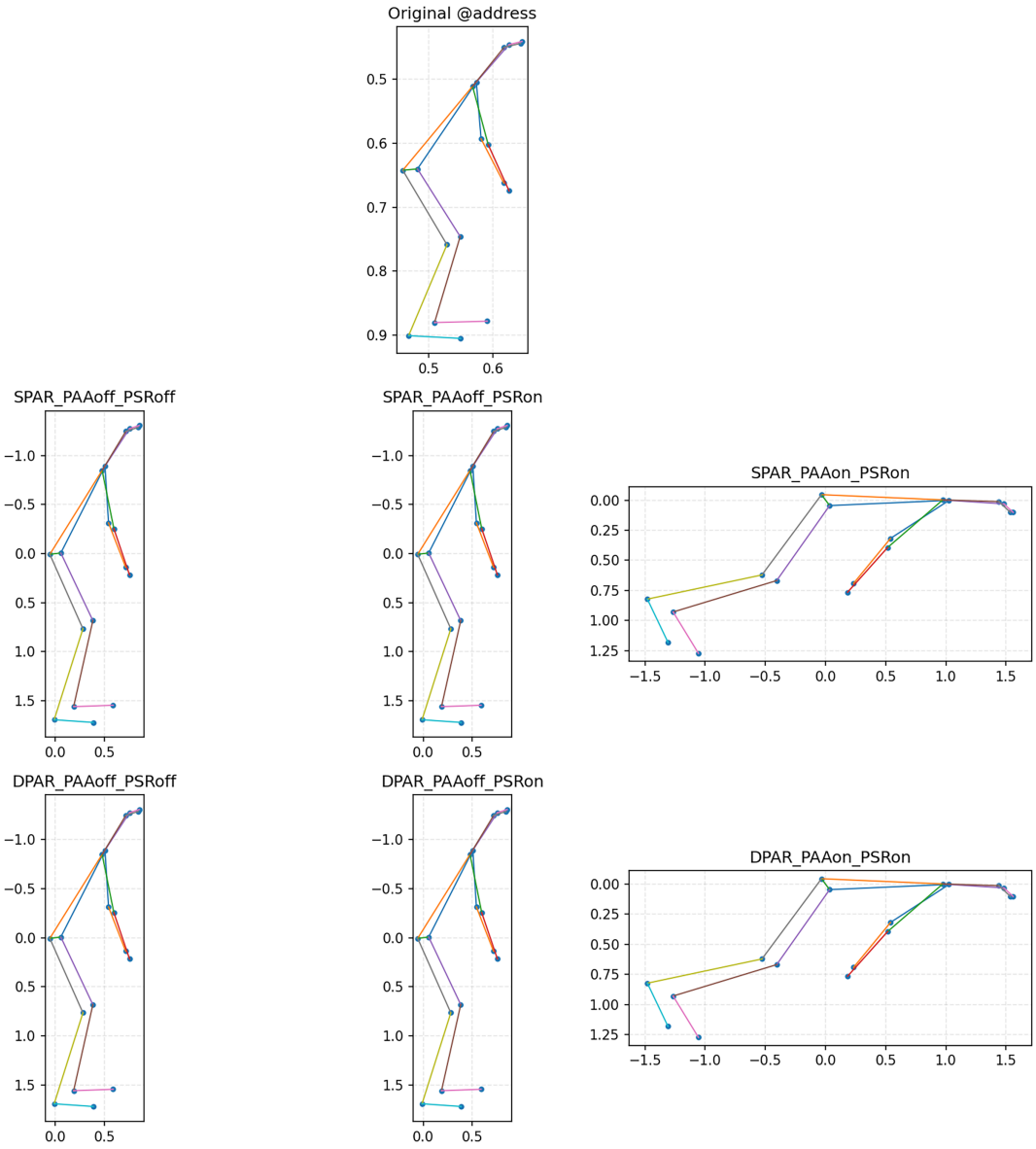

Figure 3: Effect of posture-axis alignment (PAA) under different normalization combinations. SPAR and DPAR denote static and dynamic pelvis recentering, and PSR indicates phase-synchronized resampling. Without PAA, orientation inconsistencies remain despite translation removal. Enabling PAA aligns the torso axis to a canonical direction, substantially reducing rotational variation and improving geometric consistency.

When pelvis recentering (SPAR or DPAR) is applied without PSR (PSRoff), spatial translation effects are reduced, but temporal misalignment remains visible. Event markers occur at different relative frame indices, and peak wrist displacements are not consistently aligned across the motion cycle. Although posture-axis alignment (PAA) improves geometric consistency, it does not correct temporal phase distortion.

Once PSR is enabled (PSRon), the wrist trajectories become temporally aligned with respect to key swing events. Critical motion phases—such as backswing apex and impact—are synchronized across sequences, producing smoother and semantically consistent waveform progression. This temporal canonicalization ensures that similar motion stages correspond to comparable embedding inputs, thereby improving downstream retrieval stability.

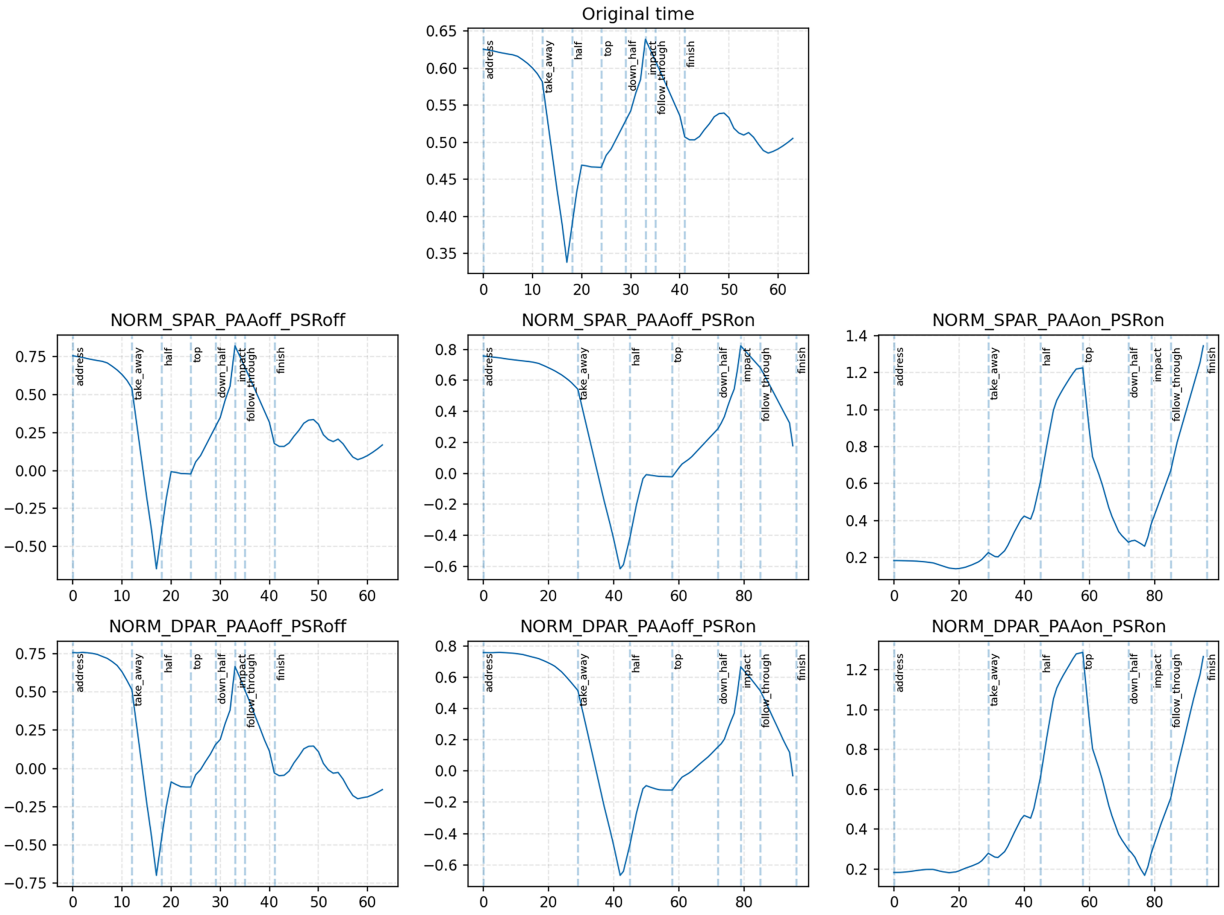

Figure 4: Wrist coordinate trajectories under different normalization configurations. SPAR and DPAR denote static and dynamic pelvis recentering; PAA indicates posture-axis alignment; PSR denotes phase-synchronized resampling. Without PSR, key swing events occur at inconsistent temporal positions. Enabling PSR aligns motion phases across sequences, producing smoother and semantically consistent temporal patterns.

The effect is consistent for both SPAR and DPAR settings, indicating that PSR primarily addresses temporal variability independently of spatial normalization strategy. Combined with spatial canonicalization (SPAR + PAA), PSR yields the most structurally coherent temporal patterns.

Phase-synchronized resampling is particularly effective for cyclic or phase-structured motions such as golf swings because semantically corresponding events can be temporally aligned across sequences. However, for stochastic or non-cyclical motions where stable semantic anchors are unavailable, the current PSR formulation may not be directly applicable. In such cases, alternative temporal normalization strategies, including uniform interpolation, fixed-length resampling, or learned temporal alignment modules, may provide more appropriate temporal canonicalization.

Importantly, the A-TCN embedding network only requires a normalized sequence of fixed temporal length as input and does not explicitly depend on the presence of manually defined phase events. Therefore, PSR should be interpreted as a task-specific temporal canonicalization module rather than a mandatory assumption of the entire framework. This modular design enables the proposed architecture to be adapted to motion domains where semantic phase anchors are absent or difficult to detect.

5 Contrastive Temporal Embedding Network

To encode normalized pose sequences into discriminative and style-consistent representations, we design a contrastive temporal embedding network based on an acausal temporal convolutional architecture. The network is optimized using a hybrid contrastive objective that combines instance-level discrimination and margin-based separation.

As illustrated in Fig. 5, the normalized pose sequence is first processed by a stack of acausal dilated temporal convolution layers. The extracted temporal features are then aggregated through adaptive average pooling and projected into a low-dimensional embedding space. Finally, L2 normalization ensures that all embeddings lie on a unit hypersphere, enabling cosine-based similarity comparison.

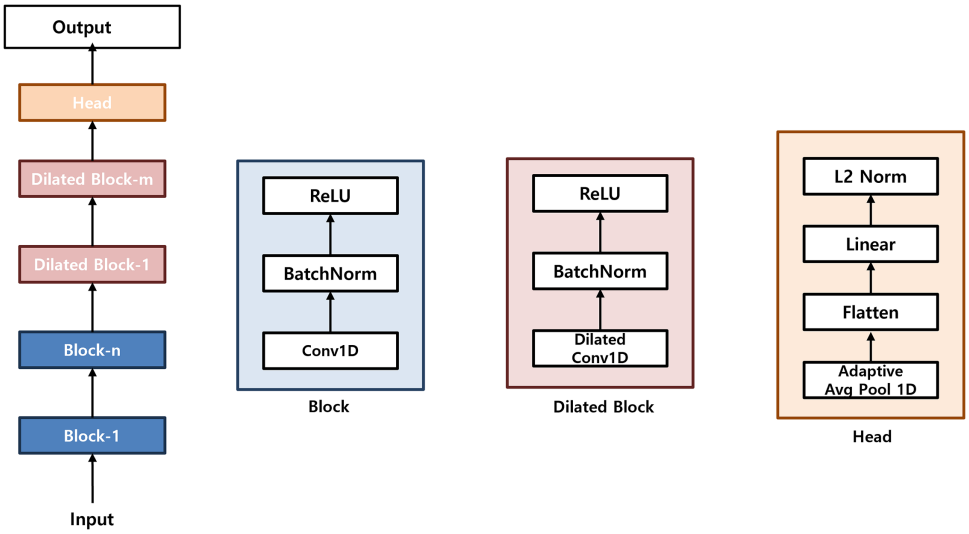

Figure 5: Architecture of the A-TCN embedding network.

5.1 Acausal Temporal Convolutional Network

The backbone of the embedding network is an acausal temporal convolutional network (A-TCN). Unlike causal TCNs used in forecasting tasks, the acausal design allows the model to access both past and future context within the sequence, which is appropriate for offline motion similarity analysis.

The architecture consists of:

- Stacked 1D convolution blocks operating along the temporal axis
- Dilated convolution layers for exponentially expanding receptive fields
- Adaptive average pooling for global temporal aggregation
- A fully connected projection head
- L2 normalization for hyperspherical embedding

Dilated convolutions allow the receptive field to grow exponentially with depth. If the kernel size is $\boldsymbol{k}$ and the network contains $\boldsymbol{L}$ layers with dilation factors $\mathbf{2}^{\boldsymbol{l}}$, the receptive field (RF) is given by:

$$RF = (\boldsymbol{k} - \mathbf{1}) \sum_{\boldsymbol{l}=\mathbf{0}}^{\boldsymbol{L}-\mathbf{1}} \mathbf{2}^{\boldsymbol{l}} + \mathbf{1} \tag{18}$$

This formulation demonstrates that long-range temporal dependencies can be modeled efficiently without significantly increasing parameter count. As a result, the network captures global swing dynamics such as backswing-to-impact transition while maintaining computational efficiency.

Compared to recurrent architectures, the convolutional structure allows full temporal parallelization during both training and inference.

5.2 Hybrid Contrastive Objective

To enforce style-consistent embedding clustering, we employ a hybrid contrastive objective that combines the normalized temperature-scaled cross-entropy (NT-Xent) loss with a semi-hard triplet margin loss.

5.2.1 NT-Xent Loss

The NT-Xent loss encourages embeddings of positive pairs to become closer while pushing away negative samples within a mini-batch. It is defined as:

$$\boldsymbol{L_{NT}} = -\mathbf{log}\,\frac{\mathbf{exp}\,(\mathrm{sim}(\boldsymbol{z_i}, \boldsymbol{z_j})/\boldsymbol{\tau})}{\sum_{\boldsymbol{k}} \mathbf{exp}\,(\mathrm{sim}(\boldsymbol{z_i}, \boldsymbol{z_k})/\boldsymbol{\tau})} \tag{19}$$

where $\boldsymbol{z_i}$ and $\boldsymbol{z_j}$ are embeddings of positive pairs, $\mathrm{sim}(\cdot)$ denotes cosine similarity, and $\boldsymbol{\tau}$ is a temperature parameter controlling distribution sharpness.

This loss operates at the instance level and encourages global separation across all samples in the batch.

5.2.2 Semi-Hard Triplet Loss

While NT-Xent promotes global separation, it does not explicitly enforce a fixed margin between positive and negative pairs. Therefore, we incorporate a semi-hard triplet loss:

$$\boldsymbol{L_{triplet}} = \mathbf{max}\,(\mathbf{0}, \boldsymbol{d}(\boldsymbol{a}, \boldsymbol{p}) - \boldsymbol{d}(\boldsymbol{a}, \boldsymbol{n}) + \boldsymbol{m}) \tag{20}$$

where $\boldsymbol{a}$, $\boldsymbol{p}$, and $\boldsymbol{n}$ represent anchor, positive, and negative samples, respectively; $\boldsymbol{d}(\cdot)$ denotes Euclidean distance; and $\boldsymbol{m}$ is a margin parameter.

Semi-hard negatives are selected such that:

$$\boldsymbol{d}(\boldsymbol{a}, \boldsymbol{p}) < \boldsymbol{d}(\boldsymbol{a}, \boldsymbol{n}) < \boldsymbol{d}(\boldsymbol{a}, \boldsymbol{p}) + \boldsymbol{m} \tag{21}$$

This strategy stabilizes training and improves fine-grained boundary separation between stylistically similar swings.

5.2.3 Final Objective

The overall loss function is defined as a weighted combination:

$$\boldsymbol{L} = \boldsymbol{\lambda_1 L_{NT}} + \boldsymbol{\lambda_2 L_{triplet}} \tag{22}$$

where $\boldsymbol{\lambda_1}$ and $\boldsymbol{\lambda_2}$ control the relative contribution of global batch-level separation and margin-based discrimination.

The hybrid objective enables both robust global clustering and stable inter-class margin enforcement, which is particularly important in motion similarity tasks where subtle stylistic differences must be captured.

5.3 Prototype Representation

To enhance interpretability and retrieval efficiency, we introduce a prototype-based representation in the embedding space.

For each motion class or professional player style $\boldsymbol{c}$, we compute a prototype vector as the mean embedding of all sequences belonging to that class:

$$\boldsymbol{p_c} = \frac{\mathbf{1}}{|\boldsymbol{D_c}|} \sum_{\boldsymbol{x} \in \boldsymbol{D_c}} \boldsymbol{f_\theta}(\boldsymbol{g}(\boldsymbol{X})) \tag{23}$$

The prototype is then L2-normalized:

$$\boldsymbol{p_c} \leftarrow \frac{\boldsymbol{p_c}}{\| \boldsymbol{p_c} \|_2} \tag{24}$$

Given a query embedding $\boldsymbol{z_q}$, similarity-based retrieval is performed by:

$$\hat{\boldsymbol{c}} = \arg\max_{c} \mathrm{sim}(\boldsymbol{z_q}, \boldsymbol{p_c}) \tag{25}$$

This prototype mechanism provides two key advantages:

1. It reduces retrieval latency by replacing large gallery comparisons with class-level similarity checks.
2. It enables interpretable amateur-to-professional mapping, where a user's swing can be directly associated with the closest professional prototype.

Because embeddings lie on a unit hypersphere, prototype comparison reduces to cosine similarity computation, ensuring efficient inference.

6 Computational Analysis

In this section, we analyze the computational efficiency, temporal modeling capacity, and robustness characteristics of the proposed framework. The analysis focuses on three aspects: time complexity, receptive field expansion, and stability under perturbation.

6.1 Time Complexity

The computational complexity of the proposed A-TCN backbone primarily depends on the temporal length $\boldsymbol{T}$, channel dimension $\boldsymbol{C}$, and convolution kernel size $\boldsymbol{k}$.

For a 1D temporal convolution layer, the time complexity is approximately:

$$\boldsymbol{O}(\boldsymbol{T} \cdot \boldsymbol{C} \cdot \boldsymbol{k}) \tag{26}$$

Since convolution operations can be parallelized across the temporal axis, all time steps can be processed simultaneously on modern GPU architectures. This property enables efficient batch processing and low inference latency.

In contrast, a bidirectional LSTM with hidden dimension $\boldsymbol{H}$ exhibits time complexity:

$$\boldsymbol{O}(\boldsymbol{T} \cdot \boldsymbol{H}^2) \tag{27}$$

Recurrent models require sequential computation over time steps due to hidden state dependency. As a result, parallelization across temporal dimensions is limited, and inference latency increases proportionally with sequence length.

Therefore, compared to BiLSTM-based motion encoders, the proposed TCN-based embedding network offers:

- Full temporal parallelization
- Reduced inference latency
- Better scalability for long sequences
- Improved suitability for real-time or large-scale retrieval systems

This computational advantage becomes particularly important when processing high-frame-rate motion sequences or large-scale retrieval databases.

To further assess practical efficiency, we measured the average inference time of the proposed A-TCN encoder under the experimental environment described in Table 1. For a normalized sequence of 128 frames, the proposed model required approximately 2–5 ms per sequence on an RTX 4090 GPU, excluding pose extraction time. Since the embedding is computed only once for each gallery sequence, retrieval can be performed efficiently using cosine similarity in the learned embedding space.

Table 1: Experimental environment configuration.

Category	Component	Specification
Hardware Platform	CPU	Intel Core i9-13900K
	GPU	NVIDIA RTX 4090 (24 GB VRAM)
	Memory	64 GB DDR5 RAM
Software Environment	Operating System	Ubuntu 22.04 LTS
	Python Version	Python 3.10
	Deep Learning Framework	PyTorch 2.x
	CUDA Version	CUDA 12.x
Training Setup	Optimizer	Adam
	Initial Learning Rate	1e−4
	Batch Size	64
	Number of Epochs	200
	Weight Decay	1e−5
	NT-Xent Temperature	0.07
	Triplet Margin	0.2
Evaluation Setting	Similarity Metric	Cosine Similarity
	Retrieval Metric	Top-1 Accuracy
	Dimensionality Reduction	UMAP, t-SNE

In contrast, DTW requires pairwise sequence alignment with quadratic temporal complexity, which becomes computationally expensive as the number of query-gallery comparisons increases. Therefore, the proposed embedding-based approach is more suitable for large-scale retrieval scenarios, where gallery embeddings can be precomputed and indexed for efficient nearest-neighbor search.

6.2 Receptive Field Analysis

The receptive field determines how many temporal frames influence the embedding at a given output position. In motion analysis tasks such as golf swing similarity retrieval, long-range dependencies are crucial because stylistic features span the entire swing cycle.

In a dilated convolutional network with kernel size $\boldsymbol{k}$ and $\boldsymbol{L}$ layers, where the dilation factor doubles at each layer (i.e., $\mathbf{2}^{\boldsymbol{l}}$), the receptive field (RF) is given by:

$$\boldsymbol{RF} = (\boldsymbol{k}-\mathbf{1})\sum_{\boldsymbol{l}=\mathbf{0}}^{\boldsymbol{L}-\mathbf{1}}\mathbf{2}^{\boldsymbol{l}}+\mathbf{1} \tag{28}$$

Since

$$\sum_{\boldsymbol{l}=\mathbf{0}}^{\boldsymbol{L}-\mathbf{1}}\mathbf{2}^{\boldsymbol{l}} = \mathbf{2}^{\boldsymbol{L}}-\mathbf{1}, \tag{29}$$

the receptive field grows exponentially with network depth:

$$\boldsymbol{RF} = (\boldsymbol{k}-\mathbf{1})(\mathbf{2}^{\boldsymbol{L}}-\mathbf{1})+\mathbf{1} \tag{30}$$

This exponential expansion allows the model to capture global temporal patterns with relatively shallow architectures. Unlike recurrent networks that accumulate information sequentially, dilated TCN layers aggregate long-range context through hierarchical convolution.

The kernel size and dilation configuration determine the balance between local articulation modeling and global trajectory modeling. A small convolution kernel captures short-term joint variations such as wrist movement, shoulder rotation, and local limb articulation. Meanwhile, exponentially increasing dilation factors expand the receptive field without a proportional increase in model parameters, allowing the network to integrate long-range temporal information across the entire swing cycle.

In the golf swing task, this design is particularly important because stylistic differences are not limited to isolated frames. Instead, they emerge from coordinated transitions across phases, such as address-to-backswing, backswing-to-top, top-to-impact, and impact-to-follow-through. The receptive field of the proposed A-TCN covers the full normalized sequence length, enabling the encoder to capture both fine-grained skeletal articulation and global temporal structure.

For example, with $\boldsymbol{k} = \mathbf{3}$ and $\boldsymbol{L} = \mathbf{6}$, the receptive field becomes:

$$\boldsymbol{RF} = \mathbf{2}(\mathbf{2}^{\mathbf{6}}-\mathbf{1})+\mathbf{1} = \mathbf{2}(\mathbf{63})+\mathbf{1} = \mathbf{127} \tag{31}$$

This means the network can effectively model dependencies across 127 frames, covering an entire swing sequence in most practical scenarios.

Such coverage ensures that the embedding encodes holistic motion dynamics rather than short-term joint fluctuations.

6.3 Stability under Perturbation

Robust motion similarity retrieval requires stability against small perturbations in input pose coordinates. Such perturbations may arise from pose estimation noise, occlusion artifacts, or minor tracking errors.

Let $\boldsymbol{f}(\boldsymbol{X})$ denote the embedding function and consider a perturbed input $\boldsymbol{X}+\boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon}$represents bounded noise. A stable embedding function should satisfy a Lipschitz continuity condition:

$$\| \boldsymbol{f}(\boldsymbol{X}) - \boldsymbol{f}(\boldsymbol{X}+\boldsymbol{\varepsilon}) \|_2 \leq \boldsymbol{L} \| \boldsymbol{\varepsilon} \|_2 \tag{32}$$

where $\boldsymbol{L}$ is the Lipschitz constant of the network.

The proposed framework improves stability in two ways:

1. *Normalization reduces input variance by suppressing nuisance factors prior to representation learning.*

Spatial normalization removes scale, translation, and rotation variations, effectively reducing the magnitude of nuisance perturbations before temporal modeling.

2. *Convolutional architecture promotes smoothness.*

Convolution operations act as local aggregators, averaging nearby temporal features and attenuating high-frequency noise.

Because normalization constrains the input domain and the A-TCN uses bounded activation functions and weight regularization, the effective Lipschitz constant is reduced compared to raw-coordinate modeling.

Consequently, small perturbations in joint coordinates do not cause large embedding shifts, improving retrieval consistency under noisy conditions.

A potential concern of deterministic normalization is that errors in estimating anatomical reference points, such as the pelvis center or shoulder axis, may propagate to subsequent processing stages. To mitigate this issue, the proposed framework uses anatomically stable reference structures and applies normalization at the sequence level rather than relying on a single isolated joint coordinate. In addition, temporal convolution aggregates neighboring temporal features, which helps attenuate frame-level pose noise.

The normalization module can also be interpreted as a structural regularizer that reduces nuisance variance before representation learning. By suppressing scale, translation, and in-plane rotation variations, the input distribution becomes more compact and stable, allowing the embedding model to focus on motion-relevant differences. Therefore, normalization does not merely introduce an additional preprocessing step; it reduces the burden on the learning model and improves embedding consistency under realistic pose perturbations.

Nevertheless, severe pose estimation failures or occlusion-induced errors may still affect the normalized sequence. This limitation is acknowledged, and future work will incorporate confidence-aware pose filtering, temporal smoothing, and uncertainty-aware embedding learning to further improve robustness.

7 Experimental Setup

To comprehensively evaluate the proposed spatial-temporal normalization and contrastive embedding framework, we conducted systematic experiments under controlled computational conditions. This section describes the experimental environment, dataset characteristics, implementation configuration, evaluation metrics, and the golf swing–specific evaluation protocol.

Table 1 summarizes the hardware and software environment used for all experiments, including GPU model, CUDA version, deep learning framework, optimizer configuration, and training parameters. All comparative experiments between A-TCN and BiLSTM were conducted under identical computational conditions to ensure fairness.

7.1 Dataset

The dataset consists of professional and amateur golf swing videos collected under semi-controlled indoor conditions. Each video captures a full swing cycle from address to follow-through.

All videos were converted into 2D pose sequences using a state-of-the-art pose estimation model. Each sequence contains approximately 90 to 140 frames sampled at 30 FPS, with **17** skeletal keypoints extracted per frame.

Professional swings were treated as style anchors. Amateur swings were reserved for similarity retrieval and prototype mapping experiments. To avoid subject leakage, the dataset was split at the subject level into training, validation, and testing sets.

7.2 Implementation Details

Before embedding, all sequences were normalized using the proposed spatial-temporal normalization framework and resampled to a fixed length of $\boldsymbol{T'} = \mathbf{128}$ frames.

The architecture of the proposed A-TCN model is summarized in Table 2.

Table 2: Detailed architecture configuration of the proposed A-TCN embedding network.

Component	Configuration
Input dimension	34 (17 joints × 2 coordinates)
Temporal length	128
Conv Blocks	6 Dilated Conv Layers
Kernel size	3
Channel dimension	128
Dilation rates	1, 2, 4, 8, 16, 32
Dropout	0.2
Projection head	Fully connected (128-dim)
Output embedding	128-dim (L2 normalized)

The model was optimized using the Adam optimizer with an initial learning rate of $\mathbf{1} \times \mathbf{10^{-3}}$. The hybrid contrastive loss was applied with weighting parameters $\boldsymbol{\lambda_1} = \mathbf{1.0}$ and $\boldsymbol{\lambda_2} = \mathbf{0.5}$. Training was performed for 150 epochs with early stopping based on validation Top-1 accuracy. The temperature parameter τ in the NT-Xent loss and the margin parameter m in the triplet loss were selected based on validation performance. The default values were set to $\tau = 0.07$ and m = 0.2, which provided stable convergence and strong retrieval performance. The loss weights λ_1 and λ_2 were initialized to balance global batch-level discrimination and margin-based local separation. A detailed sensitivity analysis of these hyperparameters is provided in Section 8.6.

For comparison, a 2-layer BiLSTM with hidden dimension 128 was implemented using the same embedding dimension and projection head to ensure architectural fairness.

7.3 Evaluation Metrics

To evaluate retrieval and clustering performance, we employed multiple quantitative metrics.

Top-1 accuracy measures the percentage of query sequences whose nearest neighbor belongs to the same professional class. Top-5 accuracy evaluates whether the correct class appears within the five nearest retrieved results.

For clustering analysis, cosine similarity statistics and silhouette scores were computed. All similarity calculations were performed in the L2-normalized embedding space using cosine similarity.

7.4 Golf Swing Evaluation Protocol

The evaluation protocol was designed to simulate realistic swing similarity retrieval scenarios.

In professional-to-professional evaluation, each professional swing was used as a query against the remaining professional swings as gallery samples.

In amateur-to-professional mapping, amateur swing embeddings were compared against professional class prototypes. The most similar prototype was interpreted as the closest professional style.

This protocol ensures that retrieval performance reflects stylistic similarity rather than direct frame-level alignment.

8 Experimental Results

This section presents quantitative and qualitative experimental results demonstrating the effectiveness of the proposed spatial-temporal normalization framework and contrastive temporal embedding network. The results are analyzed from multiple perspectives, including ablation study, embedding separability, retrieval comparison, clustering consistency, and amateur-to-professional style mapping.

8.1 Normalization Ablation

To evaluate the contribution of each normalization component described in Section 4, we conducted an ablation study by systematically enabling or disabling torso-scale normalization, pelvis alignment strategy, posture-axis alignment, and phase-synchronized temporal resampling.

The best-performing configuration consists of static pelvis alignment recentering, posture-axis alignment, and phase-synchronized temporal resampling. Under this configuration, the A-TCN model achieved a Top-1 accuracy of 91.3%, indicating that spatial canonicalization and temporal synchronization are both essential for discriminative motion embedding.

Table 3 presents a unified comparison of normalization ablation results across the A-TCN and BiLSTM backbones. Without normalization, BiLSTM exhibits severe performance degradation (13.4%) compared to A-TCN (56.5%), highlighting the stronger robustness of convolutional temporal modeling to raw pose variability.

For both backbones, incorporating spatial-temporal normalization significantly improves retrieval accuracy. Under static pelvis alignment recentering (SPAR), enabling posture-axis alignment (PAA) consistently yields performance gains, confirming that angular canonicalization reduces inter-sequence orientation variance. Phase-synchronized resampling (PSR) further enhances performance, particularly when combined with PAA, demonstrating the importance of temporal phase alignment.

Across all configurations, SPAR consistently outperforms DPAR for both backbones. The full normalization configuration (SPAR + PAA + PSR) achieves the best performance for both A-TCN (91.3%) and BiLSTM (82.6%). However, A-TCN maintains a consistent performance advantage, indicating that acausal dilated convolution captures long-range motion dependencies more effectively than recurrent modeling.

These results confirm that normalization contributes backbone-independent gains, while A-TCN better leverages the canonicalized motion representation.

These findings empirically validate the theoretical motivation presented in Section 4, demonstrating that spatial-temporal canonicalization directly improves embedding separability.

Table 3: Comparison of normalization ablation results between A-TCN and BiLSTM (Top-1 accuracy, %).

Pelvis Alignment	PAA	PSR	A-TCN (%)	BiLSTM (%)
None (Original)	✗	✗	56.5	13.4
Basic Normalization	✗	✗	61.2	39.7
SPAR	✗	✗	78.2	60.8
SPAR	✗	✓	69.5	69.5
SPAR	✓	✗	82.6	73.9
SPAR	✓	✓	**91.3**	**82.6**
DPAR	✗	✗	69.5	61.1
DPAR	✗	✓	73.9	78.2
DPAR	✓	✗	79.1	65.2
DPAR	✓	✓	86.9	78.8

Note: The bold values indicate the best-performing results.

8.2 Embedding Model Comparison

Figs. 6 and 7 present two-dimensional projections of the learned embeddings using UMAP and t-SNE for A-TCN and BiLSTM, respectively. Clear structural differences emerge between the two backbones.

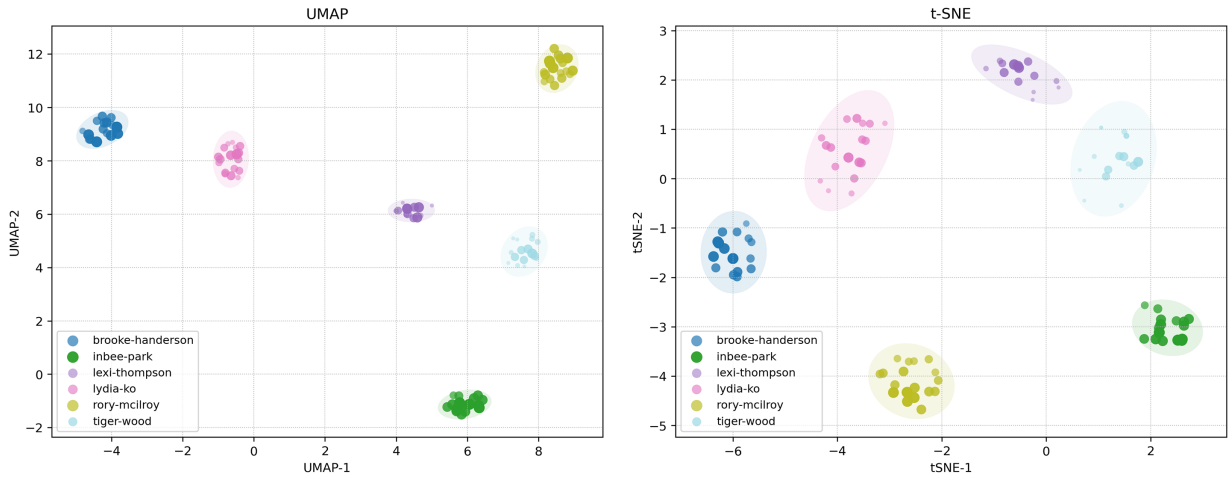

Figure 6: UMAP (**left**) and t-SNE (**right**) projections of motion embeddings learned by the proposed A-TCN under full normalization. The embeddings form compact intra-class clusters with clear inter-class separation, indicating strong discriminative capability and stable manifold structure.

For A-TCN (Fig. 6), both UMAP and t-SNE visualizations exhibit compact intra-class clusters and well-separated inter-class boundaries. Each player's swing style forms a distinct and geometrically isolated region in the embedding space, indicating strong discriminative capability. The cluster centroids are clearly spaced, and minimal overlap is observed across classes. This structural separability aligns with the superior retrieval accuracy reported in Table 3.

In contrast, BiLSTM embeddings (Fig. 7) show comparatively elongated and partially overlapping clusters. While normalization significantly improves grouping consistency, cluster compactness remains weaker than that of A-TCN. In particular, boundary regions between certain players exhibit increased dispersion, suggesting that recurrent temporal modeling is less effective in capturing global motion structure.

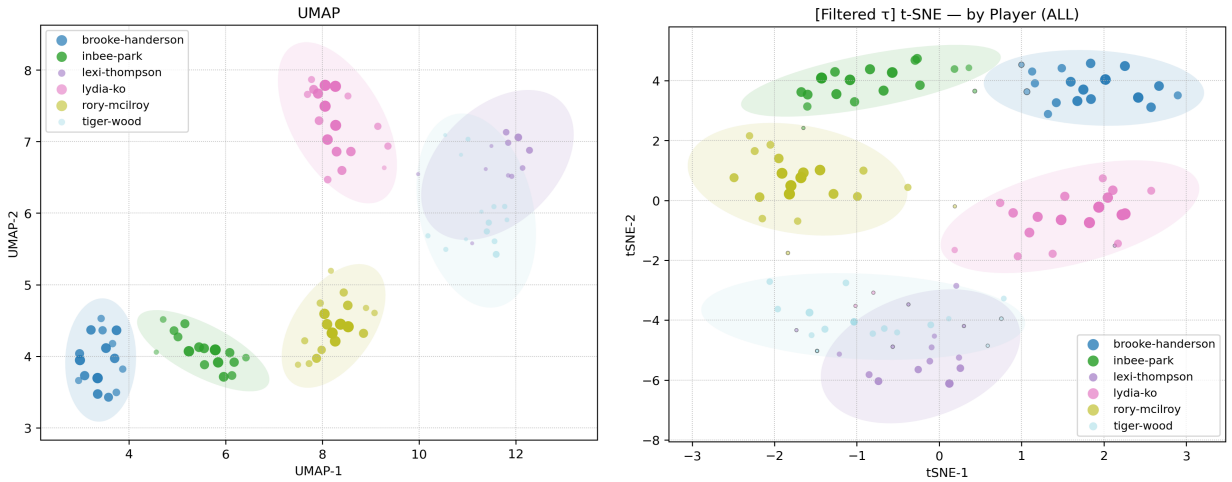

Figure 7: UMAP (**left**) and t-SNE (**right**) projections of motion embeddings learned by the BiLSTM baseline under full normalization. Although normalization improves grouping consistency, clusters exhibit greater dispersion and partial overlap compared to A-TCN, reflecting reduced structural separability.

The consistency between UMAP and t-SNE projections for A-TCN further indicates that the learned representation preserves both local neighborhood relationships and global structural geometry. For BiLSTM, although general grouping patterns are visible, the embedding manifold appears less structured and more sensitive to dimensionality reduction.

Overall, these results confirm that the acausal dilated convolution in A-TCN provides a more stable and discriminative embedding space compared to recurrent modeling, particularly under spatial-temporal canonicalization.

Fig. 8 illustrates the training dynamics of the proposed A-TCN model under the full normalization configuration. The Top-1 and Top-3 retrieval accuracies increase rapidly during the early epochs and gradually stabilize after approximately 80 epochs, indicating efficient convergence. The validation performance closely follows the training trend without severe fluctuation, suggesting stable generalization behavior.

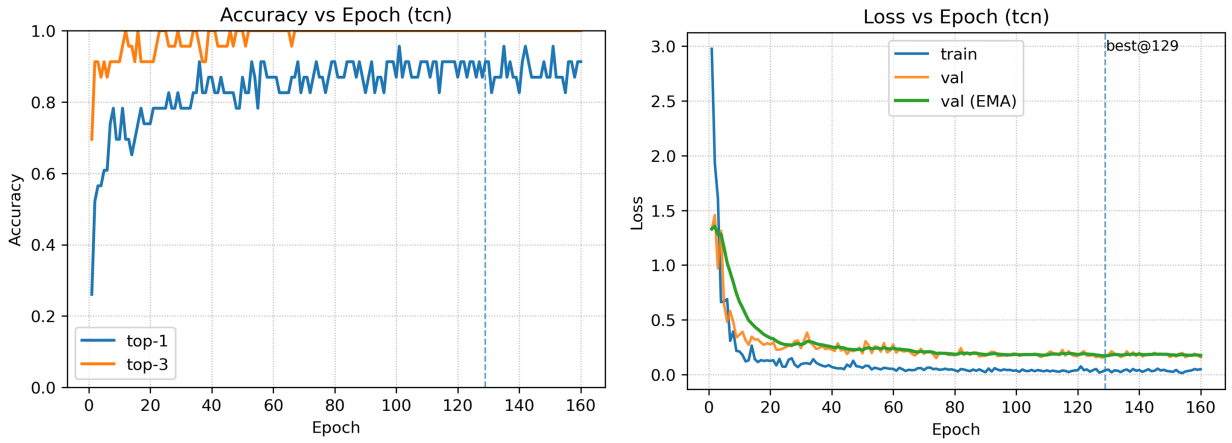

Figure 8: Training accuracy (**left**) and loss convergence (**right**) of the proposed A-TCN model under full normalization. The model demonstrates rapid convergence, stable validation performance, and smooth optimization dynamics, with the best validation performance observed at epoch 129.

The loss curves further demonstrate smooth optimization dynamics. The training loss decreases sharply during initial epochs and gradually converges to a low plateau, while the validation loss remains stable without noticeable divergence. The exponential moving average (EMA) of validation loss exhibits minimal oscillation, confirming robust convergence behavior. The best validation performance is achieved around epoch 129, as indicated in the figure.

These observations suggest that the acausal dilated convolution structure facilitates stable optimization and effective representation learning, supporting the superior embedding separability shown in Figs. 6 and 7.

8.3 Retrieval Performance: Embedding vs. DTW

To assess retrieval effectiveness, we compare the proposed embedding-based framework with Dynamic Time Warping (DTW), a classical sequence alignment baseline. While DTW directly measures frame-wise temporal similarity, the proposed approach leverages a learned spatial-temporal representation optimized for discriminative retrieval.

(a) *File-Level Retrieval*

As shown in Table 4a, the embedding method significantly outperforms DTW across all evaluation metrics at the file level. The proposed model achieves an MRR of 0.986 compared to 0.918 for DTW, indicating more accurate ranking of relevant sequences. The mAP improves substantially from 0.636 to 0.847, reflecting enhanced global retrieval quality. Most notably, Recall@1 increases from 0.864 to 0.982, demonstrating a strong improvement in top-ranked retrieval performance. These results indicate that the learned embedding captures discriminative motion characteristics beyond simple temporal alignment.

Table 4: Retrieval performance comparison between DTW and the proposed embedding method. (File and player levels). (**a**) File-Level Retrieval; (**b**) Player-Level Retrieval.

(**a**)

Method	MRR ↑	mAP ↑	Recall@1 ↑	Recall@3 ↑
DTW	0.918	0.636	0.864	0.964
Proposed (Embedding)	0.986	0.847	0.982	0.991

(**b**)

Method	MRR ↑	Recall@1 ↑	Recall@3 ↑
DTW	0.927	0.873	0.982
Proposed (Embedding)	0.976	0.955	1.000

(b) *Player-Level Retrieval*

At the player level (Table 4b), the performance advantage remains consistent. The embedding approach achieves an MRR of 0.976 compared to 0.927 for DTW. Recall@1 improves from 0.873 to 0.955, while Recall@3 reaches 1.000, meaning that the correct player is always retrieved within the top three candidates. This demonstrates that the embedding space preserves stylistic consistency at a higher semantic level.

The performance gap between the two methods can be attributed to their fundamentally different similarity mechanisms. DTW relies on local temporal alignment and remains sensitive to spatial distortions and phase variability. In contrast, the proposed method integrates spatial-temporal normalization with contrastive representation learning, producing a compact and invariant embedding space. This learned representation enables robust similarity measurement and scalable retrieval.

Overall, the results confirm that embedding-based retrieval substantially surpasses direct sequence alignment methods such as DTW, both at fine-grained file-level comparison and higher-level stylistic player identification.

Although the present experimental comparison focuses on DTW and BiLSTM as representative sequence alignment and temporal neural baselines, recent skeleton-based architectures such as ST-GCN and attention-based motion models provide important reference points for future extensions. These models are effective in action recognition or motion prediction tasks because they explicitly model joint connectivity and long-range temporal dependencies. However, they are not directly optimized for retrieval-oriented metric consistency.

The proposed framework differs from these approaches in that it explicitly formulates motion comparison as an embedding-based retrieval problem. Rather than predicting future motion or classifying action categories, the objective is to construct a metric space in which stylistically similar motions are close and dissimilar motions are separated. Therefore, the comparison with DTW and BiLSTM establishes the effectiveness of the proposed spatial-temporal normalization and temporal embedding design under retrieval-specific evaluation metrics. Future work will include direct experimental comparison with graph-based and transformer-based retrieval architectures such as ST-GCN.

8.4 Clustering Analysis

To further investigate the structural properties of the learned embedding space, we conduct a clustering analysis using cosine similarity and K-means clustering. While retrieval performance evaluates ranking behavior, clustering analysis examines whether stylistic motion patterns naturally form compact and separable groups in the embedding manifold.

(1) *Similarity Structure*

Fig. 9 presents the cosine similarity matrix of all sequences, sorted by player identity. Clear block-diagonal structures are observed, where high intra-player similarity (bright yellow regions) and relatively lower inter-player similarity (green to blue regions) form distinct partitions.

This structured similarity pattern indicates that the proposed embedding successfully preserves intra-style coherence while maintaining inter-style separability. The sharp contrast between diagonal and off-diagonal regions confirms that stylistic motion characteristics are encoded in a discriminative manner.

(2) *Optimal Number of Clusters*

To determine the appropriate number of clusters, we compute the mean cosine-based silhouette score for K ranging from 2 to 12. As shown in Fig. 10, the silhouette score peaks at **K = 6**, which corresponds exactly to the number of professional players in the dataset.

The maximum silhouette score (≈0.70) indicates strong cluster compactness and separation. The monotonic decrease beyond K = 6 suggests that over-segmentation degrades structural coherence. This result validates that the embedding space naturally forms six well-separated motion style clusters without supervision.

(3) *Cluster Visualization*

Fig. 11 visualizes K-means clustering results (K = 6) using UMAP and t-SNE projections. Each cluster corresponds almost perfectly to a distinct player, demonstrating strong consistency between unsupervised clustering and ground-truth identity.

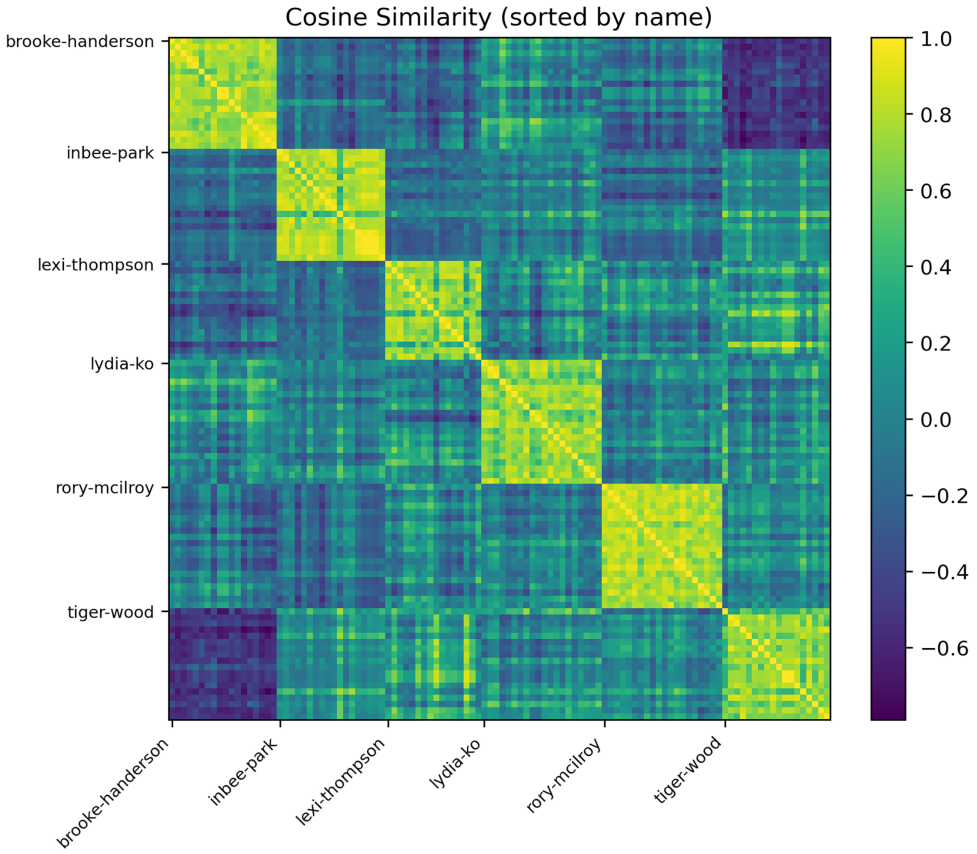

Figure 9: Cosine similarity matrix of embedding vectors sorted by player identity. Block-diagonal structures indicate high intra-player similarity and clear inter-player separation, demonstrating strong style-aware representation learning.

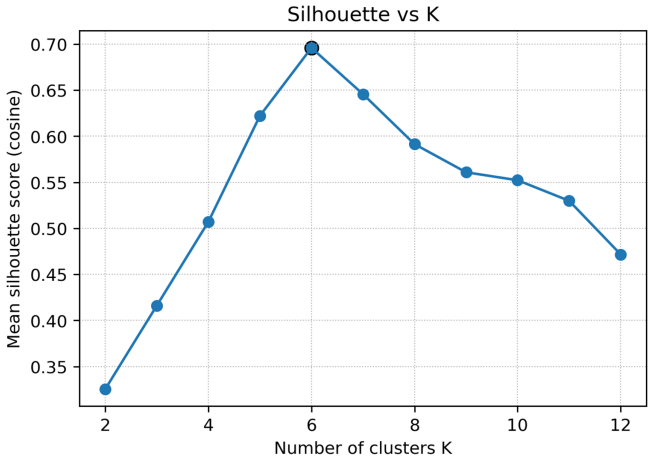

Figure 10: Mean silhouette score (cosine distance) as a function of cluster number K. The score peaaks at K = 6, corresponding to the number of players, indicating optimal cluster compactness and separation.

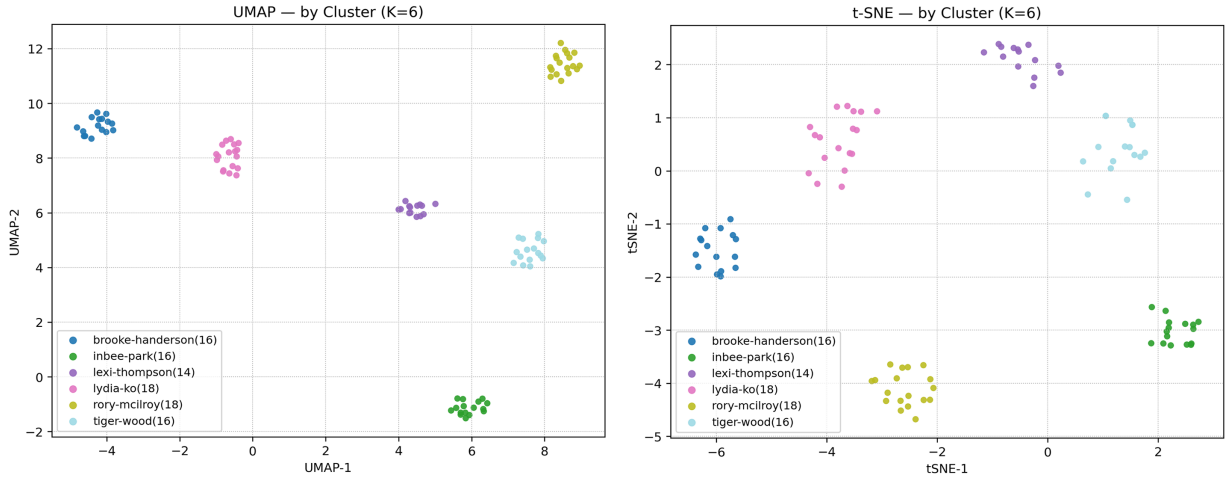

Figure 11: K-means clustering results (K = 6) visualized using UMAP (left) and t-SNE (right). Clusters align closely with player identities, confirming that the learned embedding space naturally organizes stylistic motion patterns.

The UMAP projection highlights global structural separation, while the t-SNE visualization emphasizes local cluster compactness. In both representations, clusters exhibit minimal overlap, confirming that the embedding manifold is geometrically organized in a style-aware manner.

Overall, these findings indicate that the proposed spatial-temporal normalized embedding not only improves retrieval accuracy but also produces a well-structured and semantically meaningful embedding space.

8.5 Amateur-to-Professional Mapping

To evaluate whether the learned embedding space supports cross-domain style interpretation, we analyze how amateur swing sequences (AMA1–AMA4) are positioned relative to professional player prototypes. This experiment aims to determine whether amateur swings naturally align with specific professional style clusters.

We conduct the analysis in three stages:

1. cluster visualization with fixed K = 6 including amateur samples,
2. cosine similarity comparison against professional prototypes, and
3. quantitative similarity statistics (mean ± std).

(1) *Cluster-Level Alignment (AMA1, AMA3)*

When AMA1 (Fig. 12) samples are projected into the embedding space with K = 6 fixed to professional clusters, the amateur samples are positioned near the Lydia Ko cluster in both UMAP and t-SNE projections. The proximity is visually consistent across both manifold representations, suggesting that AMA1 exhibits stylistic characteristics similar to Lydia Ko.

In contrast, AMA3 (Fig. 13) samples are primarily aligned with the Rory McIlroy cluster. The clustering result remains stable across UMAP and t-SNE, indicating that the embedding manifold preserves consistent geometric relationships even under different nonlinear projections.

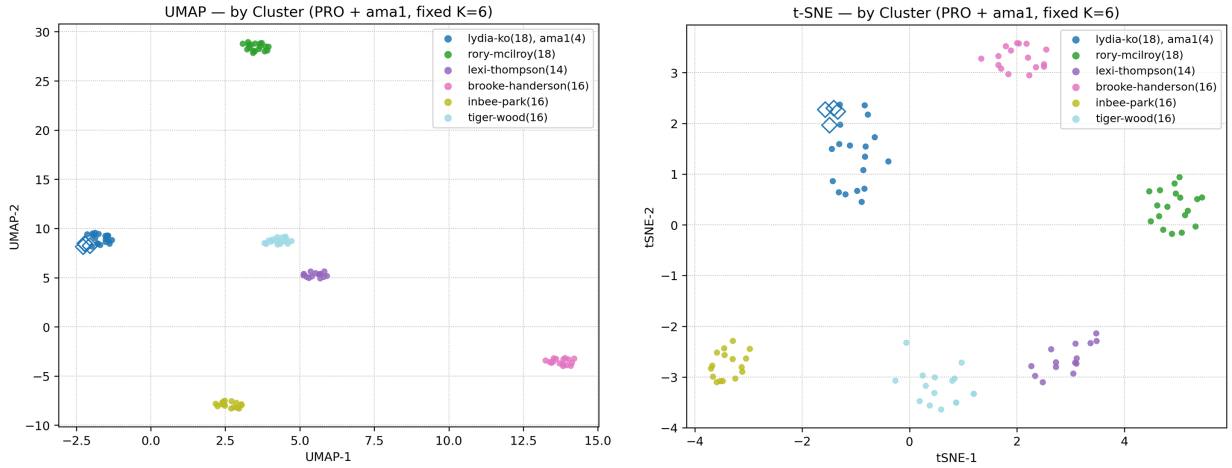

Figure 12: Clustering visualization of professional and AMA1 samples (K = 6) using UMAP (**left**) and t-SNE (**right**). AMA1 samples are embedded near the Lydia Ko cluster, indicating strong stylistic similarity.

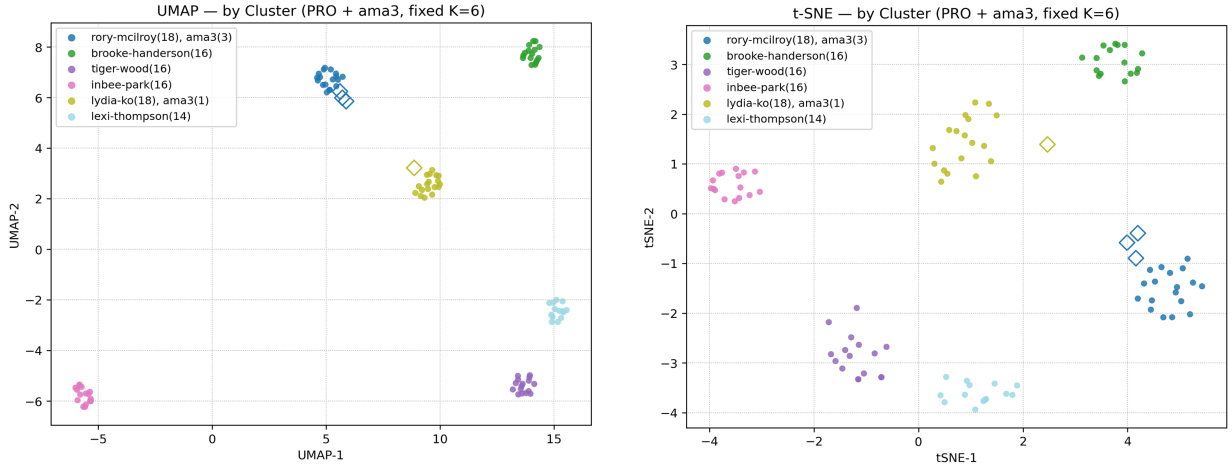

Figure 13: Clustering visualization of professional and AMA3 samples (K = 6) using UMAP (**left**) and t-SNE (**right**). AMA3 samples align predominantly with the Rory McIlroy cluster, demonstrating cross-domain style consistency.

These observations confirm that the learned representation generalizes beyond trained professional samples and enables meaningful amateur-to-prototype mapping.

(2) *Cosine Similarity to Professional Prototypes*

To quantify alignment strength, we compute cosine similarity between each amateur sample and professional cluster prototypes (mean embedding per player).

Fig. 14 shows the mean ± standard deviation of cosine similarity between AMA1 and each professional prototype. AMA1 exhibits the highest similarity to Lydia Ko (0.742 ± 0.043), while similarities to other players are substantially lower or even negative (e.g., −0.349 for Lexi Thompson). This confirms that AMA1 is stylistically closest to Lydia Ko.

(3) *Full Similarity Matrix (All Amateur Sets)*

Table 5 summarizes the cosine similarity statistics (mean ± std) between each amateur group (AMA1–AMA4) and professional prototypes.

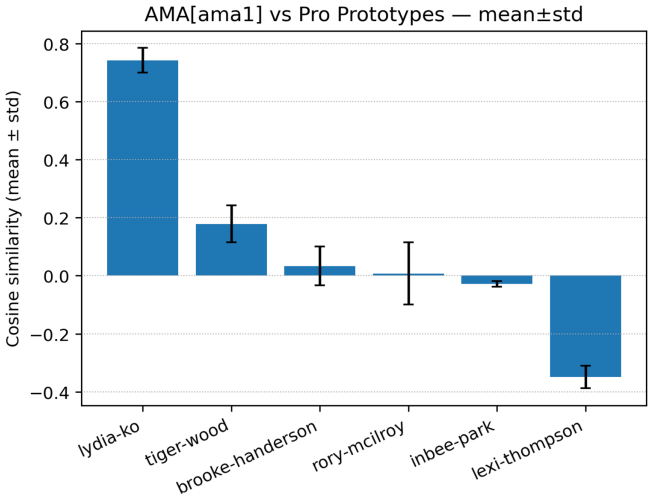

Figure 14: Cosine similarity (mean ± std) between AMA1 and professional prototypes. AMA1 shows the highest similarity to Lydia Ko, with clear separation from other professional styles.

Table 5: Cosine similarity (mean ± std) between amateur groups (AMA1–AMA4) and professional player prototypes.

Amateur	Lydia Ko	Tiger Woods	Brooke Henderson	Rory McIlroy	Inbee Park	Lexi Thompson
AMA1	0.742 ± 0.043	0.179 ± 0.063	0.034 ± 0.067	0.008 ± 0.107	−0.029 ± 0.010	−0.349 ± 0.039
AMA2	0.199 ± 0.018	0.129 ± 0.050	−0.157 ± 0.001	0.693 ± 0.057	0.216 ± 0.101	−0.057 ± 0.064
AMA3	0.508 ± 0.096	−0.019 ± 0.188	0.104 ± 0.190	0.634 ± 0.167	−0.041 ± 0.086	−0.123 ± 0.065
AMA4	0.354 ± 0.061	0.393 ± 0.063	−0.321 ± 0.065	0.593 ± 0.046	0.008 ± 0.032	−0.344 ± 0.034

Key observations:

- AMA1 → Lydia Ko (0.742 ± 0.043)
- AMA2 → Rory McIlroy (0.693 ± 0.057)
- AMA3 → Rory McIlroy (0.634 ± 0.167)
- AMA4 → Rory McIlroy (0.593 ± 0.046)

Notably, different amateur groups align with different professional styles, demonstrating that the embedding space captures diverse stylistic tendencies rather than collapsing into a single dominant prototype.

Furthermore, negative similarity values (e.g., AMA1 vs. Lexi Thompson) indicate clear stylistic divergence, reinforcing the discriminative nature of the learned embedding.

Although the prototype mapping experiment demonstrates the feasibility of embedding-based amateur-to-professional style retrieval, several limitations remain. First, the current evaluation primarily measures embedding consistency with respect to motion labels rather than perceptual similarity judged by human experts. Second, because only four amateur groups were included, the present analysis should be interpreted as an initial feasibility study rather than statistically conclusive evidence for practical coaching or rehabilitation deployment. Future work will therefore incorporate expert-based human evaluation and larger-scale motion datasets to further validate practical applicability.

8.6 Sensitivity Analysis

To evaluate the robustness of the proposed framework with respect to key hyperparameters, we conducted a sensitivity analysis on the temperature parameter τ in the NT-Xent loss, the margin parameter m in the semi-hard triplet loss, and the loss weighting parameters λ_1 and λ_2.

First, the temperature τ was varied within $\{0.05, 0.07, 0.10\}$. A lower temperature sharpens the contrastive distribution and emphasizes hard negatives, whereas a higher temperature produces a smoother similarity distribution. The results showed that $\tau = 0.07$ provided the best overall retrieval performance, while neighboring values caused only minor performance variations. This indicates that the proposed embedding space is not overly sensitive to moderate temperature changes.

Second, the triplet margin m was varied within $\{0.1, 0.2, 0.3\}$. When the margin was too small, inter-class separation became weaker because negative samples were not pushed sufficiently far from anchors. When the margin was too large, training became less stable because excessively strict separation constraints increased the number of difficult triplets. The setting $m = 0.2$ achieved the best trade-off between stable convergence and discriminative separation.

Third, we evaluated the effect of the loss weighting parameters λ_1 and λ_2. The results showed that using both NT-Xent and triplet loss produced more stable retrieval performance than using either objective alone. NT-Xent contributes to global batch-level separation, while the triplet loss reinforces local margin-based discrimination among stylistically similar motions. This complementary behavior explains why the hybrid objective improves retrieval consistency.

Overall, the sensitivity analysis confirms that the proposed framework maintains stable performance within a reasonable range of hyperparameter settings. This stability is further supported by the L2-normalized hyperspherical embedding space, which constrains representation magnitude and enables consistent cosine-based similarity comparison.

These results indicate that the proposed framework remains stable across reasonable hybrid loss weighting configurations and does not require excessively delicate hyperparameter tuning.

Table 6 shows that the proposed framework maintains stable performance across a reasonable range of hyperparameter settings. The variation in Top-1 accuracy remains within approximately ±1.5%, indicating that the embedding space is robust to moderate changes in temperature and margin values. The hybrid loss consistently outperforms single-objective configurations, demonstrating the complementary roles of NT-Xent and triplet loss. Excessive reliance on either NT-Xent or triplet loss slightly reduces retrieval consistency, whereas balanced weighting provides the most stable embedding structure. These results indicate that the proposed framework does not require excessively delicate tuning and remains stable across reasonable hybrid loss weighting configurations.

9 Discussion

The experimental results consistently validate the central hypothesis of this work: robust motion similarity retrieval requires explicit invariance modeling prior to metric learning. The ablation study demonstrates that structured spatial-temporal normalization—particularly static pelvis alignment, posture-axis alignment, and phase-synchronized resampling—substantially improves retrieval accuracy across both convolutional and recurrent backbones. These findings confirm that removing geometric and temporal nuisance factors before embedding is not merely a preprocessing step, but a fundamental prerequisite for stable metric consistency.

Table 6: Sensitivity analysis of key hyperparameters (Top-1 Accuracy, %).

Category	Parameter	Setting	Top-1 Accuracy (%)
Temperature (τ)	τ	0.05	90.2
		0.07	91.3
		0.10	90.7
Triplet Margin (m)	m	0.10	89.8
		0.20	91.3
		0.30	90.1
Loss Configuration	Objective	NT-Xent only	88.9
		Triplet only	87.5
		Hybrid	91.3

The superiority of the A-TCN backbone further highlights the importance of long-range temporal modeling. Dilated acausal convolutions effectively capture global swing dynamics, leading to compact intra-style clustering and clear inter-style separation. In contrast, recurrent modeling exhibits weaker structural separability, suggesting that convolutional receptive field expansion is better suited for holistic motion encoding in retrieval scenarios.

Importantly, unlike action recognition tasks that focus on categorical discrimination, similarity retrieval demands fine-grained metric structure preservation. The observed block-diagonal similarity matrix, optimal silhouette score alignment with the true number of players, and stable prototype-based amateur mapping collectively indicate that the proposed framework learns a geometrically organized and semantically meaningful embedding manifold.

The prototype representation introduces an additional layer of interpretability by translating abstract embedding similarity into direct stylistic correspondence. This mechanism bridges representation learning and practical coaching applications, enabling explainable amateur-to-professional style alignment.

Although evaluated on golf swings, several normalization principles of the proposed framework may be transferable to other human motion domains. The separation between structured canonicalization and contrastive embedding offers a modular and extensible design paradigm applicable to broader domains such as rehabilitation assessment, sports analytics, and human–robot interaction.

Although PSR is effective for phase-structured motions such as golf swings, its applicability to highly stochastic motions without stable semantic anchors remains limited and requires alternative alignment strategies.

Overall, the results demonstrate that integrating geometric canonicalization with contrastive temporal modeling establishes a principled and scalable framework for invariant motion similarity retrieval.

While the proposed framework is evaluated on golf swing data, its underlying design is not restricted to golf-specific motion. The spatial normalization components—torso-scale normalization, pelvis-centered alignment, and posture-axis alignment—address general nuisance factors commonly observed in monocular 2D pose sequences, including scale variation, translation drift, and in-plane rotation. Therefore, the framework can be interpreted as a modular invariant motion embedding pipeline rather than a domain-specific golf swing model.

For non-cyclical or highly stochastic motions, the phase-synchronized resampling module may not be directly applicable when reliable semantic phase anchors are unavailable. In such cases, PSR can be replaced

with event-free temporal normalization strategies, including uniform interpolation, fixed-length resampling, dynamic temporal scaling, or learned temporal alignment. Since the A-TCN encoder only requires a fixed-length normalized sequence, the overall architecture can maintain retrieval consistency without relying on manually defined phase events.

Despite the strong retrieval performance achieved on the golf swing benchmark, several limitations remain. First, the current framework was evaluated only on phase-structured golf motions, and additional validation on broader motion domains such as gait analysis, rehabilitation exercise, and daily living activities is required. Second, the use of monocular 2D pose sequences introduces inherent limitations related to depth ambiguity, self-occlusion, and out-of-plane body rotation. Third, the current retrieval evaluation focuses primarily on embedding consistency rather than perceptual similarity assessed by human experts. Finally, although the proposed framework demonstrates stable performance across reasonable hyperparameter ranges, future comparison with graph-based and transformer-based retrieval architectures such as ST-GCN remains an important direction for further validation.

Future work will extend the proposed framework toward broader motion domains beyond golf swings, including rehabilitation and daily activity analysis. Additional directions include expert-based perceptual similarity evaluation, comparison with graph-based and transformer-based retrieval architectures, and extension to 3D or multi-view pose representations for improved geometric robustness.

10 Conclusion

This paper presented a spatial-temporal normalized contrastive embedding framework for robust motion similarity retrieval from monocular 2D pose sequences. By explicitly separating nuisance-factor removal from representation learning, the proposed approach integrates structured geometric canonicalization with acausal temporal convolutional modeling and hybrid contrastive optimization within a unified pipeline.

The four-stage normalization module effectively suppresses scale, translation, rotation, and tempo variations prior to embedding, while the A-TCN backbone captures long-range temporal dynamics through dilated convolutions. Experimental results demonstrate significant improvements over BiLSTM and DTW baselines, achieving 91.3% Top-1 accuracy and producing compact, well-separated, and semantically organized embedding clusters. The prototype-based retrieval mechanism further enhances interpretability by enabling direct amateur-to-professional style mapping.

Beyond golf swing analysis, the proposed framework establishes a general paradigm for invariant motion embedding and scalable similarity retrieval. By combining explicit canonicalization with contrastive metric learning, this work provides a principled foundation for future motion analysis systems.

Future work will further validate the proposed framework on diverse human motion domains beyond golf swings, incorporate expert-based perceptual evaluation, compare retrieval performance with graph-based architectures such as ST-GCN, and extend the framework to 3D and multi-view pose representations.

Acknowledgement: Not applicable.

Funding Statement: This research was supported by Incheon National University Research Grant (2020).

Author Contributions: Conceptualization, Seung-su Lee and Kwang-il Hwang; methodology, Seung-su Lee; software, Seung-su Lee; validation, Seung-su Lee, Young-Been Noh and HwaYoung Jeong; formal analysis, Seung-su Lee; investigation, Seung-su Lee; resources, Kwang-il Hwang; data curation, Seung-su Lee and Young-Been Noh; writing—original draft preparation, Seung-su Lee; writing—review and editing, Seung-su Lee and Kwang-il Hwang; visualization,

Seung-su Lee; supervision, Kwang-il Hwang; project administration, Kwang-il Hwang; funding acquisition, Kwang-il Hwang. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The datasets generated and analyzed during the current study are available from the corresponding author [Kwang-il Hwang] upon reasonable request. The implementation code used for training and evaluation is also available upon request.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

Abbreviation	Full Term
A-TCN	Acausal Temporal Convolutional Network
TCN	Temporal Convolutional Network
NT-Xent	Normalized Temperature-Scaled Cross-Entropy
DTW	Dynamic Time Warping
BiLSTM	Bidirectional Long Short-Term Memory
IMU	Inertial Measurement Unit
UMAP	Uniform Manifold Approximation and Projection
t-SNE	t-Distributed Stochastic Neighbor Embedding
RF	Receptive Field
FPS	Frames Per Second
SPAR	Static Pelvis Alignment Recentering
DPAR	Dynamic Pelvis Alignment Recentering
PAA	Posture-Axis Alignment
PSR	Phase-Synchronized Resampling
MRR	Mean Reciprocal Rank
mAP	Mean Average Precision
EMA	Exponential Moving Average

References

1. Kim M, Park S. Golf swing segmentation from a single IMU using machine learning. Sensors. 2020;20(16):4466. doi:10.3390/s20164466.
2. Okuda I, Gribble P, Armstrong C. Trunk rotation and weight transfer patterns between skilled and low skilled golfers. J Sports Sci Med. 2010;9(1):127–33.
3. Myers J, Lephart S, Tsai YS, Sell T, Smoliga J, Jolly J. The role of upper torso and pelvis rotation in driving performance during the golf swing. J Phys Sci. 2008;26(2):181–8. doi:10.1080/02640410701373543.
4. Callaway S, Glaws K, Mitchell M, Scerbo H, Voight M, Sells P. An analysis of peak pelvis rotation speed, gluteus *Maximus* and medius strength in high versus low handicap golfers during the golf swing. Int J Sports Phys Ther. 2012;7(3):288–95. doi:10.2519/jospt.2003.33.4.196.
5. Cao Z, Simon T, Wei SE, Sheikh Y. Realtime multi-person 2D pose estimation using part affinity fields. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017 Jul 21–26; Honolulu, HI, USA. p. 1302–10. doi:10.1109/cvpr.2017.143.
6. Bazarevsky V, Grishchenko I, Raveendran K, Zhu T, Zhang F, Grundmann M. BlazePose: on-device real-time body pose tracking. arXiv:2006.10204. 2020.
7. Ju CY, Kim JH, Lee DH. GolfMate: enhanced golf swing analysis tool through pose refinement network and explainable golf swing embedding for self-training. Appl Sci. 2023;13(20):11227. doi:10.3390/app132011227.

8. Dwibedi D, Aytar Y, Tompson J, Sermanet P, Zisserman A. Temporal cycle-consistency learning. In: Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2019 Jun 15–20; Long Beach, CA, USA. p. 1801–10. doi:10.1109/cvpr.2019.00190.
9. Haresh S, Kumar S, Coskun H, Syed SN, Konin A, Zia Z, et al. Learning by aligning videos in time. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2021 Jun 20–25; Nashville, TN, USA. p. 5548–58. doi:10.1109/CVPR46437.2021.00550.
10. Gao X, Yang Y, Zhang Y, Li M, Yu JG, Du S. Efficient spatio-temporal contrastive learning for skeleton-based 3-D action recognition. IEEE Trans Multimed. 2023;25:405–17. doi:10.1109/TMM.2021.3127040.
11. Wang P, Wen J, Si C, Qian Y, Wang L. Contrast-reconstruction representation learning for self-supervised skeleton-based action recognition. IEEE Trans Image Process. 2022;31(11):6224–38. doi:10.1109/TIP.2022.3207577.
12. Zhang Y, Nie L. Human motion similarity evaluation based on deep metric learning. Sci Rep. 2024;14(1):30908. doi:10.1038/s41598-024-81762-8.
13. Lea C, Flynn MD, Vidal R, Reiter A, Hager GD. Temporal convolutional networks for action segmentation and detection. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017 Jul 21–26; Honolulu, HI, USA. p. 1003–12. doi:10.1109/cvpr.2017.113.
14. Bai S, Kolter JZ, Koltun V. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. In: Advances in Neural Information Processing Systems; 2018 Dec 3–8; Montréal, QC, Canada.
15. Chen T, Kornblith S, Norouzi M, Hinton G. A simple framework for contrastive learning of visual representations. In: Proceedings of the 37th International Conference on Machine Learning; 2020 Jul 13–18; Virtual. p. 1597–607. doi:10.48550/arxiv.2002.05709.
16. Schroff F, Kalenichenko D, Philbin J. FaceNet: a unified embedding for face recognition and clustering. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2015 Jun 7–12; Boston, MA, USA. p. 815–23. doi:10.1109/cvpr.2015.7298682.
17. Lee SS, Choi JH, Byun J, Hwang KI. Dynamic golf swing analysis framework based on efficient similarity assessment. Sensors. 2025;25(22):7073. doi:10.3390/s25227073.
18. Lee DJ, Noh YB, Byun J, Hwang KI. Reinforcement learning-based golf swing correction framework incorporating temporal rhythm and kinematic stability. Sensors. 2026;26(2):392. doi:10.3390/s26020392.
19. Debnath B, O'Brien M, Yamaguchi M, Behera A. A review of computer vision-based approaches for physical rehabilitation and assessment. Multimed Syst. 2022;28(1):209–39. doi:10.1007/s00530-021-00815-4.
20. Tang J, Chen J, Su Y, Xing M, Zhu S. MTAN: multi-degree tail-aware attention network for human motion prediction. Internet Things. 2024;25(6):101134. doi:10.1016/j.iot.2024.101134.
21. Zhu S, Chen J, Su Y. Spatio-temporal articulation & coordination co-attention graph network for human motion prediction. Signal Process. 2024;223(1):109551. doi:10.1016/j.sigpro.2024.109551.