Journal of Computer and Knowledge Engineering, Vol. 9, No. 1, 2026. (93-110)

Ferdowsi University of Mashhad

Journal of Computer and Knowledge Engineering

https://cke.um.ac.ir

Information and Communication Technology Association of Iran

Context-aware Action Quality Assessment using Latent Regression-based Progressive Sub-action Learning *

Research Article

Marjan Mazruei[1], Ehsan Fazl-Ersi[2], Abedin Vahedian[3], Ahad Harati[4]

10.22067/cke.2026.96007.1177

Abstract *Most existing Action Quality Assessment (AQA) systems infer performance scores from holistic, video-level representations, a process that frequently obscures the contribution of individual motion segments to the final prediction. This global treatment reduces transparency and limits interpretability, particularly for complex skills composed of multiple sequential phases. The problem is further compounded by limited fine-grained supervision and the difficulty of modeling long-range temporal dependencies across an action sequence. This paper introduces a Latent Regression-based Progressive Sub-action Learning framework to address these limitations by capturing rich contextual information within the action sequence. The framework first employs a procedure-parsing module to segment each video into semantically coherent sub-actions and extract corresponding spatio-temporal features. A latent variable approach subsequently generates initial pseudo-subscores by integrating these features with overall score labels from the training set. Critically, these pseudo-subscores are progressively refined through an iterative process that leverages the cumulative contextual information from preceding sub-actions. This yields an enriched feature set that accurately encapsulates the execution history. Furthermore, a novel monotonic penalty loss is introduced to enforce a logical and consistent progression in the sub-scores, mitigating abrupt and illogical score fluctuations. The model is trained in a two-stage process. First, it generates the robust pseudo-subscores, and subsequently it predicts both the overall score and the per-substage scores. This explicit modeling of long-range dependencies, along with consistent score progression, is critical for ensuring prediction accuracy. Extensive experimental evaluation confirms that this structured modeling approach delivers consistent performance gains over contemporary AQA methods while offering substantially improved interpretability.*

Key Words Action quality assessment, Long-range dependency, Progressive learning, Pseudo-subscore learning, Regression-based AQA, Spatio-temporal features, Temporal semantic segmentation.

1- INTRODUCTION

The automated evaluation of human actions from video has become a pivotal area of research in computer vision, with applications spanning professional sports analysis, robotics, and healthcare. Within this domain, AQA aims to quantitatively evaluate the proficiency of a performance by assigning a score that reflects its overall quality. In disciplines like competitive diving or gymnastics, an overall score is a composite evaluation based on the quality of several distinct sub-actions. For a system to provide truly comprehensive and actionable feedback, it must deliver not only a final score but also granular, substage-specific insights to enable targeted performance improvement. A fundamental challenge in developing such systems is the pervasive scarcity of fine-grained, per-substage quality scores in publicly available datasets. The annotation of each individual sub-action's quality demands specialized expertise, making large-scale data collection and labeling prohibitively expensive. This lack of detailed labels has constrained most AQA research to a simplistic regression approach that maps an entire video to a single quality score. While effective for predicting a final score, this holistic method fails to provide the detailed insights necessary for effective coaching and athlete development.

This paper addresses the problem of substage-level AQA by introducing a novel framework that capitalizes on

* Manuscript received October 16, 2025, Revised November 01, 2025, Accepted February 17, 2026.

[1] Phd Candidate, Department of Computer Engineering, Ferdowsi University of Mashhad, Mashhad, Iran.

[2] Corresponding Author. Assistant Professor, Department of Computer Engineering, Faculty of Engineering, Ferdowsi University of Mashhad, Mashhad, Iran.

Email: ehsan.fazl@gmail.com

[3] Associate Professor, Department of Computer Engineering, Faculty of Engineering, Ferdowsi University of Mashhad, Mashhad, Iran.

[4] Associate Professor, Department of Computer Engineering, Faculty of Engineering, Ferdowsi University of Mashhad, Mashhad, Iran.

existing temporal segmentation annotations to perform a sophisticated, fine-grained analysis. The proposed framework employs a supervised temporal segmentation model to automatically partition videos into semantically meaningful sub-actions. This segmentation provides the structural foundation for the core innovation of this work, which is a Latent Regression-based Progressive Sub-action Learning framework that rigorously models contextual dependencies. Spatio-temporal features are extracted from the segmented sub-actions, and a latent variable approach generates initial pseudo-subscores by integrating these features with the overall score labels from the training set. A key novelty is the progressive refinement of these pseudo-subscores through an iterative augmentation process that accurately encapsulates the cumulative contextual influence of all preceding sub-actions. Furthermore, a monotonic penalty loss is introduced to enforce a logical and consistent progression in the sub-score sequence, thereby enhancing the interpretability and robustness of the model's predictions. The entire framework is trained in a two-stage process: first to generate consistent pseudo-subscores, and then fine-tuned to predict both the overall and per-substage scores. This explicit modeling of long-range dependencies is essential for accurate assessment, ensuring that the prediction for any given stage accounts for the entire preceding execution history.

In summary, the principal contributions of this research include: (1) A Latent Regression-based Progressive Sub-action Learning framework is introduced to generate pseudo-subscores, which are progressively refined by explicitly modeling the cumulative contextual influence of all preceding sub-actions. (2) A monotonic penalty loss is proposed to enforce temporal consistency, promoting logical progression in the predicted sub-scores. (3) The framework achieves superior performance on the UNLV-Diving [1] and FineDiving [2] datasets, demonstrating the efficacy of this long-range dependency modeling approach. (4) Detailed ablation studies are conducted to demonstrate the effectiveness and synergistic contribution of key components in the proposed framework, while qualitative analyses confirm that the generated pseudo-subscores serve as meaningful indicators of sub-action quality, providing actionable insights for performance enhancement.

The rest of this paper is structured as follows. Section 2 reviews related work, followed by a detailed description of the proposed framework and its components in Section 3. Section 4 presents the experimental results and analysis, and Section 5 concludes the paper with final remarks and directions for future work.

2- RELATED WORK

The primary objective of AQA is to automatically assess how well a human action is performed based on visual cues. Unlike action recognition, which focuses solely on identifying the type of action, AQA requires a nuanced understanding of subtle variations in execution, such as precision, timing, and motion dynamics. These subtleties make AQA considerably more complex, often leading to intra-clip confusion and inter-clip inconsistency, especially when uniform temporal segmentation obscures local variations in quality [3]. This section reviews the literature across core methodologies and recent advancements.

A. Early Regression-Based AQA

The earliest and most common approach to AQA involved **regression-based scoring** frameworks, where models predicted a continuous quality measure from handcrafted or pose-based features. Studies like [4] combined spatio-temporal pose features with the Discrete Cosine Transform (DCT) and Linear Support Vector Regression (L-SVR) for prediction. Later, a sequence-to-sequence autoencoder model [5] was introduced to learn expert-level motion representations. Lei et al. [6] enhanced this by integrating action classification with regression using 2D skeleton data from RGB videos. Their model incorporated self-similarity descriptors and joint displacement sequences to improve temporal coherence. Li et al. [7] extended this line of research using Recurrent Neural Networks (RNNs) for deep pose-based temporal modeling. To incorporate vital visual context and overcome the limitations of pose-driven methods, Parmar and Morris [1] incorporated 3D Convolutional Neural Networks (C3D) [8] and Long Short-Term Memory (LSTM) networks [9], leading to significant performance gains. Subsequent research leveraged C3D [8] with combined ranking and Mean Squared Error (MSE) losses [10], explored knowledge transfer across multiple actions [11], and utilized multitask frameworks combining recognition, commentary generation and scoring [12].

B. Fine-Grained and Procedural Modeling AQA

A growing body of research has since emphasized **fine-grained** frameworks, decomposing complex actions into semantically meaningful sub-actions [13] - [16]. These methods employ explicit temporal boundaries to enable granular assessment, consistently outperforming approaches that rely on uniform temporal pooling. Examples include the S3D model [13], which used a supervised temporal convolutional network [17] to segment dives and fuse segment-specific features, and the method by [15], which learned and fused multiple hidden substages to generate detailed substage scores. A key challenge within this category is the rigorous modeling of contextual dependencies between sub-actions. The Label-Reconstruction-based Pseudo-Subscore Learning (PSL) framework [16], a notable example, generates pseudo-subscores but treats these variables independently, lacking a sophisticated mechanism for capturing sequential influences. Our work builds directly upon these supervised, procedural modeling methodologies, with a key focus on enhancing the modeling of long-range dependencies between sub-actions. Uncertainty modeling [18] [19] has also been employed to enhance scoring robustness by addressing ambiguities in human-assigned labels. Recently, self-supervised sub-action parsing [20] was explored, aiming to enhance the granularity of motion understanding.

C. Transformer-Based and Contrastive-Learning AQA

More recently, the field has been dominated by methods leveraging advanced architectures to capture deeper temporal and structural relationships. **Temporal attention mechanisms** [21] - [23] improved AQA accuracy by

adaptively weighting segments. More structurally, Zeng et al. [24] designed a dual-stream architecture that fuses motion and pose features through Graph Convolutional Networks (GCNs) to model inter-segment relations. Zhang et al. [25] utilized time-aware attention to enhance temporal sensitivity, while Xu et al. [26] introduced the grade-decoupling Likert Transformer with cross-attention to extract grade-sensitive representations. Zahan and Mubashar Hassan [27] proposed a discriminative non-local attention module that enhanced the integration of skeletal and visual features in long sequences. Other works concentrated on targeted attention and structural modeling, including TSA-Net [28], which leveraged SiamMask for athlete-specific focus, and Huang et al. [29], who proposed a dual-referenced assistive network that leveraged multiple reference samples to guide feature learning for more precise modeling of action performance. Li et al. [30] developed a Gaussian-guided frame sequence encoder to better capture temporal variations. Structural modeling has also utilized Hierarchical GCNs [3], multi-scale location attention with LSTMs [31], and group-aware GCNs [32]. Zhou et al. [33] investigated continual learning in AQA and proposed the MAGR framework, a manifold-aligned graph regularization method designed to preserve feature relationships during sequential learning, thus mitigating catastrophic forgetting.

Furthermore, modern contrastive learning frameworks have emerged as highly effective techniques. Pairwise ranking methods assess the relative quality between samples [34] - [36]. Yu et al. [37] introduced a group-aware contrastive regression framework, combining pairwise comparison with group-based regression trees, while An et al. [38] designed a multi-stage contrastive regression method that refined predictions through progressive representation contrast. In [39], contrastive learning was integrated with regression through consistency regularization. Ke et al. [40] developed a two-path target-aware contrastive regression framework, integrating dual-path feature extraction to enhance scoring precision. A multi-stage model [41] further improved score prediction by emphasizing sub-action importance. Procedure-aware contrastive learning [2] [42] further emphasizes stage-wise consistency using segmentation-based supervision. A coarse-to-fine instruction alignment model was presented in [43], incorporating hierarchical guidance for detailed evaluation.

Overall, prior AQA approaches have shown the importance of temporal segmentation, attention mechanisms, and contrastive objectives. However, most existing methods often rely on simple feature aggregation or local dependencies, failing to capture the cumulative contextual influence of earlier sub-actions on later ones. The proposed framework addresses this limitation through supervised long-range dependency modeling, yielding temporally consistent and semantically interpretable score progression.

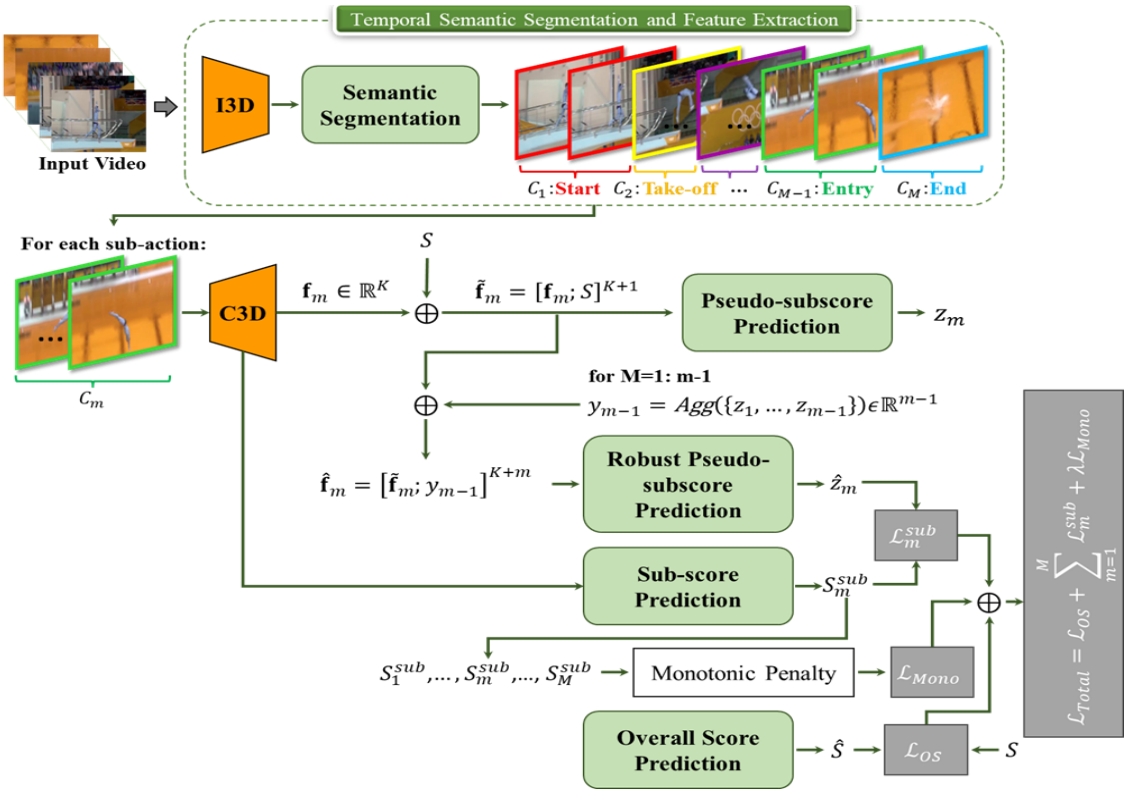

Figure 1. The architecture of the proposed Latent Regression-based Progressive Sub-action Learning framework for fine-grained AQA. The framework comprises three main stages: (1) a temporal semantic segmentation and feature extraction module, which partitions the input video into semantically coherent sub-actions and encodes their spatio-temporal features; (2) a latent regression module for progressive sub-score learning, which progressively predicts pseudo-subscores for each sub-action by integrating the overall score label with extracted features, while iteratively refining sub-action representations through contextual dependencies among preceding pseudo-subscores; and (3) a multi-substage AQA module, which aggregates the learned sub-scores to produce a comprehensive assessment of the athlete's overall performance.

3- PROPOSED METHOD

The proposed framework introduces an advanced approach for fine-grained AQA in sports videos, formulated as a regression-based problem. The design incorporates domain-specific knowledge in temporal action segmentation to establish a structured basis for precise quality evaluation. At its core lies a **Latent Regression-based Progressive Sub-action Learning** mechanism, developed to explicitly model and exploit long-range temporal dependencies across sequential sub-actions. Through a progressive learning strategy, the model incrementally integrates contextual information from preceding sub-actions to refine the assessment of subsequent ones. This hierarchical progression enables the estimation of granular sub-scores for individual sub-actions, which are subsequently aggregated to produce a comprehensive quality score representing the athlete's overall performance. As illustrated in Figure 1, the architecture consists of three main components. These include a temporal semantic segmentation and feature extraction module, a latent regression module for progressive sub-score learning, and a multi-substage AQA module. This integrated design ensures a coherent representation of both local and global temporal dynamics, providing a robust foundation for precise action quality estimation.

A. Temporal Semantic Segmentation and Feature Extraction

The first stage of the proposed framework focuses on decomposing each video into a sequence of semantically coherent sub-action segments, an essential step for achieving fine-grained performance evaluation. This process is performed through a supervised temporal semantic segmentation module designed to identify procedure-aware temporal boundaries using ground-truth annotations. Building upon established methodologies in fine-grained AQA [2], the module predicts a predefined number of transition points, denoted as *L*, which indicate the frames where the action transitions from one sub-action to the next.

Given the extracted video features, $F_{feat}(V_i)$, the segmentation component, $\mathcal{S}$, computes the probability of a step transition occurring at each frame *t* as follows:

$$[\hat{p}_1, \dots, \hat{p}_L] = \mathcal{S}\left(F_{feat}(V_i)\right) \tag{1}$$

Here, $\mathcal{S}$ outputs *L* probability distributions, $\{\hat{p}_k\}_{k=1}^{L}$, each corresponding to a potential step transition. The estimated transition time, $\hat{t}_k$, for each step is determined by selecting the frame with the highest transition probability, subject to a temporal ordering constraint that ensures consistent progression across sub-actions:

$$\hat{t}_k = \underset{(k-1)<t\leq k}{\arg\max}\ \hat{p}_k(t) \tag{2}$$

To ensure accurate temporal localization, the segmentation module is trained using a Binary Cross-Entropy (BCE) loss between the predicted probability distribution, $\hat{p}_k$, and the ground-truth binary distribution, p_k, defined as:

$$\mathcal{L}_{BCE}(\hat{p}_k, p_k) = -\sum_t p_k(t)\log\hat{p}_k(t) + \big(1-p_k(t)\big)\log\big(1-\hat{p}_k(t)\big) \tag{3}$$

Minimizing this loss function ensures that the predicted probability distribution accurately aligns with the ground truth transition frames, guiding the model to correctly parse the action into its constituent steps.

Once the video V_i is segmented, it is partitioned into *M* sub-action clips, $C = \{C_1, C_2, \dots, C_M\}$, based on the predicted transition times. Each video clip $C_m \in \mathbb{R}^{T_m \times H \times W \times 3}$ is a sequence of RGB frames with spatial resolution $H \times W$ representing a distinct sub-action. Due to inherent variations in human motion and action composition, the segment lengths T_m naturally differ among sub-action clips. Subsequently, the spatio-temporal features for each of these segments are extracted using a deep 3D convolutional network. A C3D backbone [8] is employed owing to its proven balance between computational efficiency and representational strength. The feature extraction process is formulated as:

$$\mathbf{f}_m = F_{DL}(C_m), \quad \mathbf{f}_m \in \mathbb{R}^K \tag{4}$$

where $F_{DL}(.)$ denotes the deep feature extractor that produces a *K*-dimensional feature vector $\mathbf{f}_m$ for each sub-action clip. The resulting sequence of feature representations, , $F = \{\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_M\}$, encapsulates semantically meaningful information across sub-actions, providing a robust spatio-temporal foundation for subsequent sub-score regression and progressive learning stages.

B. Latent Regression for Progressive Sub-score Learning

The core of the proposed framework is the Progressive Pseudo-subscore Learning mechanism, designed to explicitly model the cumulative temporal influence across the entire action sequence. In the absence of fine-grained sub-score annotations, this mechanism introduces sub-scores as latent intermediate variables that are iteratively refined to represent the temporal progression of execution quality. This formulation effectively captures long-range temporal dependencies, ensuring that the evaluation of any given sub-action reflects both its local performance characteristics and its relationship to the overall execution trajectory.

I. Pseudo-subscore Learning

For each sub-action segment C_m, an initial pseudo-subscore z_m, is generated by exploiting the coarse-grained overall score *S* provided in the training data. The spatio-temporal feature vector $\mathbf{f}_m$ is concatenated with the overall score *S* and subsequently fed into a latent regression network, denoted as, F_{latent}. This network is designed to infer a latent quality representation that approximates the relative contribution of each sub-action to the overall performance score. The process is formally expressed as:

$$\tilde{\mathbf{f}}_m = [\mathbf{f}_m; S] \\ z_m = F_{latent}(\tilde{\mathbf{f}}_m) \quad (5)$$

where [;] denotes the concatenation operation, The concatenation operation results in an enriched feature set $\{\tilde{\mathbf{f}}_1, \dots, \tilde{\mathbf{f}}_M\}$, where $\tilde{\mathbf{f}}_m$ has a dimension of K+1 and z_m is the initial pseudo-subscore embedding. By anchoring the initial representation to the ground-truth overall score, this step provides a strong starting point for the quality assessment. The latent regression network F_{latent}, is implemented as a fully connected network (FCN) with five layers. The initial layer of this network takes the label-augmented features as input, and the number of nodes decreases progressively in each layer until the output layer has a single node. This final layer uses a Sigmoid activation function to produce the predicted pseudo-subscore, which can be expressed as:

$$\tilde{\mathbf{f}}_m^t = FC(W^t, \tilde{\mathbf{f}}_m^{(t-1)}) \\ z_m = Sigmoid(\mathbf{f}'_m) \quad (6)$$

Here, $FC(.)$ represents the fully connected operation, $\mathbf{f}'_m$ is the output of the final layer, and W^t denotes the parameters of the *t*-th layer and z_m denotes the predicted pseudo-subscore for the *m*-th sub-action. This pseudo-subscore generator is replicated across all sub-action to independently estimate the pseudo-subscores. Once the pseudo-subscores for all sub-actions are generated, they are fed into fully connected network to predict the overall score for the video. This process is mathematically expressed as:

$$\hat{S} = sigmoid(\sum_{m=1}^{M}(w_m \times z_m)), \hat{S} \in \mathbb{R} \quad (7)$$

where $\hat{S}$ is the predicted overall score and w_m refers to the weight parameters of the fully connected layer. The training of this phase is optimized using a MSE loss, which quantifies the deviation between the predicted and ground-truth overall scores. The objective function is formulated as:

$$\mathcal{L}_\alpha = \frac{1}{N_{Tr}}\sum_{i=1}^{N_{Tr}}(\hat{S}_i - S_i)^2 \quad (8)$$

Here, N_{Tr} is the total number of training videos, $\hat{S}_i$ represents the predicted overall score for the *i*-th training sample, and S_i corresponds to its ground-truth overall score.

Upon completion of the training phase, the optimized model is employed to generate an initial set of pseudo-subscores for all training videos. These pseudo-subscores serve as latent quality indicators, reflecting the preliminary performance assessment of each sub-action segment. This stage ensures that the model effectively exploits the available supervision to produce informative pseudo-subscores, which provide a reliable foundation for the subsequent progressive refinement process.

II. Long-Range Contextual Refinement:

While embedding the overall score label into sub-action features provides valuable global context, it may constrain the model's capacity to capture fine-grained, localized variations within the sequence. To overcome this limitation and enhance temporal reasoning capability, it is essential to explicitly model dependencies across sub-actions. The quality of each sub-action is inherently influenced by the cumulative dynamics of all preceding segments. For instance, in a diving sequence, a subtle error in body posture during an early stage can propagate through subsequent movements, ultimately magnifying its impact on the final splash. This underscores the need to model the interdependencies among all preceding sub-actions for a comprehensive understanding of action execution. To capture this long-range dependency, the initial pseudo-subscore embeddings are progressively refined. For any sub-action *m*, the quality of its execution is profoundly influenced by the cumulative context of all preceding sub-actions, from 1 to *m*-1. This historical context is aggregated and then used to augment the feature vector of the current sub-action. The cumulative contextual history, y_{m-1}, is calculated by aggregating the embeddings of all preceding segments:

$$y_{m-1} = \text{Aggregate}(\{z_1, \dots, z_{m-1}\}) \quad (9)$$

This aggregated context vector, y_{m-1}, is then concatenated with the current spatio-temporal feature, $\mathbf{f}_m$, to form a progressively augmented feature vector, $\hat{\mathbf{f}}_m$:

$$\hat{\mathbf{f}}_m = [\mathbf{f}_m; y_{m-1}] \quad (10)$$

This progressively augmented feature vector is subsequently used in the training process to produce the robust pseudo-subscore for the current segment, ensuring that the prediction is conditioned on the complete execution history up to that point. At each sub-action, features are combined with the overall score label and pseudo-subscores to ensure continuous information flow across the sub-actions. The model training for this phase iteratively updates its parameters to map these enhanced feature vectors to their corresponding robust pseudo-subscores ($\hat{z}_m$), which are produced using the following process:

$$\hat{\mathbf{f}}_m^t = FC(W^t, \hat{\mathbf{f}}_m^{(t-1)}) \\ \hat{z}_m = Sigmoid(\mathbf{f}'_m) \quad (11)$$

The robust pseudo-subscores are then aggregated to predict the overall score, $\hat{S}'$:

$$\hat{S}' = sigmoid(\sum_{m=1}^{M}(\hat{w}_m \times \hat{z}_m)), \hat{S}' \in \mathbb{R} \quad (12)$$

In this subsequent phase, the optimization is also governed by a MSE criterion; however, it is applied to the refined score predictions. The corresponding error function is expressed as:

$$\mathcal{L}_\beta = \frac{1}{N_{Tr}}\sum_{i=1}^{N_{Tr}}(\hat{S}'_i - S_i)^2 \quad (13)$$

This formulation guides the model to more closely align the progressively refined predicted scores $\hat{S}'_i$ with the

ground-truth overall quality labels, promoting precise performance estimation through iterative refinement. By repeatedly integrating information from all preceding stages, the model incrementally enhances its predictions, capturing the cumulative context of the entire action sequence. Upon completion of training, the model generates robust pseudo-subscores for each training video, denoted as $\hat{Z} = \{\hat{z}_1, \hat{z}_2, \dots, \hat{z}_M\}$, which serve as reliable latent indicators of sub-action quality for subsequent processing.

C. Multi-substage AQA

The final phase of the framework focuses on fine-tuning the multi-substage AQA model using the robust pseudo-subscore labels. In this stage, the overall score label is excluded from the input feature vector. The output of the feature extraction backbone for each sub-action segment is fed into a fully connected network comprising five layers, which progressively transforms the input feature vector to generate predicted sub-scores. The initial layer of this network takes the *K*-dimensional feature vector as input and it employs a decreasing number of nodes in each layer, gradually dropping to one. This is a crucial step in preparing the model for inference, as it ensures the model relies solely on the video content to make predictions. Mathematically, this process is expressed as:

$$\begin{aligned} \mathbf{f}_m^t &= FC\left(W^t, \mathbf{f}_m^{(t-1)}\right) \\ S_m^{sub} &= sigmoid(\mathbf{f}'_m) \\ \hat{S}_i &= sigmoid\left(\sum_{m=1}^{M} w_m \times S_m^{sub}\right) \end{aligned} \tag{14}$$

Here, S_m^{sub} denotes the predicted sub-score for the *m*-th sub-action.

D. Optimization and Progressive Latent Sub-score Learning Objective

The proposed framework is trained using a composite objective that jointly constrains the overall score prediction, the latent sub-score learning process, and the temporal coherence of the predicted sub-score sequence. This total objective, formulated for the Progressive Latent Sub-score Learning (PLSL) framework, minimizes three distinct error components simultaneously:

$$\mathcal{L}_{Total} = \mathcal{L}_{OS} + \sum_{m=1}^{M} \mathcal{L}_m^{sub} + \lambda \mathcal{L}_{Mono} \tag{15}$$

Each term contributes to a distinct aspect of learning, where *M* denotes the total number of sub-actions in the action procedure and λ is a non-negative hyperparameter that controls the influence of the monotonic penalty loss, balancing temporal consistency against absolute score accuracy:

(1) Overall Score Regression Loss ($\mathcal{L}_{OS}$): This is the Mean Squared Error (MSE) between the final aggregated predicted score ($\hat{S}_i$) and the ground-truth overall score (S_i), penalizing deviation from the primary target.

$$\mathcal{L}_{OS} = \left(\hat{S}_i - S_i\right)^2, \tag{16}$$

(2) Latent Sub-score Regression Loss ($\mathcal{L}_m^{sub}$): This is the MSE between the predicted sub-score for sub-action *m* (S_m^{sub}) and the generated refined pseudo-subscore label ($\hat{z}_m$), which enforces the accurate estimation of the latent procedural quality at each stage.

$$\mathcal{L}_m^{sub} = (S_m^{sub} - \hat{z}_m)^2 \tag{17}$$

(3) Monotonic Penalty Loss ($\mathcal{L}_{Mono}$): This novel loss term designed to enforce temporal consistency across the predicted sub-score sequence. $\mathcal{L}_{Mono}$ encourages the sequential consistency of the pseudo-scores, ensuring that the model's prediction for any segment does not arbitrarily contradict the cumulative history established by preceding segments.

$$\mathcal{L}_{Mono} = \sum_{m=2}^{M} \max\left(0, \text{Avg}\left(S_{1,\dots,m-1}^{sub}\right) - S_m^{sub}\right) \tag{18}$$

A penalty is applied only when the predicted sub-score at stage *m* falls below the average of all preceding sub-scores, encouraging a non-decreasing progression over the action sequence.

Intuition and Justification for Monotonic Penalty

The $\mathcal{L}_{Mono}$ term is introduced to enhance the identifiability and stability of the learned pseudo-subscores, which are inherently ambiguous latent variables due to the lack of ground-truth labels. Skill-based procedural tasks, such as diving, exhibit a causal accumulation of execution quality: a major flaw in an early stage (e.g., poor Take-off) cannot be fully recovered later, implying that quality often persists or cascades.

Without explicit constraints, the latent sub-scores are not uniquely identifiable, allowing the model to arbitrarily distribute error across substages, leading to unstable or non-interpretable patterns. $\mathcal{L}_{Mono}$ acts as a **temporal regularization constraint** that encodes this domain-specific knowledge into the learning objective.

- **Identifiability:** By penalizing any drastic drop in a sub-score (S_m^{sub}) relative to the average preceding performance ($\text{Avg}\left(S_{1,\dots,m-1}^{sub}\right)$), the loss forces the model to distribute score reductions logically across the sequence. This prevents the model from collapsing complex temporal error patterns into a single arbitrary latent variable, thereby making the entire sequence of pseudo-subscores a more reliable and interpretable representation of procedural quality.
- **Stability:** The constraint encourages a smoother manifold representation of action quality over the temporal axis. This stability prevents local optima where the predicted sub-scores oscillate illogically, thus improving the convergence properties and robustness of the network.

This composite loss function ensures the model is optimized to minimize prediction errors while simultaneously learning a temporal sub-score progression that is both logical and representative of the dive's execution quality (see Figure 2).

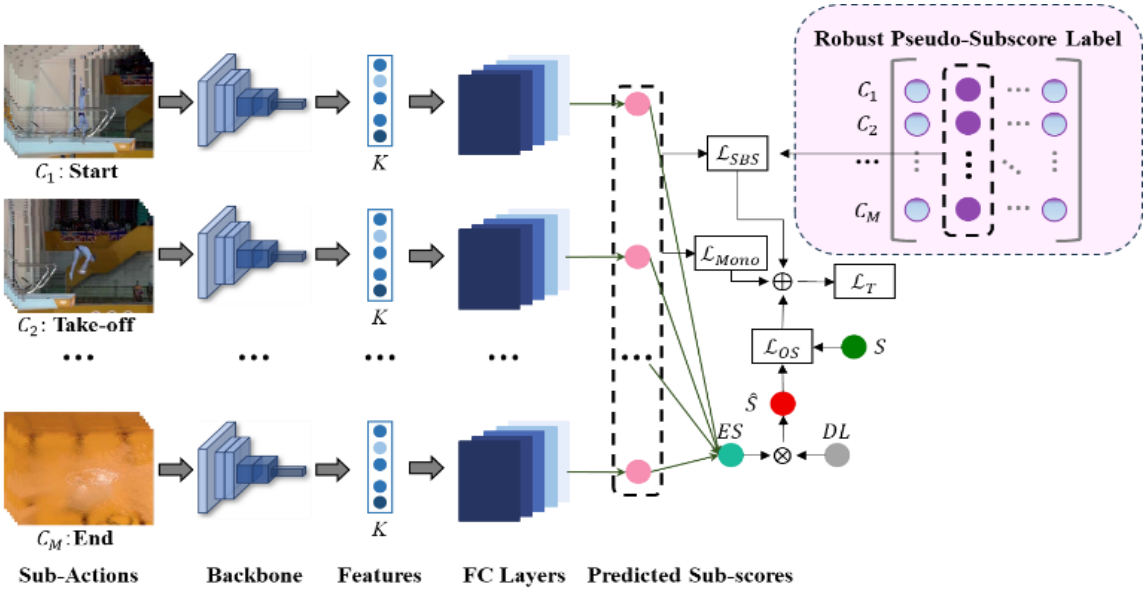

Figure 2. Overview of the multi-substage AQA

To improve clarity and provide an intuitive complement to the mathematical formulation in Eq. (5)–(14), we include Algorithm 1, which summarizes the progressive latent sub-score refinement process. This pseudocode outlines the sequential fusion of stage-wise features, contextual-history updates, and the recursive estimation of pseudo-subscores leading to the final overall score.

Algorithm 1: Progressive Latent Pseudo-Subscore Learning and Refinement

Require:
- Training segmented video clips with M sub-actions $\{C_m\}_{m=1}^{M}$
- Stage-wise features $\{\mathbf{f}_m\}_{m=1}^{M}$ (from C3D)
- Ground-truth Overall Score S. S_i.

Output:
- Refined pseudo-subscores $\{\hat{z}_m\}_{m=1}^{M}$,
- Final sub-score predictions $\{S_m^{sub}\}_{m=1}^{M}$
- Overall score prediction $\hat{S}_i$

1: **Initialize:** Latent Regression Network F_{Latent} , Parameters θ.
2: Compute initial pseudo-subscores $\{z_m\}_{m=1}^{M}$ from the label-reconstruction step (Eqs. (5)–(7)).
3: **Phase 1: Initial Pseudo-score Generation**
4: **for** each sub-action segment m=1 to M **do**
5: Concatenate feature and score: $\tilde{\mathbf{f}}_m \leftarrow [\mathbf{f}_m; S]$
6: Predict initial pseudo-score: $z_m \leftarrow F_{latent}(\tilde{\mathbf{f}}_m)$
7: **end for**
8: Train F_{latent} by minimizing $\mathcal{L}_\alpha = \frac{1}{N_{Tr}}\sum_{i=1}^{N_{Tr}}(\hat{S}_i - S_i)^2$ (Eq. 8).
9: Store initial pseudo-scores: $Z = \{z_1, z_2, \dots, z_m\}$.
10: **Phase 2: Progressive Refinement and Training**
11: **Initialize:** Progressive Network $\hat{\mathbf{f}}$, Parameters Φ.
12: **for** each sub-action segment m=1 to M **do**
13: Aggregate history: $y_{m-1} \leftarrow$ Aggregate$(\{z_1, \dots, z_{m-1}\})$ (Eq. 9)
14: Form the stage input by concatenating the local feature and cumulative context: $\hat{\mathbf{f}}_m \leftarrow [\mathbf{f}_m; y_{m-1}]$ (Eq. 10)
15: Predict refined pseudo-score: $\hat{z}_m \leftarrow Sigmoid(\mathbf{f}'_m)$ (Eq. 11)
16: **end for**
17: Train $\hat{\mathbf{f}}$ by minimize $\mathcal{L}_\beta = \frac{1}{N_{Tr}}\sum_{i=1}^{N_{Tr}}(\hat{S'}_i - S_i)^2$ (Eq. 12).
18: Store refined pseudo-scores: $\hat{Z} = \{\hat{z}_1, \hat{z}_2, \dots, \hat{z}_M\}$.
19: **Phase 3: Final prediction**
20: Take the last-iteration sub-scores $\{S_m^{sub}\}_{m=1}^{M}$ as the final stage-wise predictions.
21: Aggregate them to obtain the overall score $\hat{S}_i$.
22: Optimize all parameters jointly with $\mathcal{L}_{Total}$ (Eq. (15)), combining $\mathcal{L}_{\text{OS}}$, $\mathcal{L}_m^{\text{sub}}$, and $\mathcal{L}_{\text{Mono}}$.

4- EXPERIMENTS and RESULTS

This section presents the experimental setup and a comprehensive analysis of the results. It begins with a description of the datasets, evaluation metrics, and implementation settings used in the study. Subsequently, an extensive ablation study is conducted, followed by a comparative performance assessment against state-of-the-art AQA methods. The section concludes with a qualitative analysis that provides visualizations to demonstrate the effectiveness of the proposed framework and discusses specific failure case scenarios.

A. Datasets

The proposed framework is evaluated on two widely-used public benchmarks in AQA, the UNLV-Diving [1] and the FineDiving [2] datasets. Both are centered on the complex, sequential movements of diving competitions, making them well-suited for validating methods that perform fine-grained, multi-substage assessments.

The UNLV-Diving dataset comprises 370 video samples sourced from the 10-meter platform diving event at the 2012 London Olympics. Each video sample is annotated with an overall score and difficulty level, enabling the calculation of an execution score. A key

characteristic of this dataset is the presence of annotations that segment the dives into distinct sub-actions, facilitating substage-level analysis. The difficulty levels range from 2.7 to 4.1, and the overall scores span from 21.6 to 102.6. The videos have an average duration of 3.8 seconds, with a frame rate of 25 frames per second.

The FineDiving dataset represents a larger and more recent collection, containing 3000 videos from a variety of international diving competitions. While providing a greater volume of data, it simplifies the procedural labeling by defining primary sub-actions. This dataset offers fine-grained annotations that include action types, sub-action types, temporal boundaries, and action scores. The videos have a slightly longer average duration of 4.2 seconds and are sourced from diverse events, including the Olympics, World Cup, and World Championships. The diversity and scale of this dataset allow for a robust evaluation of the model's generalization capabilities across different action categories, scoring distributions, and competition contexts.

B. Metric Definitions

The performance of the proposed framework is evaluated using a comprehensive set of metrics that quantify both the overall prediction accuracy and the foundational temporal semantic segmentation precision. Following established practice in AQA, the effectiveness of the overall score prediction is quantified using three primary metrics that evaluate the alignment between the predicted score ($\hat{S}_i$) and the ground-truth score (S_i):

Spearman's Rank Correlation (SRC) (↑): Spearman's Rank Correlation coefficient (ρ) serves as the primary evaluation criterion in most AQA studies. As a non-parametric measure, SRC quantifies the strength of the monotonic relationship between the predicted score rankings and the ground-truth score rankings. A value closer to 1 indicates a stronger positive correlation, reflecting higher prediction accuracy in the rank-ordering of performances. Unlike linear correlation, it assesses monotonic relationships, where variables change simultaneously but not necessarily at a constant rate. The correlation is mathematically defined as:

$$\rho = \frac{cov(p,q)}{\sigma_p \sigma_q} = \frac{\sum_i (p_i - \bar{p})(q_i - \bar{q})}{\sqrt{\sum_i (p_i - \bar{p})^2 \sum_i (q_i - \bar{q})^2}} \quad (19)$$

where p and q are the ranking sequence of the predicted and ground truth scores, respectively. $cov(p,q)$ is the covariance of these ranking sequences, a measure of how the two rankings vary together. while σ_p and σ_q are the standard deviations of p and q. A larger ρ value indicates closer alignment between the predicted and actual scores, reflecting higher prediction accuracy.

Mean Squared Error (MSE) (↓): The MSE metric quantifies the average squared difference between the predicted and actual scores, providing a measure of absolute prediction error. Lower values indicate reduced discrepancy between the scores.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}\left(S_i - \hat{S}_i\right)^2 \quad (20)$$

Mean Euclidean Distance (MED) (↓): MED measures the average absolute difference between the predicted score and the actual score. Similar to MSE, a smaller MED value signifies improved prediction performance.

$$MED = \frac{1}{N}\sum_{i=1}^{N}\left|S_i - \hat{S}_i\right| \quad (21)$$

In both MSE and MED definitions, N represents the total number of samples.

Average Intersection Over Union (AIoU@d) (↑): For a comprehensive evaluation of the model's ability to perform fine-grained analysis, a metric for assessing the quality of the foundational temporal semantic segmentation step is also employed. This is essential to ensure that the model accurately identifies the boundaries of sub-actions [2]. This metric assesses the quality of the procedure-parsing module by quantifying the overlap between the predicted temporal boundaries and the ground-truth sub-action boundaries. A prediction is considered correct if its Intersection over Union (IoU) with the ground-truth segment is greater than or equal to a specified threshold, th. The $AIoU@th$ metric is formally defined as:

$$AIoU@th = \frac{1}{N}\sum_{i=1}^{N} \mathrm{I}(IoU_i \geq th) \quad (22)$$

where IoU_i is the Intersection over Union for the i-th sample's predicted and ground-truth temporal boundaries, and $\mathrm{I}(.)$ is an indicator function that returns 1 if the condition is met. A high $AIoU@th$ value confirms the precision of the segmentation, which is crucial for subsequent fine-grained sub-score prediction. The segmentation performance of the procedure-parsing module on the UNLV-Diving and FineDiving datasets is presented in Table 1.

Robustness to Segmentation Boundary Noise

The final scoring framework relies on the quality of temporal boundaries derived from the procedure-parsing model introduced in [2], rather than perfect ground-truth annotations. As shown in Table 1, the segmentation accuracy is inherently non-perfect, reaching 91.64% AIoU@0.5 on UNLV-Diving and 82.51% AIoU@0.5 on FineDiving. This indicates that the boundaries are subject to moderate noise and jitter, with frames around sub-action transitions inaccurately assigned to neighboring stages.

Despite this imperfect segmentation, the proposed PLSL framework is confirmed to be inherently robust to boundary noise. This resilience arises from several architectural properties of the model. First, the frame-sampling–based feature extraction process mitigates the impact of a few misplaced frames, because features are aggregated from uniformly sampled snippets rather than being overly influenced by noise near sub-action boundaries. In addition, the progressive latent sub-score refinement mechanism distributes structural cues across the entire sequence, which reduces the sensitivity of individual sub-scores to momentary misalignment at stage transitions. Finally, the monotonic penalty term $\mathcal{L}_{Mono}$ functions as a temporal regularizer that suppresses abrupt

inconsistencies and prevents isolated segmentation errors from propagating through the learned sub-score trajectory.

Compared to results obtained with ground-truth segmentation, the performance drop when using predicted boundaries is negligible, directly demonstrating that the framework generalizes well and is highly robust under realistic, noisy segmentation conditions without requiring additional correction mechanisms.

TABLE 1
Evaluation of the temporal semantic segmentation module using Average Intersection over Union (AIoU) on the UNLV-Diving and FineDiving datasets. The consistently high AIoU@0.5 values demonstrate that the segmentation component produces sufficiently accurate sub-action boundaries, providing reliable inputs for the subsequent sub-score learning despite moderate boundary noise.

Dataset	AIoU@0.5
UNLV-Diving	91.64%
FineDiving	82.51%

C. Experimental Setup and Data Protocol

The proposed framework is designed to evaluate athlete performance through a multi-substage scoring approach, in accordance with objective sports evaluation criteria. The proposed framework is implemented in PyTorch using Python 3.9, and all experiments are conducted on a single NVIDIA RTX 2080 GPU (11 GB). The experimental setup strictly adheres to established protocols to ensure result comparability with prior work.

I. Data Splits and Segmentation Protocol

For both benchmarks, the official dataset splits and evaluation protocols used in prior AQA works (PSL [16] and FineDiving [2]) are strictly followed to ensure fair comparability. The UNLV-Diving dataset was split into 300 training videos and 70 testing videos. The FineDiving dataset was partitioned into 2251 training and 749 testing videos. No samples were filtered or discarded in either dataset.

Temporal semantic segmentation is performed using the supervised procedure-parsing model from [2]. This model provides sub-action boundaries required for our multi-stage scoring pipeline. For the UNLV-Diving dataset, videos are partitioned into five sub-actions, including the Start, Take-off, Flight, Entry, and End, which corresponds to L=4 step transitions. In contrast, the FineDiving dataset, which defines a different set of sub-action annotations, is segmented into three primary phases: Take-off, Flight, and Entry, corresponding to L=2 transitional steps. The segmentation module is trained separately following the protocols described in [2] and is applied consistently across all experiments.

II. Feature Extraction and Preprocessing

Consistent feature extraction and preprocessing were applied across both datasets:

- **Feature Extraction:** A pre-trained I3D model [44] was adopted for the temporal segmentation component. Video clip features were extracted using a pre-trained C3D network [8]. The C3D backbone generated 4096-dimensional features.
- **Frame Sampling:** For the temporal segmentation component, 96 frames were extracted for each video, split into 9 snippets, where each snippet contained 16 continuous frames with a stride of 10 frames.
- **Segmentation Feature Processing:** After sub-action segmentation, 16 frames were uniformly sampled from each segmented sub-action clip and resized to a resolution of 112 × 112 pixels to maintain consistency with the C3D input requirements. The input was normalized using the ImageNet mean and standard deviation.
- **Data Augmentation:** No explicit data augmentation methods (e.g., flips, color jitter, cropping) are used. Only uniform sampling and resizing are applied to remain fully aligned with PSL [16] and FineDiving [2].
- **Score Normalization:** All raw scores were normalized to the range [0, 1]. Overall scores were normalized using Min-Max scaling based on the maximum and minimum scores observed in the training set, while execution scores (range 0–30) were normalized by dividing by 30.

III. Training Settings

The segmentation model was trained with Binary Cross-Entropy (BCE) loss. For the overall score prediction task, a composite loss (Eq. 15) was employed, combining MSE and the proposed Monotonic Penalty Loss ($\mathcal{L}_{Mono}$). The Adam optimizer [45] was used, with an initial learning rate of 10^{-4} for the feature extractor and 10^{-3} for the segmentation module. The learning rate decayed by a factor of 0.1 every 30 epochs. ℓ_2 regularization was applied with a weight decay of 0.0005, and a dropout layer with a probability of 0.5 was included after the average aggregation layers. All reported metrics are averaged over 10 independent runs with different random seeds to ensure statistical robustness. Metrics follow identical definitions to prior work, reporting SRC, MSE, MED.

IV. Hyperparameter Sensitivity Analysis

We evaluate the sensitivity of the proposed framework to key hyperparameters, specifically the monotonic penalty weight (λ), and clarify the structural constraints governing the number of stages and segment length.

Impact of Monotonic Penalty Weight: The monotonic penalty weight λ is the primary tunable hyperparameter in the PLSL objective, controlling the strength of temporal consistency regularization across sub-actions. To assess robustness, we evaluate performance under different λ values, as reported in Table 2. The results show that the proposed framework maintains stable performance across a wide range of λ settings, with the best results obtained around λ = 0.5. Very small values reduce the effectiveness

of temporal regularization, leading to less stable sub-score trajectories, while overly large values overconstrain the progression and degrade performance. This behavior confirms that the chosen λ provides a balanced trade-off between flexibility and consistency, and that the framework is not sensitive to precise tuning.

TABLE 2
Sensitivity analysis of the monotonic penalty weight (λ) on the FineDiving dataset. Moderate regularization (λ=0.5) yields the optimal trade-off, minimizing error while maximizing rank correlation.

λ	SRC	MSE	MED
0.0	0.93	24.32	3.70
0.1	0.93	24.21	3.65
0.3	0.93	25.77	3.79
0.5	**0.94**	**21.85**	**3.63**
1.0	0.92	29.87	4.12
2.0	0.90	32.60	4.39

Structural Constraints on Stages and Segmentation: Unlike λ, the number of sub-action stages (M) and the segment length are not free hyperparameters but are fixed by the dataset structure and backbone architecture. The number of stages corresponds directly to the manually annotated procedural taxonomy of the dataset (e.g., M=5 for UNLV-Diving, M=3 for FineDiving) and cannot be varied without redefining the semantic labels. Similarly, the feature extraction relies on a pre-trained C3D encoder which requires a fixed 16-frame input window. Consequently, segment length is constrained to 16 frames to ensure compatibility with the backbone.

Despite these constraints, the framework is evaluated on datasets with different procedural granularities, which already demonstrates robustness to varying stage definitions. Overall, the sensitivity analysis confirms that the performance gains of the proposed method are stable and not dependent on fragile hyperparameter choices.

D. Ablation Study: Analysis of Progressive Modeling and Consistency

A comprehensive ablation study was performed to rigorously quantify the contribution of the proposed framework's core mechanisms for progressive modeling and sequential consistency. To systematically evaluate these components, we compared the full framework against three distinct configurations:

Baseline: A foundational AQA model trained using only the overall score loss ($\mathcal{L}_{OS}$) on global video-level features. This configuration purposely omits both temporal segmentation and progressive sub-score learning, serving as the raw lower bound for performance.

Pseudo-Subscore Learning (PSL): This configuration incorporates the temporal semantic segmentation model to segment videos into sub-actions. It generates initial pseudo-subscores but does not include the long-range contextual refinement mechanism and the Monotonic Penalty Loss ($\mathcal{L}_{Mono}$). It represents a strong baseline that only uses sub-action segmentation.

Proposed Framework with λ=0: This configuration utilizes the full progressive architecture—including the refinement stages—but sets the monotonic penalty weight λ=0. This isolates the benefit of the progressive network structure from the explicit consistency constraint.

To ensure a clean attribution of performance gains to the PLSL mechanism and the monotonic consistency constraint, all models—including PSL baselines—are trained using the same C3D backbone. We intentionally avoid stronger modern backbones (SlowFast, Video Swin, TimeSformer) because their heterogeneous temporal receptive fields and large-scale pretraining protocols would introduce confounding factors that obscure the contribution of latent sub-score modeling. Using a fixed backbone therefore provides the fairest and most controlled setting for isolating the improvements brought specifically by PLSL.

The results presented in Table 3 (UNLV-Diving) and Table 4 (FineDiving) demonstrate the step-wise contribution of each component.

Impact of Segmentation (Baseline → PSL): The initial performance gain achieved by adopting temporal segmentation and generating pseudo-subscores is evident when comparing the Baseline and Pseudo-Subscore Learning models. Transitioning from the feature-level regression (Baseline) to the initial segmentation-based sub-score generation (PSL) yields a notable improvement, particularly in absolute error metrics. On the UNLV-Diving dataset, while SRC remains constant at 0.87±0.009, the MSE is drastically reduced from 85.24±4.12 to 38.68±2.10, representing a 54.62% improvement in absolute score accuracy. A similar trend is observed on the FineDiving dataset, where the SRC increases by 0.06 (from 0.85±0.10 to 0.91±0.005), and the MSE is reduced by 40.96% (from 56.95±3.01 to 33.62±2.10). This validates the fundamental premise that partitioning the action into meaningful segments provides a far more stable target for regression.

Impact of Progressive Architecture (PSL → λ=0): The addition of the progressive refinement module—even without the monotonic penalty—yields statistically significant gains over PSL. On UNLV-Diving, the Proposed Framework with λ=0 improves SRC to 0.89 ± 0.003 and further reduces MSE to 33.98±1.60. On FineDiving, the gain is even more pronounced, with SRC reaching 0.93±0.002 and MSE dropping to 24.32±2.66. This confirms that the progressive learning architecture itself stabilizes the training process and refines features effectively, even before explicit consistency constraints are applied.

Impact of Monotonic Consistency (λ=0 → Full Framework): Finally, enabling the Monotonic Penalty ($\mathcal{L}_{Mono}$) unlocks the full potential of the method. On UNLV-Diving, the full framework achieves the lowest error (MSE 29.42±1.79) and highest correlation (SRC 0.91±0.003), accounting for an additional 13% reduction in error relative to the λ=0 case. Similarly, on FineDiving, the MSE is further optimized to 21.85±2.07 with a peak SRC of 0.94±0.003. This final reduction demonstrates that enforcing sequential consistency across the execution history aligns the sub-score progression with logical

temporal criteria, correcting outliers that the architecture alone might miss.

The full proposed framework consistently achieves superior performance over the strong PSL baseline and exhibits low variance. On UNLV-Diving, the SRC improves from 0.87 to 0.91, and the MSE is further reduced from 38.68 to 29.42. This final step alone accounts for an additional 23.94% reduction in MSE. On FineDiving, the performance gain is even more pronounced. The SRC improves from 0.91 to 0.94, and the MSE is sharply reduced by 35.01%, dropping from 33.62 to 21.85. This reduction in error, achieved by enforcing sequential consistency across the robust pseudo-scores and modeling the cumulative execution history, aligns the sub-score progression with logical temporal criteria. In summary, the framework exhibits a synergistic effect: accurate temporal segmentation provides the foundation, the progressive architecture refines the representation, and the Monotonic Penalty Loss ensures logical consistency, delivering the most robust performance across both datasets.

TABLE 3
Ablation analysis and synergistic contribution of the proposed framework on the UNLV-Diving dataset. Each component—latent pseudo-subscore learning, progressive sub-score refinement, and the monotonic penalty—contributes to overall performance. The full configuration achieves the highest SRC and the lowest error metrics (MSE, MED), demonstrating that enforcing sequential consistency provides a strong cumulative benefit.

Method	SRC	MSE	MED
Baseline	0.87±0.003	85.24±2.02	5.66±0.15
PSL	0.87±0.002	38.68±2.10	4.80±0.18
Proposed Framework (λ=0)	0.89±0.003	33.98±1.60	4.09±0.11
Proposed Framework	0.91±0.003	29.42±1.79	4.05±0.10

TABLE 4
Ablation analysis and synergistic contribution of the proposed framework on the FineDiving dataset. Each component—latent pseudo-subscore learning, progressive sub-score refinement, and the monotonic penalty—contributes to overall performance. The full configuration achieves the highest SRC and the lowest error metrics (MSE, MED), demonstrating that enforcing sequential consistency provides a strong cumulative benefit.

Method	SRC	MSE	MED
Baseline	0.85±0.009	56.95±4.12	4.80±0.25
PSL	0.91±0.005	33.62±3.10	4.24±0.18
Proposed Framework (λ=0)	0.93±0.004	24.32±2.66	3.70±0.10
Proposed Framework	0.94±0.003	21.85±2.07	3.63±0.11

E. Comparison with State-of-the-Art Methods

The proposed framework's performance is quantitatively evaluated against existing state-of-the-art AQA methods, with results presented in Table 5 and Table 6 for the UNLV-Diving and FineDiving datasets, respectively. This comparison validates the effectiveness of our progressive long-range dependency modeling. As demonstrated in Table 5, the proposed framework achieves the highest SRC value of 0.91, indicating a strong correlation between the predicted and ground-truth rankings. Furthermore, the method achieves lower MED and MSE values of 29.42 and 4.05, respectively, confirming reduced prediction errors across both training and test datasets. Compared to [16], which also utilizes pseudo-subscores but lacks our explicit long-range dependency modeling, the proposed approach improves the SRC value by 0.04 and reduces MSE and MED values by 9.26 and 0.75, respectively, highlighting its superior accuracy in predicting both rankings and absolute scores. This underscores the importance of conditioning the assessment of each sub-action on the cumulative history of the action sequence.

TABLE 5
Comparison with state-of-the-art (SOTA) methods on the UNLV-Diving dataset. The proposed framework achieves the best performance across all reported metrics, highlighting its enhanced capability for modeling fine-grained procedural quality. ("–") means that the value was not reported in the literature.

Method	Publisher	SRC	MSE	MED
C3D-SVR [1]	CVPR 2017	0.74	–	–
C3D+CNN [10]	PCM 2018	0.80	–	7.78
S3D [13]	ICIP 2018	0.86	97.46	6.90
MUSDL [18]	CVPR 2020	0.87	129.59	–
Metric Learning [36]	TCSVT 2020	0.76	105.62	–
TAL [22]	SIVP 2021	0.86	–	–
MSRM [15]	KBS 2021	0.87	73.92	–
GDLT [26]	CVPR 2022	0.87	78.90	–
HGCN [3]	TCSVT 2023	0.88	101.48	–
PSL [16]	APPLI 2023	0.87	38.68	4.80
DAE [19]	NCAA 2024	0.84	85.30	–
T^2CR [40]	INFS 2024	0.83	96.78	–
CoFInAl [43]	IJCAI 2024	0.86	150.50	–
MAGR [33]	ECCV 2024	0.76	–	–
Proposed Framework	CKE 2026	0.91	29.42	4.05

The quantitative comparison on the FineDiving dataset, presented in Table 6, further reinforces the efficacy of our structural modeling approach. On this more comprehensive dataset, the proposed framework achieves the best performance across all three metrics (SRC, MSE, and MED).

TABLE 6
Comparison with state-of-the-art (SOTA) methods on the FineDiving dataset. The proposed framework achieves the best performance across all reported metrics, highlighting its enhanced capability for modeling fine-grained procedural quality. ("–") means that the value was not reported in the literature.

Method	Publisher	SRC	MSE	MED
MUSDL [18]	CVPR 2020	0.88	48.96	–
GDLT [26]	CVPR 2022	0.93	29.25	–
TSA [2]	CVPR 2022	0.92	–	–
MCoRe [38]	ICASSP 2024	0.92	–	–
T^2CR [40]	INFS 2024	0.92	27.21	–
STSA [42]	IJCV 2024	0.93	–	–
DAE [19]	NCAA 2024	0.93	27.17	–
CoFInAl [43]	IJCAI 2024	0.93	36.47	–
MAGR [33]	ECCV 2024	0.85	–	–
SAP-Net [20]	TIP 2024	0.84	–	–
Proposed Framework	CKE 2026	0.94	21.85	3.63

This achievement demonstrates that leveraging supervised segmentation alongside long-range progressive pseudo-score learning is highly effective in dealing with complex, multi-stage actions. The consistent outperformance of our method against strong baselines, including specialized contrastive regression methods, confirms the value of explicitly modeling procedural context for accurate AQA.

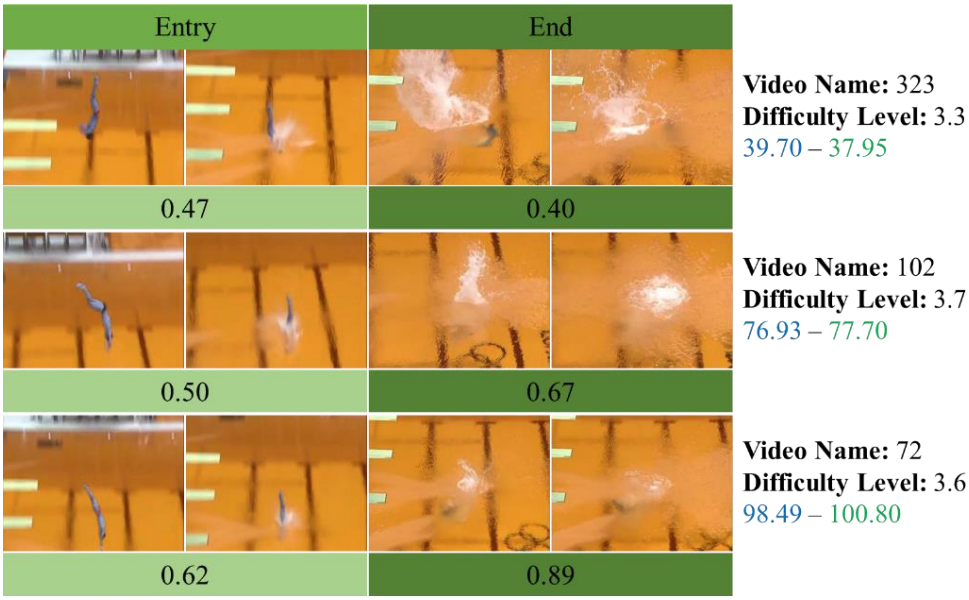

Figure 3. Visualization of the predicted sub-scores corresponding to the final two sub-actions in selected UNLV-Diving samples. The green value indicates the ground-truth overall score, while the blue value represents the overall score predicted by the proposed framework

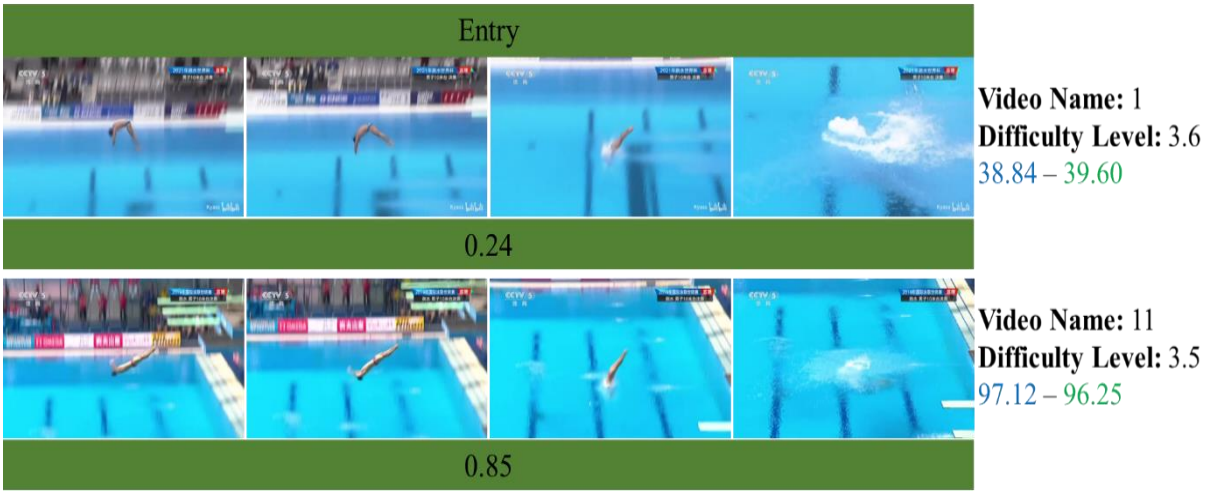

Figure 4. Visualization of the predicted sub-scores corresponding to the final sub-action in selected FineDiving samples. The green value indicates the ground-truth overall score, while the blue value represents the overall score predicted by the proposed framework

F. Qualitative Validation of Pseudo-subscores:

A qualitative examination was conducted to provide crucial empirical evidence for the meaningfulness of the generated pseudo-subscores, directly linking the model's predictions to visual execution quality. The analysis focused on the vital role of the final sub-actions, which significantly influence the overall score. Specifically, this included the last two sub-actions in the UNLV-Diving dataset and the "Entry" sub-action in the FineDiving dataset. The predicted sub-scores for specific segments show a strong correlation with expert judging criteria, such as body verticality and splash magnitude. This approach confirms that the latent variables learned by the progressive regression mechanism are not arbitrary but function as precise, interpretable indicators of stage-specific performance quality.

As shown in Figure 3, a qualitative analysis of the UNLV-Diving dataset reveals the model's ability to discriminate subtle performance differences, particularly in high-difficulty dives with an average difficulty level of 3.5. For a high-scoring instance with an overall score of 100.80, the model assigns a high sub-score of 0.62 to the "Entry" sub-action, consistent with the optimal body position and verticality observed. The "End" sub-score is also high at 0.89, reflecting the minimal splash. In a mid-scoring example (77.70 overall), the "Entry" sub-score decreases to 0.50, consistent with a slightly misaligned posture upon entry. Conversely, a low-scoring dive (37.95) yields markedly lower "Entry" (0.47) and "End" (0.40) sub-scores, effectively capturing the body misalignment and uncontrolled splash observed in the video frames.

The results on the FineDiving dataset, depicted in Figure 4, further reinforce the model's fine-grained evaluation capability. Two representative examples of a reverse armstand dive—a technically demanding maneuver involving a handstand take-off—are analyzed. For the high-scoring sample, the model predicts a high sub-score of 0.85 for the final "Entry" sub-action, accurately reflecting the clean, vertical entry into the water and resulting in a subtle splash. In contrast, for the low-scoring sample, a significantly lower sub-score of 0.24 is assigned to the same sub-action, precisely capturing the improper entry angle and the visible displacement of water.

Across both datasets, the framework consistently produces higher pseudo-subscores for well-executed motion segments and lower scores for visibly flawed stages, such as those exhibiting bent posture or excessive splash. These results confirm that the pseudo-subscores learned through the proposed progressive regression mechanism serve as interpretable and trustworthy indicators of localized performance quality. Moreover, the integration of these sub-scores enhances the overall scoring precision and introduces a valuable layer of diagnostic granularity, offering actionable feedback for performance improvement. This detailed feedback aligns with practical coaching applications, enabling athletes to concentrate on improving targeted aspects of their performance, ultimately enhancing their technical skills and competitive outcomes.

G. Interpretive Feedback and Error Analysis:

Beyond the quantitative metrics, a fine-grained analysis of the model's predictions provides valuable insight into its interpretive capabilities and a deeper understanding of its behavior in challenging scenarios. A particularly difficult aspect of AQA is the accurate assessment of low-scoring dives, which are rare in high-level competition datasets and present a unique challenge for regression models.

As illustrated in Figure 5, an example from the UNLV-Diving dataset is analyzed. This dive has a ground-truth overall score of 21.60, indicating a significant flaw. The model predicts an overall score of 44.34, which represents an overestimation. However, a closer look at the pseudo-subscores reveals the model's remarkable ability to pinpoint the source of the error. The sub-scores for the initial phases of the dive—Start, Take-off, and Flight—are 0.49, 0.49, and 0.51, respectively. The sub-scores for the critical final stages, Entry and End, drop to 0.47 and 0.31, respectively. This decline accurately corresponds to the visible, large splash upon entry and the lack of verticality in the body's position. This confirms that the model has successfully identified the primary fault in the performance, providing feedback that is highly valuable for coaching.

A comparative analysis on this specific sample further highlights the proposed framework's superior performance in a difficult case. While the overall prediction of 44.34 contains an error, it is significantly more accurate than the prediction of 73.84 reported by a previous method [15], which overestimates the score by a much larger margin. This demonstrates that the progressive modeling approach provides a more robust and reliable estimation even in challenging, low-scoring scenarios. Figure 5 visually compares the sample analyzed in [15] against the results of the proposed method.

The interpretive power of the framework is further showcased in an error analysis on the FineDiving dataset, as depicted in Figure 6. The figure presents a dive with a ground-truth score of 22.10. The model predicts a score of 45.34. Despite the predictive error, an examination of the sub-scores reveals a logical progression that aligns with the visual execution. The sub-scores for the initial Start phase is correctly estimated as high, at 0.71, indicating a strong beginning. However, a significant drop in the final sub-score for the Entry to 0.28 accurately captures the improper entry angle and the visible, large splash. These qualitative analyses confirm that while overall score predictions may have some discrepancies in rare, low-scoring cases, the framework's ability to generate interpretable sub-scores provides a powerful diagnostic tool.

I. Critical Failure Modes for Real-World Monitoring

Beyond aggregate metrics, we perform a qualitative error analysis using the interpretive feedback provided by the predicted sub-scores and the visual examples in Figures 5 and 6. This analysis reveals several recurring failure modes that are especially relevant for real-world deployment.

Based on the detailed error analysis, the primary failure modes that require monitoring are:

A first source of error is **action difficulty and label distribution bias**. The datasets are collected from

international competitions, where the majority of dives fall in the medium-to-high scoring range. Truly poor executions (e.g., overall scores below 20) are rare. Consequently, the model has limited exposure to genuine failure cases, and it tends to exhibit higher prediction error on these low-score dives. This effect is exacerbated for dives that involve very high difficulty or rarely observed configurations (e.g., certain reverse dives or uncommon dive codes), where the procedural complexity pushes the limits of the learned progressive dependencies.

A second factor involves **occlusion and camera viewpoint**. When the diver is partially occluded (by the board, clutter, or splashes) or when the camera deviates from the standard broadcast viewpoint (e.g., extreme zoom, unusual elevation, or side views), the temporal segmentation module may produce misaligned boundaries. In such situations, the pseudo-subscores for affected substages can drift, especially around Take-off and Entry, which rely heavily on precise localization. Although the overall framework remains robust to moderate boundary jitter, large segmentation errors can still propagate through the progressive learning chain and degrade both sub-score and overall-score estimates.

A third failure mode is related to **motion blur and video quality**. Fast rotations, rapid body movements, or low frame-rate recordings often introduce noticeable blur, particularly during the Flight phase. Combined with noise or compression artifacts, this reduces the discriminative power of the extracted spatio-temporal features, leading to less reliable sub-scores and, in some cases, underestimation or overestimation of the final score.

Finally, **highly atypical motion patterns**, such as incomplete rotations, aborted dives, or other unusual execution patterns, may produce out-of-distribution sub-score trajectories that are not well represented in the training data. In these cases, the model tends to produce inconsistent sub-scores across substages, reflecting its uncertainty about how to map such rare behaviors onto the learned progressive structure. While these conditions are infrequent in benchmark datasets, they highlight practical edge cases that require attention when deploying the system in unconstrained environments.

H. Practical Implications and Deployment Considerations

The primary significance of this framework lies in transforming AQA from a single-score regression task into a diagnostic and prescriptive tool for performance enhancement.

Integration into Coaching and Training Applications: A key practical benefit of the framework is its interpretable, stage-level sub-scores. Because each predicted sub-score corresponds directly to a specific procedural phase, coaches and athletes receive actionable and localized diagnostic feedback. Instead of a single opaque overall score, the system highlights exactly which sub-action (e.g., Take-off, Flight, Entry) contributed most to performance degradation. These interpretable outputs can be seamlessly integrated into coaching tools by overlaying the predicted sub-scores onto the corresponding video segments or keyframes, highlighting the temporal intervals in which execution quality declines, and linking these score reductions to specific visual cues such as body angle, rotation symmetry, or entry splash amplitude. This enables precise correction of technique, more efficient training sessions, and quantitative monitoring of progression over time. The sequential nature of the sub-scores further provides a structured, cumulative performance trajectory that mirrors human expert assessment.

Deployment Constraints: The computational profile dictates specific deployment constraints. Because the feature extraction stage (C3D or other 3D CNN backbones) is the dominant contributor to computational cost and memory usage, full-resolution video analysis is best handled on a server or cloud-based platform. This setting comfortably accommodates both the FLOP budget associated with spatio-temporal convolution and the multi-stage inference pipeline of PLSL. Although the overall inference time remains sufficiently low for near real-time analysis (~125 ms in our setup), the computational demands of 3D video backbones make direct deployment on lightweight edge devices challenging without additional optimizations. Practical deployment on mobile or embedded systems would require model quantization, frame-rate reduction, or compression of feature representations.

I. Efficiency and Scalability Analysis

To demonstrate the practical applicability of the proposed framework, we analyze its computational efficiency in terms of floating-point operations (FLOPs), parameter count, memory footprint, and training/inference latency. We compare the proposed framework against the primary baseline, PSL [16]. The proposed framework consists of three conceptual stages: (1) temporal semantic segmentation and feature extraction, (2) pseudo-subscore learning and progressive refinement, and (3) multi-substage AQA regression. The first stage functions as a generic preprocessing component. In the PSL baseline, this is implemented using ED-TCN with a P3D backbone, whereas our framework adopts the Temporal Segmentation Attention (TSA) module [2] with a C3D backbone [8]. These configurations are comparable in computational order and memory footprint. Crucially, this stage is fully modular; any reliable temporal feature extractor can be substituted without affecting the proposed PLSL formulation.

To isolate the cost introduced specifically by the proposed learning mechanism, our analysis focuses on the learning components where the methods differ. TABLE 7 reports the training time per epoch, FLOPs, parameter count, and memory usage. As shown, the proposed framework introduces only a modest computational overhead relative to PSL. The inclusion of progressive refinement and robustness-aware modeling leads to a slight increase in training time (approx. +2.0s per epoch) and FLOPs, while parameter count and memory consumption remain tightly controlled. This overhead is incurred primarily during training to support enhanced temporal reasoning and consistency modeling. The final multi-substage regression module exhibits identical parameterization in both methods.

TABLE 7
Computational efficiency comparison of learning components in PSL and the proposed framework. Metrics include training time per epoch, FLOPs, parameter count, and memory usage. The proposed framework introduces only a modest computational overhead while enabling progressive refinement and consistency-aware learning.

Method	Module	Train Time (s/epoch)	FLOPs (G)	Params (M)	Memory (MB)
Baseline PSL:	Pseudo-subscore Calculation	11.73	40.66	5.417	20.64
	Multi-substage AQA Regression	11.73	40.66	5.414	20.64
Proposed Framework:	Pseudo-subscore Calculation	13.73	44.92	5.417	20.64
	Robust Pseudo-subscore Calculation	13.73	44.92	5.420	20.70
	Multi-substage AQA Regression	13.73	44.92	5.414	20.64

Inference Latency and Throughput Stability: When aggregating all learning components, the total inference time for both PSL and the proposed framework is approximately 0.12 seconds per video. A key characteristic of our design is that inference latency is independent of the original video duration. Each video is decomposed into a fixed number of sub-actions (e.g., five for UNLV-Diving), and a fixed-length clip of 16 frames is uniformly sampled from each sub-action. Consequently, the computational cost of the learning stages depends solely on the number of sub-actions—which is fixed by the dataset annotations—rather than the raw video length. Therefore, the throughput remains constant across videos of varying durations. Based on this fixed-computational structure, the framework exhibits stable scalability without degradation for longer inputs, rendering a throughput-vs-length plot essentially constant and confirming the method's suitability for real-time applications.

J. Generalization and Cross-domain Applicability

Although the proposed framework is evaluated on diving datasets, its design is not specific to any particular sport or motion domain. The architecture does not incorporate any diving-specific cues—such as splash detectors, pose-template constraints, or domain-dependent heuristics. Instead, all components operate on generic spatio-temporal representations extracted from segmented video sequences, making the framework intrinsically applicable to a broad class of procedural skill-assessment problems.

This abstraction enables direct extension to domains such as gymnastics, figure skating, medical skill evaluation, manufacturing tasks, or other sequential operations, provided that the underlying actions can be decomposed into semantically meaningful sub-actions. The progressive sub-score refinement mechanism, as well as the temporal consistency constraint imposed by the monotonic penalty, captures cumulative execution quality in a way that naturally generalizes to any multi-stage procedural activity.

A practical limitation arises from data availability: our pipeline relies on supervised temporal semantic segmentation, which requires ground-truth temporal boundaries for each sub-action. Among currently available AQA datasets, only UNLV-Diving and FineDiving provide such annotations, and thus all empirical evaluations are necessarily restricted to these two benchmarks. Despite this constraint, the datasets differ substantially in motion structure, stage definitions, and score ranges. The consistent gains observed across both benchmarks provide strong evidence that the proposed framework can generalize effectively to other domains once datasets with suitable temporal annotations are made available.

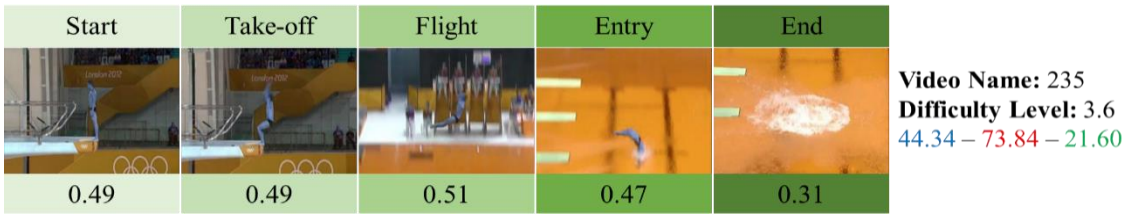

Figure 5. Error analysis on the UNLV-Diving dataset. Comparison of predicted overall scores from the MSRM method [15] and the proposed framework, along with the corresponding sub-scores for different sub-actions. The green value represents the ground-truth overall score, the red value indicates the overall score predicted by MSRM, and the blue value corresponds to the prediction obtained using the proposed framework

Figure 6. Error analysis on the FineDiving dataset. Sub-scores of different sub-actions for a challenging sample. The green value represents the ground-truth overall score, and the blue value denotes the prediction obtained using the proposed framework

5- CONCLUSION

This paper introduced a novel framework for AQA, addressing the critical need for fine-grained, substage-level feedback in a data-constrained environment. The proposed method, termed Latent Regression-based Progressive Sub-action Learning, moves beyond traditional holistic scoring by explicitly modeling the sequential dependencies within an action. The framework's key innovation lies in its ability to progressively refine pseudo-subscores, ensuring that each stage of a performance is evaluated in the full context of the preceding execution history. By combining this approach with a novel monotonic penalty loss, the model is guided to produce not only accurate scores but also a logical and interpretable progression of sub-scores. The foundation for this process is a supervised semantic segmentation technique that identifies meaningful video clips and extracts spatio-temporal features, allowing the framework to dynamically model sequential dependencies through iterative refinement.

The effectiveness of this structural modeling approach was empirically validated on two challenging public datasets, UNLV-Diving and FineDiving. The results demonstrate that the proposed framework achieves superior performance compared with existing methods, significantly improving both ranking accuracy and absolute score prediction. Ablation experiments verified the complementary roles of progressive refinement and monotonic consistency, while qualitative analyses showed that the predicted sub-scores correspond closely to human-perceived execution quality and provide actionable diagnostic insights.

The interpretive capabilities of the framework offer a new direction for sports analytics and automated coaching. Beyond sports, the sequential modeling capabilities make the method highly suitable for other procedural tasks that require detailed skill evaluation. The design does not rely on any diving-specific priors or domain-dependent modules; rather, it operates solely on temporally segmented video clips and their sequential relationships. As such, the framework could assess procedural skill and consistency in medical and surgical training (e.g., assessing specific surgical sub-tasks like suturing or endoscopy) or monitor operational quality in industrial or assembly-line tasks to detect deviations from standard workflows, thereby enhancing safety and efficiency.

Future directions will emphasize extending this robust methodology to such cross-domain contexts where procedural consistency is paramount. The core strength of the framework—its ability to model progressive, cumulative dependencies—is highly transferable. In these contexts, identifying the precise temporal failure point is critical for certification and error mitigation. To enhance practical adaptability, investigating the integration of an unsupervised or weakly supervised approach for temporal segmentation will be a priority, reducing reliance on manually annotated procedural boundaries. Another key direction will be developing datasets with standardized sub-action-level annotations to facilitate precise evaluation of predicted sub-scores.

REFERENCES

[1] P. Parmar and B. Tran Morris, "Learning to score olympic events," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2017, pp. 20–28. [Online]. Available: https://doi.org/10.1109/CVPRW.2017.16

[2] J. Xu, Y. Rao, X. Yu, G. Chen, J. Zhou, and J. Lu, "Finediving: A fine-grained dataset for procedure-aware action quality assessment," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 2949–2958. [Online]. Available: https://doi.org/10.1109/CVPR52688.2022.00296

[3] K. Zhou, Y. Ma, H. P. Shum, & X. Liang. (2023). Hierarchical graph convolutional networks for action quality assessment. *IEEE Transactions on Circuits and Systems for Video Technology*. [Online]. *33(12)*, pp. 7749–7763. Available: https://doi.org/10.1109/TCSVT.2023.3281413

[4] H. Pirsiavash, C. Vondrick, and A. Torralba, "Assessing the quality of actions," in *European conference on computer vision*, 2014, Springer, pp. 556–571. [Online]. Available: https://doi.org/10.1007/978-3-319-10599-4_36

[5] H. Jain and G. Harit, "An unsupervised sequence-to-sequence autoencoder based human action scoring model," in *2019 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, 2019: IEEE, pp. 1–5. [Online]. Available: https://doi.org/10.1109/GlobalSIP45357.2019.8969424

[6] Q. Lei, H.-B. Zhang, J.-X. Du, T.-C. Hsiao, & C.-C. Chen. (2020). Learning effective skeletal representations on RGB video for fine-grained human action quality assessment. *Electronics*. [Online]. *9(4)*, p. 568. Available: https://doi.org/10.3390/electronics9040568

[7] H. Li, Q. Lei, H. Zhang, J. Du, & S. Gao. (2022). Skeleton-based deep pose feature learning for action quality assessment on figure skating videos. *Journal of Visual Communication and Image Representation*. [Online]. *89*, p. 103625. Available: https://doi.org/10.1016/j.jvcir.2022.103625

[8] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4489–4497. [Online]. Available: https://doi.org/10.1109/ICCV.2015.510

[9] S. Hochreiter & J. Schmidhuber. (1997). Long short-term memory. *Neural Computation*. [Online]. *9(8)*, pp. 1735–1780. Available: https://doi.org/10.1162/neco.1997.9.8.1735

[10] Y. Li, X. Chai, and X. Chen, "End-to-end learning for action quality assessment," in *Pacific Rim Conference on Multimedia*, 2018: Springer, pp. 125–134. [Online]. Available: https://doi.org/10.1007/978-3-030-00767-6_12

[11] P. Parmar and B. Morris, "Action quality assessment across multiple actions," in *2019 IEEE winter conference on applications of computer vision*

(WACV), 2019: IEEE, pp. 1468–1476. [Online]. Available: https://doi.org/10.1109/WACV.2019.00161

[12] P. Parmar and B. T. Morris, "What and how well you performed? a multitask learning approach to action quality assessment," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 304–313. [Online]. Available: https://doi.org/ 10.1109/CVPR.2019.00039

[13] X. Xiang, Y. Tian, A. Reiter, G. D. Hager, and T. D. Tran, "S3d: Stacking segmental p3d for action quality assessment," in *2018 25th IEEE International conference on image processing (ICIP)*, 2018: IEEE, pp. 928–932. [Online]. Available: https://doi.org/10.1109/ICIP.2018.8451364

[14] Y. Li, X. Chai, and X. Chen, "Scoringnet: Learning key fragment for action quality assessment with ranking loss in skilled sports," in *Asian Conference on Computer Vision*, 2018: Springer, pp. 149–164. [Online]. Available:https://doi.org/10.1007/978-3-030-20876-9_10

[15] L.-J. Dong, H.-B. Zhang, Q. Shi, Q. Lei, J.-X. Du, & S. Gao. (2021). Learning and fusing multiple hidden substages for action quality assessment. *Knowledge-Based Systems*. [Online]. *229*, p. 107388. Available: https://doi.org/10.1016/j.knosys.2021.107388

[16] H.-B. Zhang, L.-J. Dong, Q. Lei, L.-J. Yang, & J.-X. Du. (2023). Label-reconstruction-based pseudo-subscore learning for action quality assessment in sporting events. *Applied Intelligence*. [Online]. *53(9)*, pp. 10053–10067. Available: https://doi.org/10.1007/s10489-022-03984-5

[17] C. Lea, M. D. Flynn, R. Vidal, A. Reiter, and G. D. Hager, "Temporal convolutional networks for action segmentation and detection," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 156–165. [Online]. Available: https://doi.org/ 10.1109/CVPR.2017.113

[18] Y. Tang *et al.*, "Uncertainty-aware score distribution learning for action quality assessment," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9839–9848. [Online]. Available: https://doi.org/ 10.1109/CVPR42600.2020.00986

[19] B. Zhang, J. Chen, Y. Xu, H. Zhang, X. Yang, & X. Geng. (2024). Auto-encoding score distribution regression for action quality assessment. *Neural Computing and Applications*. [Online]. *36(2)*, pp. 929–942. Available: https://doi.org/10.1007/s00521-023-09068-w

[20] K. Gedamu, Y. Ji, Y. Yang, J. Shao, & H. T. Shen. (2024). Self-supervised subaction parsing network for semi-supervised action quality assessment. *IEEE Transactions on Image Processing*. [Online]. *33*, pp. 6057–6070. Available: https://doi.org/10.1109/TIP.2024.3468870

[21] A. Vaswani et al. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*. [Online]. *30*. Available: https://cir.nii.ac.jp/crid/1370849946232757637

[22] Q. Lei, H. Zhang, & J. Du. (2021). Temporal attention learning for action quality assessment in sports video. *Signal, Image and Video Processing*. [Online]. *15(7)*, pp. 1575–1583. Available: https://doi.org/10.1007/s11760-021-01890-w

[23] S. Farabi, H. Himel, F. Gazzali, M. B. Hasan, M. H. Kabir, and M. Farazi, "Improving action quality assessment using weighted aggregation," in *Iberian Conference on Pattern Recognition and Image Analysis*, 2022: Springer, pp. 576–587. [Online]. Available: https://doi.org/10.1007/978-3-031-04881-4_46

[24] L.-A. Zeng *et al.*, "Hybrid dynamic-static context-aware attention network for action assessment in long videos," in *Proceedings of the 28th ACM international conference on multimedia*, 2020, pp. 2526–2534. [Online]. Available: https://doi.org/10.1145/3394171.3413560

[25] Y. Zhang, W. Xiong, & S. Mi. (2022). Learning time-aware features for action quality assessment. *Pattern Recognition Letters*. [Online]. *158*, pp. 104–110. Available: https://doi.org/10.1016/j.patrec.2022.04.015

[26] A. Xu, L.-A. Zeng, and W.-S. Zheng, "Likert scoring with grade decoupling for long-term action assessment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 3232–3241. [Online]. Available: https://doi.org/10.1109/CVPR52688.2022.00323

[27] S. Zahan, G. M. Hassan, & A. Mian. (2024). Learning sparse temporal video mapping for action quality assessment in floor gymnastics. *IEEE Transactions on Instrumentation and Measurement*. [Online]. *73*, pp. 1–11. Available: https://doi.org/10.1109/TIM.2024.3398072

[28] S. Wang, D. Yang, P. Zhai, C. Chen, and L. Zhang, "Tsa-net: Tube self-attention network for action quality assessment," in *Proceedings of the 29th ACM international conference on multimedia*, 2021, pp. 4902–4910. [Online]. Available: https://doi.org/10.1145/3474085.3475438

[29] K. Huang, Y. Tian, C. Yu, & Y. Huang. (2025). Dual-referenced assistive network for action quality assessment. *Neurocomputing*. [Online]. *614*, p. 128786. Available: https://doi.org/10.1016/j.neucom.2024.128786

[30] M.-Z. Li, H.-B. Zhang, L.-J. Dong, Q. Lei, & J.-X. Du. (2023). Gaussian guided frame sequence encoder network for action quality assessment. *Complex & Intelligent Systems*. [Online]. *9(2)*, pp. 1963–1974. Available: https://doi.org/10.1007/s40747-022-00892-6

[31] C. Han, F. Shen, L. Chen, X. Lian, H. Gou, & H. Gao. (2023). MLA-LSTM: A local and global location attention LSTM learning model for scoring figure skating. *Systems*. [Online]. *11(1)*, p. 21. Available: https://doi.org/10.3390/systems11010021

[32] S. Zhang *et al.*, "Logo: A long-form video dataset for group action quality assessment," in *Proceedings of the*

IEEE/CVF conference on computer vision and pattern recognition, 2023, pp. 2405–2414. [Online]. Available: https://doi.org/10.1109/CVPR52729.2023.00238

[33] K. Zhou *et al.*, "Magr: Manifold-aligned graph regularization for continual action quality assessment," in *European Conference on Computer Vision*, 2024: Springer, pp. 375–392. [Online]. Available: https://doi.org/10.1007/978-3-031-73247-8_22

[34] H. Doughty, D. Damen, and W. Mayol-Cuevas, "Who's better? who's best? pairwise deep ranking for skill determination," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6057–6066. [Online]. Available: https://doi.org/10.1109/CVPR.2018.00634

[35] H. Doughty, W. Mayol-Cuevas, and D. Damen, "The pros and cons: Rank-aware temporal attention for skill determination in long videos," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 7862–7871. [Online]. Available: https://doi.org/10.1109/CVPR.2019.00805

[36] H. Jain, G. Harit, & A. Sharma. (2020). Action quality assessment using Siamese network-based deep metric learning. *IEEE Transactions on Circuits and Systems for Video Technology*. [Online]. *31(6)*, pp. 2260–2273. Available: https://doi.org/10.1109/TCSVT.2020.3017727

[37] X. Yu, Y. Rao, W. Zhao, J. Lu, and J. Zhou, "Group-aware contrastive regression for action quality assessment," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 7919–7928. [Online]. Available: https://doi.org/10.1109/ICCV48922.2021.00782

[38] Q. An, M. Qi, and H. Ma, "Multi-stage contrastive regression for action quality assessment," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024: IEEE, pp. 4110–4114. [Online]. Available: https://doi.org/10.1109/ICASSP48485.2024.10447069

[39] M. Li, H.-B. Zhang, Q. Lei, Z. Fan, J. Liu, and J.-X. Du, "Pairwise contrastive learning network for action quality assessment," in *European Conference on Computer Vision*, 2022: Springer, pp. 457–473. [Online]. Available: https://doi.org/10.1007/978-3-031-19772-7_27

[40] X. Ke, H. Xu, X. Lin, & W. Guo. (2024). Two-path target-aware contrastive regression for action quality assessment. *Information Sciences*. [Online]. *664*, p. 120347. Available: https://doi.org/10.1016/j.ins.2024.120347

[41] L. Liu, P. Zhai, D. Zheng, and Y. Fang, "Multi-stage action quality assessment method," in *Proceedings of the 2023 4th International Conference on Control, Robotics and Intelligent System*, 2023, pp. 116–122. [Online]. Available: https://doi.org/10.1145/3622896.3622916

[42] J. Xu, Y. Rao, J. Zhou, & J. Lu. (2024). Procedure-aware action quality assessment: Datasets and performance evaluation. *International Journal of Computer Vision*. [Online]. *132(12)*, pp. 6069–6090. Available: https://doi.org/10.1007/s11263-024-02146-z

[43] K. Zhou, J. Li, R. Cai, L. Wang, X. Zhang, & X. Liang. (2024). Cofinal: Enhancing action quality assessment with coarse-to-fine instruction alignment. *arXiv preprint arXiv:2404.13999*. [Online]. Available: https://doi.org/10.24963/ijcai.2024/196

[44] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299–6308. [Online]. Available: https://doi.org/10.1109/CVPR.2017.502

[45] K. D. B. J. Adam. (2014). A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*. [Online]. *1412(6)*. Available: https://doi.org/10.48550/arXiv.1412.6980