Hierarchical NeuroSymbolic Approach for Comprehensive and Explainable Action Quality Assessment

Lauren Okamoto
Princeton University
lokamoto@princeton.edu

Paritosh Parmar
IHPC, A*STAR, Singapore
paritosh.parmar@alumni.unlv.edu

Abstract

Action quality assessment (AQA) applies computer vision to quantitatively assess the performance or execution of a human action. Current AQA approaches are end-to-end neural models, which lack transparency and tend to be biased because they are trained on subjective human judgements as ground-truth. To address these issues, we introduce a neuro-symbolic paradigm for AQA, which uses neural networks to abstract interpretable symbols from video data and makes quality assessments by applying rules to those symbols. We take diving as the case study. We found that domain experts prefer our system and find it more informative than purely neural approaches to AQA in diving. Our system also achieves state-of-the-art action recognition and temporal segmentation, and automatically generates a detailed report that breaks the dive down into its elements and provides objective scoring with visual evidence. As verified by a group of domain experts, this report may be used to assist judges in scoring, help train judges, and provide feedback to divers. Annotated training data and code:
https://github.com/laurenok24/NSAQA.

1. Introduction

Analyzing people's movement and actions and providing feedback has applications ranging from physical rehabilitation [1, 19, 22] to sports coaching and scoring athletic performance [11, 15, 28] to surgery skills assessment [5, 6, 14]. With the *2024 Paris Olympics* just around the corner, there is heightened interest in **action quality assessment** (AQA) in the domain of sports such as diving, gymnastics, and figure skating. These automated systems are generally deep learning-based systems. Such neural approaches analyze videos of sports performance and output a score *quantifying how well* the athlete performed. These neural models are trained using scores given out by human judges during past competitions.

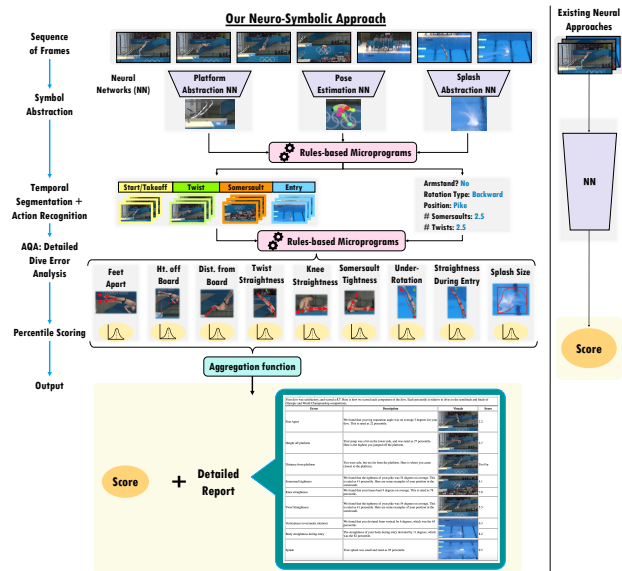The issue is that human judging itself is subjective and

Figure 1. **Neuro-Symbolic Action Quality Assessment (NS-AQA) vs Neural AQA.** Our NS-AQA approach (Left) employs neural networks to extract crucial symbolic information, such as platform location, framewise pose estimation, & splash detection. These symbols furnish objective data utilized for rules-based fine-grained action recognition, temporal segmentation, & detailed error analysis. The outcome is an objective score & a comprehensive visio-linguistic report, complete with supporting visual evidence, generated programmatically. This is much more valuable than existing AQA approaches (Right) that can only predict a single score (potentially biased) without any accompanying explanation. *Please zoom in; full-size version in supplementary.*

prone to bias. The bias may take two forms: bias for/against individual athletes; and bias in how different elements of a performative action are weighted in the overall score. An example of the latter in diving is bias towards the last phase of a dive—entry into the water and the amount of splash created. Because of this, the scores given out by the judges may not take into account all the factors/elements of the athletes' performance. This problem is further aggravated

by the fact that performance scores from human judges are never supported by an actual breakdown of the score or an analysis of what athletes did well or not. This 'black box' kind of scoring cannot guarantee that: 1) all the *factors/elements/efforts from athletes' performance are taken into account* when scoring; & 2) the *performance has been objectively scored*. This leads to a lack of trust, fairness, and explainability in the scoring. These issues are compounded when current AQA approaches, which are fully neural models [2, 4, 9, 10, 12, 17, 22, 28, 32], are trained using scores given by human judges.

To mitigate this set of problems, we propose a novel *neuro-symbolic* approach to fine-grained analysis of human movements & actions. We consider the popular and challenging sport of diving as the case study. Our neuro-symbolic approach combines the strengths of *deep neural models* and *rules-based AI*. Our hierarchical neuro-symbolic approach (see Figure 1) for fine-grained action analysis first *deconstructs* the whole performance and its surrounding environment/scene into *interpretable* "symbols" such as the platform, athletes' body poses, water splash, etc. *using neural models*. Then, using a *rules-based* approach *constructs* a *hierarchy of representations*—starting from identifying dive type/class and temporal segments of the dive phases to finally, fine-grained and detailed action quality assessment representation and score. Our system *programmatically* generates *detailed performance analysis reports* that not only check for all the performance errors, but also precisely compute their magnitudes/severities, which is hard to do even for highly-trained eyes, especially, when the actions/movements are high-speed as in case of diving. What is more, our system can sift through the entire performance and collect all the *visual evidence* to support the error detection. With our *transparent* and *objective* approach, performers can see that all their efforts have been rewarded, and bias in judging can be minimized. This is in stark contrast to current AQA approaches, or even human judges, which just give out a *single score without any explanation* of its breakdown.

The contributions of our work can be summarized as follows:

- We propose a Neuro-Symbolic paradigm for AQA that brings remarkable accountability, transparency, and trustworthiness to AQA and Sports Performance Judging. We validated our approach on competitive platform diving.
- Our system achieves state-of-the-art (SOTA) performance on fine-grained action recognition and temporal segmentation of dives.
- Our system programmatically generates a highly detailed report of the performance, including angular measurements as precise as 1 degree, & retrieving relevant images & video clips for the supporting visual evidence. This report may be used for various purposes: (a) as a potential aid to judges as they score the dive; (b) as a tool to teach judges how to score; (c) to settle disagreements between judges; (d) to encourage safety by penalizing dangerous actions (such as getting too close to hitting the platform); & (e) as a tool with an intuitive UI for coaches & divers to detect, quantify, & visualize performance errors.
- Our system generates statistically-grounded scores that compare a diver's performance relative to that of their peers. Generated scores are equally useful to all levels of diving competitions (from recreational to Olympics).
- We precisely annotate a large number of frames with splash, platform/springboard, and diver segmentation masks to be used for object detection training. We will opensource all the training data and models to help researchers.

Our neuro-symbolic approach to AQA can be extended to not only other sports, such as figure skating and gymnastics, but also to other precise complex actions, such as surgery. In each case, domain expertise is necessary to determine what symbols need to be abstracted and to create suitable rules for analysing and assessing of the action based on those symbols. We hope that further applications of our NS-AQA system can help reduce bias and objectively grade all sorts of skilled human actions.

2. Related Work

Existing approaches treat AQA as a regression problem. Thus far, efforts have focused on predicting action quality scores as closely as possible to those given by human judges. This is achieved through the use of better features [20], multitask learning [21], score distribution learning [25], more data [18], or improved regression techniques [7, 31]. However, the scores given by human judges, which serve as the 'gold standard,' are inherently biased. For instance, in diving, the last stage of the dive or the splash appears to have the most impact on judges, potentially neglecting other crucial factors of the athlete's performance. Therefore, at best, these neural approaches may assess in a biased manner.

Moreover, existing approaches being purely neural are 'black box' in nature—how they arrive at decisions & which factors of the performance are considered, as well as how each factor is scored, remain unknown. This lack of transparency results in a dearth of explainability, fairness, & trustworthiness, making it uncertain whether all the efforts from performers are genuinely rewarded. Ideally, athletes who have sacrificed a significant part of their lives should receive recognition for all their efforts. In contrast, we propose a new paradigm to AQA—a Neuro-Symbolic approach. Our approach leverages modern deep neural networks to extract useful information from performance videos into symbols. It then applies logic to these symbols to make decisions and evaluate performance without relying

Platform. The location of the platform, especially the position of its edge facing the pool, is crucial to determine when the diver leaves the platform, thus starting their dive. The platform location is also important to assess how close the diver comes to its edge, which is relevant to scoring.
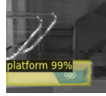
Splash at entry into the pool is a conspicuous visual feature of a dive. The size of the splash is an important element in traditional scoring of dives. A large splash mars the end of a dive and also likely indicates a flaw in form at water entry.

Pose of the diver in video frames is crucial for dive assessment. We gather 2D pose data, detailing body part locations like the head, thorax, pelvis, etc. Analyzing this data frame by frame allows identification of sub-actions (e.g., somersaults, twists) and quality assessment. Detailed pose data enables objective metrics like water entry angle.
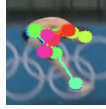
Table 1. **Symbols, their descriptions and visualizations**.

on biased ground-truths, but rather utilizes domain knowledge for objective scoring. Our system conducts a much more comprehensive and in-depth evaluation, programmatically generating an extremely detailed performance evaluation report. This makes the evaluation process transparent, fair, explainable, and trustworthy.

Neurosymbolic AI which combines the power of neural models and symbolic systems has been gaining increasing interest [3, 13, 23, 30], but has not been explored in AQA setting. In this work, we develop a hierarchical neurosymbolic approach to AQA.

3. Hierarchical Neuro-Symbolic Approach

Our neuro-symbolic action analyzer consists of two parts: 1) **Neural Action-Context Parser**; 2) **Rules-based Action Analyzer**. The Neural Action-Context Parser **deconstructs** or decomposes the action or the performance video, extracting meaningful information such as the *pose of the diver* in terms of position of joints, *diving platform segmentation*, and the *splash* created when the diver enters the water. This extracted information forms our **symbols**, which are then passed to the Rules-based Action Analyzer. The Rules-based Action Analyzer scrutinizes actions through structured and interpretable symbolic reasoning. Note that, our contribution does not lie in proposing merely a two-stage approach, but in integrating neural models, which excel at feature extraction; and symbolic systems, which provide transparent and objective analytics. We will now explain each of these parts in detail.

3.1. Neural Action-Context Parser

We implement the Neural Action-Context Parser as a collection of a dedicated: 1) *splash detector*; 2) *platform detector*, and 3) *human pose estimator*. These object detectors [27] and pose estimators [24] first extract relevant information from each frame individually. For instance, we run a splash detector on all frames; whenever a splash is detected, it is segmented out. Object segmentation provides access to its properties, such as 1) *size*, 2) *shape*, and 3) *position* within the scene. In cases where no splash is detected, all properties are filled with null values. Similarly, pose estimators record *2D positions* of all *major joints* of the diver. Direct information from these detectors forms the **primary set of symbols**. We further *derive relations* between the primary symbols, such as: 1) the *angles made by the bones at joints*, and 2) the *distance between the diver and the platform*. These form a **secondary set of symbols**. Additional information is provided in Table 1.

3.2. Rules-based Action Analyzer

This module takes in the symbols from the Neural Action-Context Parser and *analytically* processes them using the rules formed by a domain expert[1]. The rules are in accord with what is taught in the *official USA Diving Judge's Course* [8], and were further *verified by 7 other domain experts*. In practice, these rules are implemented in *Python* language [26] and termed as '**microprograms**.' Furthermore, action analysis is done in a *hierarchical* manner. Specifically, the action analyzer: 1) first detects various *class details of a dive*; 2) using these details, it then *temporally segments the dive into its phases*; and 3) lastly, using the dive and its phase details, the *quality of the dive is analyzed at a much finer granularity*. We discuss each level of this hierarchy and how they aid in the processing of the subsequent level in detail in the following.

3.2.1 Detailed Dive Recognition

Preliminary. In competitive diving, dives are identified on the basis of their: 1) **dive group** (*forward, backward, reverse, inward, twisting, armstand*); 2) **number of somersaults** (number of somersaults are chosen in *half-rotation increments*); 3) **number of twists** (chosen in *half-rotation increments*); 4) **body position** (*straight, pike, tuck, free*).

Rules-based dive recognizer. Dive recognition with our neuro-symbolic approach[2] is accomplished **analytically** from framewise pose data. Based on the abstracted symbols, the **dive group is identified** from: 1) *the direction the diver is facing* and 2) *the body movement* (*clockwise* or *counterclockwise* rotation) at the *start of the dive*. The **body position** (straight, tuck, pike, or free) is detected by the *degree of hip and knee bends*. **Somersault counts** are determined by *tracking the rotation of the vector* that points from the diver's *pelvis to thorax*. **Twists are counted** by *tracking*

---

[1]In our case, one of the authors is a diving expert with more than 10 years of experience.

[2]In contrast, recognizing a specific dive using a purely neural approach involves labeling a large number of dives to train a deep learning model.
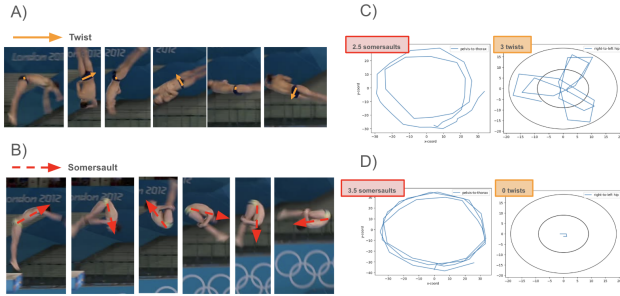
Figure 2. **Somersault and Twist Counters Visualized**. **A)** : Visualization of right-to-left hip vector. We count **twists** by counting the "**petals**" formed by this vector. **B)** : Visualization of pelvis-to-thorax vector. We count **somersaults** by *counting the rotations* of this vector over the course of the dive. **Graphical Solutions (C&D)** : We then show how the somersault and twist counters are applied to two different dives in C&D. The blue trace represents the vector we track, and the **black** circles in the twist plots represent the *boundaries* of when "petals" are counted. The petal must surpass the inner black circle while staying inside the outer black circle to count as a petal. Each petal is 0.5 twists. **C)** : A forward dive with 2.5 somersaults (2.5 revolutions from the initial vector vertically up ($\uparrow$) to final vector vertically down ($\downarrow$) in the somersault plot) and 3 twists (6 "petals" in the twist plot). **D)** : A forward dive with 3.5 somersaults (3.5 revolutions from initial vector vertically up ($\uparrow$) to final vector vertically down ($\downarrow$) in the somersault plot) in pike position and no twists (zero "petals" in the twist plot).

the magnitude and direction of the vector that points from the *diver's right to left hip joint*. For more details on how these rules are derived, see twisting and somersault vector plots of example dives in Figure 2. In there, we have shown how we first obtain graphical solutions to, for example, somersault and twist counting. These graphical solutions are then implemented in Python language. The pseudocode for the somersault counting microprogram is provided in **Algorithm 1** for reference. Further details on the microprograms that execute dive recognition and additional implementation details are provided in the supplementary material.

3.2.2 Temporal Segmentation

In this stage, the *primary symbolic information*, along with *dive class details* from the action recognition stage, is leveraged to perform the temporal segmentation of the dive. Each dive may be segmented into **start/takeoff**, **twist**, **somersault**, and **entry** phases (refer to Figure 3). Similar to detailed dive recognition, temporal segmentation is also accomplished **analytically** using **microprograms** that operate on *symbols* and *dive class details*—without the need to train a deep learning model with labeled video data for this purpose. For example, to *detect if a diver is in the*

Algorithm 1 Somersault Counter Microprogram

$half\_som\_count \leftarrow 0$
$armstand \leftarrow is\_armstand(dive\_frames)$
$curr\_pose \leftarrow None$
$vertical\_up\_vector \leftarrow [0, 1]$
$vertical\_down\_vector \leftarrow [0, -1]$
for each $frame$ in $dive\_frames$ **do**
    $curr\_pose \leftarrow get\_pose(frame)$
    if $curr\_pose$ is not None and $frame$ not in takeoff phase
then
        $thorax \leftarrow curr\_pose[thorax]$
        $pelvis \leftarrow curr\_pose[pelvis]$
        $vector1 \leftarrow thorax - pelvis$
        if ($armstand$ and $half\_som\_count$ is odd) or (not
$armstand$ and $half\_som\_count$ is even) **then**
            $vector2 \leftarrow vertical\_down\_vector$
            $angle \leftarrow get\_angle\_degrees(vector1, vector2)$
        else
            $vector2 \leftarrow vertical\_up\_vector$
            $angle \leftarrow get\_angle\_degrees(vector1, vector2)$
        end if
        if $angle \leq 75$ **then**
            $half\_som\_count \leftarrow half\_som\_count + 1$
        end if
    end if
end for
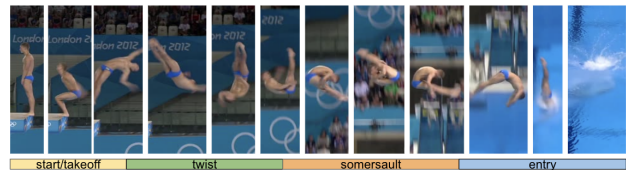**return** $half\_som\_count$

Figure 3. **Temporal Segmentation Visualization.** A dive is segmented into 4 *phases*: *start/takeoff*, *twist*, *somersault*, and *entry*.

start/takeoff phase, we *track the diver in relation to the end of the platform*. If the diver is *beyond a threshold distance away from the platform*, this signifies the *end of the start/takeoff phase* in that the diver has *jumped* and *started* their dive. The *other phases are detected* using similar *logic-based algorithms*. Refer to the supplementary materials for specific implementation details.

3.2.3 Fine-grained Dive (Action) Quality Assessment

Now, we utilize the *primary symbols* and output from the *dive recognition and segmentation* to analyze them in extreme detail, as if "under a microscope." Similar to previous stages, action quality assessment is done **analytically** using our **AQA-micro-programs**. *Output from previous stages* enables our **AQA-meta-program** to **select the correct microprogram** to execute to assess the quality of an element.

Distance-from-platform of a diver is crucial for safety and aesthetic reasons. Divers risk injury or even death by hitting the concrete platform by being too close to it, leading to penalties. Conversely, being too far from the platform results in an unaesthetic trajectory and is also penalized.

Feet-apart. A streamlined and graceful form throughout a dive is considered ideal. Taking a cue from gymnastics, one aspect of an ideal diving form is maintaining one's feet together. A diver is penalized for having their feet apart.
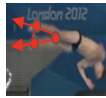
Somersault tightness. During a somersault, it is desirable for a diver to have as tight a tuck or pike position as possible. A very tight somersault position allows a diver to spin faster and is considered ideal.

Over/under-rotation. Each particular dive has a specified number of rotations of somersaults to be performed. The number of rotations is in half-rotation increments. A diver preferably performs the exact number of rotations (to the degree) that are specified for the dive–no more and no less. Over rotation and under rotation are penalized.
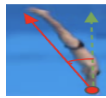
Splash size. Ideally, a diver has little or no splash upon entry into the water. Little or no splash after a dive from 10 meters high is difficult to achieve and is indicative of very good form upon entry.
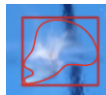
Table 2. **Examples of elements of dive quality, their descriptions and visualizations.** Image annotations shown in Red. See all of the element descriptions/visualizations in supplementary.

For example, if the diver is detected to be in *twisting phase*, then *feet apart AQA microprogram* is executed to assess the performance w.r.t. feet apart aspect. As another example, *meta-program* will *not* execute *Knee Straightness microprogram* if the diver is performing *tuck dive*, because *tuck position* does *not* have *Knee Straightness criteria*—an example is illustrated in last row of Table 9. We have discussed various elements of action/dive quality in Table 2. Further implementation details are provided in supplementary.

Quantification of quality of each element. We then quantify each of nine aspects of a dive as percentage scores, where 100% is best (no error), and 0% is worst (largest error). In other words, each error is given a score equal to its percentile ranking relative to all platform dives in the dataset. For example, for feet apart, if the average angle is smaller than 80% of the dives, then the feet apart error would be scored 80%. Advantageously, the *percentile rankings inform the diver of their performance relative to their peers* which is the most relevant statistical information to have for a competitor trying to improve their results. Furthermore, the scoring *automatically adjusts to different levels of competition*. For example, scoring for high school divers would give percentiles relative to a dataset of high school divers. Note that, instead of percentiles, scores could be given on an absolute basis. For example, a fixed scale could convert the angle of feet apart to a score based

on thresholds or a formula. Advantageously, such a fixed scale would allow direct comparison across different levels of competition. However, such a fixed scale would be somewhat arbitrary and likely require value judgements to be made which are beyond the scope of this paper. Hence, an absolute scoring system is not presented here.

Overall Score. An overall score is obtained by **aggregating** the *percentiles* of all the various aspects of the dive. We use *uniform weighted averaging* as the aggregation function[3]. Other suitable functions may be used. By using *percentiles*, we see how good the dive was compared to the level of dives in the semifinals and finals of Olympic and World Championship competitions. Using our system, it is also possible to use a weighted average instead of a uniformed average to emphasize certain aspects over others. While each phase is supposed to be weighted equally, errors that can occur within each phase have differing levels of importance depending on the personal preferences of the judges themselves. For example, some judges will weigh splash size more than verticalness, and some will do the exact opposite. Our system potentially allows for the **personalization** of weights for each error of the dive, giving each judge more flexibility in their scoring. Unlike conventional human judging, this personalization would be transparently presented, rather than hidden within a single overall score.

Visio-Linguistic Report generation. Our NS-AQA system programmatically generates detailed performance reports in *natural language*, based on the outputs discussed above. To do so, we take a simple approach without needing any large language models or deep learning models for language generation. We create a *small library of template-sentences* with *blanks to be filled in*. Again, using a *Rules-based approach*, a *microprogram* fetches an *appropriate template-sentence* from the library. For example, for feet apart error, it would select to use the following template-sentence: `We found that your leg separation angle was on average <insert average feet apart angle> degrees for your dive.` The `<insert average feet apart angle>` is replaced with the average feet separation determined from the process discussed above. This process is repeated for all the errors. These errors and description sentences are *compiled in the form of HTML page* as shown in Table 6. Furthermore, our NS-AQA system also maintains a *lookup-table* of error types with corresponding frame numbers and values of error severities. Using this lookup-table, it

---

[3]According to the USA Diving Judge's Course [8], each phase of the dive is supposed to be weighted equally. In particular, a judge can deduct up to 2 out of 10 points for each phase of the dive. Thus, we choose to average each aspect uniformly in order to give fairly equal importance to all of the aspects and phases of a dive.

Action Element	Nibali et al. [16]	MSCADC [21]	C3D-MTL [21]	Ours NS-AQA
Armstand	98.30	97.45	99.72	**99.79**
Rotation Type	78.75	84.70	97.45	**99.37**
Position	74.79	78.47	96.32	**97.28**
No. of SS	77.34	76.20	96.88	**97.31**
No. of TW	79.89	82.72	93.20	**93.27**

Table 3. **Performance evaluation on the task of fine-grained action recognition.** Accuracy (%) is used as the performance metric. *Higher is better.* SS: Somersaults; TW: Twists.

Model	AIoU@0.5	AIoU@0.75
TSA [29]	82.51	34.31
Ours NS-AQA	**93.92**	**77.17**

Table 4. **Performance evaluation on the task of temporal segmentation.** Average Intersection over Union (AIoU) is used as the performance metric. *Higher is better.*

can *retrieve* the frames contributing to a particular type of error and include them in the report as *supporting visual feedback*. This detailed report is helpful for a number of reasons including as a support to human judges and as an educational tool to teach coaches, athletes, and judges how to score. More specific implementation details are provided in the supplementary.

4. Results

We evaluate our system's performance on dive action recognition, temporal segmentation, and AQA. We compare temporal segmentation and action recognition performance to that of SOTA neural models. We evaluate AQA performance by surveying domain experts (Expert Survey) and comparing its output to a high performing purely neural model, C3D-MTL [21].

4.1. NS Action Recognition

We first evaluate the performance of our NS approach on detailed dive (action) recognition on MTL-AQA dataset [21]; & compare its performance with SOTA methods [21], which are purely neural in nature. Table 3 shows that our NS action recognition approach outperforms C3D-MTL [21] on all categories. Note that, unlike neural approaches, *our system also has the advantage of not requiring extra training specific to action recognition or temporal segmentation.*

4.2. NS Action Temporal Segmentation

Next, we evaluate the performance of our NS approach on temporal segmentation on FineDiving dataset [29], and compare it with SOTA Temporal Segmentation Attention

Element	Agreement (%)
Overall Score	92.5
Feet Apart	96.0
Height off Board	97.3
Distance from Board	92.1
Somersault Tightness	94.5
Knee Straightness	100.0
Twist Tightness	97.6
Verticalness (over/under-rotation)	90.5
Body Straightness During Entry	94.5
Splash Size	98.0

Table 5. **Expert Survey Results** *Higher is better.*

(TSA) model [29] in Table 4. We observe that Our NS temporal segmentation model outperforms TSA model [29]. Furthermore, the performance gap widens as the segmentation precision is increased from Average Intersection over Union (AIoU)@0.5 to 0.75.

4.3. NS Action Quality Assessment

Unlike temporal action segmentation & recognition, we do not evaluate our AQA scores on "ground-truth" labels because our system proposes a new objective, and comprehensive way of scoring dives. Thus, our predicted scores are expected to be different than the subjective scores given by human judges of the current scoring system (as mentioned in Sec. 1). To evaluate model's performance in the absence of groundtruth labels, we ask a group of domain experts to evaluate the accuracy of our model & give their feedback.

4.3.1 Expert Survey

We surveyed *6 domain experts* to verify the accuracy of our system's outputted scores & reports. These domain experts include trained judges, divers, & coaches who have judged/competed/coached on the international & national level. We showed each expert 50 randomly chosen dives from MTL-AQA dataset [21] & asked whether they agreed or disagreed with each dive's overall score & individual error scores. The results are shown in Table 5. Each error (feet apart, distance from board, etc.) was approved by all experts as objective criteria that should be used to score the dive. Experts agreed with our system on all categories at least 90% of the time.

4.3.2 Expert Feedback

To validate the practical usefulness of our system in relation to diving competitions & training, we presented our system & its automatically generated reports (explained in Approach) to an *Olympic diving coach* (referred to as Coach

Your dive was satisfactory, and scored a **5.7**. Here is how we scored each component of the dive. Each percentile is relative to dives in the semifinals and finals of Olympic and World Championship competitions.

Error	Description	Visuals	Score
Feet Apart	We found that your leg separation angle was on average 5 degrees for your dive. This is rated as 22 percentile.		2.2
Height off platform	Your jump was a bit on the lower side, and was rated as 27 percentile. Here is the highest you jumped off the platform.		2.7
Distance from platform	You were safe, but too far from the platform. Here is where you came closest to the platform.		Too Far
Somersault tightness	We found that the tightness of your pike was 54 degrees on average. This is rated as 41 percentile. Here are some examples of your position in the somersault.		4.1
Knee straightness	We found that your knees bent 9 degrees on average. This is rated as 78 percentile.		7.8
Twist Straightness	We found that the tightness of your pike was 54 degrees on average. This is rated as 41 percentile. Here are some examples of your position in the...		5.5
Verticalness (over/under rotation)	We found that you deviated from vertical by 6 degrees, which was the 63 percentile.		6.3
Body straightness during entry	The straightness of your body during entry deviated by 11 degrees, which was the 82 percentile.		8.2
Splash	Your splash was small and rated as 95 percentile.		9.5

Your dive was deficient, and scored a **3.4**. Here is how we scored each component of the dive. Each percentile is relative to dives in the semifinals and finals of Olympic and World Championship competitions.

Error	Description	Visuals	Score
Feet Apart	We found that your leg separation angle was on average 5 degrees for your dive. This is rated as 25 percentile.		2.5
Height off platform	Your jump was a bit on the lower side, and was rated as 12 percentile. Here is the highest you jumped off the platform.		1.2
Distance from platform	You were safe, but too far from the platform. Here is where you came closest to the platform.		Too Far
Somersault tightness	We found that the tightness of your pike was 49 degrees on average. This is rated as 60 percentile. Here are some examples of your position in the somersault.		6.0
Knee straightness	We found that your knees bent 33 degrees on average. This is rated as 58 percentile.		5.8
Twist Straightness	We found that the tightness of your pike was 49 degrees on average. This is rated as 60 percentile. Here are some examples of your position in the...		5.1
Verticalness (over/under rotation)	We found that you deviated from vertical by 15 degrees, which was the 24 percentile.		2.5
Body straightness during entry	The straightness of your body during entry deviated by 66 degrees, which was the 2 percentile.		0.2
Splash	Your splash was on the larger side and rated as 21 percentile.		2.1

Your dive was good, and scored a **7.5**. Here is how we scored each component of the dive. Each percentile is relative to dives in the semifinals and finals of Olympic and World Championship competitions.

Error	Description	Visuals	Score
Feet Apart	We found that your leg separation angle was on average 1 degrees for your dive. This is rated as 71 percentile.	There were no particular instances to show where your feet came apart.	7.1
Height off platform	Your jump was good, and was rated as 58 percentile. Here is the highest you jumped off the platform.		5.8
Distance from platform	You were a good distance from the platform. Here is where you came closest to the platform.		Good
Somersault tightness	We found that the tightness of your tuck was 36 degrees on average. This is rated as 54 percentile. Here are some examples of your position in the somersault.		5.4
Verticalness (over/under rotation)	We found that you deviated from vertical by 5 degrees, which was the 69 percentile.		6.9
Body straightness during entry	The straightness of your body during entry deviated by 2 degrees, which was the 98 percentile.		9.8
Splash	Your splash was small and rated as 77 percentile.		7.7

Table 6. **Detailed and accurate reports (three samples shown here) programmatically generated by our NS-AQA approach.** *Please zoom in.* More samples in supplementary.

A for anonymity), & a *a certified judge on the USA Diving Judges Commission* (referred to as Judge B). Both of them have generously offered valuable feedback for our system.

Expert Opinion 1: Coach A. According to Coach A, human judges emphasize overall impression when scoring a dive. This is because they cannot always see all of the diver's mistakes in the short amount of time watching the dive so as to come up with a score based on specific errors. Even if judges try to be fair to all aspects and phases of a dive, as humans, it is almost impossible to see the dive perfectly at all times. Moreover, judging based on overall impression is inherently subjective and vulnerable to bias. Coach A believes that with the support of our NS-AQA system, scoring in diving can be "*more precise and have less errors.*" Coach A believes that this system, with its generated detailed report, would be helpful in a number of ways. In particular, it would be useful as a tool to (1) teach judges how to score, (2) catch any errors that a human judge may miss, (3) settle disagreements between judges, and (4) encourage safety by penalizing dangerous actions (e.g. getting too close to hitting the board) that lots of human judges overlook. See full opinion in supplementary.

Expert Opinion 2: Judge B. Judge B envisioned our system as an educational tool for teaching judges, coaches, and divers how to break down the dive into all of its components. "*We're humans, we can't get it all right,*" Judge B says, especially since judges only get 3 seconds to score a dive. Judge B particularly emphasized our system's potential to improve diving safety due to its automated "distance from board" measurement. In diving, hitting the platform can be the difference between life and death, so it is very important to prioritize the diver's safety. Judge B is "*very excited for this system to be able to make that call for us,*" as it moves the *villain*-role from the human judges to the AI

Model Output	Expert Preference (%)
Fully Neural Score	3.9
Our Neuro-Symbolic Report	**96.1**

Table 7. **Comparison of experts' preference for Our NS model vs. fully neural models.** *Higher is better.*

Successes	Failures

Table 8. **Pose estimation qualitative results.**

to make those difficult decisions. See full opinion in supplementary.

4.3.3 Comparison to Neural Approach C3D-MTL

We additionally surveyed domain experts to compare the predicted scores outputted by our system and neural models. We particularly compare our output to that of C3D-MTL. Experts were each shown 50 randomly chosen dives from MTL-AQA dataset [21], and were asked to indicate which model output they agreed with more (either our NS model or the C3D-MTL model). We found that experts chose our system **96.1%** of the time[4] (Table 7). A qualitative comparison of C3D-MTL and our system can be seen in Table 9.

---

[4]For most of the dives when experts chose the neural output over ours, our system had failed pose estimation on some of the dive's frames. This affected the final score that was outputted.
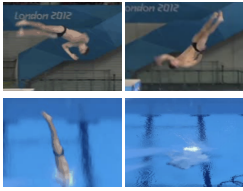
Key Frames of Dive	Our Output (Summarized, not full)	Neural Model Score	Comparison
	Overall Score: 5.7 Feet Apart: 2.2 Height off Platform: 2.7 Distance from Platform: Too Far Somersault Tightness: 4.1 Knee Straightness: 7.8 Twist Straightness: 5.5 Verticalness: 6.3 Body Straightness during Entry: 8.2 Splash: 9.5	**Overall Score: 9.7** *Note that, existing neural models only output a final score, unlike our model, which outputs a detailed report along with the final score.*	The score outputted by C3D-MTL was significantly higher than our system. We see by looking at the dive's key frames that the **start/takeoff and flight was loose and sloppy**, but the entry was extremely good. This suggests that the **neural model was only factoring the entry** into the overall score instead of also accounting for the beginning parts of the dive.
	Overall Score: 5.4 Feet Apart: 3.9 Height off Platform: 0.1 Distance from Platform: Good Somersault Tightness: 0.9 Knee Straightness: 7.8 Twist Straightness: 6.0 Verticalness: 4.8 Body Straightness during Entry: 9.5 Splash: 5.9	**Overall Score: 7.8**	The score outputted by C3D-MTL was much higher than our system. We see by looking at the dive's key frames that the height off the platform was **very low** and the **pike in the somersault was not tight**. However, the entry was with **little splash**. This suggests that the **neural model was mostly factoring the splash** into the overall score instead of also accounting for factors like height and position tightness.
	Overall Score: 6.4 Feet Apart: 8.8 Height off Platform: 8.5 Distance from Platform: Good Somersault Tightness: 6.0 Knee Straightness: N/A Twist Straightness: N/A Verticalness: 1.3 Body Straightness during Entry: 9.8 Splash: 0.0	**Overall Score: 3.4**	The score outputted by C3D-MTL was much lower than our system. Every aspect of the dive including the **height and position tightness was very good** except for a slight mistiming when the diver came out of the dive, resulting in **over-rotation**. The diver should be rewarded for the other components of the dive they did well, instead of only penalized for their poor entry.

Table 9. **Qualitative Comparison of Our Neural-Symbolic System to the Neural Model C3D-MTL.** As taught in the official USA Diving Judge's Course, "Dives should be judged as a whole, not emphasizing any single element. This approach in diving is especially true when it comes to the entry. It is easy to forgive earlier flaws if the dive has a good entry." [8] In all three examples, we see that the neural model emphasizes the entry of the dive far more than any other aspect, while our system more fairly accounts for each aspect.

4.3.4 Failure Modes of Our System

While our results are promising, our system occasionally fails to output a score that experts agree with. A main contributing factor to this is when the pose estimation on the diver is inaccurate. Particularly, pose estimation struggles when the diver is halfway underwater, and when the motion blur on the image is very strong. Examples of failed pose estimation compared with good pose are shown in Table 8. We believe that as pose estimation technology progresses, so will the accuracy of our system.

5. Discussion

We propose a neuro-symbolic (NS) paradigm for AQA, & implement an NS-AQA system specifically for scoring competitive diving. Our NS system achieves SOTA results for fine-grained action recognition & temporal segmentation, & automatically generates a detailed & comprehensive report scoring the performance of the dive. Our system was validated by domain experts including an Olympic div-

ing coach & a certified judge on the Judges Commission. Our approach ensures fairness and transparency, minimizing biases that often plague traditional assessment methods. Unlike traditional computer vision-based systems & human judges that offer single scores without explanation, our NS approach provides a breakdown of assessments, empowering athletes & coaches with actionable insights.

Extension to other actions and skills assessment. Although this paper is geared towards diving, our NS approach is general in nature, readily extensible to other sports like the vault in gymnastics & figure skating. Domain expertise would be required to formulate rules. Our approach can be extended to other critical domains like surgery skills assessment as well. In surgery skills assessment, for example, microprograms to detect bleeding (analogous to splash in our case), respect for tissue, etc. could be developed. Final rating would be an aggregation of all such factors. We hope to inspire this work's application across diverse domains.

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] Amirhossein Dadashzadeh, Shuchao Duan, Alan Whone, and Majid Mirmehdi. Pecop: Parameter efficient continual pretraining for action quality assessment. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 42–52, 2024. 2

[3] Artur S. d'Avila Garcez, Krysia Broda, and Dov M. Gabbay. Neural-symbolic learning systems - foundations and applications. In *Perspectives in Neural Computing*, 2012. 3

[4] Hazel Doughty, Dima Damen, and Walterio Mayol-Cuevas. Who's better, who's best: Skill determination in video using deep ranking. 2

[5] 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*, pages 6057–6066, 2018. 1

[6] 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*, pages 7862–7871, 2019. 1

[7] Hiteshi Jain, Gaurav Harit, and Avinash Sharma. Action quality assessment using siamese network-based deep metric learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(6):2260–2273, 2020. 2

[8] Amy Kwan. Usa diving elearning center. 3, 5, 8

[9] Qing Lei, Ji-Xiang Du, Hong-Bo Zhang, Shuang Ye, and Duan-Sheng Chen. A survey of vision-based human action evaluation methods. *Sensors*, 19(19):4129, 2019. 2

[10] Yongjun Li, Xiujuan Chai, and Xilin Chen. End-to-end learning for action quality assessment. In *Pacific Rim Conference on Multimedia*, pages 125–134. Springer, 2018. 2

[11] Yongjun Li, Xiujuan Chai, and Xilin Chen. Scoringnet: Learning key fragment for action quality assessment with ranking loss in skilled sports. In *Asian Conference on Computer Vision*, pages 149–164. Springer, 2018. 1

[12] 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*, pages 0–0, 2019. 2

[13] Chen Liang, Wenguan Wang, Jiaxu Miao, and Yi Yang. Logic-induced diagnostic reasoning for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16197–16208, 2023. 3

[14] Daochang Liu, Qiyue Li, Tingting Jiang, Yizhou Wang, Rulin Miao, Fei Shan, and Ziyu Li. Towards unified surgical skill assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9522–9531, 2021. 1

[15] Shenglan Liu, Xiang Liu, Gao Huang, Hong Qiao, Lianyu Hu, Dong Jiang, Aibin Zhang, Yang Liu, and Ge Guo. Fsd-10: A fine-grained classification dataset for figure skating. *Neurocomputing*, 413:360–367, 2020. 1

[16] Aiden Nibali, Zhen He, Stuart Morgan, and Daniel Greenwood. Extraction and classification of diving clips from continuous video footage. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 38–48, 2017. 6

[17] 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. 2

[18] Paritosh Parmar and Brendan Morris. Action quality assessment across multiple actions. In *2019 IEEE winter conference on applications of computer vision (WACV)*, pages 1468–1476. IEEE, 2019. 2

[19] Paritosh Parmar and Brendan Tran Morris. Measuring the quality of exercises. In *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 2241–2244. IEEE, 2016. 1

[20] 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

[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*, pages 304–313, 2019. 2, 6, 7

[22] Faegheh Sardari, Adeline Paiement, Sion Hannuna, and Majid Mirmehdi. Vi-net—view-invariant quality of human movement assessment. *Sensors*, 20(18):5258, 2020. 1, 2

[23] Jennifer J Sun, Megan Tjandrasuwita, Atharva Sehgal, Armando Solar-Lezama, Swarat Chaudhuri, Yisong Yue, and Omar Costilla-Reyes. Neurosymbolic programming for science. *arXiv preprint arXiv:2210.05050*, 2022. 3

[24] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation, 2019. 3

[25] 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)*, 2020. 2

[26] Guido Van Rossum and Fred L Drake Jr. *Python tutorial*. Centrum voor Wiskunde en Informatica Amsterdam, The Netherlands, 1995. 3

[27] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019. 3

[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, 2

[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. 6

[30] Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Josh Tenenbaum. Neural-symbolic vqa: Disentangling reasoning from vision and language understanding. *Advances in neural information processing systems*, 31, 2018. 3

[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. 2

[32] Yun Zhong and Yiannis Demiris. Dancemvp: Self-supervised learning for multi-task primitive-based dance performance assessment via transformer text prompting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10270–10278, 2024. 2