Interpretable Long-term Action Quality Assessment

Xu Dong, **Xinran Liu**, **Wanqing Li**, **Anthony Adeyemi-Ejeye**, **Andrew Gilbert**

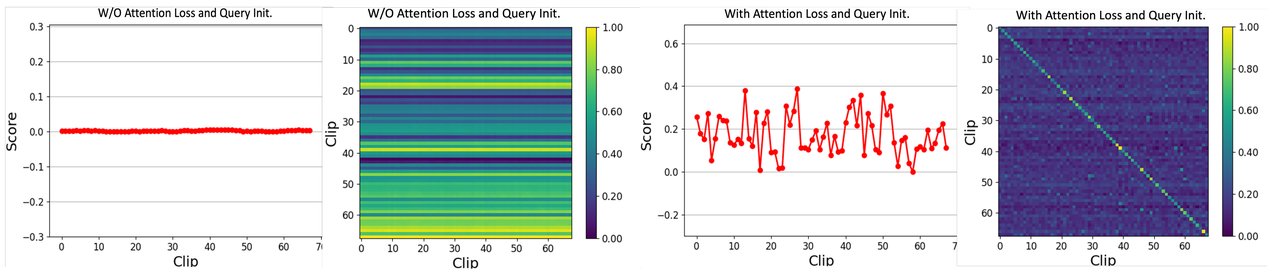
Github Page: https://github.com/dx199771/Interpretability-AQA

Scan To Read

TL;DR

AQA: Evaluate how well an action is performed in a video.
Problem1: Limited Understanding of Long Temporal Sequences.
Problem2: Existing AQA models regress single score and lack interpretability.
Solution: Develop an interpretable AQA network with the ability to handle long-term videos.

Temporal Skipping Problem

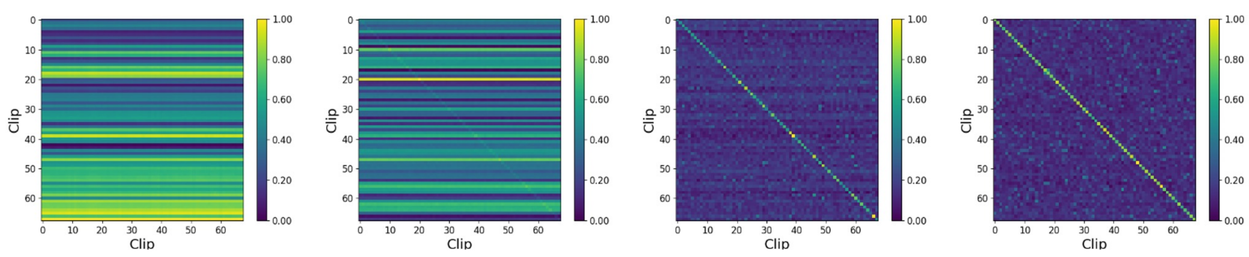


- Temporal sequences lead the model to select shortcuts and skip decoder self-attention, thus preventing output degradation.

Methodology

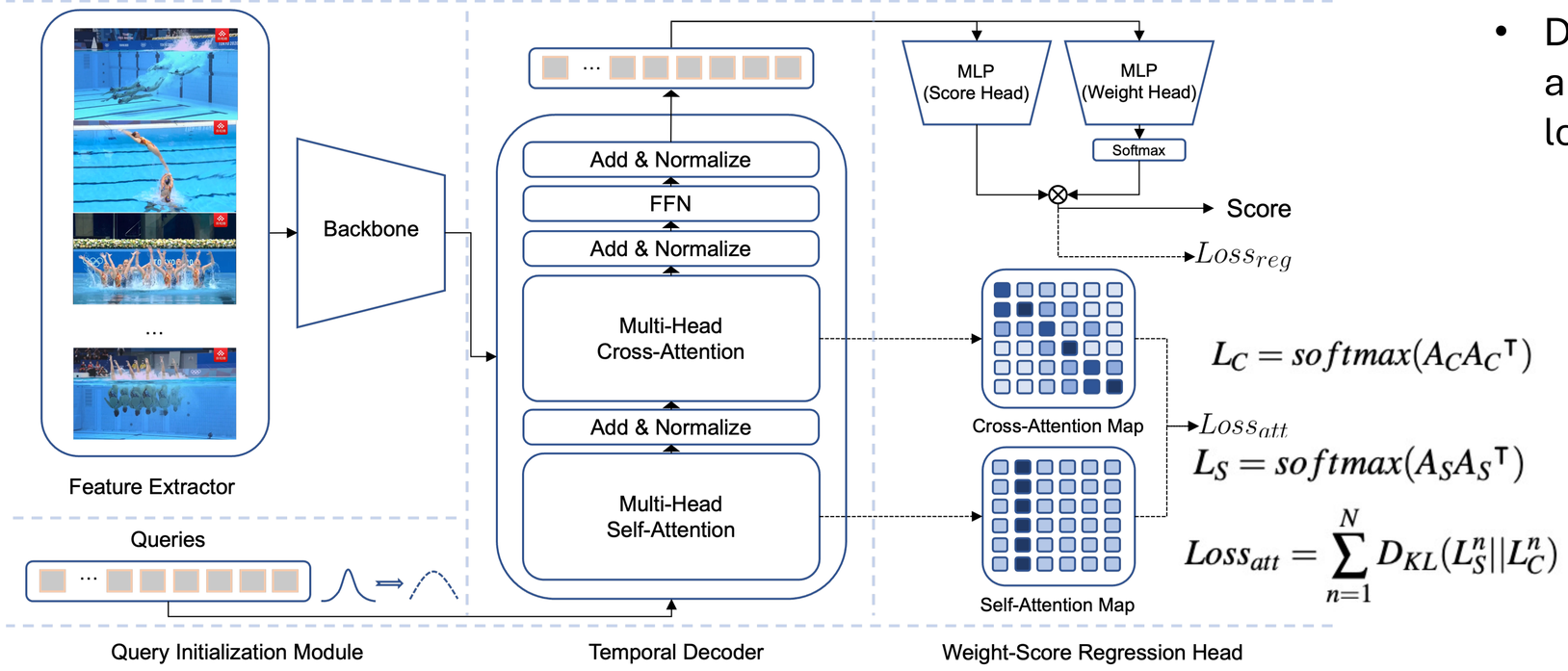
Temporal Decoder: Transformer based decoder uses learnable clip queries as input with semantic meaning to learn clip score.
Attention Loss: Using KL divergence to constrain Self-attention and Cross-attention outputs and eliminate *Temporal Skipping* Problem.
Query Init: Using different variance to initialize query boost performance.
Two Head Regression: Weight-Score regression head provides interpretability to the network.

Query Initialization Module



- Using different variance to initialize the query embedding can improve the query correlation of self-attention map.
- Using **larger** variance outperforms lower variance.

Our Pipeline



Weight-Score regression head

- Decoupling the output the decoder into **weight** and **score** branches to align with the scoring logic of human judges in the real world.

Experiment

Performance comparison on Rhythmic Gymnastics (RG) and Figure Skating Video (Fis-V) dataset

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 [32]	I3D [3]	0.515	0.621	0.540	0.522	0.551	-	-	-
	VST [17]	0.621	0.661	0.670	0.695	0.663	0.660	0.809	0.744
ACTION-NET [35]	I3D[3]+ResNet[11]	0.528	0.652	0.708	0.578	0.623	-	-	-
	VST[17]+ResNet[11]	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**

Performance comparison on LOGO dataset

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**

Ablation Studies

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

Visualization of clip-level weight-score regression