Noname manuscript No.
(will be inserted by the editor)

UIL-AQA: Uncertainty-aware Clip-level Interpretable Action Quality Assessment

Xu Dong[1] · Xinran Liu[1] · Wanqing Li[2] · Anthony Adeyemi-Ejeye[1] · Andrew Gilbert[1]

Received: date / Accepted: date

Abstract This work proposes UIL-AQA for long-term Action Quality Assessment AQA designed to be clip-level interpretable and uncertainty-aware. AQA evaluates the execution quality of actions in videos. However, the complexity and diversity of actions, especially in long videos, increase the difficulty of AQA. Existing AQA methods solve this by limiting themselves generally to short-term videos. These approaches lack detailed semantic interpretation for individual clips and fail to account for the impact of human biases and subjectivity in the data during model training. Moreover, although query-based Transformer networks demonstrate strong capabilities in long-term modelling, their interpretability in AQA remains insufficient. This is primarily due to a phenomenon we identified, termed *Temporal Skipping* , where the model skips self-attention layers to prevent output degradation. We introduce an Attention Loss function and a Query Initialization Module to enhance the modelling capability of query-based Transformer networks. Additionally, we incorporate a Gaussian Noise Injection Module to simulate biases in human scoring, mitigating the influence of uncertainty and improving model reliability. Furthermore, we propose a Difficulty-Quality Regression Module, which decomposes each clip's action score into independent difficulty and quality components, enabling a more fine-grained and interpretable evaluation. Our extensive quantitative and qualitative analysis demonstrates that our proposed method achieves state-of-the-art performance on three long-term real-world AQA datasets. Our code is available at: GitHub Repository

Keywords Action Quality Assessment · Long-term Video Understanding · Interpretability · Uncertainty

1 Introduction

AQA aims to automatically evaluate human action quality in videos by assigning numerical scores based on predefined criteria. AQA methods strive to eliminate the subjectivity of human judges, offering a consistent and unbiased evaluation of action quality and aiming to enhance accuracy, robustness, and generalisation by leveraging past performance data. Recently, this problem has attracted growing attention from the computer vision research community due to its broad applicability in real-world scenarios. It has been utilized in sports video analysis, particularly in disciplines such as synchronized swimming, figure skating, and gymnastics (Parmar and Morris, 2017; Parmar and Tran Morris, 2019; Pirsiavash et al., 2014; Venkataraman et al., 2015; Xu et al., 2019; Zeng et al., 2020; Parmar and Morris, 2019). It offers analytical support for athletes' performances by assisting judges in scoring, helping athletes analyse their performance, and enabling movement correction to improve technique and prevent errors and injury. Beyond sports, AQA also finds applications in medical care assessment, such as surgical skill training

Xu Dong
E-mail: xd00101@surrey.ac.uk

Xinran Liu
E-mail: xl01315@surrey.ac.uk

Wanqing Li
E-mail: wanqing@uow.edu.au

Anthony Adeyemi-Ejeye
E-mail: femi.ae@surrey.ac.uk

Andrew Gilbert
E-mail: a.gilbert@surrey.ac.uk

[1] University of Surrey, Guildford, UK · [2] Advanced Multimedia Research Lab, University of Wollongong, Wollongong, Australia

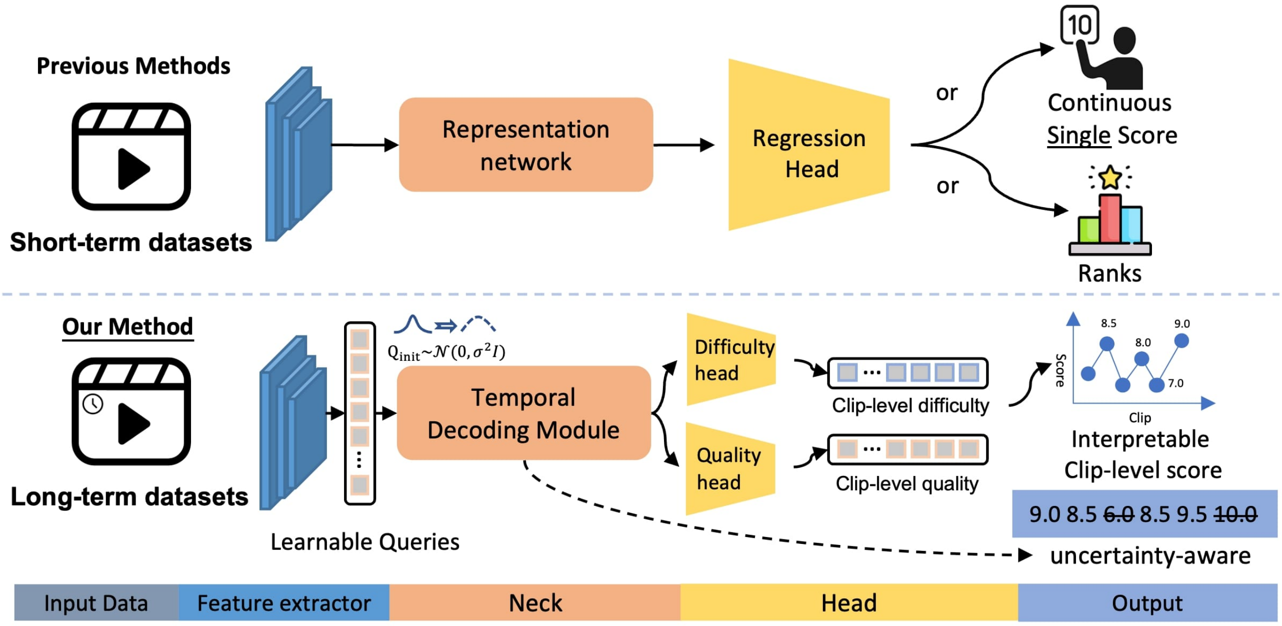


Fig. 1 Comparison with previous AQA methods, which perform single-score regression on short-term videos dataset, lacking interpretability and failing to account for uncertainty. Our proposed network extends to long-term video datasets, addressing subjectivity and scoring bias among different judges, ensuring more robust and reliable predictions. Furthermore, leveraging clip-level features and a dual difficulty-quality head enhances interpretability and improves regression performance.

(Funke et al., 2019; Wang et al., 2020; Gao et al., 2014), rehabilitation therapy (Li et al., 2024), and physical activity analysis (Parmar et al., 2022). Additionally, AQA is widely used in technical skill assessment, including professional skill training (Doughty et al., 2018, 2019) and task performance evaluation (Li et al., 2019b), to provide objective and data-driven analysis across various domains.

In contrast to our previous BMVC version, which focused on temporal skipping and Quality-Difficulty Regression, the main contribution of the current work is to introduce a novel uncertainty-aware learning mechanism that further enhances the performance of AQA. Specifically, we inject Gaussian noise into the predicted scores during training to simulate real-world ambiguity in human judging, where subjectivity and inconsistency frequently lead to variability in scoring outcomes. This simple yet effective strategy enables the model to explicitly model uncertainty, improving robustness to minor perturbations in both feature and score spaces. As a result, the model generates more stable and reliable predictions under noisy or ambiguous conditions—well aligned with real-world AQA settings where annotation noise and rater disagreement are common. We further compare our approach with existing uncertainty-aware AQA methods and demonstrate superior performance and generalisation ability, highlighting a more robust and human-aligned evaluation framework. In addition, we expand the related work section with a more comprehensive review of interpretable AQA methods, critically analysing architectural assumptions, interpretability mechanisms, and generalisation capabilities. Furthermore, extensive ablation studies validate the effectiveness of each proposed component, including different Transformer architectures, the impact of the attention loss, and the effect of query variance. These are supported by both quantitative results and attention map visualisations. We also include an efficiency analysis of FLOPs, parameter count, and inference time, showing that our method achieves comparable efficiency while offering better interpretability and performance. Finally, we conduct a user study to qualitatively assess the interpretability of our predictions, offering insights into how humans perceive difficulty and quality. In summary, the technical differences from the original BMVC paper lie in the

- Introduction of an uncertainty-aware learning mechanism via Gaussian noise injection and comparison with other uncertainty-aware methods.
- Expanded interpretability analysis, including attention visualisation and a user study.
- Comprehensive comparisons with prior interpretable AQA methods.
- Additional ablation studies on key modules such as attention loss and query variance.
- Efficiency analysis covering FLOPs, model parameters, and inference time.

- The discussion and addressing of a weakly-supervised disentanglement problem in AQA by separating difficulty and quality from video-level scores without any segment-level annotations.

1.1 AQA Challenges

AQA is generally regarded as a score regression task, with specific approaches (Parmar and Morris, 2019; Pan et al., 2019; Xu et al., 2019; Zhang et al., 2024a; Wang et al., 2021a) predicting the final score of an entire action sequence by applying simple averaging to clip features and utilizing an MLP regression head for aggregation as shown in Figure 1. However, a single score cannot offer detailed insights into individual components or subtle variations. Also, not all segments contribute equally to the final score; simple averaging assumes that all frames or segments have the same weight, which may lead to incorrect scoring. Moreover, such models inherently **lack interpretability**, making it difficult to understand how specific elements contribute to the final evaluation. For example, in diving evaluation, the size of the splash upon water entry is a key scoring factor. Still, in the previous approaches, its weight may be reduced due to inter-frame averaging. In sports scenarios where technical skills and execution proficiency are critical, such as diving, gymnastics and artistic swimming, judges determine the final score by assigning weights to each action clip based on its execution quality and difficulty level, such as in Figure 2. Furthermore, incorporating quality and difficulty-based scoring mechanisms in AQA models can significantly enhance their interpretability. Interpretability is crucial for AQA, improving transparency and trust in automated scoring models by making decision-making processes understandable. By explicitly modelling quality and difficulty factors, the system can better explain how each action contributes to the final evaluation, improving the fairness and usability of the model as shown in Figure 3. Our method uniquely integrates clip-level semantic insights, addressing both the interpretability and scoring bias challenges.

Another challenge in sports action assessment is the **subjectivity and scoring bias** among judges, which results in uncertainty, where the same action may receive varying scores. For example, individual judging styles in diving evaluation may heavily influence specific scores. If the model learns only fixed patterns from the data, it may overfit against specific judges, reducing its generalisation ability. This requires the model to possess the ability to adapt to scoring biases and the capability to model uncertainty, ensuring that it learns generalised evaluation standards rather than merely fitting the scoring patterns of specific judges.

Furthermore, previous studies on AQA (Roditakis et al., 2021; Bai et al., 2022) have primarily examined short-term videos, such as diving. These videos usually span only a few seconds, exhibit a straightforward sequential structure, and are all captured using a single camera. In contrast, **long-term AQA tasks**, which extend video beyond 120 seconds, present significantly greater challenges than short-term video (lasting 5 to 10 seconds). This increased difficulty arises from the heightened complexity, diversity of actions, and the larger amount of information that must be processed. Additionally, some datasets such as LOGO (Zhang et al., 2023), are collected using multiple cameras, which requires the model to have enhanced multi-view fusion capabilities to integrate information from different perspectives while maintaining temporal consistency effectively. Recent methods that incorporate queries into an encoder/decoder transformer architecture, such as those based on DETR models (Carion et al., 2020; Zhang et al., 2021; Bai et al., 2022; Du et al., 2023; Xu et al., 2022a), have been applied to the AQA task because of their effectiveness in long-term modelling and their decoder structure, which is well-suited for assigning temporal semantic meanings to learnable queries. However, the interpretability of these models in the context of long-term videos remains inadequate. One reason is that as the transformer processes each layer in long-term videos, the decoder's self-attention may experience a *skipping* effect and lead to the temporal collapse, as highlighted in Kim et al. (2023). The problem can be defined as *Temporal Skipping* in AQA. Specifically, *Temporal Skipping* refers to the phenomenon where, in long-term video sequences, the model tends to bypass certain steps in the decoder's self-attention mechanism, instead opting for shortcuts. This occurs because the model is influenced by the inherent temporal structure of the data, where it identifies and prioritises key moments or actions in the sequence most relevant to the task. However, this can lead to skipping intermediate frames or segments that could provide valuable context. Figure 3 demonstrates the self-attention maps and segmented scores of vanilla DETR methods compared with our proposed method. The visualisation shows that the self-attention pattern exhibits a *Temporal Skipping* issue in Figure 3(a). This leads to flattened and uniformly distributed attention weights that stray from the diagonal in the self-attention map. As a result, the interpretability is compromised in Figure 3(b) as indicated by the horizontal lines, where all clips are assigned equal weights. In contrast, the opposite results are shown in Figure 3(c) and Figure 3(d),

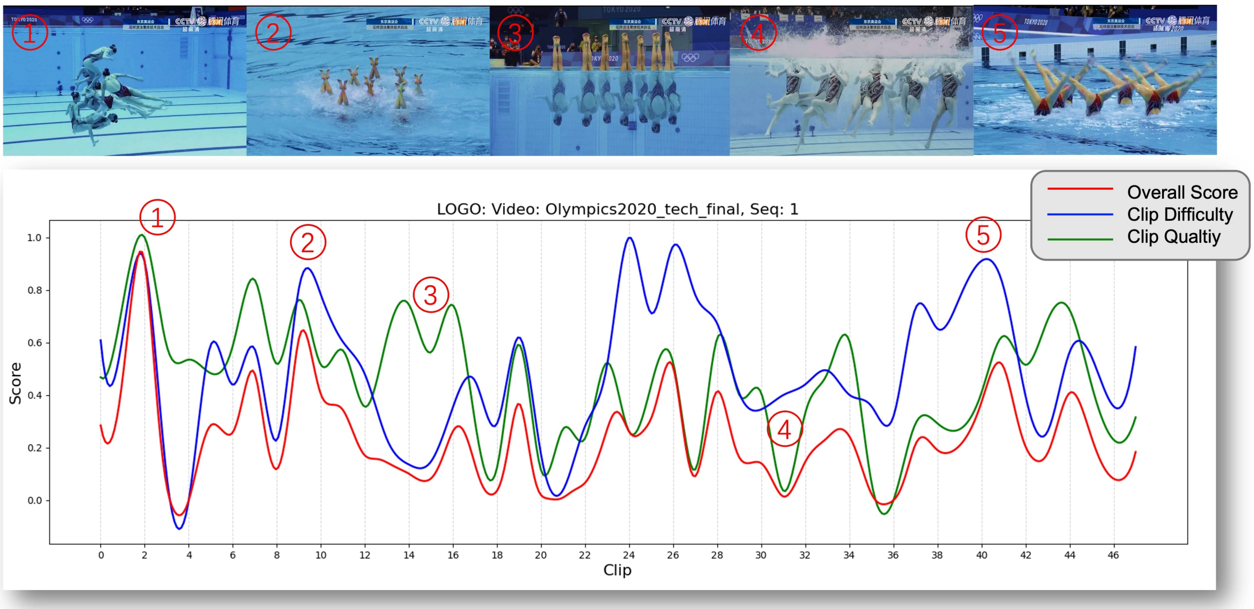
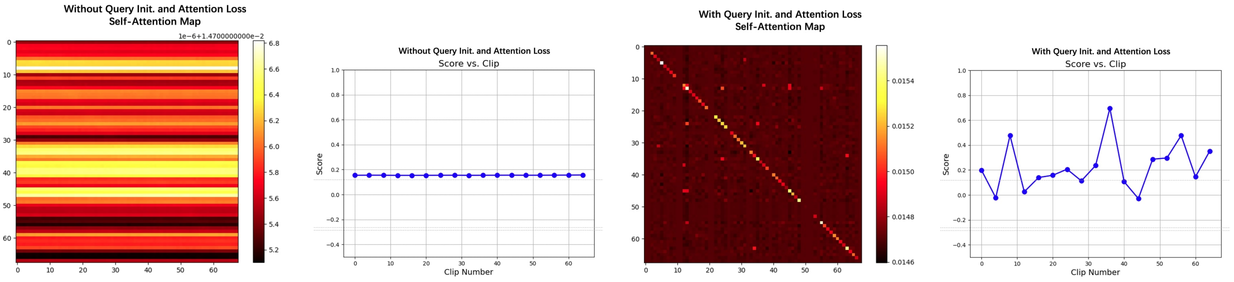


Fig. 2 The visualisation of the clip-level difficulty-quality regression method highlights that our network can mirror the evaluative framework employed by human judges in practical settings. The blue curve, representing the difficulty, signifies the relative contribution of each action clip to the final assessment. The green curve, denoting the quality, encapsulates the execution quality of the respective action. Meanwhile, the red curve conveys the aggregated overall score, integrating difficulty and quality considerations.



(a) (b) (c) (d)

Fig. 3 ***Temporal Skipping* problem of self-attention.** This figure presents the self-attention maps (3(a) and 3(c), ours) alongside the segmented score visualisations for each clip (3(b) and 3(d), ours). The pairs (3(a), 3(b)) and (3(c), 3(d)) correspond to the same action sequences. Notably, in 3(a), the self-attention map is significantly affected by the *Temporal Skipping* issue, whereas in 3(c), our approach demonstrates strong correlations between queries.

where the self-attention map exhibits a clear diagonal pattern of self-correlation in query attention. Additionally, the segment score figure illustrates that each clip is assigned a different score, highlighting the model's capability for interpretability.

1.2 Our Contributions

To address the aforementioned issues of lack of interpretability from a single score, subjectivity and scoring bias, and the challenges of long-term video sequences, we propose several modules designed to mitigate these problems.

To solve *Temporal Skipping* , we introduce an **Attention Loss** that enables mutual guidance between self-attention and cross-attention maps. This is achieved by reducing their similarity through KL divergence, ensuring that as the number of transformer decoder layers increases, the queries within the self-attention mechanism remain highly correlated. Additionally, adjusting the variance of the Gaussian distribution used for **Query embedding Initialization** enhances the self-attention correlation, as demonstrated by a more distinct diagonal pattern in the self-attention map (see Figure 3(c) and 7). Encoding positional information in queries and features is key to preserving spatial and temporal consistency and enhancing model interpretability. To achieve this, we introduce positional encoding for learnable queries to capture their temporal characteristics. Furthermore, we investigate applying positional encoding to video features but find that since the feature extractor already extracts positional information, the additional encoding provides further temporal context.

Moreover, we draw inspiration from how human judges evaluate action quality and introduce a novel ***Difficulty-Qualtiy Regression Head*** to replace the traditional single-score regression method to enable interpretability. This module disentangles the DETR decoder's output into separate difficulty and quality branches, aligning more closely with the scoring process used by human judges. The final action score is derived by computing the weighted sum of individual clip scores. As illustrated in Figure 2, our *Difficulty-Qualtiy Regression Head* effectively assigns distinct difficulty and quality to each clip.

To model the uncertainty in the judge's scores, we propose a simple yet effective ***Gaussian noise injection module*** to mitigate the impact of subjectivity and human scoring bias on the results. This module is applied to the network's output, helping to enhance robustness and stability while reducing the influence of subjective variations. Specifically, this process is achieved by incorporating noise into the network output to model human bias and enhance the network's capacity to capture uncertainty. By introducing a Gaussian Noise Injection module, the model effectively mitigates the influence of subjective variations in human scoring, promoting robustness against potential biases and improving generalisation across diverse input distributions and final performance.

In summary, our main contributions are as follows:

- We propose UIL-AQA , a Query-based Transformer decoder network for AQA, incorporating positional query encoding to extract clip-level features with temporal semantics while also addressing the *Temporal Skipping* issue, which leads to interpretability failures through an Attention Loss and a Query Initialization method. The model uses a split *Difficulty-Qualtiy Regression Head* to decouple the score into difficulty and quality, enhancing the interpretability of the model.
- We propose a *Gaussian Noise Injection Module* that adds noise to the output to model and mitigate the impact of human biases and subjectivity, enhancing the network's ability to handle uncertainty and ultimately improving the final results.
- We achieve state-of-the-art performance on three long-term AQA benchmarks, Rhythmic Gymnastics (RG), Figure Skating Video (Fis-V), and LOng-form GrOup (LOGO), through both quantitative and qualitative user-based evaluations, demonstrating the effectiveness of our proposed method.

2 Related Work

2.1 Action Quality Assessment

AQA methods can be categorised into handcrafted feature modelling and deep learning-based methods. Early statistical methods primarily relied on handcrafted features to model action features. Pirsiavash et al. (2014) was the first to investigate the AQA task by extracting spatiotemporal pose features from individuals and utilising the L-SVR model to estimate and predict action scores. Sharma et al. (2014) utilised spatial-temporal interest points (STIP) and computed the HOG (Histogram of Oriented Gradients) and HOF (Histogram of Optical Flow) on a 3D video patch around each detected STIP. Wnuk and Soatto (2010) utilised the Scale-Invariant Feature Transform (SIFT) method to extract features from videos, capturing the spatial characteristics of actions. These extracted features were then used to train a regression model for action quality assessment. However, handcrafted features struggle with complex and diverse action scenarios, limiting generalisation and robustness.

The emergence of deep learning has significantly improved AQA by enabling models to learn more robust and adaptable feature representations. This advancement has significantly benefited action quality assessment (AQA). Parmar and Morris (2017) introduced three frameworks for assessing the quality of Olympic event actions: C3D-SVR, C3D-LSTM, and LSTM-SVR, leveraged 3D convolutions (C3D) (Tran et al., 2015) and sequential modelling for AQA networks. However, while LSTMs capture temporal dependencies, it struggles with long-range relationships, and C3D, despite its effectiveness, is computationally expensive. This led to the introduction of self-attentive LSTMs and multi-scale convolutional skip LSTMs Xu et al. (2019), which aimed to balance efficiency with temporal modelling capabilities. More recently, deep learning-based AQA research has been categorised into regression-based and ranking-based methods, where regression-based methods directly predict quality scores. In contrast, ranking-based methods focus on learning relative score differences between actions. Most regression-based methods use 3D CNNs as their feature extractor because they can extract spatial and temporal features directly from video sequences. Models such as C3D (Tran et al., 2015), I3D (Carreira and Zisserman, 2017) and P3D (Qiu et al.,

2017) are widely used in AQA approaches (Xiang et al., 2018; Parmar and Tran Morris, 2019; Zhou et al., 2022).

Transformer-based architectures, such as VST (Liu et al., 2022b) and Vivit (Arnab et al., 2021), have demonstrated superior performance in long-term video modelling by capturing extended dependencies across frames. Unlike CNNs or LSTMs, transformers excel at learning hierarchical representations, making them well-suited for AQA. Studies (Xu et al., 2022a) have shown that these architectures outperform traditional CNN-LSTM approaches in AQA tasks, particularly in long-form assessments. As for ranking-based methods, Doughty et al. (2019) introduced a rank-aware loss function and trained it alongside a temporal attention module. However, they only provide overall rankings, limiting their applicability in AQA tasks requiring quantitative comparison. Yu et al. (2021) improved it and presented a group-aware regression tree (CoRe) method, which predicts relative scores while referencing others' performances. Xu et al. (2024) proposed FineParser, which can extract the fine-grained human-centric foreground action representations and achieve state-of-the-art results given a pair of query and exemplar videos. Zhou et al. (2024) proposed a coarse-to-fine alignment strategy that captures local-to-global consistency between predicted scores and instructional guidance. Zeng and Zheng (2024) proposed an adaptive multimodal fusion strategy that progressively integrates heterogeneous features, further demonstrating the effectiveness of multimodal learning in AQA tasks.

More recently, several novel works have further extended the scope of AQA. Xu et al. (2025a) introduces a language-guided audio-visual learning framework MLAVL, for long-term sports assessment, effectively modelling the correlation between actions and music to improve scoring accuracy. Xu et al. (2025b) extended AQA into the domain of group dance assessment by addressing the challenges of pose estimation errors in multi-person settings, enabling the evaluation of dance neatness and synchronisation. Chen et al. (2024) introduced a novel benchmark and methodology for assessing the quality of AI-generated videos, highlighting the importance of evaluating both natural human actions and synthetic content.

Uncertainty Some studies have aimed to tackle the uncertainty issue in AQA. Tang et al. (2020) introduced the Uncertainty Score Distribution Learning (USDL) method to improve action quality representation by treating each action as an instance linked to a score distribution, thereby mitigating the effects of inherent label ambiguity. UD-AQA Zhou et al. (2022) proposed a CVAE-based uncertainty modelling framework to capture the subjectivity in human judgment. It learns latent score distributions via a posterior and prior network, aligned through KL divergence, and uses an uncertainty-aware reweighting strategy to suppress ambiguous samples during training. One drawback of CVAE-based uncertainty modelling is its susceptibility to posterior collapse and training instability, especially under limited data conditions common in AQA tasks. LUSD-Net Ji et al. (2023) adopts uncertainty-aware modelling to improve the reliability of AQA under subjective scoring conditions. Inspired by Kendall and Gal (2017), it models aleatoric uncertainty by learning input-dependent variance through heteroscedastic regression. This approach allows the model to quantify prediction confidence and reduce the influence of noisy labels. While effective for capturing data-level ambiguity, it does not account for epistemic uncertainty, which may arise from limited or unfamiliar training data.

Interpretability The above AQA methods focus on single-score regression, which lacks clip-level temporal semantic representations. Thus, research on AQA interpretability Roditakis et al. (2021) introduced a self-supervised training technique and a differential cycle consistency loss to enhance temporal alignment and interpretability. Similarly, Farabi et al. (2022) argued that simply averaging clip-level features fails to capture their relative importance, proposing a weighted-averaging technique instead. While these approaches emphasise clip-level semantics, they do not align with the real-world scoring logic of human judges. In contrast, our method decouples clip-level features into difficulty and quality, further improving AQA interpretability. Similar to our goal of enhancing fine-grained interpretability through stage-wise scoring, Li et al. (2023) proposed a pseudo-subscore learning (PSL) framework for substage modelling without requiring subscore annotations. However, their method is only tested on the UNLV-Diving dataset, where actions follow a fixed temporal structure. In contrast, complex sports involve diverse and non-deterministic action flows, posing greater challenges for substage modelling and alignment due to long durations and multi-view dynamics. Han et al. (2025) introduces FineCausal: a causal-based framework that models stage-level feature–score relationships, enhancing interpretability and achieving strong performance in fine-grained AQA. However, its reliance on expert priors for causal graph construction may introduce additional annotation and modelling costs.

In recent years, with the rapid advancement of language models in understanding and generating natural language, there has been a growing interest in integrating language-based descriptions into action quality assessment tasks. Zhang et al. (2024b) proposed a new task, Narrative Action Evaluation (NAE), which shifts the focus from scalar regression to generating expert-style narrative feedback conditioned on video content and coarse scores. NAE enables a more interpretable and human-aligned evaluation by producing natural language explanations that resemble how judges justify their decisions. Okamoto and Parmar (2024) proposes a hierarchical neuro-symbolic framework that combines Platform Abstraction, Pose Estimation, and Splash Abstraction to perform structured and fine-grained evaluation of action executions, ultimately generating interpretable, stage-aware textual and visual reports. To move beyond score regression of AQA, several works have begun exploring descriptive and actionable feedback to improve interpretability. ExpertAF Ashutosh et al. (2025) introduces a multimodal coaching framework that generates actionable natural language feedback. TechCoach Li et al. (2025) presents a keypoint-aware coaching system that analyses the quality of individual body parts and produces fine-grained, interpretable textual feedback. While both ExpertAF and TechCoach aim to enhance interpretability and provide natural language feedback beyond scalar scores, they differ in focus: ExpertAF emphasises actionable coaching by providing both commentary and visual demonstrations, whereas TechCoach focuses on technical-point-aware reasoning, offering fine-grained explanations grounded in body part performance.

2.2 DETR in Video Understanding

With the rise of transformer-based architectures, researchers have explored DETR-style models for video understanding and AQA tasks due to their temporal reasoning capability by leveraging learnable queries effectively. DEtection TRansformer (DETR) was initially proposed by Carion et al. (2020), leveraging a transformer-based architecture to model complex relationships and dependencies in data through learnable queries. DETR has been extended to video understanding tasks such as object detection, tracking, and action recognition due to its remarkable ability in temporal modelling. For example, Liu et al. (2022a) introduced a method that applied Deformable DETR (Zhu et al., 2021) to temporal action detection (TAR), effectively removing the need for a proposal generation stage. Moon et al. (2023) proposed a Query-Dependent DETR (QD-DETR) for moment retrieval and highlight detection (MR/HD) tasks, which explicitly inject the context of the text query into the video representation. Kim et al. (2023) was the first to identify the temporal collapse problem in temporal action detection when using a DETR-like structure and introduced a self-feedback method to mitigate it. Temporal Collapse refers to the degradation of temporal representations in sequence modelling, where the model fails to capture long-term dependencies, causing feature redundancy and misalignment in tasks. Similarly, our work tackles temporal collapse within the AQA task but instead focuses on modifying the decoder's self-attention and cross-attention maps representation. Zhang et al. (2021) proposed a Temporal Query Network for fine-grained action classification on untrimmed video, which uses a query-response mechanism to regard each action clip in a video as a query. Regarding AQA tasks, Bai et al. (2022) proposed a Temporal Parsing Network (TPN) based on the DETR decoder for decomposing global features into temporal levels, which allows the network to parse temporal semantic meanings for enhanced action quality assessment. However, TPN only evaluated short-term datasets, which lacked long-term video modelling capabilities. Our network is based on DETR and incorporates a query initialization module and attention loss on self-attention and cross-attention to prevent from *Temporal Skipping* .

2.3 Long-term Video Understanding

Unlike short-term AQA, where actions occur in a compact time frame (5- 10s), long-term video understanding requires capturing sparsely distributed action cues over extended durations. This introduces challenges in maintaining temporal consistency, preventing feature redundancy, and ensuring accurate segmentation of meaningful events. Early studies (Li et al., 2019a; Gao et al., 2017; Srivastava et al., 2016) employed RNNs and LSTMs for long-term video modelling. More recently, numerous works have shifted towards transformers, leveraging their effectiveness in capturing long-range dependencies in video modelling (Wang et al., 2021b; Wu et al., 2022; Arnab et al., 2021; Liu et al., 2022b). This transition has benefited various video understanding tasks, such as temporal action detection (Zhang et al., 2022), video large language models (Song et al., 2024; Ren et al., 2023), dense video captioning (Yang et al., 2023; Wang et al., 2021c), etc. In the AQA task, early research primarily focused on short-term videos, typically lasting 5–10 seconds (Parmar and Tran Morris, 2019; Parmar and Morris, 2019). However, this is insufficient for real-world scenarios, where actions often span longer durations and involve complex temporal dependencies. Recent works have introduced long-term action quality assessment (AQA) datasets and methods, including LOGO (Zhang et al., 2023), Rhythmic Gymnastic (RG) (Zeng et al., 2020) and Figure Skating Video (Fis-V) (Xu et al., 2019). Furthermore, Zeng et al. (2020)

proposed ACTION-NET, which employs a GCN with a context-aware attention module for temporal feature modelling. However, ACTION-NET demonstrated suboptimal performance on long-term datasets. To address this, Xu et al. (2022a) improved upon prior work by incorporating learnable queries as grade prototypes and introducing a Likert Scoring Module for grade decoupling in long-term video assessment. Skating-Mixer (Xia et al., 2023) proposed an MLP-based framework for long-term audio-visual modelling in sports assessment, capturing correlations between motion and music for more accurate performance evaluation. More recently, QGVL (Xu et al., 2025c) has been proposed for long-term AQA, which leverages cross-modal alignment between video content and quality-related textual cues to enhance temporal modelling. In addition, PHI (Zhou et al., 2025) introduced a progressive instruction mechanism to bridge domain shifts in long-term AQA, improving adaptability across diverse datasets. In contrast to these approaches, our previous work (Dong et al., 2024) proposed a Transformer-based architecture well-suited for long-term video modelling, directly addressing the interpretability gap in long-term AQA. In this paper, we further enhance our previous work by introducing an uncertainty-aware module to mitigate subjectivity and human biases, thereby further improving the final results, interpretability, and robustness.

3 Methodology

To achieve an accurate and interpretable action quality assessment model that can also model the judge's uncertainty, we proposed our network UIL-AQA as illustrated in Figure 4, which has three key modules. Before the backbone feature extractor is processed, our long-term input video is split into equal-sized clips. A backbone feature extractor then processes the input video and extracts clip-level features. Next, a *Temporal Decoder* captures attention relationships between these features and a set of learnable position-encoded queries initialised by our proposed query Initialization module to model temporal semantic representations using a transformer decoder structure. The backbone extracts spatial and short-term temporal features, while the transformer-based decoder learns long-range dependencies and high-level temporal representations, improving sequence modelling. The extracted clip features are then passed to a *Difficulty-Qualtiy Regression Head* , which separately predicts the difficulty and quality for each action clip. Before computing the loss function, the network output is processed through a Gaussian Noise Injection Module to add noise, simulating human biases. The final action score is computed by weighting each clip's score and summing the results. The network is trained using two loss functions: Attention Loss ($Loss_{att}$), which minimises the KL divergence between attention maps at each layer, and Regression Loss ($Loss_{reg}$), which measures the Gaussian noise injection mean squared error.

3.1 Feature Extractor

To extract the sequence or clip features from the input video V. we divide the video into L non-overlapping clips, each containing M consecutive frames, $V = \{F^i\}_{i=1}^{L}$. We use two common feature extractors, Inflated 3D ConvNet (I3D) (Carreira and Zisserman, 2017) and Video Swin Transformer (VST) (Liu et al., 2022b, 2021) as our *Feature Extractor* . I3D is a 3D convolutional network designed to model spatiotemporal relationships in video using inflated 3D convolutions, where 2D convolutional filters are expanded into 3D by adding a temporal dimension. It processes video clips by applying 3D convolutions and 3D max pooling to learn motion patterns, followed by an Inception-based multi-scale feature extraction module (Szegedy et al., 2015), and ends with global average pooling and a classification head. VST (Video Swin Transformer) is a transformer-based video model designed to capture spatiotemporal relationships using shifted window self-attention, which computes attention locally within non-overlapping spatial windows while enabling cross-window communication through a shifting mechanism. It processes video clips by applying hierarchical patch embedding and window-based multi-head self-attention to extract spatial features efficiently, followed by temporal window attention, which extends self-attention across frames to model motion dynamics. The extracted features are progressively downsampled through patch merging layers, enabling multi-scale representation learning, and finally passed through a classification head. The *Feature Extractor* network, including I3D and VST as the feature extractors, are pre-trained on the Kinetics dataset and are frozen during training. The features obtained from L clips are denoted as ${f^i}_{i=1}^{L}$, where each $f^i \in \mathbb{R}^d$.

3.2 Temporal Decoder

Once clip-level features ${f^i}_{i=1}^{L}$ are extracted, the next step is to model temporal dependencies across these clips. We use a transformer-based decoder inspired by DETR (Carion et al., 2020) to achieve this, as shown in Figure 5. DETR is an end-to-end image object detection framework that utilises a transformer structure to

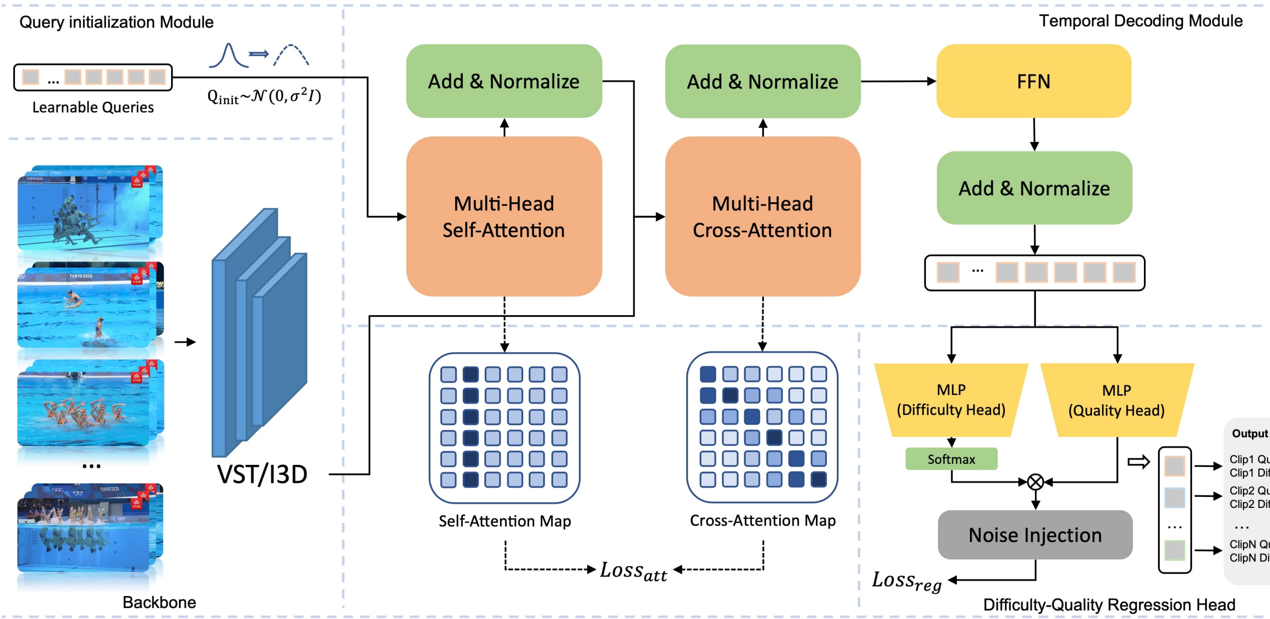


Fig. 4 The overall architecture of UIL-AQA is illustrated as above. The input video is segmented into multiple clips and processed by a feature extractor. A temporal decoder utilises query-initialised learnable positionally encoded queries to transform clip-level features into temporal representations. The difficulty-quality regression head computes the final score to ensure interpretability by taking the product of each clip's difficulty and quality. Additionally, a Gaussian noise injection module is applied before computing the loss function to mitigate the impact of subjectivity and human biases. By minimising the similarity between the self-attention and cross-attention maps, along with an optimised query Initialization strategy, our approach effectively mitigates the temporal collapse issue commonly observed in long-term video sequences, enhancing human interpretability. Furthermore, our Gaussian Noise Injection Mean Square Error Loss can minimise the uncertainty and improve robustness and final performance.

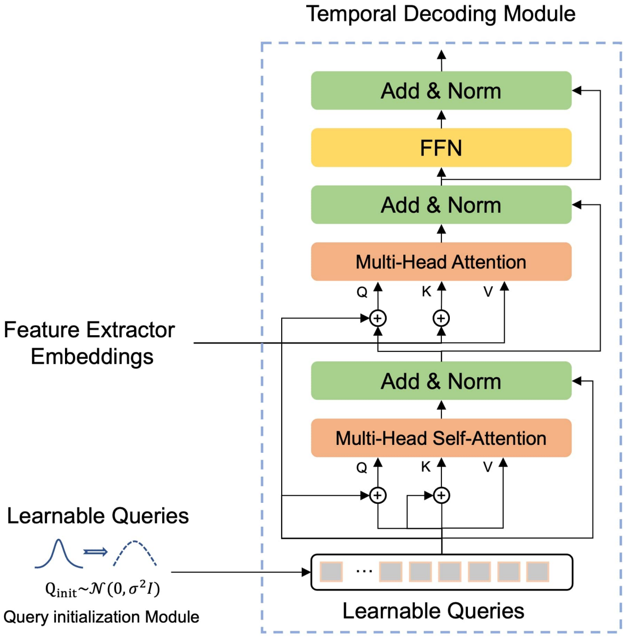

Temporal Decoding Module
Add & Norm
FFN
Add & Norm
Multi-Head Attention
Q
K
V
Feature Extractor Embeddings
Add & Norm
Multi-Head Self-Attention
Q
K
V
Learnable Queries
$Q_{init} \sim \mathcal{N}(0, \sigma^2 I)$
Query initialization Module
Learnable Queries

Fig. 5 Detailed structure of our Temporal Decoding Module and Query Initialization Module.

extract global image features and learnable queries that interact with these features via cross-attention to predict object classes and bounding boxes directly. It offers a simplified, end-to-end approach, allowing it to capture complex relationships between objects in an image. Where learnable queries interact with clip-level video features through cross-attention, adaptively aggregating information instead of uniformly processing all clips while simultaneously capturing long-range temporal dependencies. Unlike vanilla DETR, which includes both an encoder and a decoder, our model uses only a decoder. Prior research (Bai et al., 2022) indicates that an encoder can reduce efficiency without adding significant benefits. This may be due to two factors. First, the pre-trained feature extractor (e.g., I3D or VST) has already extracted high-dimensional spatiotemporal features, which inherently capture local motion patterns and short-term dependencies; an additional encoder may introduce redundant information. Second, the aggregation of clip-level features can lead to excessive smoothing of the temporal representation, weakening the information at the critical query and thus causing the *Temporal Skipping* issue. The decoder consists of layers incorporating self-attention for handling query inputs and cross-attention, enabling queries to interact with encoded clip features. It is designed with two layers, as increasing the number of layers negatively impacts *Temporal Skipping* . With the decoder capturing temporal dependencies, the next step is to compute the final action score. Instead of a simple regression, we introduce a Difficulty-Quality Regression Head that better aligns with human scoring principles. The encoded clip features ${f^i}_{i=1}^L$ are extracted from *Feature Extractor* , while a set of learnable queries is employed. The model is designed to associate each query with a specific clip and its corresponding clip features. The cross-attention mechanism is trained to associate each action query with the memory by computing attention scores between the query features and all memory features. This process enables the model to identify and concentrate on the spatial and temporal information most relevant to each query.

Generally, the queries are initialised with a Gaussian distribution (variance = 1). However, this can weaken correlations in self-attention, reducing temporal coherence. In our query Initialization module, we adjust this variance to ensure stronger query interactions, improving video consistency. Additionally, while vanilla DETR employs *sin* or *cos* positional encodings for queries and memory to encode relational position information, our decoder utilises only query positional encoding. This design choice removes reliance on complex positional encodings or prior knowledge, such as temporal features extracted from the feature extractor. The detailed experiments for this part will be analysed and validated in Section 4.

3.3 Difficulty-Quality Regression Head

Our method essentially addresses a weakly-supervised disentanglement problem of difficulty and quality. As shown by Locatello et al. (2019), unsupervised disentanglement without structural assumptions is theoretically impossible. Without introducing inductive biases or supervision, the factorisation learned by the model cannot be guaranteed to align with meaningful semantic concepts. In practice, the final score of an action video often results from multiple semantic factors: for example, artistic swimming emphasises aesthetic appeal and team coordination, figure skating focuses on fluidity and smooth transitions, while gymnastics highlights difficulty execution and technical precision. However, across different actions, a common principle in scoring lies in a combination of difficulty and execution quality.

Therefore, in our method, we explicitly hypothesise that action quality depends on two latent components—difficulty and execution quality—and design the model accordingly to guide the decoupling process. Therefore, we propose the following assumption: **Assumption:** The overall video score $\hat{y}$ is formulated as a weighted sum of per-clip difficulty and quality, under the assumption that both dimensions jointly contribute to long-term AQA performance.

Based on this assumption, we design a transferable and data-agnostic *Difficulty-Qualtiy Regression Head* module that separates the difficulty and quality of each action clip. This is implemented through two parallel regression heads, a difficulty head and a quality head, designed as MLP layers. The difficulty head applies a softmax function to ensure that the sum of all clips' weights equals 1, representing each clip's relative contribution to the final score. The final quality score Q is computed as the weighted sum of each clip's quality, where the difficulty score D is assigned by the difficulty head, as defined in Equation 1, with L representing the total number of clips.

$$\hat{y} = \sum_{l=1}^{L} D_l \cdot Q_l \tag{1}$$

3.4 Temporal Skipping

Ideally, the difficulty score D and quality score Q should vary across different clips, reflecting the differences in action difficulty and execution quality. However, without constraint, the model tends to skip the self-attention module after several training epochs. Creating the attention map as shown in Figure 3(a), which shows a horizontally uniform distribution rather than a diagonal pattern. The attention map should highlight temporal dependencies by focusing more on relevant clips. As a result, the correlation between queries diminishes, leading to each query receiving a similar weight in the self-attention mechanism. This meant that the difficulty and quality for each clip in a video tended to be averaged to a single score, as shown in Figure 3(b), which is unreasonable. This phenomenon we termed *Temporal Skipping* , which leads to the collapse of meaningful temporal interactions, reducing interpretability. Our proposed Attention Loss and Query Initialization Module explicitly mitigate this issue, ensuring each clip retains its unique temporal contribution.

Attention Loss To solve the *Temporal Skipping* , we introduce the *Attention Loss* $Loss_{att}$. We define an Attention Loss function to ensure consistency between self-attention and cross-attention. We first compute self-attention maps L_S and cross-attention maps L_C by applying a softmax function to their respective attention

Datasets	Year	Modality	#Class	#Sample	#Average Frames
MIT Olympic(Pirsiavash et al., 2014)	2014	Video & Skeleton	2	309	Dive:150, Figure Skate:4200
MTL-AQA(Parmar and Tran Morris, 2019)	2019	Video	16	1412	96
FineDiving(Xu et al., 2022b)	2022	Video	52	3000	105
Fis-V(Xu et al., 2019)	2020	Video	1	500	**4300 (Long-term)**
RG(Zeng et al., 2020)	2020	Video	1	250	**2375 (Long-term)**
LOGO(Zhang et al., 2023)	2023	Video	12	200	**5100 (Long-term)**

Table 1 Comparison of widely used AQA datasets, including publication year, modality, number of classes, sample size, and frame length. underlined denotes the long-term datasets used in our paper.

matrices as shown in Equation 2 and 3 where A_S is the output of self-attention module, and $A_S^\intercal$ is its transpose, and A_C represents the output of cross-attention, while $A_C^\intercal$ is its transpose.

$$L_S = softmax(A_S A_S^\intercal) \tag{2}$$

$$L_C = softmax(A_C A_C^\intercal) \tag{3}$$

We then minimize the KL divergence between these maps across all decoder layers. The formal definition of Attention Loss $Loss_{att}$ is formulated as Equation 4, where D_{KL} represents the Kullback-Leibler (KL) divergence, and N indicates the number of decoder layers in the *Temporal Decoder* . Attention Loss enforces consistency between self-attention and cross-attention across decoder layers. This prevents the model from disregarding temporal dependencies and helps retain fine-grained motion details, helping to alleviate *Temporal Skipping* and improving the interpretability of the model.

$$Loss_{att} = \sum_{n=1}^{N} D_{KL}(L_S^n || L_C^n) \tag{4}$$

3.5 Gaussian Noise Injection

Subjective variations often influence AQA scores in human judgment, and our Gaussian Noise Injection module counteracts this by introducing controlled randomness during training, improving model robustness, and reducing overfitting to specific scoring biases. This prevents overfitting and ensures stable predictions across varied input conditions. This module introduces random perturbations during training, enhancing the model's robustness to varying input conditions and effectively reducing instability caused by data noise or subjective scoring. Specifically, during network training, we inject noise into the model's output score using a Gaussian noise mechanism, where a Gaussian noise term with mean $\mu = 0$ and adjustable variance $\sigma^2 = 0.05$ is added to the predicted score, as shown in Equation 5 where $\hat{y}$ represents the model's original prediction before noise injection, and ϵ is the added Gaussian noise.

$$y = \hat{y} + \epsilon, \quad \epsilon \sim \mathcal{N}(\mu, \sigma^2), \tag{5}$$

3.6 Overall Training Loss

The overall training loss is made up of two terms, Attention Loss $Loss_{att}$ and Gaussian Noise Injection Mean Square Error Loss $Loss_{reg}$, where we improve upon previous AQA research (Yu et al., 2021; Wang et al., 2021a; Zhang et al., 2023; Xu et al., 2022a) by enhancing the vanilla Mean Square Error Loss (MSE Loss) approach with Gaussian Noise Injection. Gaussian Noise Injection Mean Square Error Loss minimizes the difference between the noise-injected regressed score and human judge scores as formulated in Equation 6, where y is the predicted value and $\bar{y}$ is the ground truth value. The overall training loss function is then defined in Equation 7, where λ_{reg} and λ_{Att} represent the constant assigned to the MSE loss and attention loss, respectively. The loss function demonstrates robustness to the selection of constant values, and they are generally kept the same.

$$Loss_{reg} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y}_i)^2 \tag{6}$$

$$Loss_{all} = \lambda_{reg} Loss_{reg} + \lambda_{att} Loss_{att} \tag{7}$$

In summary, our proposed UIL-AQA framework introduces three key innovations: (1) a DETR-inspired transformer decoder tailored for AQA, (2) an interpretable Difficulty-Quality Regression Head for human-aligned scoring, and (3) Attention Loss, Query Initialization Module and Gaussian Noise Injection to mitigate *Temporal Skipping* and to model judge uncertainty and improve robustness. Together, these innovations enhance the accuracy, robustness, and interpretability of AQA, making it more suitable for complex, long-term video analysis.

4 Experiment

4.1 Datasets

To evaluate the effectiveness of our proposed model, we conduct experiments on three widely used long-term AQA benchmarks: Rhythmic Gymnastics (RG) (Zeng

Methods	Feature Extractor	RG (SRCC↑)					RG (MSE)↓				
		Ball	Clubs	Hoop	Ribbon	Avg.	Ball	Clubs	Hoop	Ribbon	Avg.
SVR (Parmar and Morris, 2017)	C3D	0.357	0.551	0.495	0.516	0.483	-	-	-	-	-
MS-LSTM (Xu et al., 2019)	I3D	0.515	0.621	0.540	0.522	0.551	10.55	6.94	5.85	12.56	8.97
	VST	0.621	0.661	0.670	0.695	0.663	7.52	6.04	6.16	**5.78**	6.37
ACTION-NET (Zeng et al., 2020)	I3D+ResNet	0.528	0.652	0.708	0.578	0.623	9.09	6.40	5.93	10.23	7.91
	VST+ResNet	0.684	0.737	0.733	0.754	0.728	9.55	6.36	5.56	8.15	7.41
GDLT (Xu et al., 2022a)	VST	0.746	0.802	0.765	0.741	0.765	5.90	**4.34**	5.70	6.16	**5.53**
T^{2}CR (Ke et al., 2024)	I3D	0.537	0.730	0.645	0.698	0.658	8.67	5.49	6.78	7.07	7.00
CoFInAl (Zhou et al., 2024)	VST	0.809	0.806	0.804	0.810	0.807	**5.07**	5.19	6.37	6.30	5.72
Inter-AQA (Dong et al., 2024)	VST	0.823	0.852	0.837	0.857	0.842	7.94	5.66	7.95	8.87	7.61
Ours	VST	**0.833**	**0.881**	**0.855**	**0.862**	**0.858**	5.90	4.89	**5.05**	6.76	5.65

Table 2 Comparison of Spearman's rank correlation coefficient (SRCC, higher is better) and Mean Squared Error (MSE, lower is better) performance on the **Rhythmic Gymnastics (RG)** dataset. "Avg." denotes the average score across all subclasses (Ball, Clubs, Hoop, and Ribbon). The best results are highlighted in **bold**, while the second-best results are underlined. Missing results are denoted by "-".

Methods	Feature Extractor	Fis-V (SRCC↑)			Fis-V (MSE↓)		
		TES	PCS	Avg.	TES	PCS	Avg.
SVR (Parmar and Morris, 2017)	C3D	0.400	0.590	0.501	-	-	-
MS-LSTM (Xu et al., 2019)	VST	0.660	0.809	0.744	-	-	-
ACTION-NET (Zeng et al., 2020)	VST+ResNet	0.694	0.809	0.757	-	-	-
CoRe (Yu et al., 2021)	VST	0.660	0.820	0.751	23.50	9.25	16.38
GDLT (Xu et al., 2022a)	VST	0.685	0.820	0.761	20.99	8.75	14.87
MLP-Mixer (Xia et al., 2023)	VST	0.680	0.820	0.759	19.57	7.96	13.77
T^{2}CR (Ke et al., 2024)	I3D	0.809	0.702	0.761	-	-	-
SGN (Du et al., 2024)	VST	0.700	0.830	0.773	**19.05**	7.96	**13.51**
CoFInAl (Zhou et al., 2024)	VST	0.716	0.843	0.780	20.76	**7.91**	14.34
Inter-AQA (Dong et al., 2024)	VST	0.717	0.858	0.788	26.97	10.89	18.93
Ours	VST	**0.721**	**0.862**	**0.792**	20.22	8.56	14.39

Table 3 Comparison of Spearman's rank correlation coefficient (SRCC, higher is better) and Mean Squared Error (MSE, lower is better) performance on the **Figure Skating Video (Fis-V)** dataset. "Avg." denotes the average score across all subclasses (TES and PCS). The best results are highlighted in **bold**, while the second-best results are underlined. Missing results are denoted by "-".

et al., 2020), LOng-form GrOup (LOGO) (Zhang et al., 2023), and Figure Skating Video (Fis-V) (Xu et al., 2019) as shown in Table 1. Compared to short-term video datasets, long-term video presents more significant challenges due to extended temporal dependencies, complex action sequences and sparse features.

Rhythmic Gymnastics (RG). (Zeng et al., 2020) The RG dataset is collected from high-standard international competition videos, including footage from Artistic Gymnastics Competitions. The dataset includes video sequences of four gymnastics routines: ball, clubs, hoop, and ribbon. Each action category consists of 200 training and 50 evaluation samples, each lasting approximately 1 minute and 35 seconds. Each sample is assigned three scores: a difficulty score, an execution score, and a total score, given by the referee following the official scoring system. In our paper, each category is trained as an individual model, following the methodology outlined in (Zeng et al., 2020; Xu et al., 2022a).

Figure Skating Video (Fis-V). (Xu et al., 2019) The Fis-V dataset was collected from official high-standard international skating competitions. The dataset contains 500 figure skating videos, each averaging 2 minutes and 50 seconds in length. Following prior work, we use the same train/test split with 400 videos for training and 100 for testing. Fis-V includes two types of labels: Total Element Scores (TES) and Total Program Component Scores (PCS). TES represents the score calculation method for an athlete's technical elements during the competition. PCS represents artistic performance and program composition quality, evaluating a skater's skating skills, musical expression, choreography, and overall presentation. Two models are trained to predict these scores, each focusing on one category.

LOng-form GrOup (LOGO). (Zhang et al., 2023) LOGO is a multi-person, long-term video dataset featuring frame-wise annotations for action procedures and formations designed for artistic swimming scenarios. The LOGO dataset comprises 150 training samples and 50 testing samples. Each video sequence lasts approximately 3 minutes and 30 seconds. To our knowledge, LOGO features the most extended video durations among existing AQA datasets. The long-term and multi-person nature of the LOGO dataset introduces significant challenges for AQA.

4.1.1 Dataset Selection

These datasets exhibit diverse and complementary characteristics, collectively forming a representative and challenging benchmark for AQA. Below, we summarise their key differences and the motivations behind their selection:

- **Action diversity**: The three datasets collectively span a wide range of movement types, including aquatic (e.g., diving), terrestrial (e.g., gymnastics), individual, and group performances. This helps evaluate the model's ability to generalise across different action categories.
- **Viewpoint and scene complexity**: LOGO is a typical multi-person action dataset featuring multi-camera views and higher visual complexity, whereas RG and Fis-V consist of single-person performances captured from fixed viewpoints. This setup allows us to evaluate the model's robustness under varying spatial structures and observation conditions.
- **Temporal range**: Clip lengths range from 2,375 to 5,100 frames, with LOGO providing long sequences and RG/Fis-V offering medium-length clips. This variation helps examine the model's performance across different temporal modelling scales.
- **Scoring schemes**: The datasets employ different evaluation protocols—RG focuses on execution quality, Fis-V incorporates both technical and artistic components, and LOGO uses a composite score that includes group coordination. This allows us to assess whether the model can adapt to task-relevant scoring nuances and fine-grained distinctions.
- **Uniqueness**: To the best of our knowledge, these are the only publicly available long-form AQA datasets, making them essential and representative resources for evaluating long-sequence quality prediction models.

4.2 Evaluation Metrics

To maintain consistency with prior research (Doughty et al., 2019; Yu et al., 2021; Roditakis et al., 2021; Farabi et al., 2022; Zhang et al., 2021; Xu et al., 2025a), we adopt three widely used evaluation metrics: Spearman's rank correlation (SRCC), Relative L2 distance (R-ℓ2), and Mean Squared Error (MSE). These metrics respectively assess the ranking consistency, normalized prediction error, and absolute prediction error of the proposed model.

Spearman's Rank Correlation (SRCC) measures the correlation between the predicted and ground truth scores, ranging from -1 to 1. SRCC reflects how well the model preserves the relative ranking of actions rather than their absolute values. The reason for using SRCC is that many scoring tasks, such as figure skating, diving, and gymnastics, involve a certain degree of subjectivity. Different judges may assign varying scores, but the overall ranking trend should remain consistent, which better aligns with the judging practices in real-world competitions. A higher SRCC indicates a more substantial agreement between the predicted rankings and human-assigned scores, as shown in Equation 8.

$$\rho = \frac{\sum_i (p_i - \bar{p})(q_i - \bar{q})}{\sqrt{\sum_i (p_i - \bar{p})^2 \sum_i (q_i - \bar{q})^2}} \tag{8}$$

In contrast, Relative L2 Distance (R-ℓ2) quantifies the normalised discrepancy between the predicted and ground truth scores. This metric evaluates the absolute Euclidean distance while accounting for variations in score distribution, providing a robust measure of prediction accuracy. Lower values indicate better similarity, as described in Equation 9.

$$R\text{-}\ell 2 = \frac{1}{N} \sum_{N}^{n=1} \left(\frac{|y_n - \hat{y}_n|}{y_{max} - y_{min}}\right)^2 \tag{9}$$

In addition, we also report the Mean Squared Error (MSE), which measures the average squared difference between the predicted and ground truth scores. Unlike R-ℓ2, MSE does not apply normalisation by score range, making it directly reflect the absolute prediction error. Lower values indicate better accuracy, as shown in Equation 10.

$$\text{MSE} = \frac{1}{N} \sum_{n=1}^{N} (y_n - \hat{y}_n)^2 \tag{10}$$

4.3 Implementation Details

We adopt Inflated 3D ConvNet (I3D) (Carreira and Zisserman, 2017), and Video Swin Transformer (VST) (Liu et al., 2022b, 2021) pretrained on Kinetics as video feature extractors. It is important to note that we only use the pretrained *Feature Extractor* for feature extraction and do not train it. For the RG, FIS-V, and LOGO tasks, the clip and query sizes are set to 68, 136, and 48, respectively. These parameters are determined based on video length and extensive comparative experiments demonstrating optimal performance. We adopt the Adam optimiser with an initial learning rate of 1×10^{-4}, a batch size of 48, and a weight decay of 1×10^{-5}. The Transformer decoder has an output dimension of 1024, with each decoder containing 4 attention heads and 2 layers, and each layer applying a dropout rate of 0.8 to effectively mitigate overfitting, thereby improving the model's generalisation capability. Regarding the difficulty-quality regression head, each module consists of a three-layer MLP network, which progressively reduces the feature dimension from 1024 to 512, then to 256 as the final output dimension. Each MLP layer applies ReLU activation to introduce non-linearity. The final layer includes a softmax activation in the difficulty branch, ensuring that the outputs form a probability distribution. This facilitates the proper weighting of different action components during the scoring process. For the Gaussian noise injection module, we empirically determine the optimal variance, with $\sigma = 0.05$ yielding the best performance. The weight of $Loss_{reg}$: λ_{reg} and $Loss_{att}$: λ_{att} in our final objective are both set to 1.

We train our network for 1000 epochs on a single NVIDIA RTX 3090 GPU, with a total training time of approximately 20 minutes for one action class. We employ StepLR as the learning rate scheduler to adjust the learning rate dynamically, where the learning rate is decayed by a factor of 0.1 every 400 epochs.

4.4 Results and Analysis

Our proposed model achieves state-of-the-art performance across all three datasets, significantly improving upon previous methods. In particular, we emphasise SRCC as the primary evaluation metric, as it reflects the ranking consistency between predicted and ground-truth scores, which is especially critical for long-term AQA tasks. Under this metric, we observe substantial gains, demonstrating the effectiveness of our interpretability-driven approach. For completeness, we also report absolute error metrics such as R-$\ell 2$ and MSE as complementary measures.

For the Rhythmic Gymnastics (RG) dataset, as shown in Table 2. We compared with our previous proposed model Interpretable-AQA (Dong et al., 2024), the adoption of our proposed Gaussian noise injection method and new network settings has led to improvements in the SRCC scores across all four sub-labels and the overall averaged SRCC. Furthermore, compared to the previous state-of-the-art method, CoFInAI (Zhou et al., 2024), which focuses on fine-grained local interactions but lacks an explicit difficulty-quality scoring mechanism, our approach not only benefits from interpretable clip-level regression but also achieves a 6.32% improvement in overall SRCC on RG, along with gains of 2.97%, 9.31%, 6.34%, and 6.42% across the four sub-labels, respectively. In addition to SRCC, we also compare the Mean Squared Error (MSE) values on the RG dataset. This provides a complementary perspective by directly measuring the absolute prediction error. Our method achieves consistently lower MSE across most sub-classes and the averaged score, further demonstrating the robustness of our approach.

On the Figure Skating Video (Fis-V) dataset, as shown in Table 3, our model outperforms our previous model Inter-AQA (Dong et al., 2024) by 0.5%. Compared to the prior state-of-the-art method CoFInAl Zhou et al. (2024), the results demonstrate improvements of 0.5% and 1.9% on the TES and PCS labels, respectively, achieving an average enhancement of 1.2% on SRCC. On the Fis-V dataset, we also report the MSE results. The results show that our method achieves comparable performance on TES, PCS, and the averaged scores, further indicating that the improvements in SRCC are not achieved at the expense of higher prediction errors.

For the LOGO dataset, as shown in Table 4, our model achieves state-of-the-art results, surpassing our previous work (Dong et al., 2024) by 2.1% and outperforming the previous state-of-the-art method by 26.5% when using I3D as the feature extractor. Furthermore, with VST as the feature extractor, our model exceeds our previous method by 2% and outperforms the previous state-of-the-art method by 14%. This strong performance validates our UIL-AQA framework and suggests its applicability to broader action quality assessment scenarios. For the LOGO dataset, as shown in Table 4, our model achieves state-of-the-art SRCC results, surpassing our previous work (Dong et al., 2024) by 2.1% and outperforming the previous state-of-the-art method by 26.5% when using I3D as the feature extractor. Furthermore, with VST as the feature extractor, our model exceeds our previous method by 2% and outperforms the previous state-of-the-art method by 14%. This strong performance validates our UIL-AQA framework and suggests its applicability to broader action quality assessment scenarios. We also observe that SRCC exhibits a different trend from R-$\ell 2$ in Table 4. This is because SRCC reflects the monotonic consistency between predicted and ground-truth scores, while

Methods	Feature Extractor			
	I3D		VST	
	SRCC ↑	R-ℓ2 ↓	SRCC ↑	R-ℓ2 ↓
USDL (Tang et al., 2020)	0.426	5.736	0.473	5.076
CoRe (Yu et al., 2021)	0.471	5.402	0.500	5.960
TSA (Xu et al., 2022b)	0.452	5.533	0.475	4.778
ACTION-NET (Zeng et al., 2020)	0.306	5.858	0.410	5.569
USDL-GOAT (Zhang et al., 2023)	0.462	4.874	0.535	5.022
TSA-GOAT (Zhang et al., 2023)	0.486	5.394	0.484	5.409
CoRe-GOAT (Zhang et al., 2023)	0.494	5.072	0.560	4.763
CoFInAl (Zhou et al., 2024)	-	-	0.698	4.019
Inter-AQA (Dong et al., 2024)	0.593	**1.220**	0.780	**1.745**
Ours	**0.625**	4.107	**0.796**	3.084

Table 4 Performance comparison on the **LOGO** dataset using feature extractors including I3D (Carreira and Zisserman, 2017) and VST (Liu et al., 2022b). A higher SRCC and a lower R-ℓ2 indicate better performance. The highest results are highlighted in **bold**, while the second-highest results are marked with an underline.

R-ℓ2 penalises absolute deviations. As a result, some methods may achieve lower R-ℓ2 if their predictions align more closely with the absolute scale of the data, even if the relative ordering is less reliable. In contrast, SRCC highlights improvements in ranking quality, where our approach demonstrates consistent advantages, both in Table 4 and previously in Tables 2 and 3.

4.4.1 Ablation Study

Experiments were conducted in four different settings to evaluate the impact of attention loss, query positional encoding, query initialization and Gaussian noise injection on model performance. Our ablation study in Table 5 reveals that attention loss alone contributes a significant 28.5% improvement in SRCC, highlighting the impact of mitigating temporal skipping. Adding query positional encoding further refines long-term dependencies, while Gaussian noise injection stabilises predictions against subjective scoring biases. With the baseline approach of a vanilla DETR decoder without incorporating any of the other proposed modules, the SRCC of the model was 0.628. The introduction of attention loss resulted in a significant performance improvement, boosting the SRCC to 0.807 (28.5%), which demonstrates the effectiveness of our proposed attention loss in enhancing both interpretability and the overall SRCC results. Adding query positional encoding alone further improved the results, bringing the SRCC to 0.810. Furthermore, including the query initialization module provided an additional performance boost of approximately 4%. This enhancement suggests that utilising high variance in the initialization process promotes greater diversity and dispersion of query vectors, improving the model's overall performance. Lastly, our proposed noise injection module slightly enhanced the final results by 2% while eliminating the influence of subjectivity and uncertainty, ensuring that the outcomes were more consistent and reliable. In summary, each module contributes to the overall performance improvement of the model. Attention loss plays a crucial role in mitigating the temporal skipping issue, while query initialization and noise injection respectively enhance long-term dependency modelling and prediction robustness. More detailed ablation study results can be found in Table 6.

4.4.2 Ablation study of Gaussian Noise Injection

We compare the performance of various uncertainty-aware modules in AQA with our proposed Gaussian noise injection strategy. As shown in Table 7, our simple yet effective approach achieves better performance across correlation-based metrics SRCC, demonstrating its ability to model subjective uncertainty in action quality assessment.

4.4.3 Effect of position encoding

We compare different positional encoding methods in the *Temporal Decoder* in Table 8. Positional encoding is used to incorporate spatial and sequential information into the transformer architecture, enabling the model to distinguish between different positions in the input clips. In our study, we compare the effects of applying positional encoding at different locations, including transformer query and memory. We observe that using only query positional encoding outperforms all other

Module	#Attention Loss	#Query PE	#Query Initialization	#Noise Injection	SRCC ↑
Baseline	×	×	×	×	0.628
	✓	×	×	×	0.807
	✓	✓	×	×	0.810
	✓	✓	✓	×	0.842
Ours	✓	✓	✓	✓	**0.858**

Table 5 **Ablation study** on the average performance of four labels in the Rhythmic Gymnastics (RG) dataset across various modules.

Ball					
Noise Injection	×	×	×	×	✓
Query Init.	×	×	×	✓	✓
Query PE	×	×	✓	✓	✓
Attention Loss	×	✓	✓	✓	✓
Results	0.483	0.820	0.823	0.823	**0.833**

Clubs					
Noise Injection	×	×	×	×	✓
Query Init.	×	×	×	✓	✓
Query PE	×	×	✓	✓	✓
Attention Loss	×	✓	✓	✓	✓
Results	0.6346	0.813	0.809	0.852	**0.881**

Hoop					
Noise Injection	×	×	×	×	✓
Query Init.	×	×	×	✓	✓
Query PE	×	×	✓	✓	✓
Attention Loss	×	✓	✓	✓	✓
Results	0.626	0.761	0.770	0.837	**0.855**

Ribbon					
Noise Injection	×	×	×	×	✓
Query Init.	×	×	×	✓	✓
Query PE	×	×	✓	✓	✓
Attention Loss	×	✓	✓	✓	✓
Results	0.7669	0.835	0.837	0.857	**0.862**

Table 6 Detailed Ablation study on the individual SRCC performance of four labels in the Rhythmic Gymnastics (RG) dataset across various modules.

approaches, while combining query and memory positional encoding negatively impacts performance. This behaviour can be attributed to our main focus in the AQA task, modelling learnable queries through the DETR decoder and assigning temporal semantic meanings to these queries within the decoder structure. Since the *Feature Extractor* has already extracted the temporal information in the memory, incorporating memory positional encoding introduces unnecessary computations and potential redundancy. We streamline the process by utilising only the query positional encoding, avoiding this redundancy. This strategy allows the queries to capture and represent key action quality indicators within the video more effectively, ultimately improving both scoring accuracy and the model's interpretability.

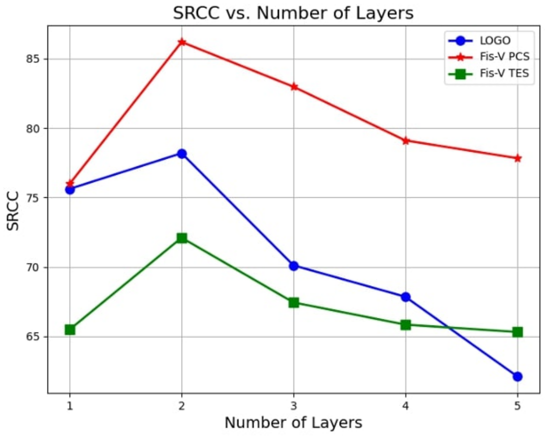


Fig. 6 Effect of different transformer layers on the LOGO and Fis-V dataset results.

4.4.4 Effect of Transformer Layers

In our network, Transformer decoder layers are key to capturing long-term dependencies and refining outputs through self-attention and cross-attention. More layers help aggregate information over time, potentially improving performance, but excessive layers may cause temporal skipping, where useful self-attention interactions are bypassed, leading to degradation. We evaluate the effectiveness using the different number of decoder layers. Figure 6 presents the impact of varying decoder depths on the LOGO and Fis-V dataset. On the LOGO dataset, the optimal performance is obtained with a depth of 2, yielding a result of 78.19, while increasing the number of decoder layers may result in deteriorated temporal skipping. A reduced number of layers might limit the model's ability to fully leverage the available temporal information, resulting in suboptimal data representations. Similarly, the PCS and TES labels achieve the highest SRCC on the Fis-V dataset using two depth layers.

4.4.5 Effect of variance in query initialization module

The Query Initialization Module is proposed for initializing *Temporal Decoder* query embeddings. We found

Model	Uncertainty Module	SRCC	Rl-2
LUSD-NET (Kendall and Gal (2017); Ji et al. (2023))	Aleatoric uncertainty modelling	0.738	**1.130**
MUSDL (Tang et al. (2020))	Distribution learning	0.709	2.840
UD-AQA (Zhou et al. (2022))	CVAE based Module	0.754	11.055
Ours	Gaussian noise injection	**0.796**	3.084

Table 7 Comparison of different uncertainty modelling modules adapted to our network on LOGO dataset. We evaluate performance using SRCC and Rl-2. Best results are shown in bold.

Methods	Query	Memory	SRCC
Baseline	×	×	0.783
	×	✓	0.751
	✓	✓	0.629
Ours	✓	×	**0.858**

Table 8 Effect of Positional Encoding on RG dataset, where SRCC results take the average of the four labels.

Variance	Ball	Clubs	Hoop	Ribbon	Avg.
0.1	**0.833**	0.821	0.821	0.838	0.828
0.5	0.814	0.814	0.826	0.845	0.825
1 (Default)	0.823	0.774	0.788	0.849	0.809
2	0.814	0.856	0.779	**0.862**	0.828
3	0.796	0.857	0.765	0.861	0.820
5	0.785	**0.881**	0.782	0.839	0.822
10	0.794	0.798	0.812	0.812	0.804
20	0.778	0.798	**0.855**	0.849	0.820

Table 9 Effect of Query Variance Initialization on Rhythmic Gymnastics (RG) dataset, where Avg. SRCC results take the average of the four labels.

Variance	PCS	TES	Avg.
0.1	0.830	**0.721**	0.776
0.5	0.809	0.703	0.756
1 (Default)	0.784	0.690	0.737
2	**0.862**	0.685	0.774
3	0.830	0.707	0.767
5	0.795	0.708	0.752
10	0.780	0.681	0.731

Table 10 Effect of Query Variance Initialization on Figure Skating Video (Fis-V) dataset, where Avg. SRCC results take the average of the two labels.

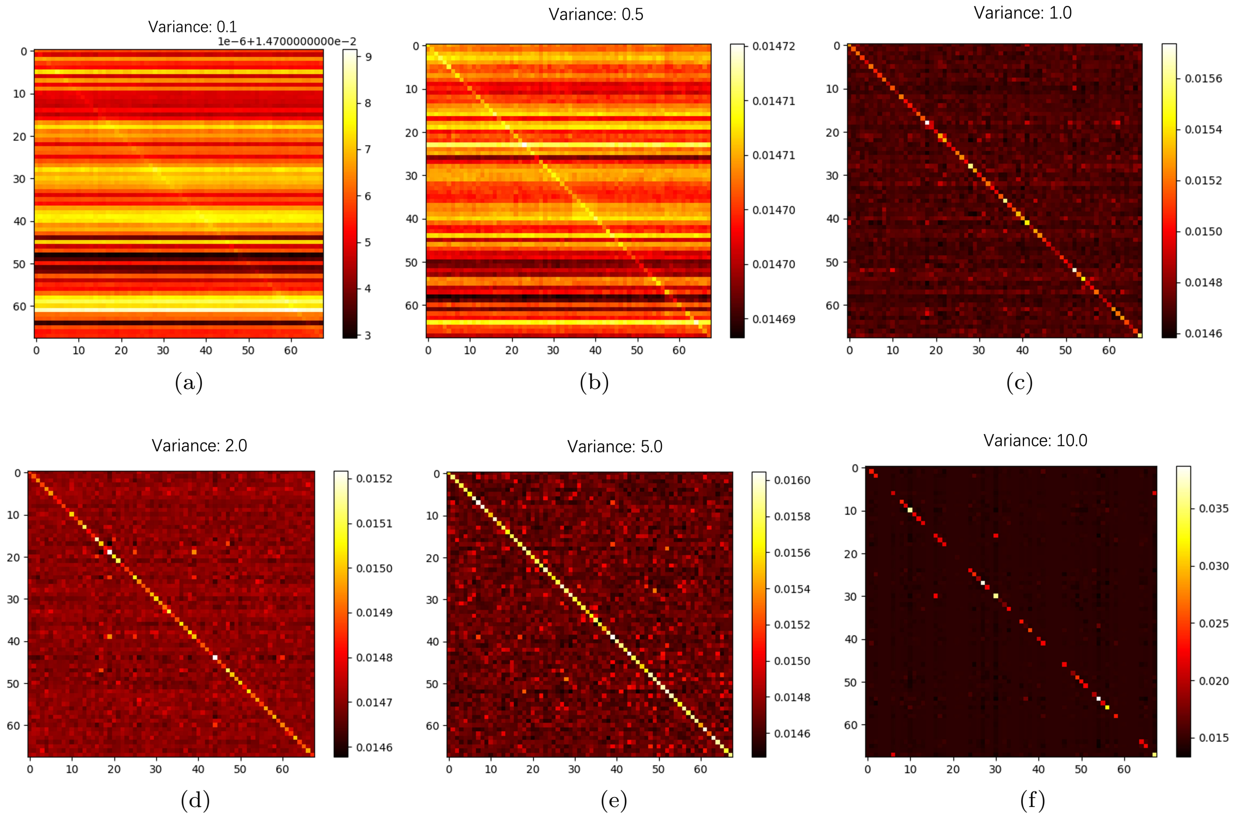
Variance	SRCC
0.1	0.782
0.5	**0.796**
1 (Default)	0.708
2	0.727
3	0.650
5	0.618
10	0.624

Table 11 Effect of Query Variance Initialization on LOGO dataset, where SRCC results take the average of the four labels.

that using different variances to initialize query embedding can effectively mitigate the temporal skipping issue and enhance interpretability. We analyse the impact of query initialization variance and the final SRCC results. As shown in Figure 7, for the self-attention map and mitigating temporal skipping, initializing the query embedding with a larger variance, results in a more compact diagonal pattern in the self-attention map, indicating a stronger correlation between action queries. This is because a higher variance injects greater semantic diversity into the initial query embeddings, preventing them from collapsing into strictly time-aligned (diagonal) attention. When the variance is too small, all queries start nearly identical and remain confined to local temporal regions, making the model prone to temporal skipping. In effect, a large variance also acts as a regulariser that encourages cross-time exploration and richer long-range associations. This ensures that each clip is assigned meaningful features rather than simple mean values, thereby improving the interpretability of the network. Regarding SRCC performance, experimental comparisons with different variance values, as presented in Tables 9, 10, and 11, show the results on the RG, Fis-V, and LOGO datasets, respectively. The results indicate that using a relatively larger variance normally yields the highest SRCC scores across all three datasets.

4.4.6 Interpretability

To improve the single-score regression method and follow the scoring logic of human judges, we disentangled each clip's score into difficulty and quality. As shown in Figure 8, we visualise the clip-level difficulty-quality regression results for synchronised swimmers in the LOGO dataset. In this visualisation, the blue curve represents the difficulty of the current frame, the green curve represents the quality, and the red curve indicates the overall score. Although the action has a high diffi-



(a) (b) (c)

(d) (e) (f)

Fig. 7 Self-attention maps for queries initialised with varying variances. Figures 7(a), 7(b), 7(c), 7(d), 7(e), and 7(f) depict the self-attention maps corresponding to variances of 0.1, 0.5, 1, 2, 5 and 10, respectively.

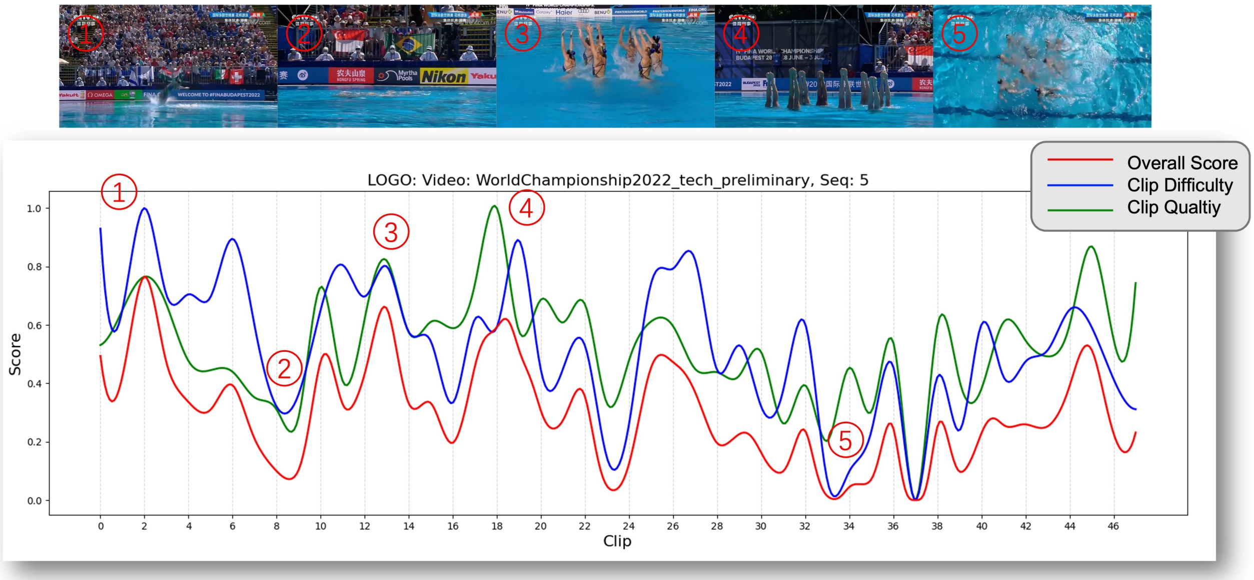


Fig. 8 Visualisation of our clip-level difficulty-quality regression method on LOGO dataset.

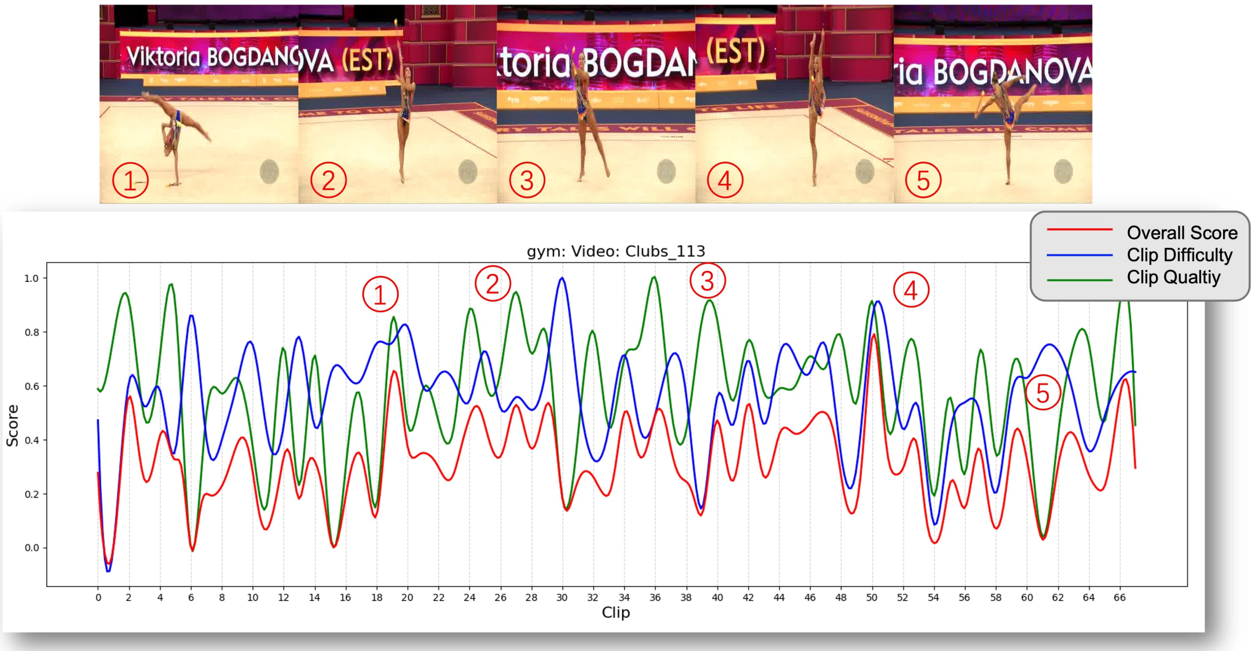


Fig. 9 Visualisation of our clip-level difficulty-quality regression method on Rhythmic Gymnastics (RG) dataset.

culty in the first clip, its execution quality is average. In the second clip, no significant action is observed in the video. In the fourth clip, the athlete performs the ballet leg movement in synchronised swimming, demonstrating high difficulty and execution quality. In summary, Figure 8 shows that our model correctly assigns higher difficulty weights to complex synchronised swimming moves while down-weighting simple transitions. This aligns well with human scoring logic and improves interpretability over black-box single-score regression models. Similarly, as shown in Figure 9, we present the performance of the *Clubs* action category in the RG dataset. In the first clip, the athlete executes a *Valdez flip*, a movement with high difficulty and high execution quality. In the third clip, the athlete performs a simple movement with low difficulty but high execution quality. In the fifth clip, a mistake occurs, leading to a significantly low score.

Empirical results demonstrate that our difficulty-quality regression module can provide clip-level scores and enhance the model's interpretability. Furthermore, by separately modelling difficulty and quality factors, this module enables a more fine-grained evaluation of action performance and offers assessments that align with the scoring logic of human judges.

In addition to providing explicit difficulty-quality decomposition for each clip, our model also exhibits temporal interpretability through attention mechanisms. We visualise the cross-attention weights from the Temporal Decoding Module to illustrate how the model distributes attention across different temporal segments during decoding. The visualisation reveals that the model dynamically allocates higher attention scores to more discriminative and semantically important clips in the input sequence. Specifically, as shown in Figure 8, in examples 1, 3, and 5, when the action segments involve high-difficulty or high-weight movements (e.g., ballet leg or lift movements in synchronised swimming), the Transformer cross-attention module assigns higher attention scores. In contrast, for less informative segments such as preparatory movements (examples 2 and 4), the attention scores are significantly lower. This suggests that the model attends to semantically important regions in a manner that is aligned with human judgment, contributing to post-hoc interpretability and human-aligned reasoning.

4.5 User Study

To rigorously evaluate the effectiveness of our UIL-AQA interpretability in assessing difficulty and quality in video sequences, we conducted a user study in which participants compared pairs of motion video clips. Each pair consisted of two clips, scored by our model based on their predicted difficulty and quality levels. Based on their perception, participants were then asked to determine which video in each pair exhibited a higher level of difficulty or quality. Their responses were compared against our model's predictions to assess the de-

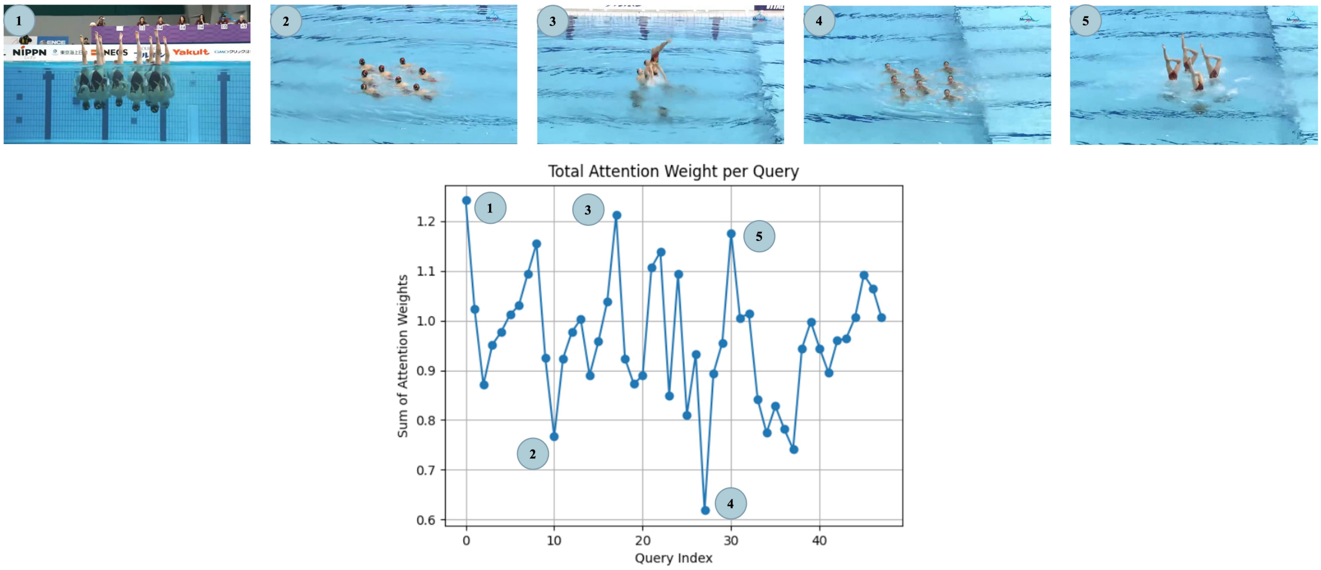


Fig. 10 Cross attention weight visualisation with video clip on LOGO dataset. Examples 1, 3, and 5 show high attention actions, whereas examples 2 and 4 show low attention actions.

gree of alignment between human judgment and our automated scoring system.

This study aimed to validate our model's ability to provide interpretable, human-aligned assessments of video difficulty and quality within an activity. Notably, the model was never explicitly shown the labels for individual clips, only the overall activity and corresponding score. A total of 15 participants took part in the study, each evaluating 10 pairs of motion video clips. The experiment (as summarised in Table 12) measured difficulty agreement, quality agreement, and the overall model-human agreement.

Specifically, difficulty agreement quantified the alignment between the model's difficulty scores and human judgments, yielding an agreement rate of 66.67%. Similarly, the quality agreement measured the alignment between the model's quality scores and human judgments, achieving a rate of 79.17%. The final Model-Human Agreement, which combines both quality and difficulty assessments, was 73.33%, representing the overall agreement result. These results indicate a strong correlation between our model's predictions and human perception, demonstrating that the model's interpretable scoring mechanism produces intuitive and meaningful outputs. This reinforces its reliability as a tool for evaluating video difficulty and quality in a manner consistent with human intuition.

Number of Participants	15
Total Video Pairs Evaluated	10*15=150
Difficulty Agreement (%)	66.67%
Quality Agreement (%)	79.17%
Model-human Agreement (%)	73.33%

Table 12 User Study Results on the Effectiveness of Our Network Interpretability. Model-human agreement is the overall consistency between difficulty and quality.

4.6 Efficiency Analysis

As shown in Table 13, we evaluate our model using a single input sample with 48 tokens of 1024-dimensional features, and measure the average inference time over 100 runs on a single NVIDIA RTX 3090 GPU. While our model has slightly higher FLOPs (0.27G vs. 0.05G) and parameter count (5.52M vs. 1.73M) compared to the baseline CoFInAl (Zhou et al. (2024)), it achieves a faster inference speed (1.37ms vs. 1.66ms) — a 0.29ms improvement. More importantly, our model yields substantial performance gains: a +0.098 improvement in SRCC and a +0.935 improvement in Rl-2. In addition, we conducted an ablation analysis of resource trade-offs by reducing the number of transformer layers and the input token length. Even with only 1 transformer layer or 24 input tokens, our model still achieves a high SRCC of 0.779 and 0.752, respectively, clearly demonstrating that the architecture maintains strong performance under constrained resources.

5 Conclusion

Our UIL-AQA advances long-term AQA by integrating interpretability into temporal modelling, significantly outperforming existing methods. By explicitly captur-

Model	FLOPs (G)	Params (M)	Inference Time (ms)	SRCC	Rl-2
ACTION-NET (Zeng et al. (2020))	2.00	3.54	**0.20**	0.410	5.569
CoRe-GOAT (Zhang et al. (2023))	109.00	25.21	27.56	0.560	4.763
USDL-GOAT (Zhang et al. (2023))	116.94	40.21	28.95	0.535	5.022
TSA-GOAT (Zhang et al. (2023))	1.85	37.95	2.14	0.560	5.409
T^{2}CR (Ke et al. (2024))	446.03	12.90	83.11	0.607	4.254
CoFInAl (Zhou et al. (2024))	**0.05**	**1.73**	1.66	0.698	4.019
Ours (48 tokens)	0.27	5.52	1.37	**0.796**	**3.084**
Ours (36 tokens)	0.20	5.52	2.42	0.724	5.601
Ours (24 tokens)	0.13	5.52	2.53	0.752	7.006
Ours (1 transformer layer)	0.16	3.42	0.88	0.779	9.767

Table 13 Comparison of model efficiency on the LOGO dataset. All models use VST as the backbone. We report FLOPs, number of parameters, and inference time per sample. **Bold** numbers indicate the best performance in each column.

ing clip-level difficulty and quality, we provide a more transparent and reliable scoring mechanism, setting a new benchmark for future AQA research. We introduce a new Attention Loss function and a Query Initialization Module while exploring the impact of different positional encodings. To further reduce the effect of uncertainty on scoring stability, we propose a simple Gaussian noise injection module, which simulates human biases. In addition, we introduce a Difficulty-Quality Regression Module that decouples each clip's action score into difficulty and quality, enabling a fine-grained and interpretable assessment and making AQA scoring more meaningful and informative. Experimental results demonstrate that our approach achieves state-of-the-art performance on three AQA benchmark datasets, validating from both qualitative and quantitative perspectives that our model effectively parses clip-level semantic meanings.

5.1 Future Work

In future work, we plan to invite expert judges to evaluate the interpretability of our model, making the assessment process more comprehensive and accessible for analysis while increasing the fine-grained nature of the results. Furthermore, we plan to address the issue of fragmented evaluation from fixed-length segmentation. Possible directions include using adaptive clip lengths to match semantic subactions better and applying temporal smoothness regularisation to ensure continuity across segments. We believe these improvements can further enhance the interpretability and robustness of clip-level predictions.

References

Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., and Schmid, C. (2021). Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)*, pages 6836–6846.

Ashutosh, K., Nagarajan, T., Pavlakos, G., Kitani, K., and Grauman, K. (2025). Expertaf: Expert actionable feedback from video.

Bai, Y., Zhou, D., Zhang, S., Wang, J., Ding, E., Guan, Y., Long, Y., and Wang, J. (2022). Action quality assessment with temporal parsing transformer. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 422–438.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., and Zagoruyko, S. (2020). End-to-end object detection with transformers. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 213–229.

Carreira, J. and Zisserman, A. (2017). Quo vadis, action recognition? a new model and the kinetics dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4733.

Chen, Z., Sun, W., Tian, Y., Jia, J., Zhang, Z., Wang, J., Huang, R., Min, X., Zhai, G., and Zhang, W. (2024). Gaia: Rethinking action quality assessment for ai-generated videos. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Dong, X., Liu, X., Li, W., Adeyemi-Ejeye, A., and Gilbert, A. (2024). Interpretable long-term action quality assessment. In *Proceedings of the British Machine Vision Conference (BMVC)*.

Doughty, H., Damen, D., and Mayol-Cuevas, W. (2018). Who's better? who's best? pairwise deep ranking for skill determination. In *Proceedings of the IEEE Con-*

ference on Computer Vision and Pattern Recognition (CVPR).

Doughty, H., Mayol-Cuevas, W., and Damen, D. (2019). 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 (CVPR).*

Du, Z., He, D., Wang, X., and Wang, Q. (2023). Learning semantics-guided representations for scoring figure skating. *IEEE Transactions on Multimedia.*

Du, Z., He, D., Wang, X., and Wang, Q. (2024). Learning semantics-guided representations for scoring figure skating. In *IEEE Transactions on Multimedia (TMM)*, pages 4987–4997.

Farabi, S., Himel, H., Gazzali, F., Hasan, M. B., Kabir, M. H., and Farazi, M. (2022). Improving action quality assessment using weighted aggregation. In *Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA)*, pages 576–587.

Funke, I., Mees, S. T., Weitz, J., and Speidel, S. (2019). Video-based surgical skill assessment using 3d convolutional neural networks. *International journal of computer assisted radiology and surgery*, 14:1217–1225.

Gao, J., Sun, C., Yang, Z., and Nevatia, R. (2017). Tall: Temporal activity localization via language query. In *Proceedings of the IEEE international conference on computer vision (ICCV)*, pages 5267–5275.

Gao, Y., Vedula, S. S., Reiley, C. E., Ahmidi, N., Varadarajan, B., Lin, H. C., Tao, L., Zappella, L., Béjar, B., Yuh, D. D., et al. (2014). Jhu-isi gesture and skill assessment working set (jigsaws): A surgical activity dataset for human motion modeling. In *MICCAI workshop*, volume 3, page 3.

Han, R., Zhou, K., Atapour-Abarghouei, A., Liang, X., and Shum, H. P. H. (2025). Finecausal: A causal-based framework for interpretable fine-grained action quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW).*

Ji, Y., Ye, L., Huang, H., Mao, L., Zhou, Y., and Gao, L. (2023). Localization-assisted uncertainty score disentanglement network for action quality assessment. In *Proceedings of ACM International Conference on Multimedia (ACM MM)*, pages 8590–8597.

Ke, X., Xu, H., Lin, X., and Guo, W. (2024). Two-path target-aware contrastive regression for action quality assessment. *Information Sciences*, 664:120347.

Kendall, A. and Gal, Y. (2017). What uncertainties do we need in bayesian deep learning for computer vision? In *Proceedings of Neural Information Processing Systems (NIPS).*

Kim, J., Lee, M., and Heo, J.-P. (2023). Self-feedback detr for temporal action detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10286–10296.

Li, J., Xue, J., Cao, R., Du, X., Mo, S., Ran, K., and Zhang, Z. (2024). Finerehab: A multi-modality and multi-task dataset for rehabilitation analysis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3184–3193.

Li, W., Li, X., Li, B., Wang, S., Ma, L., Liu, Y., and Shi, Z. (2023). Label-reconstruction-based pseudo-subscore learning for action quality assessment in sporting events. *Applied Intelligence*, 53:16191–16207.

Li, X., Song, J., Gao, L., Liu, X., Huang, W., He, X., and Gan, C. (2019a). Beyond rnns: Positional self-attention with co-attention for video question answering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 8658–8665.

Li, Y.-M., Wang, A.-L., Lin, K.-Y., Tang, Y.-M., Zeng, L.-A., Hu, J.-F., and Zheng, W.-S. (2025). Techcoach: Towards technical-point-aware descriptive action coaching.

Li, Z., Huang, Y., Cai, M., and Sato, Y. (2019b). Manipulation-skill assessment from videos with spatial attention network. In *Proceedings of the IEEE/CVF international conference on computer vision workshops (ICCVW)*, pages 4385–4395.

Liu, X., Wang, Q., Hu, Y., Tang, X., Zhang, S., Bai, S., and Bai, X. (2022a). End-to-end temporal action detection with transformer. *IEEE Transactions on Image Processing*, 31:5427–5441.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9992–10002.

Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., and Hu, H. (2022b). Video swin transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3202–3211.

Locatello, F., Bauer, S., Lucic, M., Rätsch, G., Gelly, S., Schölkopf, B., and Bachem, O. (2019). Challenging common assumptions in the unsupervised learning of disentangled representations.

Moon, W., Hyun, S., Park, S., Park, D., and Heo, J.-P. (2023). Query-dependent video representation for moment retrieval and highlight detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23023–23033.

Okamoto, L. and Parmar, P. (2024). Hierarchical neurosymbolic approach for comprehensive and explainable action quality assessment.

Pan, J.-H., Gao, J., and Zheng, W.-S. (2019). Action assessment by joint relation graphs. In *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)*, pages 6331–6340.

Parmar, P., Gharat, A., and Rhodin, H. (2022). Domain knowledge-informed self-supervised representations for workout form assessment. In *Proceedings of the European Conference on Computer Vision (ECCV),*, pages 105–123. Springer.

Parmar, P. and Morris, B. (2017). Learning to score olympic events. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 76–84.

Parmar, P. and Morris, B. (2019). Action quality assessment across multiple actions. In *Proceedings of the IEEE winter conference on applications of computer vision (WACV)*, pages 1468–1476.

Parmar, P. and Tran Morris, B. (2019). What and how well you performed? a multitask learning approach to action quality assessment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 304–313.

Pirsiavash, H., Vondrick, C., and Torralba, A. (2014). Assessing the quality of actions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 556–571.

Qiu, Z., Yao, T., and Mei, T. (2017). Learning spatio-temporal representation with pseudo-3d residual networks. In *proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 5533–5541.

Ren, S., Yao, L., Li, S., Sun, X., and Hou, L. (2023). Timechat: A time-sensitive multimodal large language model for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, volume abs/2312.02051.

Roditakis, K., Makris, A., and Argyros, A. (2021). Towards improved and interpretable action quality assessment with self-supervised alignment. In *Proceedings of the PErvasive Technologies Related to Assistive Environments Conference*, pages 507–513.

Sharma, Y., Bettadapura, V., Hammerla, N., Mellor, S., McNaney, R., Olivier, P., Deshmukh, S., McCaskie, A., Essa, I., et al. (2014). Video based assessment of osats using sequential motion textures. In *Workshop on Modeling and Monitoring of Computer Assisted Interventions 2014*. Springer.

Song, E., Chai, W., Wang, G., Zhang, Y., Zhou, H., Wu, F., Chi, H., Guo, X., Ye, T., Zhang, Y., Lu, Y., Hwang, J.-N., and Wang, G. (2024). Moviechat: From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).*

Srivastava, N., Mansimov, E., and Salakhutdinov, R. (2016). Unsupervised learning of video representations using lstms. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, volume 37, pages 843–852.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).*

Tang, Y., Ni, Z., Zhou, J., Zhang, D., Lu, J., Wu, Y., and Zhou, J. (2020). Uncertainty-aware score distribution learning for action quality assessment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 9839–9848.

Tran, D., Bourdev, L., Fergus, R., Torresani, L., and Paluri, M. (2015). Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision (ICCV)*, pages 4489–4497.

Venkataraman, V., Vlachos, I., and Turaga, P. K. (2015). Dynamical regularity for action analysis. In *Proceedings of the British Machine Vision Conference (BMVC)*, volume 67, pages 1–12.

Wang, S., Yang, D., Zhai, P., Chen, C., and Zhang, L. (2021a). Tsa-net: Tube self-attention network for action quality assessment. In *Proceedings of the 29th ACM international conference on multimedia (ACM MM)*, pages 4902–4910.

Wang, T., Wang, Y., and Li, M. (2020). Towards accurate and interpretable surgical skill assessment: A video-based method incorporating recognized surgical gestures and skill levels. In *Proceedings of Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 668–678. Springer.

Wang, T., Zhang, R., Lu, Z., Zheng, F., Cheng, R., and Luo, P. (2021b). End-to-end dense video captioning with parallel decoding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6847–6857.

Wang, T., Zhang, R., Lu, Z., Zheng, F., Cheng, R., and Luo, P. (2021c). End-to-end dense video captioning with parallel decoding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6847–6857.

Wnuk, K. and Soatto, S. (2010). Analyzing diving: A dataset for judging action quality. In *Asian con-*

ference on computer vision (ACCV), pages 266–276. Springer.

Wu, C.-Y., Li, Y., Mangalam, K., Fan, H., Xiong, B., Malik, J., and Feichtenhofer, C. (2022). Memvit: Memory-augmented multiscale vision transformer for efficient long-term video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13587–13597.

Xia, J., Zhuge, M., Geng, T., Fan, S., Wei, Y., He, Z., and Zheng, F. (2023). Skating-mixer: Long-term sport audio-visual modeling with mlps. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 5140–5148.

Xiang, X., Tian, Y., Reiter, A., Hager, G. D., and Tran, T. D. (2018). S3d: Stacking segmental p3d for action quality assessment. In *25th IEEE international conference on image processing (ICIP)*, pages 928–932. IEEE.

Xu, A., Zeng, L.-A., and Zheng, W.-S. (2022a). Likert scoring with grade decoupling for long-term action assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3232–3241.

Xu, C., Fu, Y., Zhang, B., Chen, Z., Jiang, Y.-G., and Xue, X. (2019). Learning to score figure skating sport videos. *IEEE transactions on circuits and systems for video technology*, 30(12):4578–4590.

Xu, H., Ke, X., Wu, H., Xu, R., Li, Y., and Guo, W. (2025a). Language-guided audio-visual learning for long-term sports assessment. In *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23967–23977.

Xu, H., Ke, X., Wu, H., Xu, R., Li, Y., Xu, P., and Guo, W. (2025b). Dancefix: An exploration in group dance neatness assessment through fixing abnormal challenges of human pose. In *In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 8869–8877.

Xu, H., Wu, H., Ke, X., Li, Y., Xu, R., and Guo, W. (2025c). Quality-guided vision-language learning for long-term action quality assessment. In *IEEE Transactions on Multimedia (TMM)*, pages 1–13.

Xu, J., Rao, Y., Yu, X., Chen, G., Zhou, J., and Lu, J. (2022b). Finediving: A fine-grained dataset for procedure-aware action quality assessment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 2949–2958.

Xu, J., Yin, S., Zhao, G., Wang, Z., and Peng, Y. (2024). Fineparser: A fine-grained spatio-temporal action parser for human-centric action quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14628–14637.

Yang, A., Nagrani, A., Seo, P. H., Miech, A., Pont-Tuset, J., Laptev, I., Sivic, J., and Schmid, C. (2023). Vid2seq: Large-scale pretraining of a visual language model for dense video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10714–10726.

Yu, X., Rao, Y., Zhao, W., Lu, J., and Zhou, J. (2021). Group-aware contrastive regression for action quality assessment. In *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)*, pages 7919–7928.

Zeng, L.-A., Hong, F.-T., Zheng, W.-S., Yu, Q.-Z., Zeng, W., Wang, Y.-W., and Lai, J.-H. (2020). Hybrid dynamic-static context-aware attention network for action assessment in long videos. In *Proceedings of ACM International Conference on Multimedia (ACM MM)*.

Zeng, L.-A. and Zheng, W.-S. (2024). Multimodal action quality assessment. In *IEEE Transactions on Image Processing (TIP)*.

Zhang, B., Chen, J., Xu, Y., Zhang, H., Yang, X., and Geng, X. (2024a). Auto-encoding score distribution regression for action quality assessment. *Neural Computing and Applications*, 36(2):929–942.

Zhang, C., Gupta, A., and Zisserman, A. (2021). Temporal query networks for fine-grained video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4486–4496.

Zhang, C., Wu, J., and Li, Y. (2022). Actionformer: Localizing moments of actions with transformers. In *Proceedings of the European Conference on Computer Vision (ECCV)*, volume 13664, pages 492–510.

Zhang, S., Bai, S., Chen, G., Chen, L., Lu, J., Wang, J., and Tang, Y. (2024b). Narrative action evaluation with prompt-guided multimodal interaction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition.*

Zhang, S., Dai, W., Wang, S., Shen, X., Lu, J., Zhou, J., and Tang, Y. (2023). Logo: A long-form video dataset for group action quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2405–2414.

Zhou, C., Huang, Y., and Ling, H. (2022). Uncertainty-driven action quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).*

Zhou, K., Li, J., Cai, R., Wang, L., Zhang, X., and Liang, X. (2024). Cofinal: Enhancing action quality assessment with coarse-to-fine instruction alignment. In *Proceedings of the Thirty-Third Interna-*

tional Joint Conference on Artificial Intelligence (IJ-CAI), page 1771–1779. International Joint Conferences on Artificial Intelligence Organization.

Zhou, K., Shum, H. P. H., Li, F. W. B., Zhang, X., and Liang, X. (2025). Phi: Bridging domain shift in long-term action quality assessment via progressive hierarchical instruction. pages 3718–3732.

Zhu, X., Su, W., Lu, L., Li, B., Wang, X., and Dai, J. (2021). Deformable DETR: deformable transformers for end-to-end object detection. In *9th International Conference on Learning Representations (ICLR)*.