FineCausal: A Causal-Based Framework for Interpretable Fine-Grained Action Quality Assessment

Ruisheng Han[1], Kanglei Zhou[2], Amir Atapour-Abarghouei[1], Xiaohui Liang[2], Hubert P. H. Shum[1†]
[1]Durham University [2]Beihang University
{ruisheng.han, amir.atapour-abarghouei, hubert.shum}@durham.ac.uk,
{zhoukanglei, liang_xiaohui}@buaa.edu.cn

Abstract

Action quality assessment (AQA) is critical for evaluating athletic performance, informing training strategies, and ensuring safety in competitive sports. However, existing deep learning approaches often operate as black boxes and are vulnerable to spurious correlations, limiting both their reliability and interpretability. In this paper, we introduce ***FineCausal****, a novel causal-based framework that achieves state-of-the-art performance on the FineDiving-HM dataset. Our approach leverages a Graph Attention Network–based causal intervention module to disentangle human-centric foreground cues from background confounders, and incorporates a temporal causal attention module to capture fine-grained temporal dependencies across action stages. This dual-module strategy enables FineCausal to generate detailed spatio-temporal representations that not only achieve state-of-the-art scoring performance but also provide transparent, interpretable feedback on which features drive the assessment. Despite its strong performance, FineCausal requires extensive expert knowledge to define causal structures and depends on high-quality annotations, challenges that we discuss and address as future research directions. Code is available at* https://github.com/Harrison21/FineCausal.

1. Introduction

Action Quality Assessment (AQA) has emerged as a pivotal research area for objectively evaluating the quality of performed actions, offering an alternative to subjective human judgment [38]. Since its early development by Gordon *et al.* [6], AQA has become indispensable in addressing challenges related to bias, reliability, and cost in expert evaluations. Its applications span diverse fields such as skill evaluation [20, 32], medical rehabilitation [1, 5, 36, 40], and sports analysis [12, 13, 28, 34, 39]. In sports, precise AQA is essential for evaluating athlete performance, designing targeted training programs, and preventing injuries.

Sports videos, unlike general videos [41], are sequential in nature and encapsulate explicit procedural knowledge. For example, in competitive diving, athletes perform a series of rapid and complex movements, ranging from decisive take-off, through intricate somersaults and twists, to the precise entry into the water. Even minor variations in take-off angle, body posture during somersaults, or water entry can significantly influence the final score. However, subtle differences are often difficult to discern accurately by the human eye, which limits the reliability of traditional judgment-based assessments.

Existing AQA methods typically face two major challenges [15, 22, 29–31]. First, many approaches either disregard valuable background context by focusing solely on masked foreground regions or, when they attempt to incorporate background information, they rely on simplistic fusion techniques (e.g., sigmoid-based fusion of video and mask features). Such approaches can inadvertently introduce spurious correlations from irrelevant environmental cues, thereby undermining the robustness of the assessment. Second, current methods generally lack interpretability. They fail to provide clear insights into how different stages of an action individually contribute to the overall quality score. This deficiency not only hampers the trustworthiness of the system but also limits its practical utility for athletes and coaches, who require detailed feedback to identify specific areas for improvement. In short, without a fine-grained understanding of the action stages, it becomes difficult to pinpoint which aspects of performance are lacking and how to optimize them effectively.

Motivated by these challenges, we propose **FineCausal**, a novel causal-based framework that enhances both the performance and interpretability of AQA models. Our approach leverages a causal graph to explicitly model the relationships among original video features, fused features, stage features, and the final action score. Instead of relying on rudimentary fusion techniques, FineCausal employs

† Corresponding Author

a Graph Attention Network (GAT)-based causal intervention to dynamically balance the contributions of foreground and background information. In addition, we introduce a temporal causal attention module to capture the semantic and temporal dependencies among different action stages. This temporal module provides fine-grained insights into how each sub-action (e.g., take-off, somersault, twist, and entry) contributes to the overall quality score, thereby offering actionable feedback for performance improvement. Overall, these designs not only improve prediction accuracy but also enhance interpretability by revealing which spatial regions and temporal stages are most influential in determining action quality (see Tab. 2). Our contributions are:

1. We introduce **FineCausal**, the first causal-based AQA framework that enhances interpretability by explicitly modeling the causal dependencies among video features.
2. We propose a GAT-based intervention module that adaptively integrates both human-centric foreground cues and valuable background context.
3. We develop a temporal causal attention module to capture fine-grained, stage-wise relationships across time.

2. Related Work

2.1. Action Quality Assessment

Action Quality Assessment has evolved considerably over the years. Early work by Pirsiavash *et al.* [21] treated AQA as a regression problem from action representations to scores, while Parisi *et al.* [17] evaluated quality based on the correctness of action matches. Later, Parmar *et al.* [19] demonstrated the effectiveness of leveraging spatio-temporal features for score estimation in competitive sports. Recently, Tang *et al.* [22] introduced an uncertainty-aware score distribution learning approach to mitigate the ambiguity in judges' scores, and Yu *et al.* [31] developed a contrastive regression method that leverages video-level features for accurate ranking and prediction. In parallel, Wang *et al.* [24] proposed TSA-Net, which utilizes outputs from a VOT tracker to generate more informative action representations, while Xu *et al.* [29] introduced an action procedure-aware method alongside a fine-grained sports video dataset to further boost AQA performance. Zhang *et al.* [33] also enriched clip-wise representations by integrating contextual group information through a plug-and-play attention module. More recently, Xu *et al.* [30] propose a fine-grained spatio-temporal action parser that explicitly parses actions in both space and time to focus on human-centric foreground regions; importantly, they also introduce the FineDiving-HM dataset, which provides fine-grained annotations of human-centric foreground action masks for the FineDiving dataset, promoting the development of real-world AQA systems. Complementing this, Okamoto *et al.* [15] introduce a hierarchical neuro-symbolic approach that uses neural networks to abstract interpretable symbols from video data and applies rule-based reasoning to assess action quality, ultimately generating detailed visio-linguistic reports that offer transparent feedback on what aspects of performance were good or bad. However, their framework does not explicitly explore the causal relationships that explain *why* certain execution elements lead to success or failure. By contrast, our work explicitly models these causal dependencies, both among visual features and across different sub-action stages, enabling a more comprehensive understanding of how specific cues and temporal phases interact to determine the final outcome.

2.2. Causal Inference

In deep learning, confounding factors can lead to the capture of spurious correlations between inputs and outputs. Causal inference techniques provide a theoretical framework to disentangle correlation from causation, thereby enhancing model generalizability, robustness, and interpretability. For instance, Nie *et al.* [14] adopt a causal perspective for chest X-ray classification by constructing structural causal models and applying backdoor adjustments to mitigate the influence of spurious correlations. Similarly, Carloni *et al.* [2] leverage causal inference via contrastive disentangled learning to foster robustness against domain shifts. In the domain of gaze estimation, Liang *et al.* [10] propose a de-confounded approach that separates gaze-relevant features from irrelevant factors using a dynamic confounder bank strategy, leading to significant cross-domain improvements. Moreover, Liu *et al.* [11] introduce cross-modal causal relational reasoning for event-level visual question answering, employing both front-door and back-door interventions alongside spatial-temporal transformers to capture fine-grained visual-linguistic interactions. Collectively, these studies demonstrate that incorporating causal inference not only improves the reliability and fairness of computer vision models but also provides a clearer, more interpretable basis for understanding the underlying mechanisms that drive predictions.

3. Methodology

In this section, we present our causal-based Action Quality Assessment (AQA) framework, termed **FineCausal**, as illustrated in Fig. 1. Our approach is based on the causal inference methodology proposed in [10], with the primary goal of effectively utilizing both the foreground (athlete action) and contextual background information (e.g., diving board, audience) to improve prediction accuracy. Unlike prior methods [30] that apply a straightforward sigmoid-based fusion of video and foreground mask features, which may result in the loss of valuable background information, our proposed GAT-based causal intervention module explicitly identifies and preserves beneficial background features

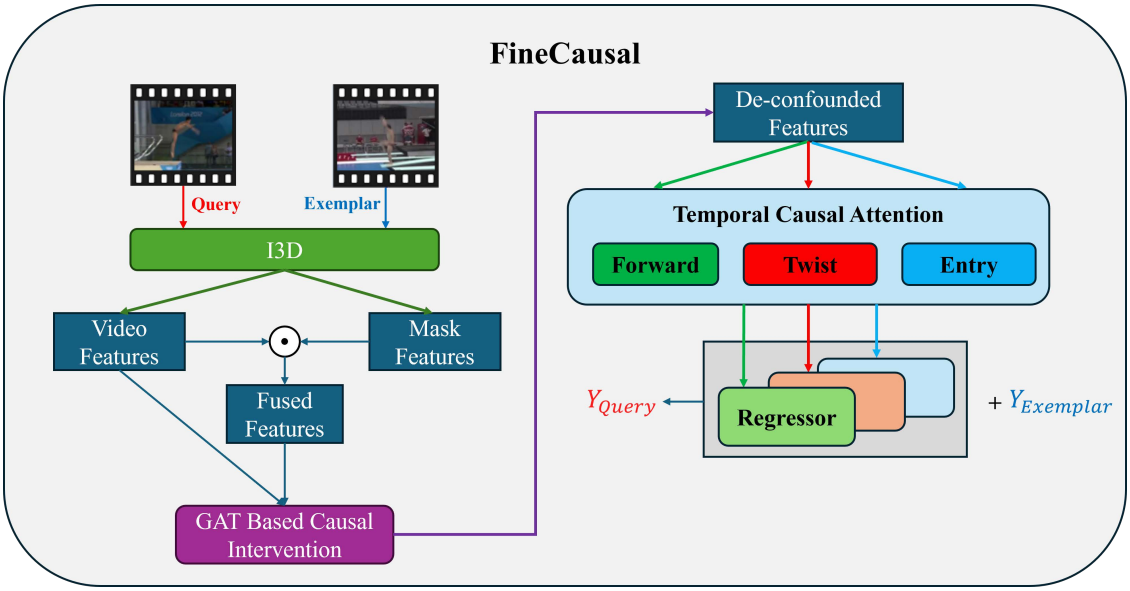


Figure 1. The architecture of FineCausal. The model takes a query and an exemplar video as input, extracting video and mask features through an I3D backbone. These features are fused and refined using a GAT-based causal intervention module to remove spurious correlations, producing deconfounded features. The refined features are then processed through a temporal causal attention mechanism, which decomposes the action into forward, twist, and entry stages. A regressor aggregates stage-wise contributions to predict the query action score Y_{Query}, adjusted based on the exemplar score Y_{Exemplar}, ensuring robust AQA.

while suppressing spurious correlations.

We first construct a causal graph capturing the relationships among four central variables: **O** (Original video features), **F** (Fused video features), **S** (Stage features), and **Y** (Action score). Next, we elaborate on how our GAT-based intervention mechanism effectively disentangles genuine causal relationships from spurious correlations. As a convention, we use solid arrows to indicate genuine causal effects, while dashed arrows represent spurious causal effects. The following section provides a detailed explanation of the constructed causal graph.

3.1. Causal Graph Construction

We consider four key variables in the AQA pipeline:

- **O**: *Original video features* extracted from both the query and exemplar videos (e.g., via a backbone network).
- **F**: *Fused video features*, obtained by combining **O** with human-centric mask information to emphasize the athlete's body and reduce irrelevant background.
- **S**: *Stage features*, encoding the sub-action phases (forward, twist, entry) from **F**.
- **Y**: *Action score*, the final assessment of action quality.

Following [10], we depict these variables in a directed acyclic graph (DAG), illustrated in Fig. 2. The intended causal flow follows:

$$\mathbf{O} \longrightarrow \mathbf{F} \longrightarrow \mathbf{S} \longrightarrow \mathbf{Y}.$$

However, prior methods typically rely on a simple sigmoid function to fuse video and mask features. This approach focuses almost exclusively on the masked foreground regions, thereby neglecting background information that could be valuable for AQA. Our hypothesis is that background elements (e.g., diving board structure, audience reaction) also carry informative cues for assessing action quality. While ignoring these cues may limit model performance, their naive incorporation could also introduce misleading dependencies, as the model might inadvertently overfit to environmental biases like lighting variations or audience arrangement. The fundamental causal relationships underpinning AQA are:

- **O** $\rightarrow$ **F**. The original video features **O** (extracted frame-

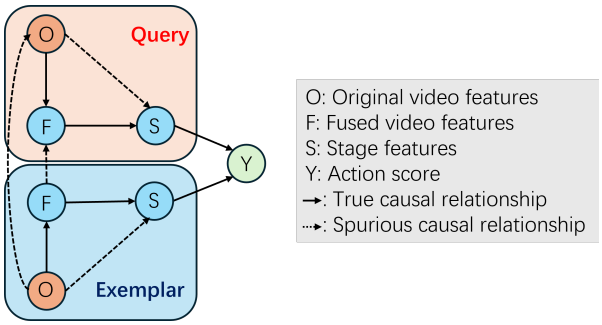


Figure 2. The causal graph of our AQA framework. Nodes represent variables: **O** for original video features, **F** for fused video features, **S** for stage features, and **Y** for final action score. Solid arrows indicate true causal relationships, whereas dashed arrows represent spurious correlations.

wise) determine the fused features $\mathbf{F}$. In principle, $\mathbf{F}$ should highlight the relevant foreground regions (e.g., the diver) and discard irrelevant background.

- $\mathbf{F} \rightarrow \mathbf{S}$. From the fused representation, we segment the action into sub-stages, producing $\mathbf{S}$. These stage features capture the evolving posture of the athlete during forward, twist, and entry phases.
- $\mathbf{S} \rightarrow \mathbf{Y}$. Finally, each sub-action's quality affects the overall action score. If a twist phase is poorly executed, it lowers $\mathbf{S}$ quality and thus reduces $\mathbf{Y}$.

Beyond the intended causal paths, spurious correlations arise due to the influence of shared environmental factors between the query and exemplar videos. These external influences include elements such as background, audience presence, and lighting conditions, which are not directly related to the athlete's performance but can nonetheless impact the final AQA. Key spurious relationships in our framework include:

- $\mathbf{O}_{\text{Exemplar}} \dashrightarrow \mathbf{O}_{\text{Query}}$: The original video features of the exemplar video may influence the original features of the query video due to shared environmental elements, such as the diving board structure or audience positioning, leading to unintended dependencies in feature extraction.
- $\mathbf{F}_{\text{Exemplar}} \dashrightarrow \mathbf{F}_{\text{Query}}$: Similar to the original features, fused video features may also inherit biases from the exemplar video, particularly if the background segmentation is imperfect or if shared contextual elements are mistakenly considered as action-relevant information.
- $\mathbf{O} \dashrightarrow \mathbf{S}$ (for both query and exemplar videos): External factors such as lighting variations or shadows may directly affect the stage-wise decomposition, introducing correlations between the original video features and the extracted stage features that do not genuinely reflect the execution quality.

These spurious correlations can mislead contrastive learning-based assessments, where the final action score of the query video is computed based on differences in feature representations between the query and exemplar videos. If these representations are influenced by irrelevant environmental factors, the resulting score may be biased away from the true action quality.

3.2. Causal Intervention via Graph Attention Networks (GAT)

To mitigate the impact of spurious correlations and enhance the robustness of AQA, we employ Graph Attention Networks (GAT) [8, 9, 23, 26] to refine the fused video features through structured feature aggregation. Unlike traditional causal adjustments such as the front-door or back-door criterion [10, 14], which require explicit mediators or confounder control variables, the GAT mechanism enables dynamic and adaptive aggregation of node features. In this context, each node represents a fused feature corresponding to a video feature. By computing attention weights conditioned on each node's feature, the model selectively integrates relevant contextual information, such as beneficial background cues, while suppressing less informative or misleading signals.

3.2.1. Problem Formulation

In our framework, we model the causal dependencies among the key variables as follows:

$$P(\mathbf{Y} \mid \mathbf{O}, \mathbf{F}, \mathbf{S}) = P(\mathbf{Y} \mid \mathbf{S}) P(\mathbf{S} \mid \mathbf{F}) P(\mathbf{F} \mid \mathbf{O}). \quad (1)$$

However, shared environmental elements between query and exemplar videos can introduce unintended correlations in the raw fused features $\mathbf{F}$ (see Fig. 2). Such spurious correlations may lead to overfitting to background biases, even as potentially valuable contextual information is inadvertently discarded. To address this, we introduce a refined feature representation:

$$\widetilde{\mathbf{F}} = \mathrm{GAT}(\mathbf{O}, \mathbf{F}). \quad (2)$$

where the GAT-based intervention is designed to disentangle genuine causal effects from spurious influences. This process ensures that $\widetilde{\mathbf{F}}$ retains beneficial background cues that contribute to a more accurate assessment of action quality while suppressing irrelevant dependencies induced by shared environmental factors.

3.2.2. Implementation via GAT

Our GAT-based module consists of two key layers:

Initial Feature Propagation: The first GAT layer computes the refined representation of $\mathbf{F}$ by attending to feature dependencies across video frames. The updated feature representation is given by:

$$\mathbf{F}^{(1)} = \sigma\left(\sum_{j \in \mathcal{N}(i)} \alpha_{ij} \Theta \mathbf{F}_j\right), \quad (3)$$

where Θ is a trainable weight matrix, α_{ij} represents the learned attention weight between nodes i and j, and $\mathcal{N}(i)$ denotes the neighborhood of node i in the feature graph. The attention coefficient α_{ij} is computed as:

$$\alpha_{i,j} = \frac{\exp\left(\mathbf{a}^\top \mathrm{LeakyReLU}\left(\mathbf{\Theta}_s \mathbf{x}_i + \mathbf{\Theta}_t \mathbf{x}_j\right)\right)}{\sum_{k \in \mathcal{N}(i) \cup \{i\}} \exp\left(\mathbf{a}^\top \mathrm{LeakyReLU}\left(\mathbf{\Theta}_s \mathbf{x}_i + \mathbf{\Theta}_t \mathbf{x}_k\right)\right)}, \quad (4)$$

where $\mathbf{\Theta}_s$ and $\mathbf{\Theta}_t$ are trainable weight matrices for the source and target nodes, respectively, and $\mathbf{a}$ is a learnable attention vector.

Residual Feature Correction: The second GAT layer further refines the feature representations by incorporating a residual connection to preserve the original structure of $\mathbf{F}$:

$$\mathbf{F}^{(2)} = \sigma\left(\sum_{j \in \mathcal{N}(i)} \alpha'_{ij} W' \mathbf{F}_j\right) + \lambda \mathbf{F}^{(1)}, \quad (5)$$

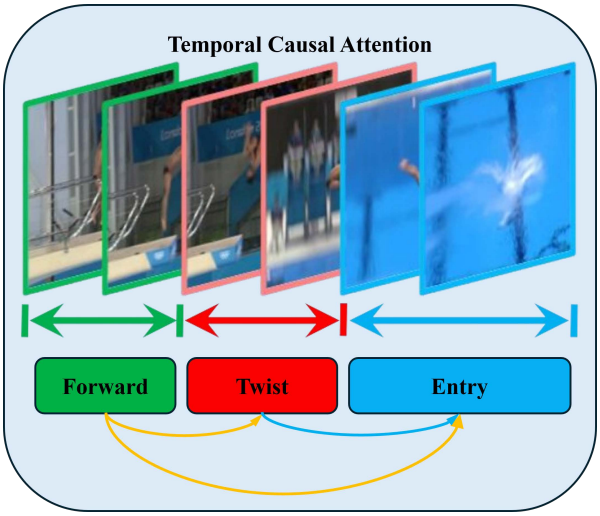


Figure 3. Illustration of stage-wise decomposition in action sequences. The movement is split into three stages: forward, twist, and entry. Temporal causal attention models the influence of each stage on the next.

where λ is a learnable residual weight that balances the contribution of the refined features and the original input. The final deconfounded representation is then obtained as:

$$\widetilde{\mathbf{F}} = \mathbf{F}^{(2)}. \tag{6}$$

By leveraging the GAT-based causal intervention, our framework ensures that $\widetilde{\mathbf{F}}$ reflect genuine action execution quality. This refined feature representation captures both the critical foreground actions and the useful background context, while suppressing irrelevant dependencies due to environmental biases. As a result, our model achieves improved generalization across diverse scenarios and more robust AQA.

3.3. Temporal Causal Attention on Stage Features (TCA)

After performing causal intervention on the fused features $\mathbf{F}$, the features are further decomposed into three distinct stage representations: $\mathbf{S}_{\text{forward}}$, $\mathbf{S}_{\text{twist}}$, and $\mathbf{S}_{\text{entry}}$. These stages represent different phases of an athlete's movement, such as forward, twist, and water entry. As shown in Fig. 3, to enhance the interpretability of the model and provide actionable insights for athletes, we introduce **Temporal Causal Attention** to quantify the influence of one stage on the next. This allows us to determine how much each stage affects subsequent stages, guiding athletes to focus on areas that need improvement.

3.3.1. Formulation of Temporal Causal Attention

The causal effect between different stages can be represented as:

$$P(\mathbf{S}_{t+1} \mid \text{do}(\mathbf{S}_t)), \tag{7}$$

which measures how an intervention on one stage influences the next. Since these relationships are temporally ordered, we model them using a *Causal Attention Mechanism* that captures directed dependencies while enforcing a strict forward temporal constraint.

For each stage feature representation $\mathbf{S}_t$, we apply a masked self-attention mechanism, ensuring that each stage can only attend to itself and previous stages. Given stage features $\mathbf{S} = [\mathbf{S}_{\text{forward}}, \mathbf{S}_{\text{twist}}, \mathbf{S}_{\text{entry}}]$, the temporal attention scores are computed as:

$$A_{ij} = \frac{\exp(\mathbf{Q}_i \cdot \mathbf{K}_j / \sqrt{d})}{\sum_{k \le j} \exp(\mathbf{Q}_i \cdot \mathbf{K}_k / \sqrt{d})}, \tag{8}$$

where $\mathbf{Q}$ and $\mathbf{K}$ are query and key representations of the stage features, and the denominator ensures causal masking by summing only over past and current stages.

3.3.2. Implementation via Transformer-Based Causal Attention

We implement temporal causal attention using a transformer encoder with masked self-attention layers. The causal attention module restricts information flow to prevent future stages from influencing earlier ones while capturing dependencies across multiple representation subspaces. Additionally, residual connections and layer normalization stabilize training and maintain gradient flow. The forward computation is given by:

$$\mathbf{S}_{t+1}^{(\text{refined})} = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}} + \mathbf{M}\right)\mathbf{V}, \tag{9}$$

where $\mathbf{M}$ is the causal mask ensuring $A_{ij} = 0$ for $j < i$.

By analyzing the learned attention weights A_{ij}, we can interpret the contribution of each stage to the subsequent one. If a low attention weight is assigned to a transition (e.g., $A_{\text{twist, entry}}$), it suggests that poor execution in the twist phase minimally impacts entry, or that the model fails to capture a meaningful connection. This allows us to provide targeted feedback by highlighting which stages significantly influence the final action score, enabling athletes to refine their techniques effectively.

4. Experiments

4.1. Datasets

FineDiving-HM. The FineDiving-HM [30] dataset comprises 3,000 videos encompassing 52 action types, 29 subaction types, and 23 difficulty levels. Each video is annotated with fine-grained temporal boundaries and official action scores. To enhance the reliability and interpretability of our model, human-centric action mask annotations are incorporated to distinguish target action regions from the background. FineDiving-HM includes 312,256 annotated

Methods	AQA Metrics	
	$\rho \uparrow$	$R\text{-}\ell_2 \downarrow (\times 100)$
C3D-LSTM [19]	0.6969	1.0767
C3D-AVG [18]	0.6801	0.6251
MSCADC [18]	0.7688	0.9372
I3D+MLP [22]	0.8227	0.4878
USDL [22]	0.8351	0.5104
MUSDL [22]	0.8240	0.4212
CoRe [31]	0.9308	0.3068
TSA [29]	0.9361	0.2746
HGCN [37]	0.9381	0.2321
FineParser [30]	0.9435	0.2602
FineCausal (Ours)	**0.9447**	**0.2338**

Methods	TAP Metrics	
	AIoU@0.5 $\uparrow$	AIoU@0.75 $\uparrow$
TSA [29]	0.9239	0.5007
FineParser [30]	**0.9946**	**0.9467**
FineCausal (Ours)	<u>0.9937</u>	<u>0.9453</u>

Table 1. Comparisons of performance with state-of-the-art AQA methods on the FineDiving-HM Dataset. The best result is highlighted in **bold**, and the second-best result is <u>underlined</u>.

mask frames across 3,000 videos, with each mask segmenting the relevant foreground action. The dataset mitigates the challenges associated with frame-level annotation requirements for recognizing human-centric actions in fine-grained spatial and temporal contexts. Among the 312,256 foreground action masks, 248,713 correspond to individual action frames, while 63,543 represent synchronized diving events. These statistics provide valuable insights for athletes and coaches to analyze competition strategies and assess the prevalence of specific actions in diving.

4.2. Evaluation Metrics

Action Quality Assessment. In line with prior research [16, 18, 19, 22, 29–31, 41], we evaluate our model using Spearman's rank correlation (ρ, where higher is better) and Relative ℓ_2 distance (R-ℓ_2, where lower is better), which quantify the model's ability to predict action scores.

Temporal Action Parsing. To assess the model's ability to segment action sequences, we utilize the Average Intersection over Union (AIoU) metric [29, 30]. AIoU measures the alignment between predicted and ground-truth temporal action boundaries, with higher values indicating superior segmentation accuracy.

4.3. Implementation Details

The training objective of our model is defined as the sum of three losses, $L = L_{\text{SAP}} + L_{\text{TAP}} + L_{\text{Reg}}$, where L_{SAP} is a focal loss applied to optimize the predicted human-centric action masks in the spatial action parser, ensuring accurate foreground extraction; L_{TAP} is a binary cross-entropy loss supervising the temporal action parser by comparing the predicted step-transition probabilities with the ground-truth transitions; and L_{Reg} is a mean squared error loss that refines the regression in the fine-grained contrastive module by minimizing the discrepancy between the predicted and ground-truth action scores. To further enhance the stability of multitask learning, we incorporate an automated loss weighting mechanism [7] that dynamically adjusts the contribution of each task loss based on its respective uncertainty, thereby preventing any single objective from dominating the training process. The overall training time of our model is approximately one day. We adhere to the experimental framework established in [30], employing the I3D [3] model pre-trained on the Kinetics dataset as the backbone for both the SAP and TAP modules. The learning rates for these modules are set to 10^{-3} for SAP and 10^{-4} for TAP, ensuring distinct parameter optimization. Additionally, the backbone network for shared feature extraction is initialized with a learning rate of 10^{-3}. Optimization is performed using the Adam optimizer with weight decay as 0. Following prior work [22, 29, 31], we extract 96 frames from each video and divide them into 9 snippets, where each snippet consists of 16 consecutive frames sampled at a stride of 10. The scale factor L' is set to 3, while the weighting parameters λ_l are assigned values 3, 5, 2. For training and evaluation, we adopt the exemplar selection strategy outlined in FineDiving-HM [30]. Consistent with previous studies [22, 29, 31], we allocate 75% of the dataset for training and 25% for testing.

4.4. Comparison with the State-of-the-Arts

To evaluate the effectiveness of our proposed method, we compare our approach, FineCausal, with state-of-the-art action quality assessment (AQA) models on the FineDiving-HM dataset. The results, presented in Tab. 1, demonstrate that FineCausal achieves competitive performance across multiple evaluation metrics.

In terms of AQA metrics, FineCausal attains a Spearman's rank correlation (ρ) of 0.9447, surpassing the previous state-of-the-art (FineParser) performance (0.9435) and establishing a new benchmark in the field. Notably, FineCausal achieves the best performance in Relative ℓ_2-distance (R_{ℓ_2}), significantly outperforming FineParser with a relative improvement of 10.16% (0.2338 vs. 0.2602). Compared to other models, FineCausal demonstrates substantial improvements, reducing the R_{ℓ_2} by 52.05% over I3D+MLP, 54.18% over USDL, and 23.73% over CoRe.

These improvements can be attributed to the GAT-based causal intervention, which effectively disentangles spurious correlations from genuine performance cues by selectively aggregating relevant features, including useful background information, and filtering out confounding signals.

For temporal action parsing (TAP), FineCausal attains an AIoU@0.5 score of 0.9907 and AIoU@0.75 of 0.9453, closely aligning with FineParser (0.9946 and 0.9467, respectively). While FineParser slightly outperforms our approach, FineCausal maintains strong performance with only a minor decrease of 0.39% in AIoU@0.5 and 0.15% in AIoU@0.75. We attribute this slight drop to the GAT-based causal intervention module: while it effectively disentangles causal features by integrating foreground and background information, the dynamic re-weighting process can introduce minor temporal inconsistencies, making the temporal boundaries of features slightly more vague. Despite this, the overall robustness in predicting temporal boundaries remains high.

Overall, our FineCausal method exhibits superior generalization capabilities, particularly in reducing prediction errors, as evidenced by the lowest R_{ℓ_2} score. This improvement highlights the advantage of leveraging causal attention mechanisms to mitigate spurious correlations, ultimately enhancing model reliability in real-world AQA tasks.

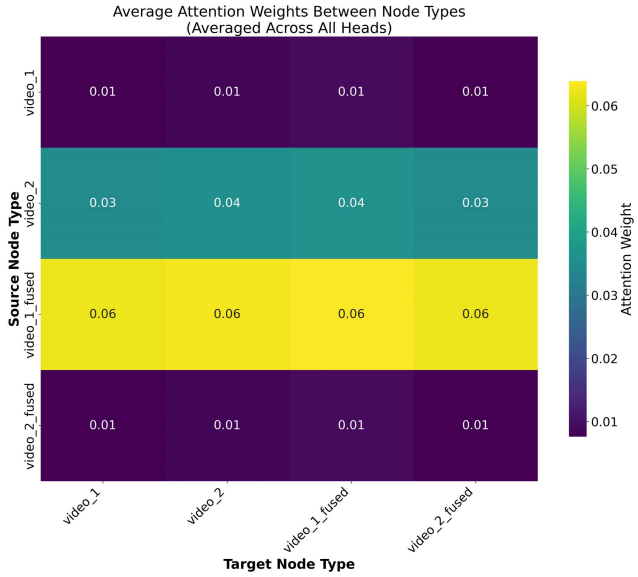
4.5. Ablation Studies

We conducted an ablation study on the FineDiving-HM dataset to demonstrate the effectiveness of individual parts of **FineCausal** by designing different modules that selectively incorporate GAT and TCA. Tab. 2 summarizes the experimental results.

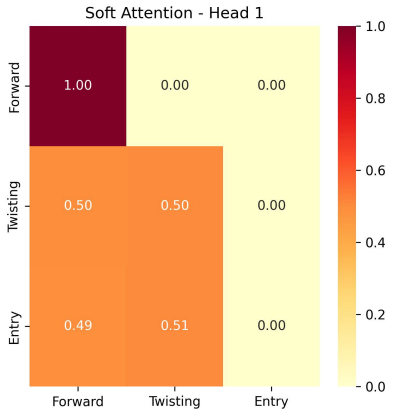
As shown in Tab. 2, we consider three variants: Method A (GAT only), Method B (TCA only), and Method C (GAT + TCA). Under Spearman's rank correlation, the AQA performance of the model with GAT (Method A) is 0.9425, while using only TCA (Method B) yields 0.9382. Incorporating both modules (Method C) further improves the correlation to 0.9447. Furthermore, the R-ℓ_2 error decreases from 0.2572 in Method A and 0.2707 in Method B to 0.2338 in Method C, indicating that the combined model provides a more precise AQA.

On the temporal action parsing side, significant improvements on AIoU@0.5 and AIoU@0.75 are closely related to the accuracy of the AQA task. Method B (TCA only) achieves the highest AIoU@0.5 of 0.9973 and AIoU@0.75 of 0.9666. Notably, Method C obtains AIoU@0.5 of 0.9937 and AIoU@0.75 of 0.9453, striking a balance between precise temporal segmentation and robust action scoring. These results demonstrate that TCA alone excels in temporal parsing, but when combined with GAT, the model still maintains strong TAP metrics while achieving the best AQA correlation.

In summary, using only GAT or only TCA leads to suboptimal performance in either action quality assessment or temporal action parsing. By integrating both modules, our final FineCausal model achieves the highest Spearman's rank correlation ($\rho = 0.9447$) and the lowest R-ℓ_2 error (0.2338), while maintaining competitive TAP accuracy. This confirms the complementary benefits of the GAT and TCA modules for comprehensive AQA.



(a) GAT attention map.



(b) Temporal sub-action attention map.

Figure 4. Visualization of attention mechanisms in our framework. (a) GAT attention weights between original and fused video features. (b) Temporal attention weights across different sub-action phases.

4.6. Visulisation

To intuitively illustrate how our framework leverages attention mechanisms for robust AQA, we present two sets of attention maps.

Figure 4a shows the GAT attention weights between original and fused video features. Although the fused

Methods	Modules		AQA Metrics		TAP Metrics	
	GAT	TCA	$\rho \uparrow$	$R\text{-}\ell_2 \downarrow (\times 100)$	AIoU@0.5 ↑	AIoU@0.75 ↑
A	✓	–	0.9425	0.2572	0.9920	0.9069
B	–	✓	0.9382	0.2707	**0.9973**	**0.9666**
C	✓	✓	**0.9447**	**0.2338**	0.9937	0.9453

Table 2. Merged ablation study results on different modules (GAT and TCA) in FineCausal on FineDiving-HM. Unavailable methods are omitted.

features of query and exemnplar video (*video_1_fused*, *video_2_fused*) generally receive higher attention, the original features of query and exemplar video (*video_1*, *video_2*) still exhibit non-negligible weights (around 0.01–0.04). This indicates that GAT adaptively integrates both masked (foreground) and raw (background) cues, reinforcing our premise that valuable contextual information in the original video frames should not be discarded. Consequently, the model benefits from a balanced representation that captures the athlete's motion and relevant environmental context, leading to a more comprehensive AQA.

Figure 4b provides a closer look at the attention distribution across different sub-action phases (e.g., *Forward*, *Twist*, *Entry*). The TSA module selectively highlights important temporal segments, assigning higher weights to sub-actions that significantly impact the overall execution quality. As shown, certain phases (e.g., *Forward*) can dominate the attention map in scenarios where the initial phase is critical to the athlete's performance. Conversely, when multiple phases (e.g., *Twist* and *Entry*) are equally crucial, the model balances its attention accordingly.

In addition, Figure 5 illustrates a real-world diving failure where the athlete ultimately receives a zero score. The initial mistake made in the *Forward* stage propagates through the *Twist* and *Entry* stages, leading to a completely unsuccessful dive. The TSA module in our framework is designed to capture such causal dependencies, revealing how early errors can influence later sub-actions. By assigning heightened attention to the stages most affected by the athlete's initial misstep, our model highlights the chain of events that results in the overall failure. This demonstration underscores the importance of robust temporal modeling in AQA, where a single error can have cascading effects on the final outcome.

5. Conclusion and Discussion

We presented **FineCausal**, a novel causal-based framework for AQA that integrates a Graph Attention Network–based causal intervention module and a temporal causal attention module. FineCausal effectively disentangles and balances human-centric foreground features with valuable background context, while the temporal causal attention module captures fine-grained temporal dependencies across action stages. Together, these modules enable the model to learn detailed spatio-temporal representations and provide interpretable feedback on action quality.

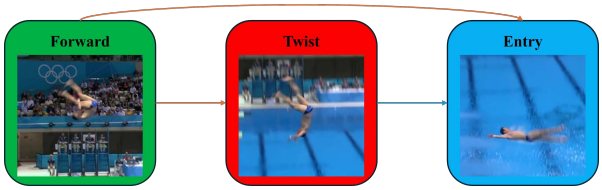


Figure 5. A failure case in which the athlete scores 0 due to an initial mistake in the *Forward* phase, negatively influencing subsequent *Twist* and *Entry* stages. The temporal causal attention highlights how an early misstep can propagate across phases, underscoring the importance of capturing causal dependencies in complex action sequences.

Despite achieving state-of-the-art performance and enhanced interpretability, our approach has some macro-level limitations. First, the causal framework often requires substantial expert knowledge to define the underlying causal relationships and design effective interventions, which can be a barrier for applications in domains with less well-established expert guidelines. Second, the high-quality, fine-grained annotations needed, such as those provided in the FineDiving-HM dataset, are critical for the success of causal-based methods, yet such datasets are expensive and time-consuming to create. We believe that addressing these challenges through semi-supervised learning [27] or automated annotation techniques, such as diffusion and Gaussian models [4, 25, 35], will be key to broadening the applicability of causality-based AQA systems. Overall, FineCausal establishes a promising baseline for interpretable, causal-based AQA and lays the groundwork for future research in both methodological innovation and dataset development.

Acknowledgment

This project is supported in part by the EPSRC NorthFutures project (ref: EP/X031012/1).

References

[1] Marianna Capecci, Maria Gabriella Ceravolo, Francesco Ferracuti, Sabrina Iarlori, Andrea Monteriu, Luca Romeo,

and Federica Verdini. The kimore dataset: Kinematic assessment of movement and clinical scores for remote monitoring of physical rehabilitation. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 27(7):1436–1448, 2019. 1

[2] Gianluca Carloni, Sotirios A Tsaftaris, and Sara Colantonio. Crocodile: Causality aids robustness via contrastive disentangled learning. In *International Workshop on Uncertainty for Safe Utilization of Machine Learning in Medical Imaging*, pages 105–116. Springer, 2024. 2

[3] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 6

[4] Ziyi Chang, George Alex Koulieris, and Hubert P. H. Shum. On the design fundamentals of diffusion models: A survey. *arXiv*, 2023. 8

[5] Mihai Fieraru, Mihai Zanfir, Silviu Cristian Pirlea, Vlad Olaru, and Cristian Sminchisescu. Aifit: Automatic 3d human-interpretable feedback models for fitness training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9919–9928, 2021. 1

[6] Andrew S Gordon. Automated video assessment of human performance. In *Proceedings of AI-ED*, page 10, 1995. 1

[7] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7482–7491, 2018. 6

[8] Ruochen Li, Stamos Katsigiannis, Tae-Kyun Kim, and Hubert P. H. Shum. Bp-sgcn: Behavioral pseudo-label informed sparse graph convolution network for pedestrian and heterogeneous trajectory prediction. *IEEE Transactions on Neural Networks and Learning Systems*, 2025. 4

[9] Ruochen Li, Tanqiu Qiao, Stamos Katsigiannis, Zhanxing Zhu, and Hubert P. H. Shum. Unified spatial-temporal edge-enhanced graph networks for pedestrian trajectory prediction. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025. 4

[10] Ziyang Liang, Yiwei Bao, and Feng Lu. De-confounded gaze estimation. In *European Conference on Computer Vision*, pages 219–235. Springer, 2024. 2, 3, 4

[11] Yang Liu, Guanbin Li, and Liang Lin. Cross-modal causal relational reasoning for event-level visual question answering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):11624–11641, 2023. 2

[12] Yanchao LIU, Xina CHENG, and Takeshi IKENAGA. A hierarchical joint training based replay-guided contrastive transformer for action quality assessment of figure skating. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, 2024. 1

[13] Mahdiar Nekoui, Fidel Omar Tito Cruz, and Li Cheng. Eagle-eye: Extreme-pose action grader using detail bird's-eye view. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 394–402, 2021. 1

[14] Weizhi Nie, Chen Zhang, Dan Song, Yunpeng Bai, Keliang Xie, and An-An Liu. Chest x-ray image classification: a causal perspective. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 25–35. Springer, 2023. 2, 4

[15] Lauren Okamoto and Paritosh Parmar. Hierarchical neurosymbolic approach for comprehensive and explainable action quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3204–3213, 2024. 1, 2

[16] 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*, pages 6331–6340, 2019. 6

[17] German I Parisi, Sven Magg, and Stefan Wermter. Human motion assessment in real time using recurrent self-organization. In *2016 25th IEEE international symposium on robot and human interactive communication (RO-MAN)*, pages 71–76. IEEE, 2016. 2

[18] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 304–313, 2019. 6

[19] Paritosh Parmar and Brendan Tran Morris. Learning to score olympic events. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 20–28, 2017. 2, 6

[20] Paritosh Parmar, Jaiden Reddy, and Brendan Morris. Piano skills assessment. In *2021 IEEE 23rd international workshop on multimedia signal processing (MMSP)*, pages 1–5. IEEE, 2021. 1

[21] Hamed Pirsiavash, Carl Vondrick, and Antonio Torralba. Assessing the quality of actions. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI 13*, pages 556–571. Springer, 2014. 2

[22] 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*, pages 9839–9848, 2020. 1, 2, 6

[23] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, Yoshua Bengio, et al. Graph attention networks. *stat*, 1050(20):10–48550, 2017. 4

[24] 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. 2

[25] Yin Wang, Mu Li, Jiapeng Liu, Zhiying Leng, Frederick WB Li, Ziyao Zhang, and Xiaohui Liang. Fg-t2m++: Llms-augmented fine-grained text driven human motion generation. *International Journal of Computer Vision*, 2025. 8

[26] Yutong Xia, Yuxuan Liang, Haomin Wen, Xu Liu, Kun Wang, Zhengyang Zhou, and Roger Zimmermann. Deciphering spatio-temporal graph forecasting: A causal lens and treatment. *Advances in Neural Information Processing Systems*, 36:37068–37088, 2023. 4

[27] Xuri Xin, Kezhong Liu, Huanhuan Li, and Zaili Yang. Maritime traffic partitioning: An adaptive semi-supervised spectral regularization approach for leveraging multi-graph evolutionary traffic interactions. *Transportation Research Part C: Emerging Technologies*, 164:104670, 2024. 8

[28] 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. 1

[29] 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*, pages 2949–2958, 2022. 1, 2, 6

[30] Jinglin Xu, Sibo Yin, Guohao Zhao, Zishuo Wang, and Yuxin Peng. 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*, pages 14628–14637, 2024. 2, 5, 6

[31] 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*, pages 7919–7928, 2021. 1, 2, 6

[32] Qiang Zhang and Baoxin Li. Relative hidden markov models for video-based evaluation of motion skills in surgical training. *IEEE transactions on pattern analysis and machine intelligence*, 37(6):1206–1218, 2014. 1

[33] 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*, pages 2405–2414, 2023. 2

[34] Yu Zhang, Wei Xiong, and Siya Mi. Learning time-aware features for action quality assessment. *Pattern Recognition Letters*, 158:104–110, 2022. 1

[35] Lizhi Zhao, Xuequan Lu, Runze Fan, Sio Kei Im, and Lili Wang. Gaussianhand: Real-time 3d gaussian rendering for hand avatar animation. *IEEE Transactions on Visualization and Computer Graphics*, 2024. 8

[36] Kanglei Zhou, Ruizhi Cai, Yue Ma, Qingqing Tan, Xinning Wang, Jianguo Li, Hubert PH Shum, Frederick WB Li, Song Jin, and Xiaohui Liang. A video-based augmented reality system for human-in-the-loop muscle strength assessment of juvenile dermatomyositis. *IEEE Transactions on Visualization and Computer Graphics*, 29(5):2456–2466, 2023. 1

[37] Kanglei Zhou, Yue Ma, Hubert PH Shum, and Xiaohui Liang. Hierarchical graph convolutional networks for action quality assessment. *IEEE TCSVT*, 33(12):7749–7763, 2023. 6

[38] Kanglei Zhou, Ruizhi Cai, Liyuan Wang, Hubert P. H. Shum, and Xiaohui Liang. A comprehensive survey of action quality assessment: Method and benchmark. *arXiv preprint arXiv:2412.11149*, 2024. 1

[39] Kanglei Zhou, Junlin Li, Ruizhi Cai, Liyuan Wang, Xingxing Zhang, and Xiaohui Liang. Cofinal: Enhancing action quality assessment with coarse-to-fine instruction alignment. In *IJCAI*, pages 1771–1779, 2024. 1

[40] Kanglei Zhou, Liyuan Wang, Xingxing Zhang, Hubert PH Shum, Frederick WB Li, Jianguo Li, and Xiaohui Liang. Magr: Manifold-aligned graph regularization for continual action quality assessment. In *ECCV*, pages 375–392, 2024. 1

[41] Kanglei Zhou, Zikai Hao, Liyuan Wang, and Xiaohui Liang. Adaptive score alignment learning for continual perceptual quality assessment of 360-degree videos in virtual reality. *IEEE Transactions on Visualization and Computer Graphics*, 2025. 1, 6