Interpretable Long-term Action Quality Assessment

Xu Dong[1]
xd00101@surrey.ac.uk
Xinran Liu[1]
xl01315@surrey.ac.uk
Wanqing Li[2]
wanqing@uow.edu.au
Anthony Adeyemi-Ejeye[1]
femi.ae@surrey.ac.uk
Andrew Gilbert[1]
a.gilbert@surrey.ac.uk

[1] University of Surrey
Guildford, UK

[2] Advanced Multimedia Research Lab
University of Wollongong
Wollongong, Australia

Abstract

Long-term Action Quality Assessment (AQA) evaluates the execution of activities in videos. However, the length presents challenges in fine-grained interpretability, with current AQA methods typically producing a single score by averaging clip features, lacking detailed semantic meanings of individual clips. Long-term videos pose additional difficulty due to the complexity and diversity of actions, exacerbating interpretability challenges. While query-based transformer networks offer promising long-term modelling capabilities, their interpretability in AQA remains unsatisfactory due to a phenomenon we term *Temporal Skipping* , where the model skips self-attention layers to prevent output degradation. To address this, we propose an attention loss function and a query initialization method to enhance performance and interpretability. Additionally, we introduce a weight-score regression module designed to approximate the scoring patterns observed in human judgments and replace conventional single-score regression, improving the rationality of interpretability. Our approach achieves state-of-the-art results on three real-world, long-term AQA benchmarks. Our code is available at: https://github.com/dx199771/Interpretability-AQA

1 Introduction

Action quality assessment (AQA) is a task to evaluate how well a particular action is performed. Recently, this problem has garnered increasing interest within the computer vision research community and has a wide range of applications across different real-world scenarios. It is widely used in sports video analysis, including synchronized swimming, figure skating, and gymnastics [19, 20, 21, 22, 26, 32, 35]. Providing assistive analysis of athletes' performances and offers objective and precise scoring. Furthermore, AQA can also be used in healthcare [8, 10, 29], technical skill training and education purposes [4, 5, 14].

AQA is typically a score regression task, [18, 20, 27, 32, 36] directly predicting the final scores for entire action video by aggregating clip features through simple averaging and an

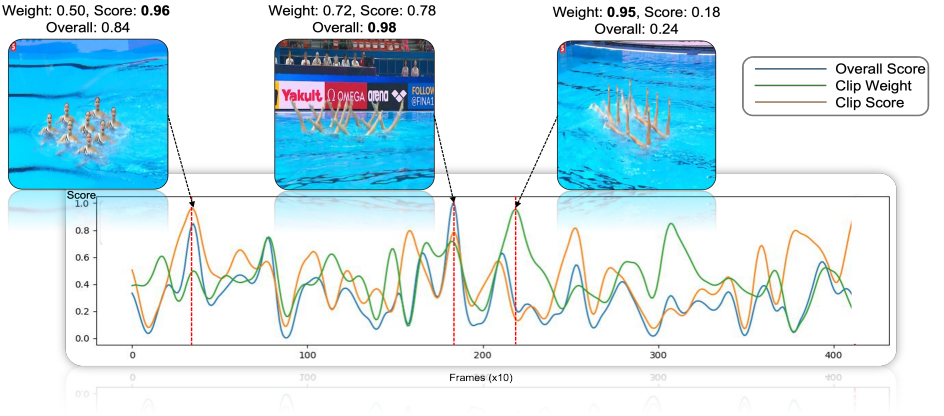
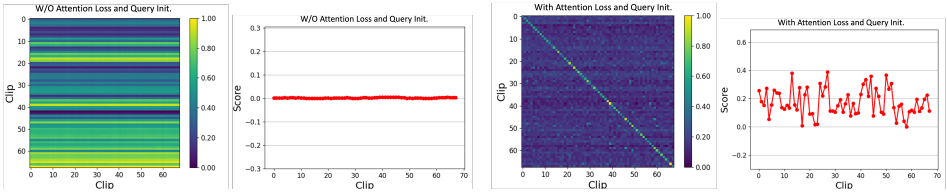


Figure 1: The visualization of the clip-level weight-score regression method illustrates that our network can adhere to the same evaluative logic as human judges in real-world scenarios. The green curve representing weight delineates the significance of the respective action clip, whereas the orange curve for score quantifies the execution quality of the action, the overall score is shown by the blue curve. All scores are normalized to a range of 0 to 1 for easier comparison.

MLP regression head. However, a single score fails to provide detailed feedback on the individual components or subtle differences, lacking fine-grained interpretation of temporal sequences. In sports contexts where technical skills and proficiency are crucial, such as gymnastics and artistic swimming, judges typically calculate the final score by weighting each action based on its execution quality and the difficulty factor. Moreover, Prior research on the interpretability of AQA [1, 23] focused on short-term actions such as diving. These action videos typically last only a few seconds and follow a sequential pattern. Compared to short-term AQA (5 to 10 seconds), long-term AQA (over 120 seconds) tasks are more challenging due to the complexity and diversity of the information and actions involved.

Recent approaches that have used queries within an encoder/decoder transformer architecture network [1, 2, 6, 31, 37] have been introduced in the AQA task due to their capabilities for long-term modelling, and because their decoder architecture is ideal for endowing the learnable queries with temporal semantic meanings. However, the model interpretability in long-term videos has not yet been satisfactory. One reason is that as each layer of the transformer is processed in long-term video, the decoder's self-attention may exhibit a *skipping* phenomenon, as stated in [12]. Temporal sequences lead the model to select shortcuts and skip decoder self-attention, thus preventing output degradation. The problem can be defined as *Temporal Skipping* in AQA, as shown in Figure 2; We present our solution with the self-attention map in 2(c) and the segmented scores in 2(d). In contrast, the opposite results are shown in 2(a) and 2(b). Compared to 2(c), which displays a distinct self-correlation diagonal of queries attention, Figure 2(a) exhibits a *Temporal Skipping* issue, resulting in smooth and averaged attention weights that deviate from the diagonal in the self-attention map. This results in a failure of interpretability in Figure 2(b) 's graph, where each clip has the same weight, whereas Figure 2(d) shows the opposite.

We propose introducing an Attention loss that facilitates mutual guidance between the self-attention and cross-attention maps to solve this issue. This is achieved by minimizing the similarity between the two attention maps using KL divergence, ensuring that as the number



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

Figure 2: ***Temporal Skipping* problem of self-attention.** This figure shows the self-attention map 2(a) and 2(c) (ours) and visualization of segmented score of each clip 2(b) and 2(d) (ours). 2(a) and 2(b) represent the same action sequences, as do 2(c) and 2(d). We can observe that in 2(a), the self-attention map severely suffers from *Temporal Skipping* problem where 2(c) shows high correlations between queries.

of layers in the transformer decoder increases, the queries within the self-attention maintain a high correlation. Additionally, by altering the variance of the Gaussian distribution used to initialize the query embedding, the correlation of the self-attention map increases, as evidenced by a clearer diagonal in the self-attention map as in 2(c). Encoding the position of the query and features is crucial to maintaining spatial and temporal encoding for interpretable sequences. Therefore, we propose to add positional encoding to the learnable queries to encode their temporal nature. We also explore positional encoding on the video features; however, we discover that because positional information has already been extracted in the backbone, providing only minimal additional temporal information. Moreover, complex positional encoding makes the features redundant, hindering network training and convergence.

Furthermore, inspired by how humans judge action quality and to enhance the rationality of the interpretability of the AQA model, we have developed a *Weight-Score Regression Head* to substitute the conventional single-score regression approach. This module decouples the output of the DETR decoder into weight and score branches to align with the scoring logic of human judges in the real world. The final action score is obtained by calculating the weighted sum of the scores for each clip. Figure 1 shows that our *Weight-Score Regression Head* can parse clip-level scores and weights. In summary, our main contributions are as follows:

- We propose a Query-based transformer decoder network for AQA, with positional query encoding to extract the clip-level feature with temporal semantic meanings.
- We identify a *Temporal Skipping* issue in self-attention that causes interpretability failures and propose an Attention Loss and query initialization method to address it.
- To decouple the score into weight and score, We propose a split *Weight-Score Regression Head* to improve the interpretability further.
- Our extensive experiments show that our network can extract interpretable features of temporal clips and achieve new state-of-the-art results on standard three long-term AQA benchmarks: Rhythmic Gymnastics (RG) [35], Figure Skating Video (Fis-V) [32] and LOng-form GrOup (LOGO) [38].

2 Related Work

Action Quality Assessment Current research on AQA regard as a regression task [18, 20, 27, 32, 36]. [22] first explored the AQA task by extracting spatiotemporal pose features

of individuals and employing the L-SVR model to estimate and predict action scores. [19] proposed three frameworks for evaluating the quality of Olympic event actions: C3D-SVR, C3D-LSTM, and LSTM-SVR. However, using only human pose information lacks modelling of external appearance information, such as splash size in diving. Additionally, LSTM lacks modeling of global relationships. Some other research focuses on addressing the uncertainty problem of AQA, [24] proposed an Uncertainty Score Distribution Learning (USDL) method for enhancing action quality representation, which regards each action as an instance associated with a score distribution, thereby reducing the impact of inherent ambiguity in the score label. Another branch developed an AQA task as a ranking problem. [5] proposed a novel rank-aware loss function and trained it with a temporal attention module. [34] proposed a group-aware regression tree (CoRe) method, which regresses the relative score and refers to other videos having similar attributes. However, these methods only regress a single score for the video sequences, which lack clip-level temporal semantic meanings. Regarding the interpretability of AQA, Roditakis et al. [23] utilized a self-supervised training technique and a differential cycle consistency loss to improve the temporal alignment and interpretability. [7] stated that averagely aggregating the clip-level feature cannot capture the relative importance of clip-level features and proposed a weighted-averaging technique. Although these works focus on clip-level semantic meanings, they do not follow the scoring logic of human judges in the real world. Our work decouples clip-level features into weight and score, further enhancing the interpretability of AQA.

DETR in Video Understanding DEtection TRansformer (DETR) was first introduced by Carion et al. [2], which uses transformer architecture to capture complex relationships and dependencies in a set of data via learnable queries. Many works adapted DETR to video understanding tasks and demonstrated its ability in temporal modelling. [37] proposed a Temporal Query Network, which uses a query-response mechanism to regard each action clip in a video as a query. [15] proposed a method that adopted Deformable DETR [39] for temporal action detection tasks, eliminating the proposal generation stage. [12] first identified the temporal collapse problem in the temporal action detection task using a DETR-like structure and proposed a self-feedback method. Our work addresses a similar situation of temporal collapse but in the context of the AQA task and modifies the representation of the self-attention map and cross-attention map in the decoder. [1] 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. 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* .

Long-term Video Understanding Early work [9, 13] using RNNs for long-term video modelling. Recently, many works have adopted transformers due to their capabilities in long-term video modelling [28, 30]. In the AQA task, [35] proposed a long-term video dataset and an ACTION-NET using GCN with a Context-aware attention module for temporal feature modelling. However, the performance of ACTION-NET is not satisfied on long-term datasets. [31] improved previous work by adopting the learnable query as a grade prototype and utilized a Likert Scoring Module for grade decoupling in the long-term video. Unlike other networks, our model adopts a transformer decoder structure, which can model long-term videos. Additionally, our attention loss and query initialization modules address the *Temporal Skipping* issue in modelling long-term videos, enhancing fine-grained feature extraction.

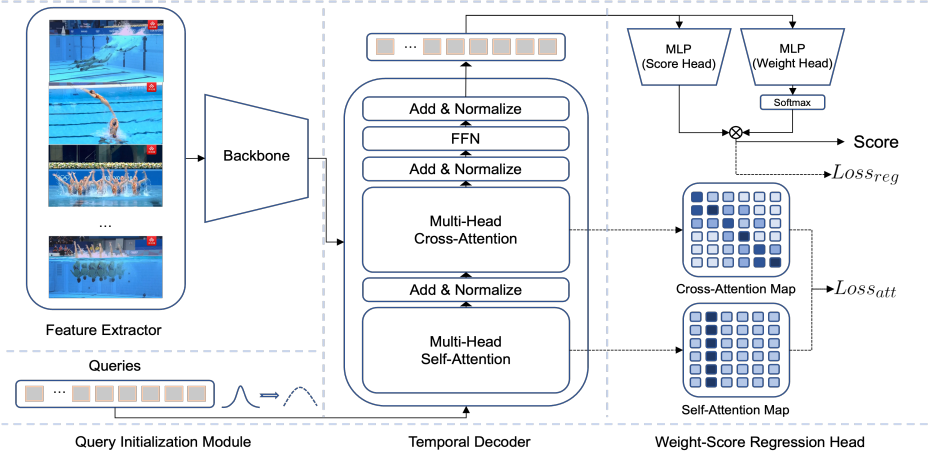


Figure 3: The overview architecture of our *Query-based transformer decoder* . The input video is divided into clips and fed into a backbone network. A temporal decoder models the clip-level features into temporal representations via learnable positionally encoded queries. The interpretable weight-score regression head can regress the final score by multiplying the weight and score of each clip. By minimizing the similarity between the self-attention map and cross-attention map, as well as query initialization, the problem of temporal collapse common in longer-term video sequences disappears and improves human interpretability.

3 Method

Our network, shown in Figure 3, consists of three modules: a *Feature Extractor* used to extract the clip-level features of the input video, followed by a *Temporal Decoder* that encodes the attention relation between these features and a set of learnable positional encoded queries to extract temporal semantic features. Subsequently, the clip features from the temporal decoder are fed into a regression head module that decouples the weight and score of each action clip. The final score is obtained by multiplying the score of each clip by its weight and summing the results across all clips. The network is optimized using two loss functions: Attention Loss $Loss_{Att}$, the sum of the KL divergence between two attention maps at each layer, and an MSE Loss $Loss_{MSE}$ to measure the mean squared error.

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) [3] and Video Swin Transformer (VST) [16, 17] as our *Feature Extractor* . The *Feature Extractor* network is frozen, using the pre-extracted *Feature Extractor* features as input to the *Temporal Decoder* and weight-score regression head network. The features obtained from L clips are denoted as $f^i{}_{i=1}^{L}$, where each $f^i \in \mathbb{R}^d$.

Temporal Decoder To encode the relationship of the temporal features, we use the decoder of a query-based positional transformer architecture [2], as previously employing an encoder has been shown to reduce performance Bai et al. [1]. The decoder has layers with self-attention for processing query inputs and cross-attention, where queries interact with encoded clip features. The decoder has two layers, as having more layers leads to a deterioration of temporal skipping. A skip connection across the cross-attention layer is used to directly propagate the output of each self-attention layer to subsequent layers. The en-

coded clip features $f_{i=1}^{iL}$ are taken from *Feature Extractor* , and a set of learnable queries is used; the model is forced to set each query to correspond to a clip and assorted clip features. The cross-attention mechanism learns to match each action query with the memory. This matching process is achieved by computing attention scores between the query vector and all memory vectors, effectively selecting and focusing on the spatial and temporal information relevant to the query. However, different to the standard query initialization method in DETR, which takes 1 as variance, we adjusted the variance of the Gaussian distribution used to obtain embedding to increase the correlation between the queries in the self-attention map. Furthermore, in the vanilla DETR, sin or cos positional encodings for queries and memory are used to provide relational position information. However, in our decoder, only query positional encoding is used to eliminate the dependence on complex positional encodings or any prior knowledge, such as extracted temporal features from the backbone.

Attention Loss We observe that the model tends to skip the self-attention module after several training epochs. Consequently, the correlation between queries is lost, and each query carries a similar weight in self-attention. This phenomenon *Temporal Skipping* is illustrated in Figure 2(a), where the self-attention map deviates from a desired diagonal line, which indicates a collapse of temporal information and a failure of temporal interpretability. Therefore, we propose the following *Attention Loss* $Loss_{Att}$ to solve this. A_S and A_C are defined as the outputs of the self-attention and cross-attention layers, respectively. The self-attention map L_S and L_C are obtained by taking the matrix product of A_S, A_C and its transpose. Finally, the Attention Loss $Loss_{Att}$ is formulated as equation 3, where D_{KL} is the Kullback-Leibler (KL) divergence, and N denotes the number of decoder layers in the *Temporal Decoder* . By using Attention Loss, each layer's self-attention and cross-attention in the transformer decoder maintain a high correlation with the previous layer until the output of the final layer, thereby mitigating the problem of temporal smoothing and resolving issues of interpretability failure.

$$L_S = softmax(A_S A_S^\intercal) \quad (1) \qquad L_C = softmax(A_C A_C^\intercal) \quad (2)$$

$$Loss_{att} = \sum_{n=1}^{N} D_{KL}(L_S^n || L_C^n) \quad (3)$$

Weight-Score Regression Head To mimic human-like scoring in our AQA model. We propose a *Weight-Score Regression Head* to decouple each action clip's weight and quality. This is achieved by using parallel weight and score heads, each implemented as an MLP layer, with the weight head employing a softmax function to output values between 0 and 1. The final *score* is the sum of each clip's score multiplied by its *weight*, shown in equation 4, where K is the number of clips.

Model Training We follow the prior work of AQA [27, 31, 34, 38] alongside our Attention Loss to utilize Mean Square Error Loss (MSE) to minimize the predicted value and ground truth value as shown in equation 5, where y is the predicted value and $\hat{y}$ is the ground truth value. The final loss can then be described as in equation 6 where λ_{reg} and λ_{Att} are the weights for MSE loss and attention loss.

$$score = \sum_{k=1}^{K} weight_k \cdot score_k \quad (4) \qquad Loss_{reg} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (5)$$

$$Loss_{all} = \lambda_{reg} Loss_{reg} + \lambda_{att} Loss_{att} \quad (6)$$

4 Experiment

4.1 Datasets and Metrics

Experiments are conducted on three widely used long-term AQA benchmarks, including Rhythmic Gymnastics [35], LOng-form GrOup [38] and Figure Skating Video [32] to evaluate our model. To be consistent with prior research [5, 7, 23, 34, 37], we adopted two metrics in our experiment, the **Spearman's rank correlation (SRCC)** and **Relative L2 distance (R-ℓ2)**. Spearman's rank correlation measures the correlation between two sequences containing ordinal or numerical data, with values ranging from -1 to 1 (higher values indicating stronger correlation) as shown in 7. In contrast, Relative L2 distance calculates the Euclidean distance between corresponding elements of two sequences, providing a measure of their dissimilarity (lower values indicating better similarity) as shown in 8.

$$\rho = \frac{\sum_i (p_i - \bar{p})(q_i - \bar{q})}{\sqrt{\sum_i (p_i - \bar{p})^2 \sum_i (q_i - \bar{q})^2}} \quad (7) \qquad R\text{-}\ell 2 = \frac{1}{N} \sum_{N}^{n=1} \left(\frac{|y_n - \hat{y}_n|}{y_{max} - y_{min}}\right)^2 \quad (8)$$

Rhythmic Gymnastics (RG). [35] The RG dataset contains video sequences of four distinct types of gymnastics routines: ball, clubs, hoop, and ribbon. Each action class comprises 200 training samples and 50 evaluation samples, with each sample approximately 1 minute and 35 seconds in duration. Each class is trained as a model, adhering to the practices described in [31, 35].

Figure Skating Video (Fis-V). [32] The Fis-V dataset contains 500 videos of figure skating videos with an average length of 2 minutes and 50 seconds. We followed the previous work, which had 400 videos for training and 100 videos for testing. Fis-V provides two labels: Total Element Scores (TES) and Total Program Component Score (PCS). Two individual models are trained to predict scores for two classes.

LOng-form GrOup (LOGO). [38] The LOGO dataset is a multi-person long-term video dataset with 150 samples for training and 50 for testing. Each video sequence is approximately 3 and a half minutes in length. To our knowledge, LOGO has the longest video lengths among existing AQA datasets.

4.2 Implementation Details

We adopt Inflated 3D ConvNet (I3D) [3], and Video Swin Transformer (VST) [16, 17] pretrained on Kinetics as our video feature extractor. Note that we only utilize the *Feature Extractor* to extract features and do not train it. The clip and query sizes for RG, FIS-V, and LOGO are set to 68, 136, and 48, respectively. These parameters are determined based on the length of the videos and extensive comparison experiments, demonstrating optimal performance. The Adam optimizer is adopted with a learning rate 1×10^{-4} and a batch size of 48. The output dimension of the transformer decoder is 1024, with each decoder having 4 heads and 2 layers, each layer applying a dropout rate of 0.7. For the weight-score regression head, each module consists of three MLP layers and a softmax layer after the weight branch.

4.3 Results and Analysis

The *Query-based transformer decoder* is compared with state-of-the-art methods on three benchmarks. Results on the RG and Fis-V dataset are shown in Table 1; our model outperforms the current state-of-the-art method GDLT [31] on RG, which uses the same image features on the four subclasses, achieving an average improvement of 0.077. Furthermore, our

model outperforms the current state-of-the-art method GDLT [31] by an average of 0.027 on the Fis-V dataset of two subclasses, TES and PCS. On the LOGO dataset, our model achieves state-of-the-art as shown in Table 2 and outperforms prior methods [38] by 0.220 using VST as the *Feature Extractor* and 0.099 using I3D as the *Feature Extractor* on SRCC.

Methods	Feature Extractor	RG (SRCC↑)					Fis-V (SRCC↑)		
		Ball	Clubs	Hoop	Ribbon	Avg.	TES	PCS	Avg.
SVR [19]	C3D [25]	0.357	0.551	0.495	0.516	0.483	0.400	0.590	0.501
MS-LSTM	I3D [3]	0.515	0.621	0.540	0.522	0.551	-	-	-
[32]	VST [17]	0.621	0.661	0.670	0.695	0.663	0.660	0.809	0.744
ACTION-NET	I3D[3]+ResNet[13]	0.528	0.652	0.708	0.578	0.623	-	-	-
[35]	VST[17]+ResNet[13]	0.684	0.737	0.733	0.754	0.728	0.694	0.809	0.757
GDLT [31]	VST [17]	0.746	0.802	0.765	0.741	0.765	0.685	0.820	0.761
Ours	VST [17]	**0.823**	**0.852**	**0.837**	**0.857**	**0.842**	**0.717**	**0.858**	**0.788**

Table 1: Spearman's rank correlation coefficient performance comparison on **Rhythmic Gymnastics (RG)** and **Figure Skating Video (Fis-V)** dataset. Avg. is the average SRCC for all subclasses. The higher SRCC suggests better performance.

Methods	I3D [3]		VST [17]	
	SRCC ↑	R-ℓ2(×100) ↓	SRCC ↑	R-ℓ2(×100) ↓
USDL [24]	0.426	5.736	0.473	5.076
CoRe [34]	0.471	5.402	0.500	5.960
TSA [33]	0.452	5.533	0.475	4.778
ACTION-NET [35]	0.306	5.858	0.410	5.569
USDL-GOAT [38]	0.462	4.874	0.535	5.022
TSA-GOAT [38]	0.486	5.394	0.484	5.409
CoRe-GOAT [38]	0.494	5.072	0.560	4.763
Ours	**0.593**	**1.220**	**0.780**	**1.745**

Table 2: Performance comparison on **LOGO** dataset. The higher the SRCC, the lower R-ℓ2, it suggests better performance.

Ablation Study Experiments were conducted in three settings to compare the effects of attention loss, query positional encoding, and query initialization. As shown in table 3, only use the data-decoder without any other proposed modules. Our proposal has only 0.628 SRCC as the baseline. Adding the attention loss significantly improved the performance by 28.5%, which verifies that our proposed attention loss improves the interpretability and the SRCC results. With the addition of only query positional encoding, the results improved to 0.810. Finally, our query initialization module further enhanced the performance by around 4%, showing that using high variance allows query vectors to have a wider diversity and dispersion of initialization states and improve the overall performance.

Effect of position encoding Different positional encoding methods in the *Temporal Decoder* are compared in Table 4. We find that only adopting query positional encoding outperformed other methods while incorporating query and memory positional encoding harms the final performance. This phenomenon is probably because, in the AQA task, we primarily focus on modelling learnable queries using the DETR decoder and endowing these queries with temporal semantic meanings through the decoder structure. However, the temporal information of the memory is already extracted by the *Feature Extractor* , and using only the

Module	Attention Loss	Query PE	Query Init.	SRCC ↑
Baseline	×	×	×	0.628
	✓	×	×	0.807
	✓	✓	×	0.810
Ours	✓	✓	✓	**0.842**

Table 3: **Ablation study** on the average performance of four labels in the Rhythmic Gymnastics (RG) dataset across various modules.

Methods	Query	Memory	SRCC
Baseline	×	×	0.758
	×	✓	0.778
	✓	✓	0.751
Ours	✓	×	**0.824**

Table 4: Effect of Positional Encoding on RG dataset, where SRCC results take the average of the four labels.

Variance Init.	SRCC
0.5	0.810
1	0.810
3	0.811
5	**0.820**

Table 5: Effect of Query Variance Initialization on RG dataset, where SRCC results take the average of the four labels

query positional encoding avoids unnecessary computation and potential information redundancy. This approach ensures that these queries can effectively capture and express the key action quality indicators in the video, thereby enhancing the scoring performance and the model's interpretability.

Effect of variance in query initialization module In the *Temporal Decoder* , the impact of different variances in query initialization on the *Temporal Skipping* problem and final SRCC results is compared as shown in Table 5 specifically, using a larger variance to initialize the query embedding results in a more compact diagonal pattern in the self-attention map, which represents higher correlation between action queries. Furthermore, initializing variance boosts the final SRCC results, as shown in Table 5. Different initialization variance values are compared in the experiment, and it is concluded that the SRCC is highest when using a larger variance.

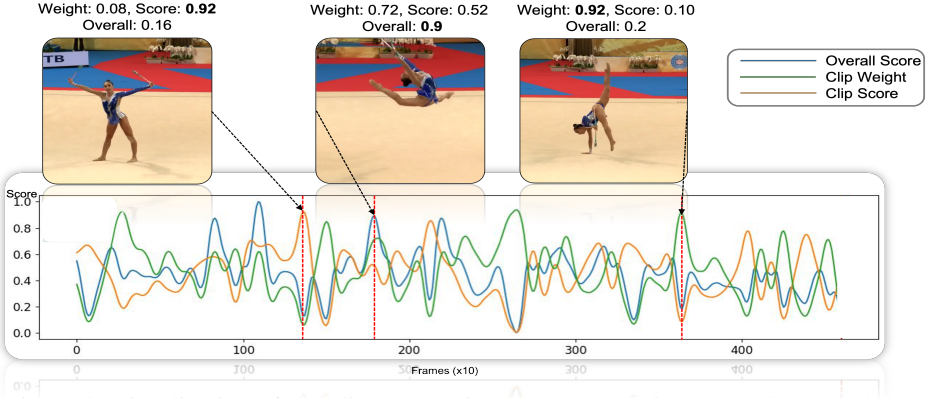


Figure 4: Visualization of our clip-level weight-score regression method on RG dataset.

Sequence Interpretability To replace the single-score regression method and follow the scoring logic of human judges, we decoupled each clip's score into weight and score. Figure

1 (a) shows that in the first clip, despite high completion quality and synchronization, the weight (difficulty) is low. In the second clip, the action has a high weight and score. However, the movement quality is low in the last clip, although the difficulty of the action is high (high weight). Figure 4 shows another visualization of the clip-level weight-score regression method on the RG dataset, where the green line represents the weight (difficulty) of the current frame, the yellow line represents the score (quality) of the current frame, and the blue line represents the combined score. Empirical evidence demonstrates that our weight-score regression module is effective in enhancing the interpretability of AQA.

4.4 Conclusion

In this work, we propose a novel framework to enhance interpretability in long-term AQA tasks, addressing the *Temporal Skipping* issue which fails interpretability. A novel attention loss function and a query initialization module are proposed, and the impact of different positional encodings is explored. Our approach also includes a weight-score regression module that decouples each clip's action score into weight and score, facilitating a fine-grained and interpretable assessment that makes AQA scoring more meaningful and informative. Demonstrating effectiveness, our model achieves state-of-the-art results on three AQA benchmarks and effectively parses clip-level semantic meanings through interpretability results. In our future work, we will explore interpretability evaluation methods from qualitative and quantitative perspectives. We aim to enhance the evaluation process in interpretable AQA tasks by making them more comprehensive and accessible for analysis.

References

[1] Yang Bai, Desen Zhou, Songyang Zhang, Jian Wang, Errui Ding, Yu Guan, Yang Long, and Jingdong Wang. Action quality assessment with temporal parsing transformer. In *Proceedings of the IEEE International Conference on Computer Vision (ECCV)*, pages 422–438, 2022.

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Proceedings of the IEEE International Conference on Computer Vision (ECCV)*, pages 213–229, 2020.

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

[4] Hazel Doughty, Dima Damen, and Walterio Mayol-Cuevas. Who's better? who's best? pairwise deep ranking for skill determination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[5] Hazel Doughty, Walterio Mayol-Cuevas, and Dima Damen. The pros and cons: Rank-aware temporal attention for skill determination in long videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[6] Zexing Du, Di He, Xue Wang, and Qing Wang. Learning semantics-guided representations for scoring figure skating. *IEEE Transactions on Multimedia*, 2023.

[7] Shafkat Farabi, Hasibul Himel, Fakhruddin Gazzali, Md Bakhtiar Hasan, Md Hasanul Kabir, and Moshiur Farazi. Improving action quality assessment using weighted aggregation. In *Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA)*, pages 576–587, 2022.

[8] Isabel Funke, Sören Torge Mees, Jürgen Weitz, and Stefanie Speidel. Video-based surgical skill assessment using 3d convolutional neural networks. *International journal of computer assisted radiology and surgery*, 14:1217–1225, 2019.

[9] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE international conference on computer vision (ICCV)*, pages 5267–5275, 2017.

[10] Yixin Gao, S Swaroop Vedula, Carol E Reiley, Narges Ahmidi, Balakrishnan Varadarajan, Henry C Lin, Lingling Tao, Luca Zappella, Benjamın Béjar, David D Yuh, et al. Jhu-isi gesture and skill assessment working set (jigsaws): A surgical activity dataset for human motion modeling. In *MICCAI workshop*, volume 3, page 3, 2014.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 770–778, 2016.

[12] Jihwan Kim, Miso Lee, and Jae-Pil Heo. Self-feedback detr for temporal action detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10286–10296, 2023.

[13] Xiangpeng Li, Jingkuan Song, Lianli Gao, Xianglong Liu, Wenbing Huang, Xiangnan He, and Chuang Gan. 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, 2019.

[14] Zhenqiang Li, Yifei Huang, Minjie Cai, and Yoichi Sato. 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, 2019.

[15] Xiaolong Liu, Qimeng Wang, Yao Hu, Xu Tang, Shiwei Zhang, Song Bai, and Xiang Bai. End-to-end temporal action detection with transformer. *IEEE Transactions on Image Processing*, 31:5427–5441, 2022.

[16] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. pages 9992–10002, 2021.

[17] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3202–3211, 2022.

[18] Jia-Hui Pan, Jibin Gao, and Wei-Shi Zheng. Action assessment by joint relation graphs. In *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)*, pages 6331–6340, 2019.

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

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

[21] Paritosh Parmar and Brendan Tran Morris. 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, 2019.

[22] Hamed Pirsiavash, Carl Vondrick, and Antonio Torralba. Assessing the quality of actions. In *Proceedings of the IEEE International Conference on Computer Vision (ECCV)*, pages 556–571, 2014.

[23] Konstantinos Roditakis, Alexandros Makris, and Antonis Argyros. 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, 2021.

[24] Yansong Tang, Zanlin Ni, Jiahuan Zhou, Danyang Zhang, Jiwen Lu, Ying Wu, and Jie Zhou. 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, 2020.

[25] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision (ICCV)*, pages 4489–4497, 2015.

[26] Vinay Venkataraman, Ioannis Vlachos, and Pavan K Turaga. Dynamical regularity for action analysis. In *Proceedings of the British Machine Vision Conference (BMVC)*, volume 67, pages 1–12, 2015.

[27] Shunli Wang, Dingkang Yang, Peng Zhai, Chixiao Chen, and Lihua Zhang. Tsa-net: Tube self-attention network for action quality assessment. In *Proceedings of the 29th ACM international conference on multimedia*, pages 4902–4910, 2021.

[28] Teng Wang, Ruimao Zhang, Zhichao Lu, Feng Zheng, Ran Cheng, and Ping Luo. End-to-end dense video captioning with parallel decoding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6847–6857, 2021.

[29] Tianyu Wang, Yijie Wang, and Mian Li. 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, 2020.

[30] Chao-Yuan Wu, Yanghao Li, Karttikeya Mangalam, Haoqi Fan, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. 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, 2022.

[31] Angchi Xu, Ling-An Zeng, and Wei-Shi Zheng. 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, 2022.

[32] Chengming Xu, Yanwei Fu, Bing Zhang, Zitian Chen, Yu-Gang Jiang, and Xiangyang Xue. Learning to score figure skating sport videos. *IEEE transactions on circuits and systems for video technology*, 30(12):4578–4590, 2019.

[33] Jinglin Xu, Yongming Rao, Xumin Yu, Guangyi Chen, Jie Zhou, and Jiwen Lu. Finediving: A fine-grained dataset for procedure-aware action quality assessment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 2949–2958, 2022.

[34] Xumin Yu, Yongming Rao, Wenliang Zhao, Jiwen Lu, and Jie Zhou. Group-aware contrastive regression for action quality assessment. In *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)*, pages 7919–7928, 2021.

[35] Ling-An Zeng, Fa-Ting Hong, Wei-Shi Zheng, Qi-Zhi Yu, Wei Zeng, Yao-Wei Wang, and Jian-Huang Lai. Hybrid dynamic-static context-aware attention network for action assessment in long videos. In *Proceedings of ACM International Conference on Multimedia (ACM MM)*, 2020.

[36] Boyu Zhang, Jiayuan Chen, Yinfei Xu, Hui Zhang, Xu Yang, and Xin Geng. Auto-encoding score distribution regression for action quality assessment. *Neural Computing and Applications*, 36(2):929–942, 2024.

[37] Chuhan Zhang, Ankush Gupta, and Andrew Zisserman. Temporal query networks for fine-grained video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4486–4496, 2021.

[38] Shiyi Zhang, Wenxun Dai, Sujia Wang, Xiangwei Shen, Jiwen Lu, Jie Zhou, and Yansong Tang. 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, 2023.

[39] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: deformable transformers for end-to-end object detection. In *9th International Conference on Learning Representations (ICLR)*, 2021.