A Decade of Action Quality Assessment: Largest Systematic Survey of Trends, Challenges, and Future Directions

Hao Yin[1, 2*] · Paritosh Parmar[3*] · Daoliang Xu[1, 2] · Yang Zhang[2] · Tianyou Zheng[2 ✉] · Weiwei Fu[1, 2 ✉]

Abstract Action Quality Assessment (AQA)—the ability to quantify the quality of human motion, actions, or skill levels and provide feedback—has far-reaching implications in areas such as low-cost physiotherapy, sports training, and workforce development. As such, it has become a critical field in computer vision & video understanding over the past decade. Significant progress has been made in AQA methodologies, datasets, & applications, yet a pressing need remains for a comprehensive synthesis of this rapidly evolving field. In this paper, we present a thorough survey of the AQA landscape, systematically reviewing over 200 research papers using the preferred reporting items for systematic reviews & meta-analyses (PRISMA) framework. We begin by covering foundational concepts & definitions, then move to general frameworks & performance metrics, & finally discuss the latest advances in methodologies & datasets. This survey provides a detailed analysis of research trends, performance comparisons, challenges, & future directions. Through this work, we aim to offer a valuable resource for both newcomers & experienced researchers, promoting further exploration & progress in AQA. Data are available at https://haoyin116.github.io/Survey_of_AQA/.

*These authors contributed equally to this work.

✉ Weiwei Fu
E-mail: fuww@sibet.ac.cn
✉ Tianyou Zheng
E-mail: zhengty@sibet.ac.cn
[1] School of Biomedical Engineering (Suzhou), Division of Life Sciences and Medicine, University of Science and Technology of China, Hefei, Anhui, China
[2] Suzhou Institute of Biomedical Engineering and Technology, Chinese Academy of Science, Suzhou, Jiangsu, China
[3] Institute of High Performance Computing, Agency for Science, Technology and Research, Singapore

Keywords action quality assessment, skills assessment, action understanding, video understanding, computer vision, deep learning, survey

1 Introduction

Skills and Action Quality Assessment (**AQA**) is an *emerging* and *critical* field in video understanding, moving beyond action recognition and action prediction [1] to **evaluate how well** actions are performed and score the skill level of performers (see Fig. 1). These techniques are essential in a range of **domains**, including *sports* [2, 3], *healthcare* [4–7], *fitness* [8–10], *industrial training* [11, 12], & *AI video content generation* [13] where accurate assessment of human actions/performance is crucial.

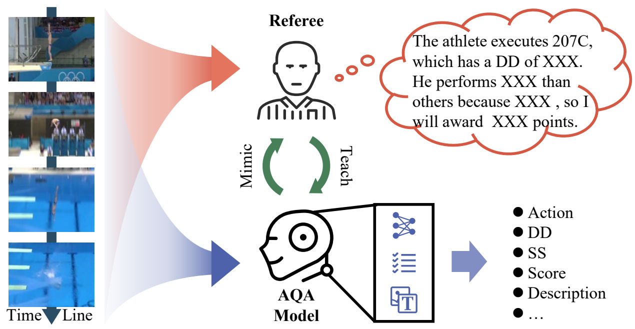


Fig. 1 AQA model plays the role of referee to evaluate how well actions are performed and score the skill level of performers.

For example, in *sports* [2, 3], it could be used to assess how well an Olympics diver performed, reporting what they did well/correctly, what they did wrong and what was the severity of these errors; take all (ideally) such factors into consideration and quantify the how well was the performance of the diver. Similarly, it could be used for other sports or *physical rehabilitation* [5, 6] or *managing diseases* like Cerebral Palsy [14]. AQA can be used to train future surgeons by assessing their *surgical performances* [15], assessing them, and giving feedback on what can be improved. In *manufacturing sector* [11], AQA can be used to assess the skill levels of workers, train them, and ensure their actions follow safety standards. Over the last few years, *generative AI* [13] has been on the rise, where AQA could serve as an objective metric for evaluating AI-generated video content.

Automated AQA systems have **far-reaching implications** across diverse sectors, from *promoting social equity* to *enhancing industrial efficiency*. On a societal level, AQA can democratize access to training and evaluation tools, particularly benefiting under-resourced communities. Low-cost, AI-driven assessments provide underserved athletes with tailored feedback, support fair sports judging [2, 3], and facilitate remote patient evaluations in healthcare [4–6], improving access for rural or low-income populations. In skilled actions/vocational training, AQA offers objective skill assessments, helping workers certify and enhance their technical skills [16, 17], which is essential for career growth. In industries like manufacturing and automation, AQA can reduce errors, ensure product consistency, and support a skilled workforce, leading to lower costs and higher productivity [11]. By assessing both human and robotic systems, AQA enables effective human-robot collaboration, boosting workplace safety and efficiency. Looking ahead, the development of ethical and unbiased AQA systems will ensure that these technologies can be equitably applied across various demographics, preventing algorithmic biases from perpetuating inequalities. As such, the continued growth and refinement of AQA technology promises to not only enhance performance and quality control but also play a pivotal role in creating more inclusive, accessible, and efficient systems that benefit society as a whole. As generative AI becomes more prevalent, AQA can also serve as an objective metric for assessing the quality of AI-generated content [13], ensuring transparency and trust in AI systems.

AQA has evolved significantly over the past decade (see Fig. 2), marking major advancements across sports [2, 3], healthcare [4–6], fitness [8, 9], industrial training [11], and generative AI evaluation [13]. With the last comprehensive survey now outdated [18, 19], there's a critical need for a current synthesis of this fast-growing field. New computer vision and machine learning methodologies have expanded the capabilities of AQA, introducing refined evaluation techniques, metrics, and cross-disciplinary applications that address diverse sector-specific needs—from precise feedback in sports performance to accessibility in remote patient healthcare.

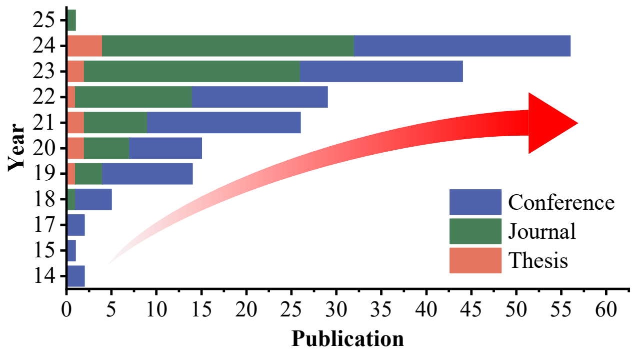


Fig. 2 The number of papers related to AQA over the past decade.

This survey paper addresses these developments by offering a structured overview of the latest AQA methodologies, benchmarks, and evaluation metrics. We also examine the field's unique ethical and societal challenges, emphasizing the importance of unbiased, inclusive assessment systems that promote equitable access to training and evaluation resources across demographics. By consolidating recent progress and illuminating future research directions, this survey aims to serve as a foundational guide for both researchers and practitioners, providing insights that drive innovation in creating accessible, robust, and fair AQA systems.

In this survey, we employed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (**PRISMA**) [20] framework to systematically select relevant papers and continuously update until December 2024. This process unfolded over four key stages: **1)** *Identification*: A search for the keyword' action quality assessment' in Web of Science and Google Scholar identified 505 potentially relevant papers. **2)** *Screening*: After initial screening based on titles and abstracts, we removed duplicates and irrelevant entries, leaving 276 papers. **3)** *Eligibility*: We reviewed the full text of these papers to assess quality, documenting each study's approach, results, and novelty. **4)** *Inclusion*: Ultimately, 195 papers, including 26 datasets, met the inclusion criteria for this survey.

The rest of the survey paper is organized as follows. To facilitate navigation of the survey's structure, Fig. 3 provides an overview of the sections and their titles. Section 2 defines and classifies the AQA problem, introducing common model evaluation metrics. Section 3 offers a compilation of popular AQA datasets, categorizing them by action scenarios and providing detailed descriptions. Section 4 reviews fundamental AQA research, summarizing papers from the past decade and dividing them into 7 principal trends, with a detailed discussion of research methodologies and performance comparisons within each trend. Section 5 analyzes ongoing challenges in current research and proposes directions for future research. In conclusion, Section 6 summarizes the findings of this survey.

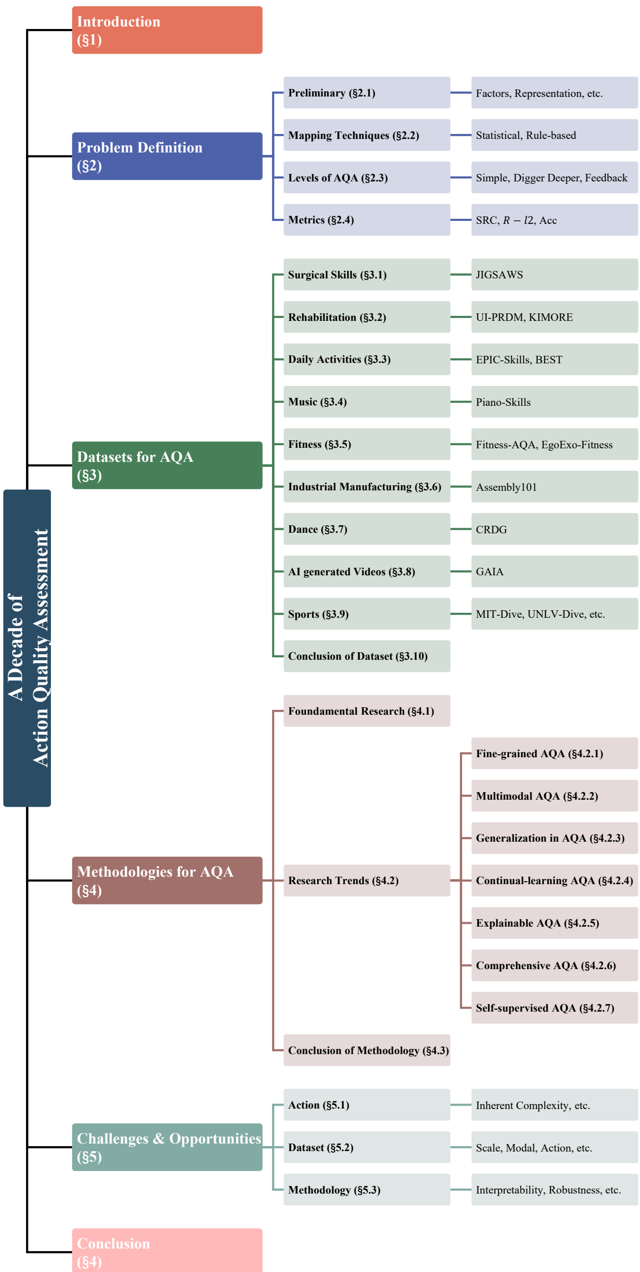


Fig. 3 Structure of the overall survey. Zoom in for the best view.

2 Problem Definition

2.1 Preliminary

Action and its Quality. First, let's develop some intuition about action quality. Action can be thought of as having two components: *difficulty level* (what action was done), and *execution quality* (how well that action was done) [21]. An example of difficulty level is that somersaulting in the Pike position (in sports like Diving or Gymnastic Vault) is more difficult than somersaulting in the Tuck position. So, if executed perfectly, somersaulting in the Pike position is of more worth than somersaulting in the Tuck position. An example of intuiting about execution quality: In the Pike position, having feet together is worthy of more points than having feet apart or crossed over. The action quality score is directly proportional to both difficulty level and execution quality. To give a better intuition regarding execution quality and difficulty level, we have contrasted them in Fig. 4.

Factors to be captured by representations. In Fig. 5, we illustrate some instances of action elements that matter in AQA. We mentioned what elements are desirable or hold a higher value. The task of AQA involves not only identifying those elements in a given action sequence but also determining/quantifying their value. E.g., let's consider Fig. 5(f); we see that the feet of the gymnast on the left are not together. A judge judging that gymnast's performance would identify that her feet were not together; secondly, they would determine how much to penalize (in terms of the points) for that error— the larger the gap, the more would be the penalty. Following this example, we can see that a CNN will have to learn to capture elements like these, determine the severity of the error, and penalize accordingly. Moreover, not all the elements may be equally important. Note that this was for just an element, usually, an action instance is composed of multiple such elements; each of those elements needs to be taken into consideration, based on which the final action quality score is calculated.

Representation formats. Several kinds of representations are used to capture actors' movements and other relevant factors, such as raw RGB video [2, 3], optical flow [22, 23], skeleton pose sequence [24–26], sound [23, 27, 28], human-object distance [26], etc.

Generic AQA pipeline. Typically, AQA methodologies follow a two-stage approach [3] (see Fig. 6): Stage 1: feature extractors are used to extract relevant features from performance recording (video, audio, etc.); Stage 2: these features are then mapped to generate AQA scores, reports, feedback using statistical or rules-based map-

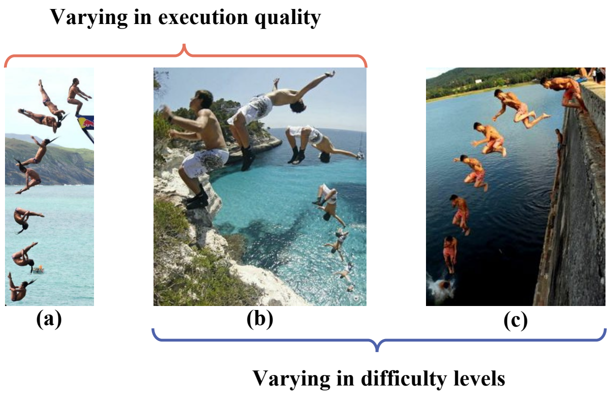


Fig. 4 Execution quality vs difficulty level. A professional cliff diver is shown doing a backflip in (a). Notice how straight his legs are, how crisp his overall form is, and how graceful his execution is, compared to an amateur diving in (b). Amateur's legs are bent at the knees, as rotating with bent legs would result in faster rotation (lesser effort on the diver's part). It generally requires a lot of training and practice to be able to attain and maintain a great form like (a). Now, we compare two amateurs (b) and (c). Even though both have bad forms, doing a backflip, as in (b), is more difficult (requires more effort and skills) than simply falling into the lake like (c).

ping modules (discussed in the following). AQA research includes improving both these stages.

2.2 Mapping Techniques

Mapping from (visual/multimodal) action representation space to output score space can be Statistical or Rules-based in nature (see Fig. 7), discussed in the following.

Statistical methodology. In this methodology, models learn to estimate the relationship between the visual action representations/features and the action quality score from datasets in a supervised manner. Mapping model parameters are adjusted during training to fit the dataset. Note that while the weights can be adjusted in a data-driven manner, control over what patterns the mapping model fits on is not entirely in our control. Hence, the model might fit shortcut correlations, which might not be reflective of the true (intended) relation between action features and action scores. There are following three formats of statistical methodologies used:

1. **Regression-based scoring.** Here, the statistical model predicts a numerical score on a continuous scale. In regression, the output score is a real number, and the model tries to find a relationship between the input features and this continuous output. For example, the score can be any real quantity between, let's say, 0 to 100. Models include linear regressors, shallow/deep neural nets, support vector regressors (SVR), etc. Regression allows or is used for fine-grained or detailed analysis of action quality. Examples of regression-based methodologies include [2, 3].

2. **Classification-based scoring.** In this case, the model predicts a discrete label or category. In classification, the output action quality score/label is a category or class (not a continuous value), and the model assigns input features to one of these classes. Classification is generally used for coarsely quantifying the action quality. Sometimes this coarse nature can allow in recruiting non-experts for data labeling. Examples of classification-based methodologies include [8, 14, 27, 29].

3. **Pairwise ranking.** In this case, models compare two or more samples to decide which sample is of the highest action quality. Note that this is not a standalone analysis like the regression and classification. Examples of pairwise ranking methodologies include [16, 17, 30].

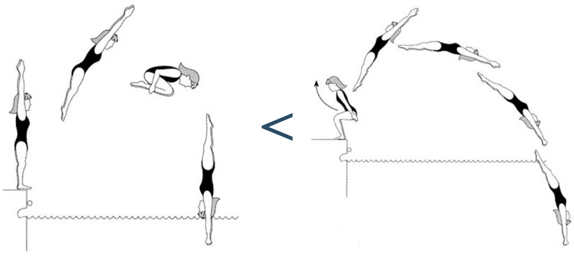
Rules-based scoring. Unlike statistical methodologies, in rules-based methodologies, features/symbols are processed using a set of rules to compute the action quality score. A set of rules can be hand-crafted incorporating human expert knowledge, or they could be learned from data. In rules-based methodologies, how each individual feature of the performance affected the score can be determined/traced back. As such, the causal nature of the rules makes rules-based interpretable and explainable. Examples of rules-based methodologies include [26].

2.3 Levels of AQA

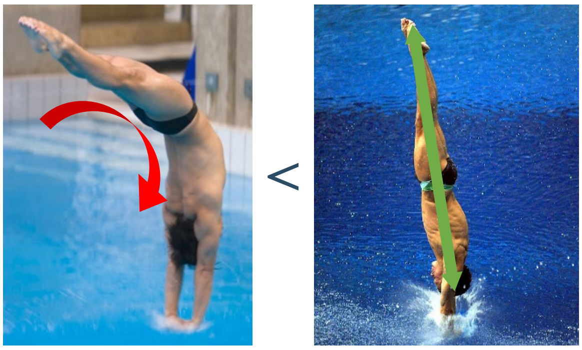
Simple AQA. In its simplest terms, AQA involves using a model for mapping an action video to a performance/action quality score. Ideally, the idea is that the model will extract all the relevant (spatiotemporal) features that contribute to action quality (score). Values are associated with these extracted features. Models learn to extract relevant action quality features and their values and, based on all the values, output a final score.

Digging Deeper. A deeper level of assessment is that the model lists out all the factors and their analysis for what the diver did well, where the diver made a mistake, assessing the severities of these errors and then computing the final score based on it.

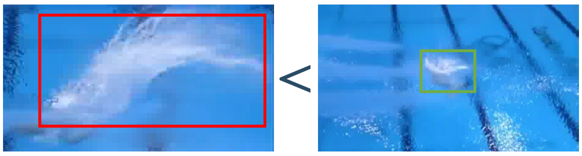
Feedback Generation. Here, the model is able to assess the performance and its good and bad points. Based on this, the model may be able to suggest what parts the actor can work on to improve. Generally, if the model has a level of transparency–has an internal record of the pros and cons of the performance/action, it should be able to suggest a preliminary level of feedback. For example, if the model identifies that a diver had bent their legs in the pike position, it may indicate that



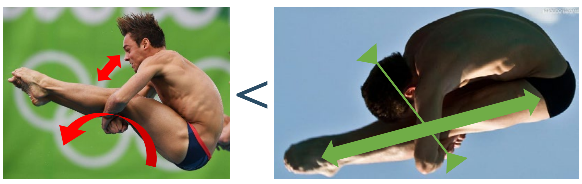
(a) A dive with backward rotation has higher base value than the same dive with forward rotation. *How much more?*



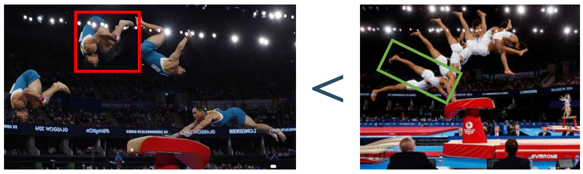
(b) Perpendicular entry is awarded more points than entering water with a bend. *How much to penalize for crooked entry?*



(c) Bigger the splash, lower the points. Rip entry into the water is desirable. *How much to penalize for that big splash?*

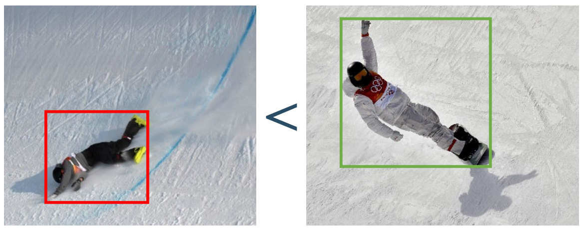


(d) Maintaining tight form while keeping legs straight earns you more points than having space between face and knees and bent legs. *How much to penalize for improper form?*

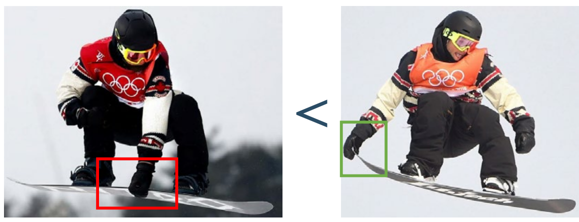


(e) Somersaulting in Pike position yields more points than doing so in a Tuck position. *How many more points?*

(f) Twisted ankles or feet apart results in losing points, while having those together earns to you more points. *How much to penalize to twisted ankles?*



(g) Stable landing without tumbling or falling helps you earn more points, while tumbling results in loss of points. *How many points to deduct for tumbling?*



(h) Different grabs earn you different no. of points, for e.g., Indy grab being a basic grab earns you fewer points than a Tail grab. *How many points for Indy grab and Tail grab?*

Fig. 5 Examples of elements that matter in AQA. Zoom in for the best view.

the diver may work on keeping their legs during the pike position to improve their score.

2.4 Metrics

The following metrics are used to quantify/measure the performance of AQA models. The metric is chosen depending on the type of mapping model used. For example, correlation between ground truth and predicted scores is used when scores are on a continuous scale, while accuracy is used when action quality labels are discrete.

Spearman's Rank Correlation. Pirsiavash et al. [2] first pioneered the assessment of action quality in 2014 and also first introduced the use of Spearman's Rank Correlation (SRC, denoted as ρ) to evaluate the performance of regression models. The SRC formula is as

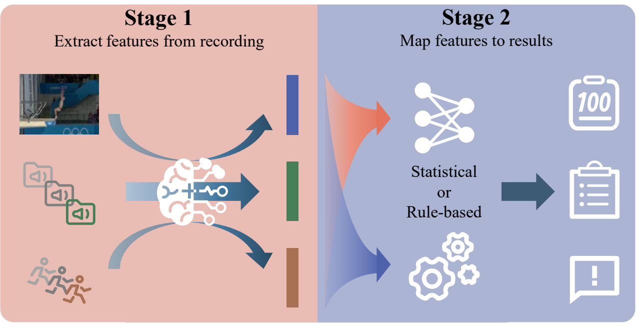


Fig. 6 The generic AQA pipeline.

follows:

$$\rho = 1 - \frac{6\sum_{i=1}^{n}(R_i - \widehat{R_i})^2}{n(n^2-1)} \tag{1}$$

where R_i and $\widehat{R}_i$ represent the ground truth and predicted rankings of the i-th sample, respectively. n is the total number of samples. The ρ range is $[-1, 1]$, with values closer to 1 indicating better performance. SRC is currently the most widely used performance metric in AQA. However, SRC can only measure the strength and direction of the monotonic relationship between ground truth and predicted rankings without measuring the differences between ground truth and predicted scores.
Relative $l2$ Distance. To address this limitation, Yu at al. [31] proposed Relative $l2$ Distance ($R-l2$) as a new metric for evaluating model performance. In contrast to SRC, which focuses on the ranking of predicted scores, $R-l2$ places more emphasis on the numerical values of the predicted score. Moreover, compared to the traditional $l2$ distance, $R-l2$ takes into account the score intervals between different categories of actions, facilitating cross-category model training. The $R-l2$ formula is as follows:

$$R-l2 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{|s_i - \widehat{s_i}|}{s_{max} - s_{min}}\right)^2 \tag{2}$$

where s_i and $\widehat{s_i}$ represent the ground truth and predicted scores of the i-th sample. s_{max} and s_{min} represent the maximum and minimum scores of this action category. n is the total number of samples. The $R-l2$ range is $[0, 1]$, closer to 0 indicating better performance.
Accuracy. For other mapping technique formats, pairwise rank and classification, accuracy (Acc) are commonly used as model performance metrics. The accuracy formula is as follows:

$$Acc = \frac{TP+TN}{TP+TN+FP+FN} \tag{3}$$

where TP, TN, FP, and FN represent true positives, true negatives, false positives, and false negatives, respectively.

3 Datasets for AQA

Datasets are a crucial component of machine learning projects, and they play a key role in AQA by enabling model training, performance evaluation, and real-world deployment. The success and advancements in AQA can be attributed to the development of novel and high-quality datasets. In this section, we systematically collect and summarize all existing datasets since 2014. Following the PRISMA guidelines, we identify 26 publicly available datasets related to skills and action quality assessment across various domains. We classify datasets into 9 domains {Surgery, Daily Activities, Rehabilitation, Music, Fitness, Industrial Manufacturing, Dance, AIGV, Sports} as illustrated in Fig. 8. These datasets are summarized in Tab. 1 and Fig. 9, and we provide a more detailed description of each dataset in the following. We have organized the datasets by domain. Finally, we conclude the section by providing a summary of the datasets.

3.1 Surgical Skills

JIGSAWS, proposed by Gao et al. [4] at MICCAI Workshop 2014, is the first surgical skill dataset for human action modeling. The dataset contains 103 samples with an average sample duration of 92 seconds. Three basic surgical operations were captured by the daVinci Surgical System from eight surgeons with different skill levels. Each operation was repeated five times, and the specific actions are illustrated in Fig. 8(a). There are 39 suturing samples, 36 knot-tying samples, and 28 needle-passing samples. Each sample consists of 76 dimensions of kinematic data (including cartesian positions, orientations, velocities, angular velocities, and gripper angle describing the motion of the manipulator) and video data. The dataset annotations contain action labels and action scores, with a total of 15 action labels corresponding to video frames, and action scores are the total scores of the six sub-elements, with a score interval of 1-5 for each element.

3.2 Rehabilitation

The datasets of rehabilitation contain UI-PRMD and KIMORE, which were proposed by Vakanski et al. [5] at Data 2018 and Capecci et al. [6] at TNSRE 2019, respectively. Detailed actions are illustrated in Fig. 8(b).
UI-PRMD, fills the gap of a comprehensive dataset of physiotherapy actions. The dataset contains 100 samples of 10 different actions, each repeated 10 times, from 10 healthy subjects simultaneously captured by two

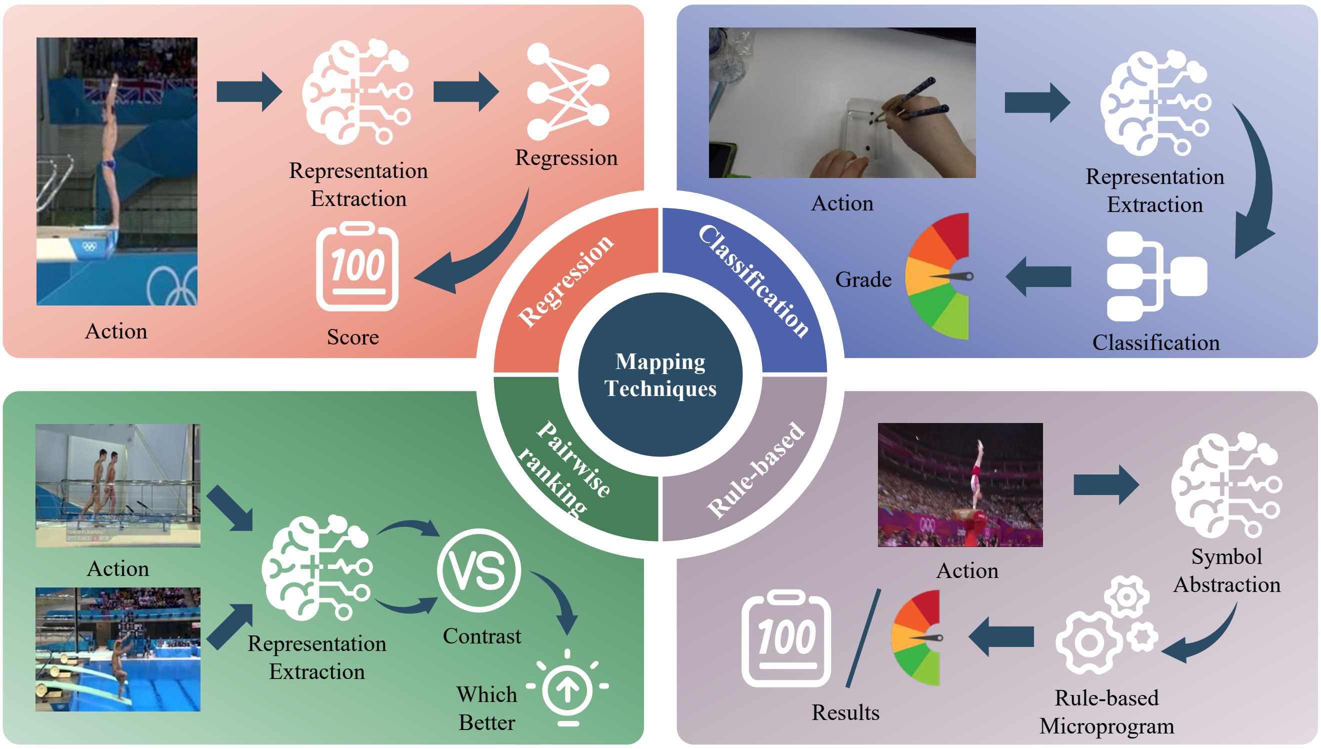


Fig. 7 Different formats of mapping techniques used in AQA.

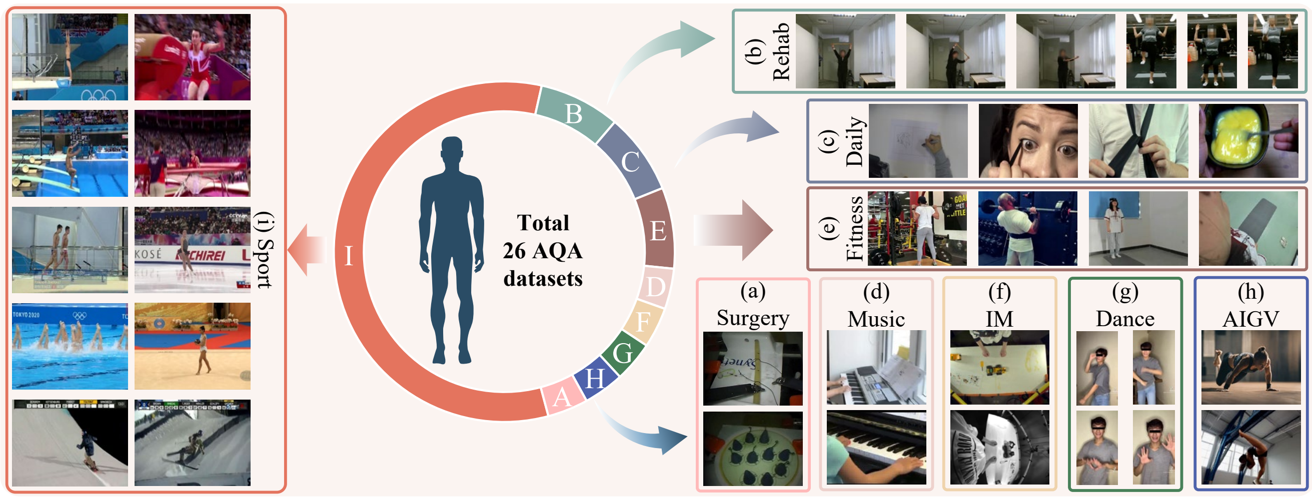


Fig. 8 Action samples of different dataset domains. Zoom in for the best view.

systems, Vicon and Kinect. Each sample consists of skeletal sequences, including position and angle. The dataset annotations are correct or incorrect binary-grade labels.

KIMORE, is the only RGB single-view rehabilitation action dataset. The dataset contains 353 samples of 5 different actions from 44 healthy subjects and 34 patients, with an average sample duration of 29.9 seconds. The dataset was annotated with scores, including PO_S and CF_S, with values in the range 0 to 50 for each action as defined by clinicians.

3.3 Daily Activities

The datasets of daily activities scenarios contain EPIC-Skill [16] and BEST [17], which were proposed by Doughty et al. at CVPR 2018 and CVPR 2019, respectively. Detailed actions are illustrated in Fig. 8(c).

EPIC-Skill contains 216 samples divided into 4 sub-datasets: surgery, dough-rolling, drawing, and chopstick-using, with an average sample duration of 86.6 seconds. The surgery sub-dataset is JIGSAWS [4], the Dough-Rolling sub-dataset was 33 samples selected from

Table 1 List of publicly available datasets used in skills and action quality assessment. AD: Average Duration, RV: Real Videos, AG: AI-generated, Desc: Description, Rehab: Rehabilitation, IM: Industrial manufacturing. *-indicates only the no. of coarse action types—no. of finegrained action types contained is much higher.

Datasets	Year	Sample	Type	AD	Annotation Type	Source	Domain
MIT-Dive [2]	2014	159	1	2.5s	Score	RV	Sport
MIT-Skate [2]	2014	150	1	175s	Score	RV	Sport
JIGSAWS [4]	2014	103	3	92s	Score, Action	RV	Surgery
UNLV-Dive [3]	2017	370	1	3.8s	Score	RV	Sport
UNLV-Vault [3]	2017	176	1	2.8s	Score	RV	Sport
UI-PRMD [5]	2018	100	10	—	Grade, Action	RV	Rehab
EPIC-Skill [16]	2018	216	4	86.6s	Rankings, Action	RV	Daily
AQA-7 [32]	2019	1189	7*	6.7s	Score, Action	RV	Sport
MTL-AQA [21]	2019	1412	52	4.1s	Score, Action, Desc	RV	Sport
Fis-V [33]	2019	500	1	170s	Score	RV	Sport
Best [17]	2019	500	5	187.6s	Grade, Rankings, Action	RV	Daily
KIMORE [6]	2019	353	5	29.9s	Score, Action	RV	Rehab
TASD-2 [34]	2020	606	2	4.1s	Score, Action	RV	Sport
Rhy.Gym. [35]	2020	1000	4	95s	Score, Action	RV	Sport
Piano-Skills [27]	2021	992	1	160fr	Grade, Song-difficulty	RV	Music
FR-FS [36]	2021	417	1	103fr	Grade, Action	RV	Sport
SMART [37]	2021	5000	10	420fr	Score, Action	RV	Sport
Fitness-AQA [8]	2022	21284	3	4.1s	Grade, Action	RV	Fitness
FineDiving [38]	2022	3000	52	4.2s	Score, Action	RV	Sport
Assembly101 [11]	2022	4321	101	426s	Score, Action	RV	IM
LOGO [39]	2023	200	12	204.2s	Score, Action, Formation	RV	Sport
FineFS [40]	2023	1167	1	215s	Score, Action	RV	Sport
PaSk [41]	2023	1018	1	10.7s	Score	RV	Sport
CDRG [28]	2023	240	12	14.7s	Rankings, Action	RV	Dance
GAIA [13]	2024	9180	510	2.8s	Score, Action	AG	AIGV
EgoExo-Fitness [9]	2024	6131	12	18.8s	Score, Action, Desc	RV	Fitness

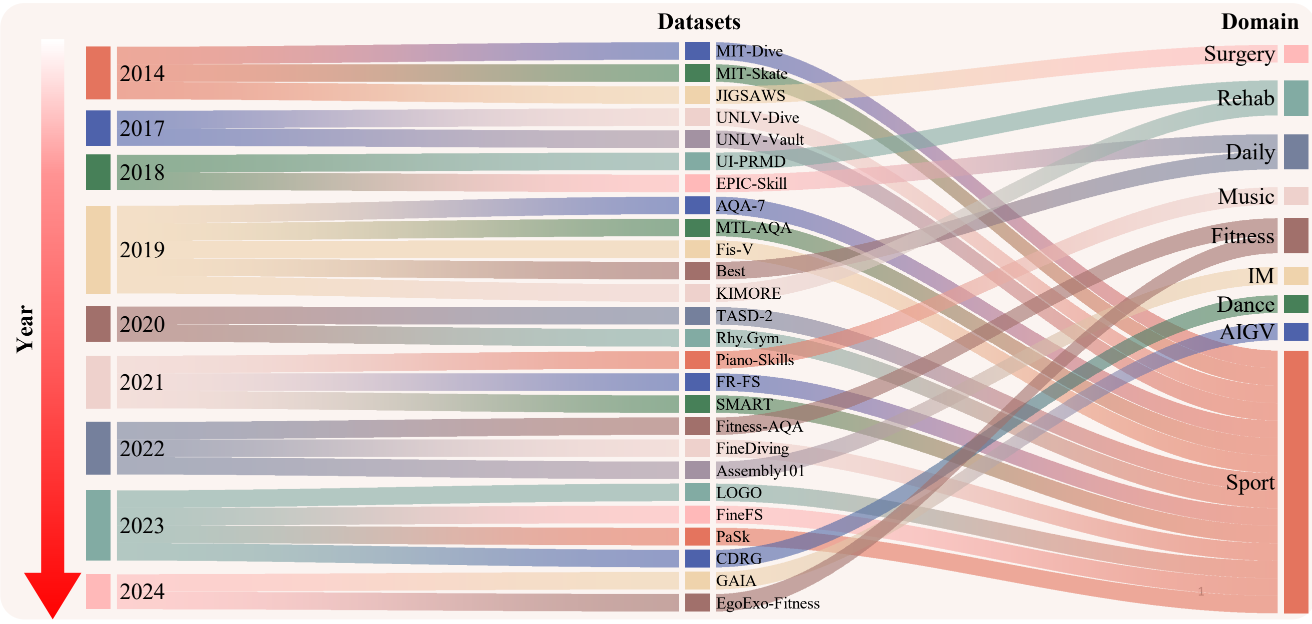


Fig. 9 An intuitive visualization of the annual distribution of publicly available datasets, including their titles and corresponding domains.

Kitchen-based CMU-MMAC [42], the Drawing sub-dataset contains videos of 4 volunteers repeating the drawing of SONIC and HAND 5 times respectively, and the Chopstick-Using sub-dataset contains videos of 8 volunteers repeating the chucking of beans in a box 5 times. Since JIGSAWS [4] was annotated, the other three sub-datasets and rankings were obtained by pairwise comparison of samples within classes.

BEST contains 500 samples consisting of 5 daily skill tasks: scrambling eggs, braiding hair, tying a tie, making

an origami crane, and applying eyeliner, with an average sample duration of 187.6 seconds. Videos are related action videos downloaded from YouTube. Dataset annotations contain the start and end times of actions in the videos, action grades (Beginner, Intermediate, or Expert), and rankings.
EgoExo4D [223] is a dataset for skills assessment for daily everyday and sports activities. It contains 1224 samples with a total of 16 hours of video.

3.4 Music

Piano-Skills, proposed by Parmar et al. [27] at MMSP 2021, is the first AQA dataset in music and introduced audio into consideration. The dataset contains 992 samples from 61 piano performances, with an average sample duration of 160 frames, as illustrated in Fig. 8(d). The dataset was annotated by a trained pianist with skill levels and song difficulty.

3.5 Fitness

Fitness-AQA, proposed by Parmar et al. [8] at ECCV 2022, is the dataset with the most samples in AQA and the first in fitness. The dataset contains 21284 samples collected from video-sharing sites, covering BackSquat, BarbellRow, and Overhead Press, with an average sample duration of 4.1 seconds, as illustrated in Fig. 8(e). The dataset annotations contain binary-grade labels by two professional gym trainers.
EgoExo-Fitness, proposed by Li et al. [9] at ECCV 2024. The dataset contains 6131 samples from 86 action sequences by combining 3 to 6 different actions of 12 types of fitness actions, constructed by the synchronized recording of 3 egocentric-view and 3 fixed exocentric-view (third-person) videos. The average duration is about 18.8 seconds. The dataset annotations contain technical key point verification of fitness action, natural language comment on subjects' action, and action quality score (range 1-5 points), which provide interpretable action judgment annotations.

3.6 Industrial Manufacturing

Assembly101, proposed by Sener et al. [11] at CVPR 2022, is currently the data set with most camera views in AQA. The dataset contains 4321 samples, constructed by capturing 362 different volunteers assembling and disassembling 101 different toy cars with 8 fixed and 4 egocentric cameras, as illustrated in Fig. 8(c). The average sample duration of 426 seconds. The dataset annotations contain over 1 million action segments, 1380 fine-grained and 202 coarse-grained action classes, and action scores.

3.7 Dance

CDRG, proposed by Hipiny et al. [28] in IJAIN 2023. Fig. 8(g). The dataset contains 240 samples from 20 subjects performing 12 TikTok dance challenges respectively. Subjects used their own devices to record videos, resulting in different levels of image quality. The average duration of samples is 14.7 seconds. The dataset was annotated by 100 annotators to mark the winning one of each dance pair.

3.8 AI Generated Videos

GAIA, proposed by Chen et al. [13] at NeurlIPS 2024, is the first AQA dataset constructed by collecting videos generated by text-to-video (T2V) models to assess their video generation quality. The dataset contains 9180 samples generated by T2V models from 18 different laboratories and commercial platforms, covering a total of 510 actions of the whole body, hand, and face, as illustrated in Fig. 8(h), with an average sample duration of 2.8 seconds. The dataset was annotated as scores of the generated videos from three perspectives, including subject quality, action completeness, and action-scene interaction, with a score interval of 0-100 for each perspective.

3.9 Sports

Derived from their clear rules and easy access, sports account for the highest percentage of AQA datasets, including 15 datasets covering 10 different types, as illustrated in Fig. 8(i). In recent years, the advancement of sports AQA datasets has been remarkable, with the publication of numerous high-quality datasets providing important benchmarks for AQA. The rest of this section provides a detailed description of sports datasets.
MIT-Dive & MIT-Skate, proposed by Pirsiavash et al. [2] at ECCV 2014. MIT-Dive dataset contains 159 samples of Olympic diving, all in slow motion from television broadcasts at a frame rate of 60 fps. The average duration of dive samples is 2.5 seconds. MIT-Dive dataset was annotated with scores awarded by referees, ranging from 20 to 100 points. Similarly, MIT-Skate dataset consists of 150 skating samples, with a frame rate of 24 fps. The average duration of skate

samples is 175 seconds. The dataset was also annotated with awarded scores, ranging from 0 to 100 points.

UNLV-Dive & UNLV-Vault, proposed by Parmar and Morris et al. [3] at CVPR Workshop 2017. UNLV-Dive contains 370 samples with an average sample length of 3.8 seconds and is an extension of the original MIT-Dive dataset. UNLV-Dive was annotated with scores calculated by multiplying the execution score (range 0-30 points) by the difficulty score (no upper limit specified in the rules, range 2.7-4.1 points in the dataset), resulting in a final dataset score. Similarly, UNLV-Vault contains 176 samples, with a sample average length of 2.8 s. UNLV-Vault was annotated with scores, which were calculated by adding the execution score (range 0-10 points) and the difficulty score (range 0-10 points), resulting in a final score.

AQA-7, proposed by Parmar and Morris et al. [32] at WACV 2017. The dataset contains 1189 samples from both the Winter and Summer Olympics, covering 7 different action types (including singles diving-10m platform, gymnastic vault, big air skiing, big air snowboarding, synchronous diving-3m springboard, synchronous diving-10m platform, and trampoline). The average duration of samples is 6.7 seconds. The dataset was annotated with scores specific to each type, reflecting the unique scoring system. For each type, the scoring intervals are as follows: singles diving-10m platform (21.6-102.6), gymnastics vault (12.3-16.87), big air skiing (8-50), big air snowboarding (8-50), synchronized diving 3m springboard (46.2-104.88), synchronized diving 10m platform (49.8-99.36) and trampoline (6.72-62.99).

MTL-AQA, proposed by Parmar and Morris et al. [21] at CVPR 2019. The dataset contains 1412 samples from 16 different competitions, including 10m platform and 3m springboard, male and female athletes, individual or pairs of synchronized divers, and different views. It also pioneered the: 1) natural language detailed qualitative description of the performance; and 2) disassembling (diving) actions into finegrained subcategories, divided into five parts: position, armstand, rotation type, somersaults, and twists. The average duration of samples is 4.1 seconds. The dataset annotations include the difficulty and action scores awarded by seven referees, along with the finegrained action breakdown and detailed qualitative descriptions of the corresponding actions.

Fis-V, proposed by Xu et al. [33] at TCSVT 2019. The dataset contains 500 samples of selected women's singles figure skating short program videos. There are 149 athletes from 20 countries, and the average duration of samples is 170 seconds, with an average frame rate of 25 FPS. The dataset annotations include the total element score (TES) and total program component score (PCS), each awarded by 9 different professional figure skating referees.

TASD-2, proposed by Gao et al. [34] at ECCV 2020, different from previous datasets containing synchronized diving with side view, uses front view of video shot. The dataset contains 606 samples with an average duration of 4.1 seconds. The dataset was annotated with the beginning and end frames of the action as well as the final score, which is the product of the execution score and the difficulty score multiplied by the execution score.

Rhythmic Gymnastics, proposed by Zeng et al. [35] at ACM MM 2020. The dataset contains 1000 samples from the 36th and 37th International Artistic Gymnastics Competitions, encompassing the four types of ball, clubs, hoop, and ribbon. Each type has 250 sample videos, with an average duration of 95 seconds and an average frame rate of 25 FPS. The dataset was annotated with scores, including a difficulty score, an execution score, and a total score, which is the sum of the execution score and difficulty score.

FR-FS, proposed by Wang et al. [36] at ACM MM 2020, unlike previous figure skating AQA dataset with duration (close to 3 minutes), this contains frame-level information and plans to gradually construct a more fine-grained figure skating AQA system, starting with the identification of the most basic errors. The dataset contains 417 samples from the Fis-V dataset and the PyeongChang Winter Olympics, including multiple athletes' key actions (take-off, rotation, and landing). Of these, 276 are smooth landing videos and 141 are fall videos. The dataset was annotated with key action and binary-grade labels.

SMART, proposed by Chen et al. [37] at IJCV 2021. The dataset contains 5000 samples of 10 different action types, including balance beam, diving, uneven bars, vault, hurdling, pole vault, high jump, boxing, keep-fit, and badminton, with an average sample duration of 420 frames. The dataset annotations contain sub-action labels and action assessment scores.

FineDiving, proposed by Xu et al. [38] at CVPR 2022. The dataset contains 3000 samples from 30 different diving competitions, including the Olympics, World Cup, World Championships, and European Aquatics Championships, and encompasses 52 action types, 29 sub-action types, and 23 difficulty levels. The average duration is 4.2 seconds. The dataset annotations contain semantic and temporal structures, both of which are two-level annotations. In the semantic structure, the action-level label describes the action category, and the step-level label describes the sub-action types of the successive actions. In the temporal structure, the action-level label describes the beginning and end time of the athlete's execution of the complete action, and the step-level label describes the beginning and end frames of the sub-action. The competition referees awarded the score. This dataset was extended in [43].

LOGO, proposed by Zhang et al. [39] at CVPR 2023, was constructed as multi-person long video AQA dataset. The dataset contains 200 samples, with 8 athletes in each frame, from 26 artistic swimming events during 2018-2022, with an average sample duration of 204.2 seconds. The dataset annotations are structured in a temporal structure, including fine-grained action and formation annotations. There are 12 types of action annotation and 17 types of formation annotation. The annotations of both action and formation were annotated through frame-by-frame analysis. Scores were awarded by referees with 3 sub-scores and a total score.

FineFS, proposed by Ji et al. [40] at ACM MM 2023. The dataset contains 1167 samples containing RGB videos and skeleton sequences from 72 A-level international events under the new ISU rules from the 2018-2019 season to the 2021-2022 season, with an average sample duration of 215 seconds. The dataset annotations contain 4 levels of scores, 2 levels of sub-action categories, and time segmentation.

PaSk, proposed by Gao et al. [41] at IJCV 2023, can be used to complement the AQA in a strong primary-secondary relation since the male performer always assists the female performer to execute various actions in pairs figure skating. The dataset contains 1018 samples from figure skating competitions organized by ISU, with an average sample duration of 10.7 seconds. The dataset annotations contain the beginning and end frames of the independent actions and the awarded scores.

3.10 Conclusion of Dataset

Since the fundamental research of the AQA dataset in 2014, 26 popular datasets have emerged within the AQA domain, encompassing a wide range of domains from surgery to sport. Of these, sport accounts for the largest proportion and is the most important application scenario of AQA. A number of researchers have made significant contributions to the construction of datasets. To illustrate, Parmar has proposed a total of 6 widely used datasets before and after [3, 8, 21, 27, 32], while Xu has proposed the most fine-grained diving action dataset [38]. It is evident that the size of the dataset is continuously expanding, the action types and modality data are constantly being enriched, and the data annotation is moving towards fine-grained annotation, all of which are driving research progress. However, compared to action recognition datasets [44, 45] and other related datasets [46–57] , AQA datasets still face challenges in terms of scale, action diversity, data modality, and annotations.

In particular, GAIA [13], as the first dataset constructed using AI-generated videos, marks a significant breakthrough. It not only broadens the application scenarios of AQA but also introduces AQA into the T2V model to assess the quality of video generation. GAIA also inspires researchers to explore the possibility of using the T2V model to generate massive videos with multiple action types and skill levels based on predefined prompts, thereby easily constructing large-scale AQA datasets.

4 Methodologies for AQA

Following the comprehensive overview of publicly available datasets in AQA, it is evident that further investigation is required. What methodologies are being employed by researchers to utilize these rich resources of datasets to advance AQA fully? In order to answer this question, this section will concentrate on the methodologies employed in the field of AQA. By categorizing the pertinent and standardized research methodologies related to AQA in the last decade, it is possible to gain insight not only into the research paradigm in this field but also into the prevailing trends in research (see Fig. 10). The preceding decade of research has been classified into 7 principal trends, with comprehensive descriptions of the employed research methodologies and performance.

4.1 Fundamental Research

To the best of our knowledge, AQA research dates back to 1995 when Gordon [59] proposed to apply computer vision technologies such as tracking to AQA. There has been some research on AQA between 1995 and 2014, such as [60–62]. However, the AQA field still lacked datasets to conduct reliable studies and evaluation metrics of AQA methodologies. Pirsiavash et al. [2] and Gao et al. [4] introduced the first AQA datasets. These include MIT-Dive, MIT-Skate, and JIGSAWS datasets. Pirsiavash et al. [2] additionally proposed one of the initial machine learning frameworks for AQA by treating AQA as a supervised regression problem. Specifically, in their methodology, spatiotemporal features (pixel gradients and athlete pose features) were fed into a linear support vector regression (**L-SVR**) model. This methodology yielded an encouraging performance (see Tab. 2), while Venkataraman et al. [58] achieved better performance one year later. Gao et al. [4] mainly collected the surgical skills assessment dataset and employed a methodology proposed by Tao et al. [63]—

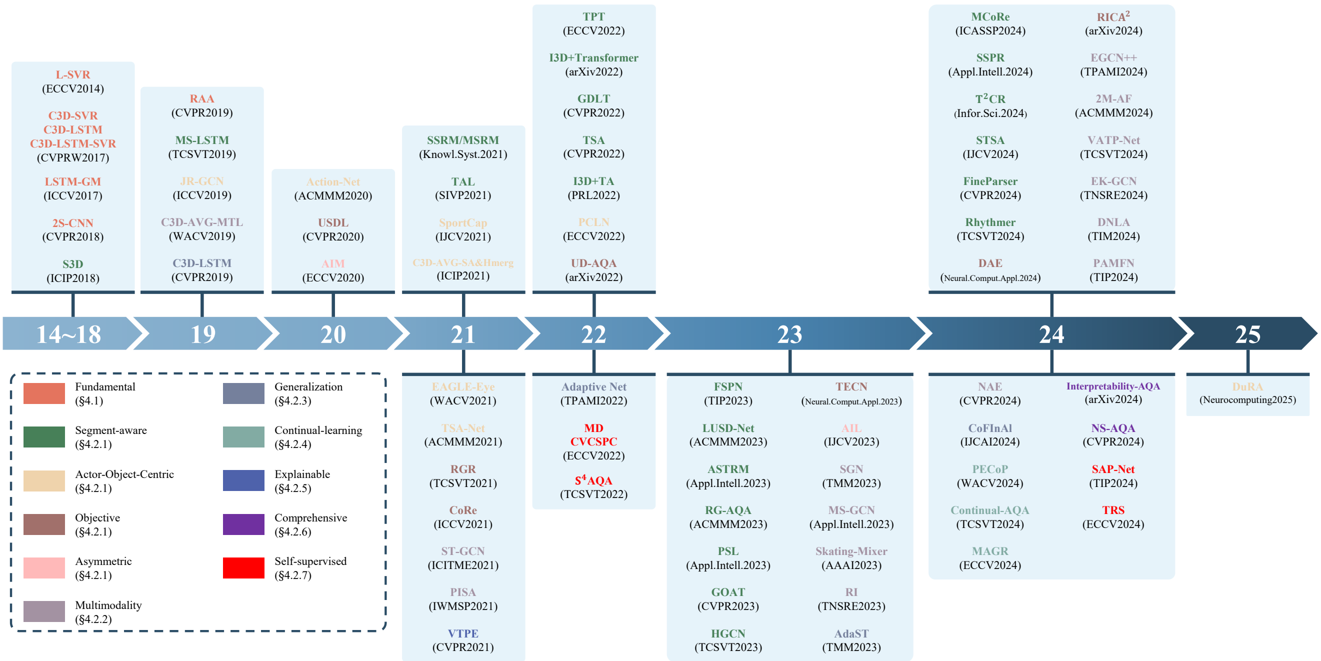


Fig. 10 Timeline of representative AQA methodologies from 2014 to 2024. Zoom in for the best view.

Table 2 Performance comparison of fundamental research methodologies. FD: Frequency domain, TD: Time domain.

Model	Year	Format	Metric	Performance	Datasets
L-SVR [2]	2014	Regression	SRC	0.4100 0.4500	MIT-Dive MIT-Skate
PCA-SVM [58]	2015	Regression	SRC	0.4500	MIT-Dive
SVM (FD) [14]	2016	Classification	Acc	0.8926	LAM
AdaBoosted-Tree (TD) [14]	2016	Classification	Acc	0.9249	LAM
C3D-SVR [3]	2017	Regression	SRC	0.7800 0.6600	UNLV-Dive UNLV-Vault
C3D-LSTM [3]	2017	Regression	SRC	0.3600 0.0500	UNLV-Dive UNLV-Vault
C3D-LSTM-SVR [3]	2017	Regression	SRC	0.6600 0.3700	UNLV-Dive UNLV-Vault
LSTM-GM [30]	2017	Pairwise Rank	Acc	0.7650	FP-Basketball
2S-CNN [16]	2018	Pairwise Rank	Acc	0.7607	EPIC-Skill
RAA [17]	2019	Pairwise Rank	Acc	0.8030 0.8120	EPIC-Skill BEST

sparse dictionary learning (SDL) combined with hidden Markov model (HMM) to realize surgical gesture recognition. AQA methodologies thus far were based on traditional/shallow machine learning.

Parmar et al. [14] introduced the classification paradigm to AQA by coarsely assessing the quality of user exercises and classifying them into erroneous vs. non-erroneous. They further tried several machine learning models on time-series and frequency-domain representations of human poses. Importantly, this work also concretely highlighted the problem of generalization of AQA models across human subjects.

Parmar and Morris et al. [3] were the first to propose utilizing deep spatiotemporal convolutional features (C3D network [64]) for AQA and proposed the UNLV-Dive and UNLV-Vault datasets based on the work of Pirsiavash et al. [2]. We believe this work marks the beginning of new age of AQA and the increased interest in AQA from the computer vision community. This work proposed three different frameworks for automated AQA: **C3D+SVR**, **C3D+LSTM**, and **C3D+LSTM+SVR**, and achieved promising performance. These frameworks differed in the way they aggregated clip-level features to obtain global video-level information. Particularly, averaging and LSTM-based [65] aggregation were explored.

Their work significantly advanced AQA in terms of both performance and datasets, with most subsequent research referencing and following this feature extraction methodology. Especially the use of pre-trained networks on large-scale, labeled action recognition datasets [66,67], such as C3D [64], Resnet [68], I3D [66], P3D [69], and VST [70], as the backbone for feature extraction is very popular in AQA.

In the same year, Bertasius et al. [30] proposed the first-person basketball dataset and contrastive learning model, which represented an expansion of the mapping techniques from regression to pairwise ranking. This work employs convolutional LSTM and Gaussian mixture models (**LSTM-GM**) to generate highly nonlinear spatiotemporal features from atomic basketball events, then compute action quality by multiplying features with linear weights learned from the data. LSTM-GM could learn the evaluation criteria and pairwise ranking from pairs of weakly labeled first-person basketball videos.

Doughty et al. [16] proposed the first skills assessment dataset for daily life activities, EPIC-Skills, in 2018. Additionally, Doughty [16] proposed a contrastive ranking model that utilizes temporal and spatial segment networks (**2S-CNN**) in combination with video segment ranking loss and similarity loss function to achieve high accuracy pairwise ranking. In 2019, Doughty et al. [17,71] also proposed a rank-aware attention model (**RAA**), in which a dual attention mechanism is used to focus on the pros and cons of action performance in long video clips and assign different weights to the clips to fuse the data. The model is trained using the combination of ranking loss, disparity loss, rank-aware loss, and diversity loss. This methodology enables accurate skill level assessment in video segments with different skill performances.

4.2 Research Trends

The fundamental research [2,3,14,16,17,30,58] established the formats of task and application scenarios of AQA, offering datasets and references for subsequent research. As computing capability and deep learning technology advanced, AQA has attracted more attention, giving rise to new methodologies. These novel methodologies consistently enhance model performance and exhibit disparate research trends, which can be categorized into multimodal methodologies, generalization, continual learning, explainable, and comprehensive AQA.

4.2.1 Fine-grained AQA

The rise of fine-grained research as the principal trend in AQA can be attributed to its capacity to address pivotal issues such as the neglect of details, the influence of noise, and low interpretability, which were prevalent in early studies. The advancements in computing capability and deep learning technology have made fine-grained research not only feasible but also mainstream. However, researchers have explored various methodologies to achieve fine-grain. To provide a more comprehensive explanation of fine-grained research, we have divided it into four sub-trends, which are elaborated respectively in the following sections.

(1) Segment-aware feature extraction

Action maybe broken down into phases/segments. For example, a diving action can be broken down into take-off, flight, entry into the water, etc. Some researchers bear the view that computing segment-level features may constitute a better methodology. Fundamental research [2,3,14,16,17,30,58] that we discussed previously did not explicitly/deliberately compute segment aware features, but were rather segment agnostic methodologies. In this section, we discuss research that computes and uses segment-aware features (see Fig. 11). The brief illustration and performance of representative research are summarized in Tab. 3 and Fig. 11.

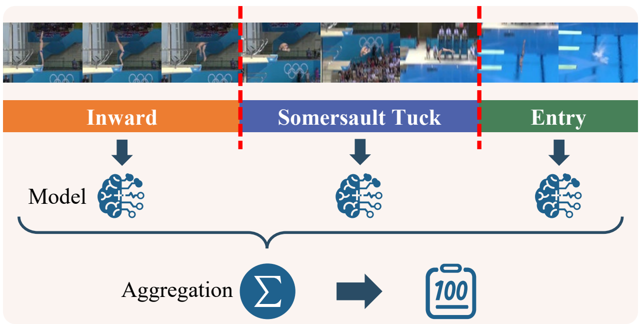


Fig. 11 Brief illustration of segment-aware feature extraction.

The first segment-aware methodology was proposed by Xiang et al. [72], utilizing multiple P3D [69] to construct the stacked 3D regressor (**S3D**). By utilizing ED-TCN [89], the entire video is segmented into video clips of different stages of dives, which are then fed into the corresponding P3D to extract features. Subsequently, the features of the separate stages are fused to predict the score. Similarly, Dong et al. [73] and Zhang et al. [82] proposed methodologies also based on ED-TCN and P3D. **SSRM/MSRM** [73] (multi-stage regression module) first segments videos into 5 segmentation by ED-TCN and inputs them into the corresponding stage regression module. Moreover, three different training strategies are proposed according to different action scenarios (the overall-score-guided scenario, execution-score-guided scenario, and difficulty-level-based overall-score-guided scenario) to achieve pertinent and rational assessment. **PSL** [82] (label-reconstruction-based

Table 3 Performance comparison of representative methodologies for segment-aware feature extraction.

Model	Year	Format	Metric	Performance	Datasets
S3D [72]	2018	Regression	SRC	0.8600	MIT-Dive
MS-LSTM [33]	2019	Regression	SRC	0.5900 0.7150	MIT-Skate Fis-V
MSRM [73]	2021	Regression	SRC	0.8752	UNLV-Dive
TAL [74]	2021	Regression	SRC	0.8649 0.7858	UNLV-Dive UNLV-Vault
GDLT [75]	2022	Regression	SRC	0.7610 0.7650	Fis-V Rhy.Gym.
I3D-Transformer [76]	2022	Regression	SRC	0.9317	MTL-AQA
TPT [77]	2022	Regression	SRC	0.8900 0.8715 0.9607	JIGSAWS AQA-7 MTL-AQA
TSA [38]	2022	Regression	SRC	0.9203	FineDiving
I3D+TA [78]	2022	Regression	SRC	0.9279	MTL-AQA
FSPN [79]	2023	Regression	SRC	0.8724 0.9601 0.9420	AQA-7 MTL-AQA FineDiving
LUSD-Net [40]	2023	Regression	SRC	0.7600 0.7920	Fis-V FineFS
ASTRM [80]	2023	Regression	SRC	0.9222	FineDiving
RG-AQA [81]	2023	Regression	SRC	0.9346	RFJS
PSL [82]	2023	Regression	SRC	0.8700	UNLV-Dive
GOAT [39]	2023	Regression	SRC	0.5599	LOGO
HGCN [83]	2023	Regression	SRC	0.9000 0.8501 0.9390	JIGSAWS AQA-7 MTL-AQA
MCoRe [84]	2024	Regression	SRC	0.9232	FineDiving
SSPR [85]	2024	Regression	SRC	0.9257 0.8538	UNLV-Dive AQA-7
T^{2}CR [86]	2024	Regression	SRC	0.9100 0.8726 0.9638 0.9275	JIGSAWS AQA-7 MTL-AQA FineDiving
STSA [43]	2024	Regression	SRC	0.9397	FineDiving
FineParser [87]	2024	Regression	SRC	0.9580 0.9435	MTL-AQA FineDiving
Rhythmer [88]	2024	Pairwise Rank	Acc	0.8845	EPIC-Skill

pseudo-subscore learning) uses the overall score as training labels and generates pseudo-score labels for each sub-stage, thereby addressing the issue of a lack of quality labels for sub-stages.

In addition to models based on CNNs like P3D [69], Transformer [90] have also been applied to AQA [75–77, 79, 91]. Alali et al. [76] simply explored different Transformer-based architectures, with **I3D-Transformer** decoder performing best. While Xu et al. [75], Bai et al. [77], and Gedamu et al. [79] implemented modifications to Transformer, mainly on the loss function. For example, **GDLT** [75] (grade-decoupling Likert Transformer) regards the score of long-term action as the combination of quantified grades and corresponding responses estimated from the video. So GDLT extracts clip features by VST [70] and feeds into the temporal context encoder to obtain the K and V, combined with a set of grades prototypes (Q) to extract grade-aware features by a grade-aware decoder (Transformer). Lastly, scores are predicted by combining the quantitative values and response intensities from grade-aware features. Similarly, Bai et al. [77, 92]proposed a temporal parsing Transformer (**TPT**), which decouples the holistic video features into local temporal features. TPT utilizes a set of learnable queries to represent the atomic temporal patterns for a specific action, combined with holistic features to decode video into a fixed number of part representations. Meanwhile, the ranking and sparsity loss are proposed on cross-attention responses to guide the queries to attend to temporally ordered clips. Finally, CoRe [31] is adopted to estimate the relative score between input and exemplar videos. **FSPN** [79] (fine-grained spatiotemporal parsing network) extract spatiotemporal action features from pairwise inputs and feed into sub-action parsing module to extract fine-grained sub-action features. Using the spatio-temporal multi-scale Transformer module to capture the long-range spatio-temporal dependencies between sub-actions. Finally, multiple scale features are concatenated to calculate the final score.

Segment-aware research also focuses on *how to segment and utilize segment information* [33, 39, 40, 74, 80, 81, 84, 86, 93–98]. Xu et al. [33] proposed **MS-LSTM** (self-attentive LSTM and multi-scale convolutional skip LSTM), combining S-LSTM for clip selection and M-LSTM for key action extraction, which together predict TES and PCS scores. Lei et al. [74] introduced temporal attention learning (**TAL**), simulating the difference in human attention to different action stages when assessing. The video features are extracted by I3D [66] before being fed into the attention learning module, which computes the prediction loss for each clip to dynamically adapt the clip weights and form weighted temporal fusion features. Similarly, **I3D-TA** [78] utilizes the time-aware attention (TA) mechanism to learn the clip relationship from adversarial loss, which compares the average aggregation result with TA. This enables I3D-TA to capture important clips.

Moreover, Lian and Shao [80] proposed an improved methodology based on cross-stage temporal inference, which feeds video features into the across-staged temporal reasoning module (**ASTRM**) to learn the temporal relationships between stages. Additionally, kernel density estimation (KDE) is employed to recalculate label density, addressing the issue of data imbalance. Liu et al. [81] proposed **RG-AQA** by introducing live, playback, and example videos with a three-stream comparative learning methodology, which is performed between live and sample videos and between live and playback videos, respectively. This ensures that the model focuses on the action itself and learns action features under varying viewpoints and scales. Similarly, An et al. [84] proposed multi-stage contrastive regression (**MCoRe**), which uses 2D CNN and gated-shift module (GSM) to extract the spatiotemporal features of the video and bidirectional GRU to divide the video into multiple consecutive stages. With the contrastive learning strategy, positive and negative sample pairs are established between different videos within the same stage. This enhances the accuracy of stage segmentation and the capacity to perceptually differentiate between features [86, 87, 99, 100]. Ke et al. [86] proposed two-path target-aware contrastive regression (**T^{2}CR**), combining direct regression and contrastive regression to model action-scoring relationships using global spatio-temporal features and local action differences. Xu et al. [87] proposed a fine-grained spatio-temporal action parser (**FineParser**) consisting of four modules: the spatial action parser (SAP) extracts the key action regions of the athlete in the video, the temporal action parser (TAP) divides actions into a series of continuous steps, the static visual encoder (SVE) enhances the representation of actions, and the fine-grained contrastive regression (FineReg) compares the action steps in the query video with those in the sample video, assesses the differences in their quality, and generates a final action score.

Most of the above methodologies tend to segment video into equally sized clips [101–114], which could lead to inter-clip confusion and inter-clip incoherence. Therefore, many researchers proposed *unequal-length segmentation* [83, 85, 88, 115–119]. **HGCN** [83] (hierarchical graph convolutional network) eliminates semantic confusion within clips through the clip refinement module, moving superfluous information to the clip where its missing part is located. Then, the scene construction module combines multiple consecutive shots into meaningful scenes to capture the dynamics of local actions. Next, the action aggregation module generates video-level features based on the dependencies between scenes and finally predicts the score through the score distribution regression module. **SSPR** [85] (semantic-sequence performance regression) extracts video features via I3D [66] and feeds into MS-TCN [120] to segment them into unequal-length clips. A feature fusion module consisting of multiple 1D convolutions to fuse semantic features of the clips, and the clip scores are predicted by regression. **Rhythmer** [88] (rhythm-aware Transformer) adaptively mines rhythm patterns related to task durations, allowing the model to dynamically adjust its focus based on the timing and sequence features of the video. Additionally, Rhythmer incorporates a co-attention module, which highlights duration-related information when comparing video segments with similar execution times. These significantly improve evaluation performance across varying task durations.

Clip-based AQA research is not only limited to methodologies but also to *datasets* [38–40, 43]. Xu et al. [38] proposed the first fine-grained sports video dataset FineDiving and temporal segmentation attention module (**TSA**), which parses the semantics and temporal structure of actions to enhance the transparency and accuracy of AQA. The quality is assessed by successively performing procedure segmentation, procedure-aware cross-attention learning, and fine-grained contrastive regression. Meanwhile, TSA is supervised by step transition labels and action score labels, which guide the model in focusing on sample regions that are consistent with the query step and quantifying their differences to predict reliable action scores. On this basis, Xu et al. [43] further proposed FineDiving+ and spatial-temporal segmentation attention module (**STAT**). In comparison to TSA, STSA incorporates a spatial motion attention module (SMA) to facilitate implicit supervision, thereby enabling the model to discern foreground action regions and filter out background noise. Different from Xu et al. [38,43], Ji et al. [40] and Zhang et al. [39] proposed datasets focusing on long video processing. Ji et al. [40] proposed FineFS and localization-assisted uncertainty score disentanglement network (**LUSD-Net**). LUSD-Net extracts video clip features through VST [39], feeds them into uncertainty score disentanglement module to decouple features into representations for PCS and TES, technical subaction localization module to locate key technical actions, and temporal interaction encoder to integrate clip features to capture contextual relationships. Finally, PCS and TES are predicted separately, and the reliability of the score is improved by uncertainty regression. Zhang et al. [39] proposed LOGO and group-aware attention (**GOAT**), using GCN to extract spatial group features and combining temporal features to focus on key action clips while ignoring irrelevant information, improving processing for long, complex videos.

(2) Actor-Object-Centric Representations

Pirsiavash et al. [2] proposed to use skeletons as features. However, just using skeletons as features would not allow to take into consideration critical non-human factors such as splash. Taking a simple and more generalized feature extraction route, which would allow to focus on human and non-human factors, Parmar and Morris et al. [3] proposed to use generalized 3DCNN features (C3D features). Following that, many of the fundamental research [2,3,14,16,17,30,58] typically extracts global image features as the input for the model. However, the global image inevitably contains irrelevant background or noise, particularly in complex scenes with diverse backgrounds. This interference could affect the model's judgment and lead to inaccurate assessments. Therefore, by focusing on the detailed analysis of the key parts and action subjects, from global image features to actor-object-centric features, it is possible to avoid interference from irrelevant factors and improve sensitivity to the details of the action. The brief illustration and performance of representative research are summarized in Fig. 12 and Tab. 4.

A different strategy focuses on local images of key parts, extracts skeletal information, and combines the two [22,37,105,121,123,126–142]. Pan et al. [121] firstly proposed joint relation GCN (**JR-GCN**), focusing on joint interactions and action. Building on the insights gained from GCN, the joint commonality module and joint difference module were employed to calculate joint and neighborhood motion features and temporal and spatial differences of the input local patches around joints, respectively. Integrating the whole-scene feature facilitates precise, fine-grained assessment with an ac-

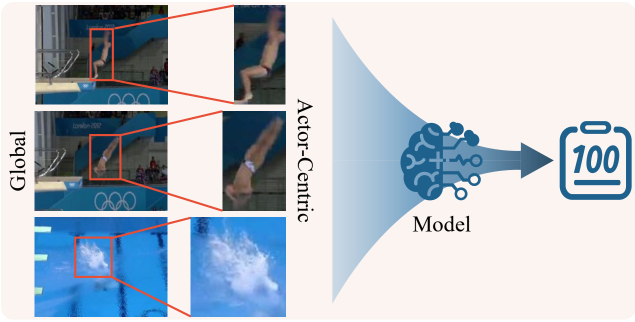


Fig. 12 Brief illustration of actor-object-centric representations.

ceptable degree of interpretability. Similarly, Chen et al. [37] proposed **SportCap**. The motion embedding module extracts implicit motion features and explicit 3D motion details, embedding them into high-dimensional space. Then, using spatial-temporal GCN (ST-GCN) to model the spatial-temporal relationships between body joints. Additionally, fine-grained action understanding is achieved through a sub-motion classifier and semantic attribute mapping module, integrating multiple action attributes into high-level action labels. Nekoui et al. [123,143] proposed (**EAGLE-Eye**). The joints coordination assessor (JCA) captures local skeletal point dynamics and interactions through multi-scale temporal convolutions, while the appearance dynamics assessor (ADA) extracts global appearance dynamics. With explicit spatiotemporal attention mechanisms, the network could focus on critical time points and spatial regions.

The other strategy focuses on the action subject [35, 36, 122, 124, 125, 144–154]. For example, Zeng et al. [35] focused on the analysis of both motion information and static pose information, proposed **ACTION-Net**. The network uses I3D [66] and ResNet [68] to extract motion and background information from video segments and the spatial information of the athlete's pose and appearance in specific frames. These features are then aggregated to produce dynamic and static features. The context-aware attention module aggregates all video segments/frames to produce dynamic/static features in each stream, then fed into a regression network to predict scores. Different from ACTION-NET [35], Nagai et al. [122], Wang et al. [36], Li et al. [124], and Haung et al. [125] proposed methodologies that ignore irrelevant backgrounds by highlighting the primary subject of actions. But Nagai et al. [122] and Wang et al. [36] achieved by direct regression, Li et al. [124], and Haung et al. [125] combined contrastive learning and regression network.

C3D-AVG-SA&HMreg [122] encompass the scene adversarial loss (SAL) and human-masked regression loss (HMRL). SAL restrains the model's dependency

Table 4 Performance comparison of representative methodologies for actor-object-centric representations.

Model	Year	Format	Metric	Performance	Datasets
JR-GCN [121]	2019	Regression	SRC	0.5700 0.7849	JIGSAWS AQA-7
ACTION-Net [35]	2020	Regression	SRC	0.6150 0.6175	MIT-Skate Rhy.Gym.
SportCap [37]	2021	Regression	SRC	0.8620 0.6170	MTL-AQA SMART
C3D-AVG-SA&Hmreg [122]	2021	Regression	SRC	0.8970	MTL-AQA
EAGLE-Eye [123]	2021	Regression	SRC	0.6010 0.9158	MIT-SKate AQA-7
TSA-Net [36]	2021	Regression	SRC	0.8476 0.9393	AQA-7 MTL-AQA
PCLN [124]	2022	Regression	SRC	0.8795 0.9014	AQA-7 MTL-AQA
NS-AQA [26]	2024	Regression	Expert Preference	0.9610	FineDiving
DuRA [125]	2025	Regression	SRC	0.8727 0.9533	AQA-7 MTL-AQA

on scene background information through adversarial training. In contrast, HMRL ensures the model's concentration on actions by regressing to a fixed score when the target action is not visible. While, **TSA-Net** [36] combines target tracking and self-attention mechanisms, using VOT trackers and I3D [66] to extract tracking results and features, which are then aggregated using the TSA module and fed into regression network to predict scores. **PCLN** [124] uses Resnet [68] to extract features of paired videos, then calculate the expected score and predicted difference score, respectively. **DuRA** [125] leverages both semantic-level grade prototypes and individual-level reference samples to enhance the focus on action details while filtering out irrelevant information. At the semantic level, the rating-guided attention (RGA) module is used to emphasize local features critical to action quality. At the individual level, consistency preserving (CP) constraints are used to refine the representation of local features and suppress distractions.

Okamoto and Parmar [26] proposed to use **dedicated human pose estimators** and **object detectors** like platform detectors and splash detectors to extract Actor and Object-centric information. And then processes this information using rules to generate detailed AQA reports, including scores.

Another way to guide model's attention to athlete and other important features like splash is using **detailed supervision** as done by Parmar and Morris [21]. By optimizing the network end-to-end for detailed action classification, detailed action description and AQA scoring, the network would be able to better characterize the action by better focusing on the actor, action, and important factors like splash.

(3) From Subjective to Objective Evaluation

"*If you want fairness, come to sports; if you want absolute fairness, don't come to sports!*" — From film "**Pegasus**"

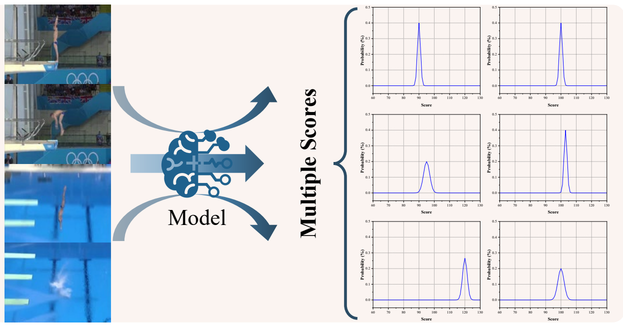


Fig. 13 Brief illustration of objective evaluation.

Fairness has always been a critical concern in competitive sports. Various measures, such as multi-judge scoring systems, have been implemented to make competitions as fair as possible. However, biases still exist, including nationalistic bias [155, 156], where judges may favor athletes from their own country or give preference to athletes performing more difficult maneuvers [157]. Fundamental research [2, 3, 14, 16, 17, 30, 58] used referee scores as ground truth without accounting for the uncertainty and potential ambiguity arising from subjective assessments. This reliance on subjective scores affected the accuracy and fairness of the models and hindered their practical application. Therefore, extracting effective information from existing scoring labels and improving the accuracy of AQA has become a research focus. Researchers have introduced methodologies such as uncertainty modeling [158–163], contrastive regression [31, 164] to reduce the subjective bias in human scoring. The brief illustration and performance of representative research are summarized in Fig. 13 and Tab. 6.

Tang et al. [158] was the first to introduce uncertainty modeling into AQA to reduce the intrinsic ambiguity in score labels caused by multiple judges and their subjective assessments, proposed uncertainty-aware score distribution learning (**USDL**), which generates a gaussian distribution of the labeled scores and calculates predicted score distribution of a given video and score label. By minimizing the KL divergence between two distributions, the network is optimized to fit the actual scoring process better. Drawing inspiration from the Olympic multi-judge system, USDL could extend to K pathways, where scores from different paths are weighted and fused to predict the final score. Similarly, several researchers have further explored uncertainty-based methodologies. Zhou et al. [159] proposed uncertainty-driven AQA (**UD-AQA**), consisting of a deterministic branch and a latent branch. The deterministic branch extracts video features by I3D [66], while the latent branch models the inherent ambiguity of the video using conditional variational auto-encoder (CVAE), combining both branches to predict scores. Li et al. [161] proposed a gaussian-guided frame sequence encoder network, which uses ResNet [68] to extract features from videos, then fed into sequence-based temporal encoder convolutional network (**TECN**) to extract the spatio-temporal semantic information of actions. Gaussian loss function is used to maximize the probability of the predicted scores fitting the distribution. Zhang et al. [163] proposed distribution auto-encoder (**DAE**), a regression model that encodes video features into a score distribution and uses a reparameterization trick to sample predictions, simultaneously learning to predict scores and uncertainty. Majeedi et al. [162] proposed a rubric-informed, calibrated assessment of actions (**RICA^2**), which represents action steps and scoring criteria as a directed acyclic graph (DAG). GNN is used to capture uncertainty in the scoring process by generating probabilistic embeddings for each action step, which are then propagated upward to calculate the final score.

In contrast to uncertainty modeling, contrastive regression [31, 154, 164–167] is used to reduce subjective scoring bias by comparing sample and reference videos. Jain et al. [164, 168] proposed reference-guided regression (**RGR**), transforming an action scoring task into an action video similarity scoring task. Videos are processed in pairs using the Siamese-LSTM network, where C3D [64] extracts clip features and feeds them into LSTM to generate global action features. Subsequently, extracted features are concatenated and fed into a regression model, which outputs a similarity judgment. Besides, RGR could calculate clip similarity between sample and reference videos, allowing for more fine-grained and interpretable scoring. Yu et al. [31] proposed the contrastive regression (**CoRe**) framework, which selects a reference video for each sample video based on action category and difficulty. Paired spatiotemporal features are extracted using I3D and combined with the reference score, then fed into a group-aware regression tree (GART). GART classifies features from coarse to fine, regressing the score difference between the two videos.

(4) Modeling Asymmetric Relationships

Fundamental research [2, 3, 14, 16, 17, 30, 58] concentrate on single-agent actions. For multiple agents in the scene, a general assumption is that the impact of different agents on the quality is equivalent. However, in daily life and competitive sports, it is evident that many behaviors comprise interactions with asymmetric relationships between different agents. Therefore, previous research fails to consider priority in real-world scenarios and lacks the modeling of asymmetric relationships between different agents. Asymmetric research facilitates more detailed feature extraction, so we consider this to be a fine-grained category. The brief illustration and performance of representative research are summarized in Fig. 14 and Tab. 5.

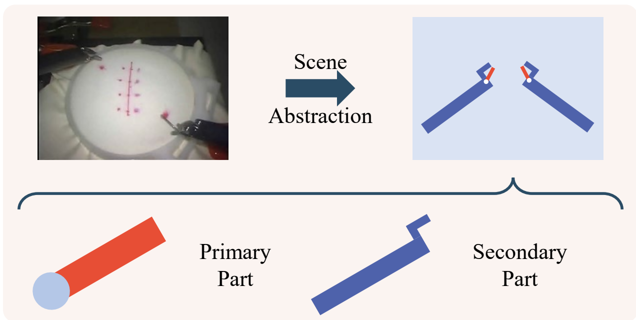


Fig. 14 Brief illustration of modeling asymmetric relationships.

The asymmetric relationship was first proposed in AQA and made major contributions by Gao et al. [34, 41]. To date, this research include two models, **AIM** [34] and **AIL** [41], and two dedicated datasets, TASD-2 [34] and PaSK [41]. AIM [34], as the first asymmetric relationship AQA model, is used to explicitly model asymmetric interactions between agents. Initially, AIM manually categorizes participating agents as primary or secondary, calculates the difference features between them and fuses the primary information to obtain the primary-secondary information. These are fed into LSTM to learn the temporal interaction, thereby obtaining complete AIM features. Additionally, global scene features are extracted using I3D [66] and fused with AIM features

Table 5 Performance comparison of representative methodologies for objective evaluation.

Model	Year	Format	Metric	Performance	Datasets
USDL [158]	2020	Regression	SRC	0.7000	JIGSAWS
				0.8102	AQA-7
				0.9273	MTL-AQA
RGR [164]	2021	Regression	SRC	0.6900	UNLV-Dive
				0.5360	UNLV-Vault
				0.7600	MTL-AQA
CoRe [31]	2021	Regression	SRC	0.8400	JIGSAWS
				0.8410	AQA-7
				0.9512	MTL-AQA
UD-AQA [159]	2022	Regression	SRC	0.8900	JIGSAWS
				0.9545	MTL-AQA
				0.9341	FineDiving
TECN [161]	2023	Regression	SRC	0.8504	AQA-7
				0.9095	MTL-AQA
DAE [163]	2024	Regression	SRC	0.7600	JIGSAWS
				0.8258	AQA-7
				0.9232	MTL-AQA
NS-AQA [26]	2024	Regression	Expert Preference	0.9610	FineDiving
RICA^2 [162]	2024	Regression	SRC	0.9200	JIGSAWS
				0.9512	MTL-AQA
				0.9421	FineDiving

via an attention mechanism to predict action scores. Building on AIM, AIL [41] introduced two key modules: the automatic assigner and the operation search module, which can automatically classify the primary and secondary agents in an action and dynamically adjust their structure for different interactive actions. Furthermore, AIL can adaptively model different action scenarios, including strong asymmetry and weak asymmetry, enhancing the flexibility and generalization ability of the model.

Table 6 Performance comparison of representative methodologies for modeling asymmetric relationships. Yr.: Year, Fmt.: Format, Met.: Metric, Perf.: Performance.

Model	Yr.	Fmt.	Met.	Perf.	Datasets
AIM [34]	'20	Reg	SRC	0.7100	JIGSAWS
				0.7789	AQA-7
				0.8831	TASD-2
AIL [41]	'23	Reg	SRC	0.8800	JIGSAWS
				0.8126	AQA-7
				0.8600	PaSk
		PR	Acc	0.8340	EPIC-Skill
				0.8190	BEST

(5) Conclusion of Fine-grained

A thorough review of the methodologies described and an analysis of the tables (Tab. 3, Tab. 4, Tab. 5, Tab. 6) above clearly demonstrates that fine-grained research has significantly improved model performance compared to [2,3,14,16,17,30,58]. In particular, this finding demonstrates the superiority of fine-grained research methods in capturing minutiae, suppressing extraneous variables, resolving temporal dependencies, and handling label ambiguity. The four sub-trends encompassed within the fine-grained research reflect the researchers' in-depth consideration of multiple dimensions, including temporal, spatial, outcome, and agent relationships. Given that AQA involves traversing the entire action sequence, research emphasizing segment-aware feature extraction has become the primary focus of fine-grained research.

4.2.2 Multitask & Multimodal AQA

Early AQA research demonstrated promising performance using RGB videos [2,3,16,17] and skeletons [5] as input modalities. However, with the advancement of time and technology, the limitations of single-modal data have become evident. Researchers now can handle more complex data modalities, leading to the rise of multimodal methodologies as a key research trend. Parmar et al. [27] introduced multimodal AQA. Multimodal methodologies offer significant advantages, compensating for single-modality limitations in complex scenarios by providing complementary information. This enhances model robustness and enables a more detailed AQA. The brief illustration and performance of representative research is summarized in Fig. 15 and Tab. 7.

The initial multi-task model in AQA came from Parmar and Morris [21,178], who proposed an end-to-end multi-task learning framework (**C3D-AVG-MTL**), marking a pioneering introduction of text into the AQA. The C3D-AVG-MTL extracts spatiotemporal features through a shared C3D [64] backbone and transmits them

Table 7 Performance comparison of representative multitask and multimodal methodologies.

Model	Year	Format	Metric	Performance	Datasets
C3D-AVG-MTL [21]	2019	Regression	SRC	0.9044	MTL-AQA
ST-GCN [25]	2021	Regression	SRC	0.6300	MIT-Skate
PISA [27]	2021	Classification	Acc	0.7460	Piano-Skills
SGN [169]	2023	Regression	SRC	0.9607 0.7650	MTL-AQA Fis-V
MS-GCN [170]	2023	Regression	SRC	0.7200	Rhy.Gym.
Skating-Mixer [171]	2023	Regression	SRC	0.7500	Fis-V
RI [172]	2023	Classification	Acc	0.9886	UI-PRDM
EGCN++ [29]	2024	Classification	Acc	0.9500 0.9282	UI-PRDM KIMORE
2M-AF [173]	2024	Regression	SRC	0.8838 0.8901	UNLV-Diving AQA-7
VATP-Net [174]	2024	Regression	SRC	0.9588 0.7960 0.8000 0.7725	MTL-AQA Fis-V Rhy.Gym. FineFS
EK-GCN [175]	2024	Regression	SRC	0.8080	KIMORE
DNLA [176]	2024	Regression	SRC	0.6800	AGF-Olympics
PAMFN [23]	2024	Regression	SRC	0.8220 0.8190	Fis-V Rhy.Gym.
NAE [177]	2024	Regression	SRC	0.9790	MTL-AQA

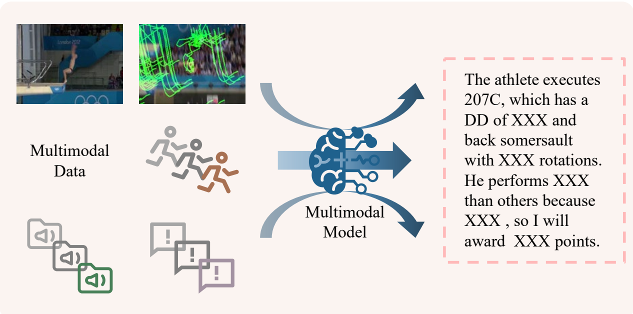


Fig. 15 Brief illustration of multimodal AQA.

to independently task-specific branches, including AQA score regressor, factorized action classifier, and caption generator. The whole model, including the backbone and individual branches, is trained end-to-end to learn powerful AQA-oriented representations. An end-to-end joint optimization network enables the realization of fine-grained action descriptions and AQA scoring.

Different from Parmar and Morris [21], Du et al. [169], Xu et al. [179], and Gedamu et al. [174] chose to amalgamate visual and semantic information rather than process them independently. **SGN** [169] (semantics-guided network) employs teacher-student architecture to amalgamate visual and semantic information, with the teacher network providing knowledge containing semantic information and the student network adapting this knowledge by a set of learnable atomic queries and attention mechanism. Three loss functions were proposed to ensure the effective alignment of visual and semantic features. Similarly, **VATP-Net** [174] (visual-semantic alignment temporal parsing network) contains two modules. The self-supervised temporal parsing module is responsible for understanding the high-level representation along with internal temporal structures of action features, while the multimodal interaction module captures visual-semantic action features and the interaction between different modalities, enabling a better understanding of scene-invariant action execution. The entire network leverages self-supervised and multimodal learning, significantly improving action quality assessment performance without relying on external labels. Following Parmar and Morris [21], Zhang et al. [177] proposed a new task called narrative action evaluation (**NAE**), transforming score regression into a video-text matching task. In this methodology, video features are extracted by a video encoder. At the same time, context-aware prompt learning employs a multi-head cross-attention mechanism to match learnable prompts with video features, allowing prompts to perceive the video's context. In the score-guided tokens learning, video features are combined with prompts containing scoring information. These combined features are fed into the multimodal-aware text generator, which generates natural language descriptions that include action details, scores, and quality assessments. A tri-token attention mask is employed throughout the process to balance the linguistic richness of the action description with the accuracy of the assessment, thereby generating detailed and professional narratives and evaluations. The interaction between different modal tasks, particularly the linkage between action assessment and text generation, facilitates enhanced mutual promotion.

In addition to the text modality, many researchers introduced skeleton into AQA [24, 25, 29, 170, 172, 173, 175, 176, 180–189]. The majority of them conducted research based on GCN but with distinct methodologies. For example, **ST-GCN** [25] (spatiotemporal GCN) extracts features from skeleton data and deep pose features in the spatiotemporal dimension, then fed into regression network to predict scores. **MS-GCN** [170] (multi-skeleton structures GCN) extracts and segments skeleton sequences in long videos into non-overlapping subsequences. Meanwhile, three types of skeleton graphs (self-connection, intra-part connection, and inter-part connection) are constructed to extract pose features. The temporal relationships between sub-sequences were learned by a temporal attention module and then fed into the regression network. **RI** [172] calculates the relative directional relationships between body joints to construct a dot product matrix that effectively removes the influence of body orientation, thereby addressing the sensitivity of traditional models to orientation changes. **EGCN++** [29] combines positional and orientation information from skeleton data at the data level for comprehensive feature integration. In the graph convolutional layers, the model captures spatial relationships between joints and extracts deep features through the graph structure. At the model level, an ensemble learning strategy is employed to combine predictions from multiple graph convolutional networks, improving accuracy and robustness. It is worth mentioning that EGCN++ relabels KIMORE [6] to change the score to a 2-level categorical label. **2M-AF** [173] (multi-modality assessment framework) employs self-supervised mask encoded GCN (SME-GCN) and I3D [66] to extract skeleton and RGB features. These are fused by a preference fusion module (PFM) and fed into a regression network to predict scores. **EK-GCN** [175] captures spatial relationships between joints by GCN, with expert knowledge guiding the network to focus on key joints and motion features through weighted graph structures. The gated pooling module selectively aggregates features, while Transformer module models long-range temporal dependencies. **DNLA** [176, 190] (discriminative nonlocal attention) extracts sparse features with cross-temporal correlations from the skeleton and video clips, which are fed into the regression network for predicting scores. DNLA eliminates superfluous data, enabling the model to concentrate on pivotal actions and frames.

Beyond text and skeleton, researchers [23, 27, 171, 191] combined visual information with audio information. **PISA** [27] utilized 3DCNN and 2DCNN to extract video and audio features, respectively, concatenated the features, and finally output the piano performance skill level through the fully connected layer. **Skating-Mixer** [171] extracts video and audio features through CNN and MFCC, respectively. These features are then fused into a joint feature vector and passed through the fully connected layer. Meanwhile, Skating-Mixer enhances traditional MLP frameworks by incorporating a memory mechanism, making the model not only process the current frame but also recall and leverage information from previous frames. Notably, Zeng and Zheng et al. [23] took optical flow into account, proposing a progressive, adaptive multimodal fusion network (**PAMFN**). The RGB, optical flow, and audio data were extracted through three independent modality-specific branches, and a mixed-modality branch progressively fused the information using a modality-specific feature decoder and an adaptive fusion module, selecting the optimal fusion strategy. The cross-modal feature decoder further transmits the cross-modal features after adaptive fusion to the mixed modality branch, leading to accurate action quality assessments.

4.2.3 Generalization in AQA

As discussed above, AQA research has typically focused on single-action scenarios. In the initial stage, due to the limitations of available datasets and technical capabilities, researchers prioritized single-action scenarios with clear assessment criteria. However, the generalization ability of models has become a major focus in AQA research. Single-action scene dataset trained models face significant limitations when applied to diverse, open environments and cross-task assessments. Researchers seek to enhance the generalization ability of models, enabling their deployment in disparate datasets, tasks, and application scenarios. The brief illustration and performance of representative research is summarized in Fig. 16 and Tab. 8.

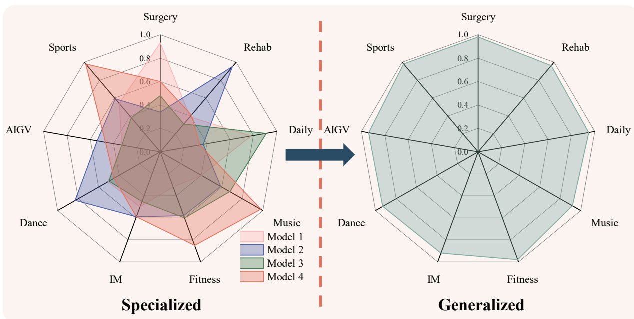


Fig. 16 Brief illustration of generalization in AQA.

The initial generalization model in AQA also came from Parmar and Morris [32], who trained a **C3D+LSTM**

Table 8 Performance comparison of representative methodologies for generalization.

Model	Year	Format	Metric	Performance	Datasets
C3D+LSTM [32]	2019	Regression	SRC	0.6478	AQA-7
Adaptive Net [192]	2022	Regression	SRC	0.7815	JIGSAWS
				0.8500	AQA-7
		Pairwise Rank	Acc	0.8406	EPIC-Skill
				0.8327	BEST
AdaST [193]	2023	Regression	SRC	0.8443	AQA-7
		Pairwise Rank	Acc	0.8832	EPIC-Skill
				0.8534	BEST
CoFInAl [194]	2024	Classification	SRC	0.7880	Fis-V
				0.8070	Rhy.Gym.

model using the multi-action scene dataset AQA-7. The experimental results demonstrated that the model trained on a multi-action scene dataset showed superior performance and generalization ability. Parmar and Morris [32] highlighted that common features can be learned across different actions, and training on multi-action scene dataset allows the model to benefit from knowledge transfer and sharing, leading to faster convergence and improved evaluation results even with a limited number of samples. This study laid the foundation for subsequent research on the generalization of multi-action scenarios and provided a widely cited dataset.

Unlike Parmar and Morris [32], who focused on training models by multi-action scenarios, some researchers [192–194] aspire to realize high generalization through advancements in model architecture. Pan et al. [192] addressed the problem of single architectures struggling to achieve high performance across all actions and the necessity for manually designing architectures for different types of actions. They proposed the initial adaptive network (**Adaptive Net**) for AQA, which is constructed by acyclic graph and inputs features extracted on JR-GCN. The network contains multiple operation selectors, each containing seven candidate operations. The optimal network architecture is searched using the differentiate architecture search (DARTS) mechanism. Adaptive Net demonstrated feasibility and superior performance in regression and ranking tasks across various datasets, including competitive sports, surgery, and daily actions. Zhang et al. [193] propose an adaptive stage-aware assessment skill transfer framework (**AdaST**), different from the existing assessment skill transfer that pre-train the model with all source actions or jointly training with source actions and target actions, AdaST models the relationship between the source and target action. Firstly, AdaST employs a genetic-based adaptive source action search scheme to identify relevant source actions. Next, the target action samples are directly fed into the encoders to extract features, after which the stage-aware assessment module transfers different assessment skills from the source actions to the corresponding stages of the target action. Similarly, Zhou et al. [194] proposed coarse-to-fine instruction alignment (**CoFInAl**) to solve issues of domain shifting and overfitting under limited data. CoFInAl mimics the referee's judging process by first providing coarse grade assessments and then fine-grain scores within each grade, aligning the AQA task with the pre-trained model.

4.2.4 Continual learning AQA

To adapt to external changes, humans and other living organisms have developed strong adaptive abilities, enabling them to acquire, update, accumulate, and utilize knowledge continuously. Naturally, we hope that AI can function similarly, which is why the concept of continuous learning has been proposed. Unlike traditional models that learn from static data distributions, continual learning involves learning from dynamic data distributions. A significant challenge identified in continual learning is catastrophic forgetting, where the model's adaptation to a new distribution typically results in a substantial decline in performance on the old distribution. Some researchers introduced continual learning to AQA [195–197] and tried to address the catastrophic forgetting. The brief illustration and performance of representative research is summarized in Fig. 12 and Tab. 9.

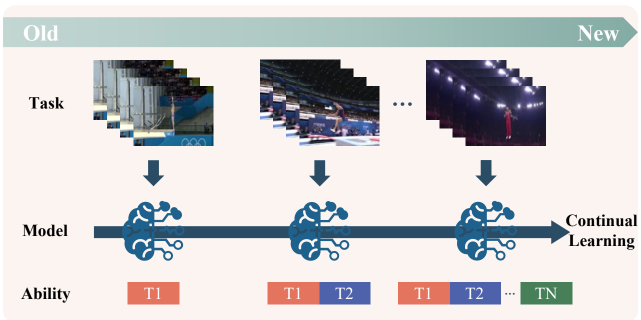


Fig. 17 Brief illustration of continual learning in AQA.

Dadashzadeh et al. [195, 198] proposed a parameter-efficient continual pre-training (**PECoP**) framework to address this problem. A lightweight convolutional bottleneck block, 3D-Adapter, is employed to conduct domain-specific self-supervised pre-training of I3D [66]. Only the parameters of the 3D-Adapter are updated, while the I3D weights are maintained in a frozen state. However, on the target dataset, I3D is fine-tuned in conjunction with 3D-Adapter to align with the requirements of AQA. **Continual-AQA** [196] contains two core components: feature-score correlation-aware rehearsal (FSCAR) and action general-specific graph (AGSG) modules. Upon receiving a new task, FSCAR first extracts representative samples from the finite storage of previous tasks and enhances the features and scores in order to maintain the continuity of the feature distribution. Then, AGSG combines the action features of the new task with the general and specific knowledge of previous tasks to extract discriminative scoring features that are consistent with the task. Finally, the model is trained on the new task in order to mitigate the issue of forgetting by integrating the features of the current and historical tasks. Similarly, Zhou et al. [197] proposed manifold-aligned graph regularization (**MAGR**), which addresses the misalignment between static old features and the dynamically changing feature manifold causes severe catastrophic forgetting. MAGR aligns the old features with the current feature manifold, constructs graph regularization to maintain the consistency of the feature distribution and mass fraction distribution, and combines feature replay to retain the memory of old data while adapting to the new data distribution.

4.2.5 Explainable AQA

Interpretability is crucial in AQA to make systems trustworthy and adaptable to the wider population. For example, a community would not adopt an AI judge if it cannot explain how it arrived at a decision, especially if its decision was different than fellow human judges, or people would be hesitant to follow an AI physiotherapist if they cannot "see" inner workings of this AI physiotherapists' brain. However, given that existing research is mainly based on end-to-end black box deep learning, interpretability is relatively challenging. Some researchers have attempted to use visualization or understandable neural symbols to increase the interpretability of models [26, 199–203]. The brief illustration and performance of representative research is summarized in Fig. 18 and Tab. 10.

Distinct from dominant black-box deep learning, Okamoto and Parmar [26] introduced novel Hierarchical **NeuroSymbolic** paradigm (**NS-AQA**) [26], which

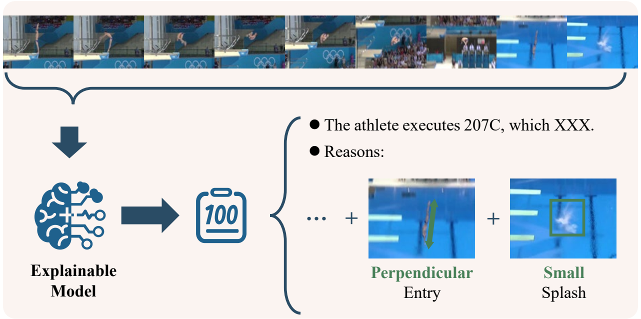


Fig. 18 Brief illustration of explainable AQA.

combines neural networks with **interpretable rule-based models**. Neural networks are used to deconstruct the visual scene of the performance by extracting interpretable symbols from video data, and rule-based models are constructed based on current rules of sport for action recognition and temporal segmentation. **Microprograms** are employed to assess fine-grained action quality at each stage, where a ranking-based percentage or fixed scoring formula is utilized to quantify the score of each action element, generating an overall score and detailed AQA reports, complete with visual evidence. Such an interpretable and flexible system also allows for personalized AQA, where users can modify the system on the fly to suit their needs. This formula provides a traceable and transparent scoring system that makes the assessment interpretable. Moreover, their interpretable system achieves SOTA performance on the tasks of finegrained **action recognition** and **temporal segmentation**.

Dong et al. [203] proposed **Interpretability-AQA** which combines a novel attention loss function with a query-based Transformer decoder network to make sure the model focuses on keyframes and segments while avoiding temporal skipping. A weight-score regression module is used to simulate human scoring patterns, allowing each video segment's score to be explained in terms of different weights.

Matsuyama et al. [202] proposed **IRIS** to address AQA in Figure Skating domain by leveraging rubrics to: 1) segment the action sequence; 2) compute technical element score differences of each segment relative to base scores; 3) compute multiple program component scores; and 4) compute the final score. They found that their approach offered interpretability while improving the performance on AQA task.

4.2.6 Comprehensive AQA

Action execution quality, in general, is composed of factors. Ideally, we want the models to take into consideration all these factors when quantifying the quality

Table 9 Performance comparison of representative continual-learning methodologies.

Model	Year	Format	Metric	Performance	Datasets
PECoP [195]	2024	Regression	SRC	0.8900	JIGSAWS
				0.9520	MTL-AQA
Continual-AQA [196]	2024	Regression	SRC	0.6492	AQA-7
				0.7895	MTL-AQA
		Pairwise Rank	Acc	0.7883	BEST
MAGR [197]	2024	Regression	SRC	0.7666	UNLV-Dive
				0.8979	MTL-AQA
				0.8580	FineDiving

Table 10 Performance comparison of representative explainable methodologies.

Model	Year	Format	Metric	Performance	Datasets
IRIS [202]	2023	Regression	SRC	0.8910	MIT-Skate
NeuroSymbolic-AQA [26]	2024	Regression	Expert Preference	0.9610	Finediving
Interpretability-AQA [203]	2024	Regression	SRC	0.7880	Fis-V
				0.8420	Rhy.Gym.

of an action. However, it was revealed that models do not take all the factors into consideration. This may be due to the fact that AQA datasets are trained on scores given out by human judges. Human judges are required to give out scores very quickly without any replays in most cases and domains. As such, the last part of the performance influences judges most, while the other factors are "forgotten". As a result, the human judges give out biased or non-comprehensive scores. Hence, machine learning models trained using such biased ground truth are also biased and non-comprehensive. Therefore, some researchers attempt to take more factors into account in addition to the action subject [204, 205]. Although their research is still far from the ideal, there has been significant progress compared to existing research. The brief illustration and performance of representative research is summarized in Fig. 19 and Tab. 11.

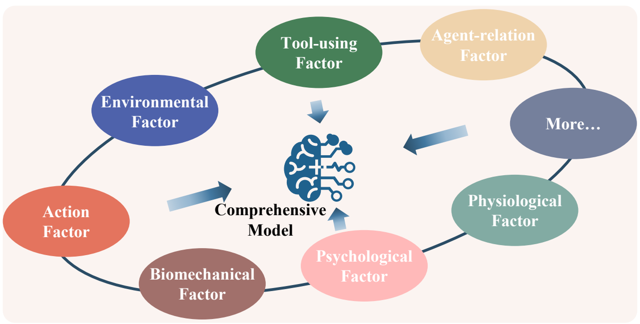


Fig. 19 Brief illustration of comprehensive AQA.

Liu et al. [204] proposed a unified multi-path framework for automatic surgical skill assessment (**VTPE**), which takes multiple aspects of surgical performance into consideration, including surgical tool usage, surgical field clearness, and surgical event pattern. Each path extracts features from different aspects and transfers them into skill score sequences. Moreover, a path dependency module models the interdependencies among these different aspects to provide weights of temporal importance for the score sequences. Lastly, the weighted score sequences are pooled over time and fused across paths as the final assessment prediction. Similarly, Wang et al. [205] proposed to take the skeletal, holistic appearance, facial and scenic factors into consideration. Meanwhile, a relational temporal convolutional network (**RTCN**) and pyramidal skeleton GCN (**PSGCN**) are utilized to exploit appearance dynamics with non-local temporal relations and hierarchically refine human skeleton graphs, respectively. Finally, the adaptively learned attention weights for different streams according to the input video are used to fuse features and predict the score.

4.2.7 Self-supervised Representation Learning for AQA

Annotating a dataset requires human efforts and financial resources. This is more so in the case of AQA because unlike in the case of object or action recognition, where untrained humans can easily annotate datasets, for AQA, trained experts are needed to assign a quality score to action samples. Trained experts/judges generally undergo elaborate training and get their licenses from governing bodies of the respective areas. This makes crowdsourcing of annotations infeasible. Datasets thus far have been created from Olympic footage. Unfortunately, a large amount of Olympic footage is unavailable for every action. Labels for action recognition samples can be obtained at a faster rate by mining video sites using keywords. However, compared to action recognition, more

Table 11 Performance comparison of representative comprehensive methodologies.

Model	Year	Format	Metric	Performance	Datasets
VTPE [204]	2021	Regression	SRC	0.8000	JIGSAWS
PSGCN-RTCN [205]	2022	Regression	SRC	0.7200	MIT-Skate
				0.8800	UNLV-Dive
				0.7600	UNLV-Vault
NS-AQA [26]	2024	Regression	Expert Preference	0.9610	FineDiving

effort is required on the annotators' part to annotate the AQA dataset as the annotators actually need to go through the whole footage and stop and record scores given out by judges. To overcome the need for large labeled datasets, some research [8, 206–214] have explored self-supervised representation learning (SSL) for AQA.

The brief illustration and performance of representative research is summarized in Fig. 20 and Tab. 12.

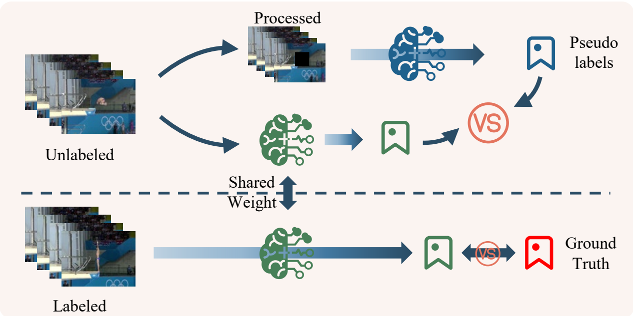


Fig. 20 Brief illustration of self-supervised representation learning for AQA.

Roditakis et al. [208] propose an SSL methodology based on temporal alignment pretext task. Parmar et al. [8] propose a **Motion Disentangling (MD)** SSL methodology, where erroneous motion is isolated from global motion. Global motion is equivalent to the ideal way of executing actions; while erroneous motion refers to harmful deviations induced as a result of lower-quality execution. They also use domain knowledge in formulating pretext tasks. They show applicability to fitness actions and diving actions. Their methodology is applicable to other domains like diving as well. Parmar et al. [8] also propose **Self-Supervised Pose-Contrastive** approach (CVCSPC) and **Self-Supervised Pose Disentangling** to learn robust pose sensitive representations from noisy in-the-wild videos. Parmar et al. [222] propose a novel SSL methodology based on connecting actions and their effects (**SSL CATE**) to learn fine-grained, pose-sensitive representations from unlabeled videos.

Zhang et al. [209] extended generalization research and took the lead in proposing self-supervised semi-supervised action quality assessment (**S^{4}AQA**), thereby alleviating the dependence of existing methods on a large amount of labeled data. S^{4}AQA employed multi-task learning by combining a self-supervised masked clip feature recovery task with a supervised score regression task. A representation distribution alignment module with a gradient inversion layer was used to align the representation distributions of labeled and unlabeled videos through adversarial training, enhancing the model's ability to generalize by efficiently utilizing unlabeled data.

Gedamu et al. [212] proposed a self-supervised sub-action parsing network (**SAP-Net**) that utilizes a teacher-student network structure to learn consistent semantic representations between labeled and unlabeled samples. Specifically, the teacher network generates high-quality pseudo-labels through actor-centric region detection, while the self-supervised sub-action parsing methodology disaggregates complex actions into fine-grained sub-action sequences. Additionally, SAP-Net employs a contrastive learning mechanism with pseudo-labels to ensure consistency in motion-oriented action features between the teacher and student branches.

Similarly, Yun et al. [213] proposed semi-supervised teacher-reference-student (**TRS**), where the teacher network generates pseudo-labels, the reference network provides supplementary supervised information, thereby facilitating the learning of unlabeled data by the student network. It is evident that introducing semi-supervised learning into AQA can mitigate the reliance of existing methodologies on extensive labeled data. By integrating limited labeled data with vast unlabeled data, models can more effectively discern the underlying structure and distribution of the data, thereby enhancing their robustness and generalization capabilities.

4.3 Conclusion of Methodology

Over the past decade, AQA has evolved from a niche research field to a thriving one, with a growing number of researchers coming into the field and proposing many novel research methodologies (see Fig. 2). Consequently, AQA has progressed significantly, gradually evolving from fundamental coarse-grained research to more fine-grained, multimodal, generalized, inter-

Table 12 Performance comparison of representative self-supervised methodologies.

Model	Year	Format	Metric	Performance	Datasets
Roditakis et al. [208]	2021	Regression	SRC	0.7700	MTL-AQA
Motion Disentangling [8]	2022	Regression	SRC	0.7763	MTL-AQA
Pose Contrastive [8]	2022	Classification	F1 Score	0.7763	MTL-AQA
S^4AQA [209]	2022	Regression	SRC	0.6550 0.7460 0.3420	JIGSAWS MTL-AQA Rhy.Gym.
SSL-CATE [222]	2024	Regression	SRC	0.7936	MTL-AQA
SAP-Net [212]	2024	Regression	SRC	0.8010 0.3930 0.8450 0.7090	MTL-AQA Rhy.Gym. FineDiving FineFS
TRS [213]	2024	Regression	SRC	0.7530 0.9010 0.5290	JIGSAWS MTL-AQA Rhy.Gym.

pretable, and comprehensive research, along with the incorporation of new research paradigms such as continuous learning and self-supervised learning. Research has gradually transitioned from a focus on single-task optimization to the integration of multi-dimensional methodologies and the expansion of application scenarios. However, there are still some limitations in existing research, while some challenges within AQA hinder its development (discussed in the following).

5 Challenges & Opportunities

As evidenced by the datasets (section 3) and methodologies (section 4), there have been dramatic advances in AQA over the past decade. Numerous datasets and publications have been proposed, particularly in 2023, with an increase of nearly 100% compared to the previous (see Fig. 2). Nevertheless, several challenges persist, hindering the advancement of AQA. Thus, to guide the community and stimulate advances in the field of AQA, in this section, we discuss challenges and corresponding future directions on three fronts: actions, datasets, and methodologies.

5.1 Action

The Challenge of Inherent complexity. Actions can be conscious or unconscious, natural or trained, but regardless of their type, they are imbued inherently with an individual intention. This means the same action command can be performed differently by various individuals. For instance, "run" can be executed fast, slow, or even with variations like jumping. In other words, a single action category may encompass numerous movement styles. Additionally, the same action can appear differently when captured from various views. This indicates significant inter- and intra-class differences in actions. Furthermore, the requirements for models vary across different action scenarios. To illustrate, competitive sports demand objective and highly interpretable models, whereas rehabilitation and medical care demand precision and long-term processing.

Consequently, the inherent complexity of actions is considered the main challenge hindering the generalization of AQA to date. Despite the numerous research methodologies proposed, relatively few researchers have attempted to take on this challenge. The inherent complexity of the action is likened to an insurmountable barrier that stands before researchers.

Gold Standard. Owing to the action's inherent complexity, this has become the main challenge hindering model generalization. Therefore, some have proposed a framework for reference comparison [31, 84, 86, 124], hoping by setting up simple comparison labels and abundant training data, the model can learn features autonomously from the latent space. Although this comparison methodology has achieved quite impressive results. However, regarding the selection of reference samples, researchers tend to choose samples with better performance within the dataset. This brings up the question: what samples can become gold standard actions?

The concept of the gold standard is widely used in medical diagnosis, referring to the most reliable diagnosis or assessment currently available and a benchmark for measuring other new methodologies. In the field of action quality assessment, what is the gold standard action? Is it the sample with the best performance in the dataset, the human with the best performance, or the sample created using generative models to be perfect? Much confusion exists in this matter, but one thing is worth mentioning: there is currently no sample that can be used as a gold standard. Therefore, a promising research direction in the future is to construct gold-standard action samples, which will become the benchmark for future research on actions, and all related research will align with this ground truth. Although this is not within the scope of computer science research, this will have a significant impact on AQA.

5.2 Dataset

Challenges. As illustrated in Tab. 1, there are presently 26 publicly available datasets in AQA encompassing 9 distinct domains. We can easily identify significant progress in existing datasets compared to pioneering datasets [2, 4]. However, challenges remain in terms of data scale, action diversity, data modality, and annotations compared to datasets in neighboring research fields [66, 67].

1. **Scant Scale.** Up to now, the biggest AQA dataset has more than 20000 samples [8]. Different from the dataset for object detection, which was annotated by searching for suitable photos in a gallery and boxing them. Real-world AQA datasets involve collecting/selecting videos in a specific action domain and labeling them by domain experts. The collection/selection and annotation consume a huge amount of time and financial costs. As a result, the construction of a big-scale, real-world AQA dataset is still a challenge that remains to be addressed. While scant data astrict the possible methodologies researchers are capable of using and the performance.

2. **Narrow Action.** As previously stated, existing datasets are limited in data sources, while construction is time-consuming and labor-intensive, which results in a narrow action category. To address this challenge, many researchers divide large limb actions into different sub-actions, hoping to enhance the inter-category differences [38]. However, this approach does not effectively expand the action categories. Additionally, due to limited data sources, the variation in action categories included in different datasets is minimal.

3. **Coarse Annotation.** The earliest datasets were limited by the data sources, and to save time on data annotation, the final scores of the athletes were usually used as labels directly. Obviously, a single score cannot allow the model to learn and utilize features sufficiently. Thus, researchers proposed to use multiple judges' scores, event commentary, etc., to improve the annotations, but a more refined annotation could not be achieved. The limitation of fine-grained labels has indeed restricted the advancement of AQA models in interpretability.

Future directions. Obviously, the challenges on the dataset front remain in terms of data scale, action diversity, data modality, and annotations. Therefore, future directions will mainly revolve around these. However, the implementation details differ from other surveys' [216, 217], which tend to employ AQA to evaluate the quality of generated actions and utilize generative models for data argumentation.

1. **Large-scale multi-action real-world dataset.** Despite the fact that we know the construction of datasets is a time-consuming and laborious process, with huge challenges. However, a large-scale multi-action real-world dataset is of great significance to AQA, which can not only promote the model to extract the commonalities between different actions to achieve higher generalization but also promote the model's ability to comprehensively assess within a specific action domain comrehensively.

2. **Large-scale multi-action high-quality AIGC dataset.** Derived from the development of the recent T2V model, there is now a new paradigm for obtaining videos beyond collection. This has also inspired researchers to consider whether the T2V model can be used to generate videos in specific action domains instead of the currently used data set construction methodology to reduce consumption. GAIA [13], as pioneering research, introduced the first non-real-world dataset. However, we can also see (Fig. 8(h)) that there are many deficiencies, such as incomplete human body generation, incoherent actions, and the absence of physical laws. However, the introduction of physical laws in Genesis [218] may indeed achieve more realistic results. Meanwhile, video prompts can also be used as a fine-grained description of the video, which further reduces the difficulty and time-consuming of annotation. Therefore, we believe that one of the future directions of the dataset is to construct a large-scale, multi-action, high-quality AIGC dataset.

5.3 Methodology

Challenges. Combine the methodologies described in section 4 with performance comparisons. A significant methodological improvement in AQA over the past decade can be found, especially in the extraction of detailed features such as spatio-temporal modeling and skeletal relationships, where various novel network structures have been applied. However, this also brings many problems, especially in feedback, interpretability, and multimodal robustness, which remain unsolved challenges.

1. **Computational Complexity & Latency.** With the continuous advancement of network structures, models can extract features with more detail and achieve higher performance. However, this also requires powerful computing power to support model training, and even during the inference stage, it requires considerable com-

puting resources. At the same time, existing datasets and methodologies are aimed at the quality assessment of the complete action process. Therefore, due to various factors, fewer researchers have proposed real-time and lightweight computing models, for example, [26, 107].
2. **Blackbox models.** Interpretability of deep learning has always been a hot topic, especially in the field of AQA. Given that the purpose of AQA since its creation has been to eliminate personal bias and achieve objective action evaluation, as well as the fact that the main application scenario of competitive sports requires fairness, interpretability is one of the critical concerns of AQA models. However, due to the scant data scale and coarse annotation, interpretability has always been a difficult challenge to solve.
3. **Fragile Robustness** Many methodologies in AQA combine multimodal data to realize more accurate quality assessment. Researchers tend to focus on how to efficiently combine data from different modalities while ignoring the performance research of some modalities that are absent due to actual environmental conditions in practical applications. Therefore, ensuring the accuracy of multimodal models when some modalities are absent is also a challenge.

Future directions. Despite the considerable challenges associated with existing models in terms of real-time feedback, interpretability, and robustness. However, adversity brings opportunities, which also indicate future directions of methodologies.
1. **Lightweight, Efficient, Real-time.** Only a handful of AQA research has focused on developing lightweight, efficient, and real-time solutions [26, 107]. One of the biggest areas in which the AQA can make a difference is improving the lives of socio-economically disadvantaged communities through providing automated AQA/physiotherapy/training solutions. Such communities cannot be expected to have high-end computational resources to run heavy AQA apps. Thus, we believe one of the future directions should focus on developing real-time AQA solutions, which are also efficient in terms of computing power and communication bandwidth requirements.
2. **Interpretability.** Interpretability is currently the main challenge at the methodological level. Researchers have tried various methods to improve the interpretability of models, most of which attempt to use visualization techniques to make the decision-making process public rather than optimizing the network structure. Given the current rise of neural symbolic computing [219], researchers have attempted to improve interpretability using symbols that humans can understand, with reasonable achievements [26]. In addition, it is believed that the study of causal [220] in medical imaging [221] can be used to construct a causal relationship between features and quality, which can make the decision-making process transparent and credible. Thus, one of the future directions is to use neural symbolic computing or causal inference to improve interpretability and make the decision-making process transparent and credible.
3. **Robustness.** Most multimodal methods ignore the performance under modal missing. Some researchers [21, 56, 108] choose to use multi-task learning to mitigate the impact of modal missing, which will lead to partial output results. The future research direction of multimodality should not only consider the integration of cross-modal data but also the cross-modal data augmentation and prediction. Existing modalities can fill in the absent modalities to reduce the impact on the model. At the same time, the data mapping relationship between different modalities can be established.

6 Conclusion

Action Quality Assessment, an emerging and critical field in video understanding, has garnered increasing attention in the last decade, and AQA has made significant progress. To consolidate the research efforts in the area thus far, we present the largest and most comprehensive survey to date, to the best of our knowledge, on AQA. Towards this end, we systematically reviewed over 200 relevant papers that met the criteria of the PRISMA framework, providing a detailed description of the definition, classification of AQA problems, model evaluation metrics, and datasets. We distilled research trends from these papers and organized them into 7 principal trends, with detailed discussions on the research methodologies and performance comparisons within each trend. Furthermore, we analyzed the challenges in the field of AQA and identified areas and research opportunities that future work could tackle.

Acknowledgement

This work is supported by Youth Innovation Promotion Association of the Chinese Academy of Sciences, Grant/Award Number: E1290301.

Data Availability

This article is a survey of existing publications related to action quality assessment. No new datasets or codes were generated during the current study. All data presented or discussed are derived from publicly available sources, which are cited throughout the manuscript.

References

1. Yu Kong and Yun Fu. Human action recognition and prediction: A survey. *International Journal of Computer Vision*, 130(5):1366–1401, 2022. 1
2. Hamed Pirsiavash, Carl Vondrick, and Antonio Torralba. Assessing the quality of actions. In *European Conference on Computer Vision*, pages 556–571. Springer, 2014. 1, 2, 3, 4, 5, 8, 9, 11, 12, 13, 16, 17, 18, 19, 27
3. 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. 1, 2, 3, 4, 8, 10, 11, 12, 13, 16, 17, 18, 19
4. Yixin Gao, S Swaroop Vedula, Carol E Reiley, Narges Ahmidi, Balakrishnan Varadarajan, Henry C Lin, Lingling Tao, Luca Zappella, Benjamın Béjar, and David D Yuh. Jhu-isi gesture and skill assessment working set (jigsaws): A surgical activity dataset for human motion modeling. In *Medical Image Computing and Computer Assisted Intervention Workshop*, volume 3, page 3, 2014. 1, 2, 6, 7, 8, 11, 27
5. Aleksandar Vakanski, Hyung-pil Jun, David Paul, and Russell Baker. A data set of human body movements for physical rehabilitation exercises. *Data*, 3(1):2, 2018. 1, 2, 6, 8, 19
6. 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, 6, 8, 21
7. Ruimin Li, Hong Fu, Yang Zheng, Wai-Lun Lo, Jane J. Yu, Cindy H. P. Sit, Zheru Chi, Zongxi Song, and Desheng Wen. Automated fine motor evaluation for developmental coordination disorder. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 27(5):963–973,2019. 1
8. Paritosh Parmar, Amol Gharat, and Helge Rhodin. Domain knowledge-informed self-supervised representations for workout form assessment. In *European Conference on Computer Vision*, pages 105–123. Springer, 2022. 1, 2, 4, 8, 9, 11, 25, 26, 27
9. Yuan-Ming Li, Wei-Jin Huang, An-Lan Wang, Ling-An Zeng, Jing-Ke Meng, and Wei-Shi Zheng. Egoexo-fitness: Towards egocentric and exocentric full-body action understanding. In *European Conference on Computer Vision*, 2024. 1, 2, 8, 9
10. Yiwen Gu, Mahir Patel, and Margrit Betke. ExeChecker: Where Did I Go Wrong? *arXiv preprint arXiv:2412.10573*, 2024. 1
11. Fadime Sener, Dibyadip Chatterjee, Daniel Shelepov, Kun He, Dipika Singhania, Robert Wang, and Angela Yao. Assembly101: A large-scale multi-view video dataset for understanding procedural activities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21096–21106, 2022. 1, 2, 8, 9
12. Luigi Seminara, Giovanni Maria, and Farinella Antonino Furnari. Differentiable Task Graph Learning: Procedural Activity Representation and Online Mistake Detection from Egocentric Videos. In *arXiv preprint arXiv:2406.01486*, 2024. 1
13. Zijian Chen, Wei Sun, Yuan Tian, Jun Jia, Zicheng Zhang, Jiarui Wang, Ru Huang, Xiongkuo Min, Guangtao Zhai, and Wenjun Zhang. Gaia: Rethinking action quality assessment for ai-generated videos. In *Advances in Neural Information Processing Systems*, 2024. 1, 2, 8, 9, 11, 27
14. Paritosh Parmar and Brendan Tran Morris. Measuring the quality of exercises. In *International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 2241–2244. IEEE, 2016. 2, 4, 12, 13, 16, 17, 18, 19
15. Rogerio Nespolo, George R Nahass, Mahtab Faraji, Darvin Yi, Yannek Isaac Leiderman. Assessing vitreoretinal surgical training experience by leveraging instrument maneuvers and visual attention with deep learning neural networks. In *Investigative Ophthalmology & Visual Science*, 65(7):900, 2024. 2
16. 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. 2, 4, 7, 8, 12, 13, 16, 17, 18, 19
17. 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. 2, 4, 7, 8, 12, 13, 16, 17, 18, 19
18. 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
19. Shunli Wang, Dingkang Yang, Peng Zhai, Qing Yu, Tao Suo, Zhan Sun, Ka Li, and Lihua Zhang. A survey of video-based action quality assessment. In *International Conference on Networking Systems of AI*, pages 1–9. IEEE, 2021. 2
20. Matthew J Page, Joanne E McKenzie, Patrick M Bossuyt, Isabelle Boutron, Tammy C Hoffmann, Cynthia D Mulrow, Larissa Shamseer, Jennifer M Tetzlaff, Elie A Akl, Sue E Brennan, et al. The prisma 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, 372, 2021. 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 304–313, 2019. 3, 8, 10, 11, 17, 19, 20, 28
22. Shunli Wang, Shuaibing Wang, Dingkang Yang, Mingcheng Li, Haopeng Kuang, Xiao Zhao, Liuzhen Su, Peng Zhai, and Lihua Zhang. Cpr-coach: Recognizing composite error actions based on single-class training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18782–18792, 2024. 3, 16
23. Ling-An Zeng and Wei-Shi Zheng. Multimodal action quality assessment. *IEEE Transactions on Image Processing*, 2024. 3, 20, 21
24. Ryoji Ogata, Edgar Simo-Serra, Satoshi Iizuka, and Hiroshi Ishikawa. Temporal distance matrices for squat classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 3, 21
25. Hui-Ying Li, Qing Lei, Hong-Bo Zhang, and Ji-Xiang Du. Skeleton based action quality assessment of figure skating videos. In *International Conference on Information Technology in Medicine and Education*, pages 196–200. IEEE, 2021. 3, 20, 21
26. Lauren Okamoto and Paritosh Parmar. Hierarchical neurosymbolic approach for comprehensive and explain-

able action quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3204–3213, 2024. 3, 4, 17, 19, 23, 24, 25, 28

27. Paritosh Parmar, Jaiden Reddy, and Brendan Morris. Piano skills assessment. In *IEEE International Workshop on Multimedia Signal Processing*, pages 1–5. IEEE, 2021. 3, 4, 8, 9, 11, 19, 20, 21
28. Irwandi Hipiny, Hamimah Ujir, Aidil Azli Alias, Musdi Shanat, and Mohamad Khairi Ishak. Who danced better? ranked tiktok dance video dataset and pairwise action quality assessment method. *International Journal of Advances in Intelligent Informatics*, 9(1):96–107, 2023. 3, 8, 9
29. XB Bruce, Yan Liu, Keith CC Chan, and Chang Wen Chen. Egcn++: A new fusion strategy for ensemble learning in skeleton-based rehabilitation exercise assessment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 4, 20, 21
30. Gedas Bertasius, Hyun Soo Park, Stella X Yu, and Jianbo Shi. Am i a baller? basketball performance assessment from first-person videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2177–2185, 2017. 4, 12, 13, 16, 17, 18, 19
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. 6, 14, 17, 18, 19, 26
32. Paritosh Parmar and Brendan Morris. Action quality assessment across multiple actions. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1468–1476. IEEE, 2019. 8, 10, 11, 21, 22
33. 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. 8, 10, 14, 15
34. Jibin Gao, Wei-Shi Zheng, Jia-Hui Pan, Chengying Gao, Yaowei Wang, Wei Zeng, and Jianhuang Lai. An asymmetric modeling for action assessment. In *European Conference on Computer Vision*, pages 222–238. Springer, 2020. 8, 10, 18, 19
35. Ling-An Zeng, Fa-Ting Hong, Wei-Shi Zheng, Qi-Zhi Yu, Wei Zeng, Yao-Wei Wang, and Jian-Huang Lai. Hybrid dynamic-static context-aware attention network for action assessment in long videos. In *Proceedings of the ACM International Conference on Multimedia*, pages 2526–2534, 2020. 8, 10, 16, 17
36. 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 ACM International Conference on Multimedia*, pages 4902–4910, 2021. 8, 10, 16, 17
37. Xin Chen, Anqi Pang, Wei Yang, Yuexin Ma, Lan Xu, and Jingyi Yu. Sportscap: Monocular 3d human motion capture and fine-grained understanding in challenging sports videos. *International Journal of Computer Vision*, 129:2846–2864, 2021. 8, 10, 16, 17
38. 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. 8, 10, 11, 14, 15, 16, 27
39. 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. 8, 11, 14, 15, 16
40. Yanli Ji, Lingfeng Ye, Huili Huang, Lijing Mao, Yang Zhou, and Lingling Gao. Localization-assisted uncertainty score disentanglement network for action quality assessment. In *Proceedings of the ACM International Conference on Multimedia*, pages 8590–8597, 2023. 8, 11, 14, 15, 16
41. Jibin Gao, Jia-Hui Pan, Shao-Jie Zhang, and Wei-Shi Zheng. Automatic modelling for interactive action assessment. *International Journal of Computer Vision*, 131(3):659–679, 2023. 8, 11, 18, 19
42. Fernando De la Torre Frade, Jessica K. Hodgins, Adam W. Bargteil, Xavier Martin Artal, Justin C. Macey, Alexandre Collado I Castells, and Josep Beltran. Guide to the carnegie mellon university multimodal activity (cmu-mmac) database. Technical Report CMU-RI-TR-08-22, Pittsburgh, PA, April 2008. 8
43. Jinglin Xu, Yongming Rao, Jie Zhou, and Jiwen Lu. Procedure-aware action quality assessment: Datasets and performance evaluation. *International Journal of Computer Vision*, pages 1–22, 2024. 11, 14, 15, 16
44. Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, 2013. 11
45. Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 11
46. 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. 11
47. Faegheh Sardari, Adeline Paiement, Sion Hannuna, and Majid Mirmehdi. Vi-net—view-invariant quality of human movement assessment. *Sensors*, 20(18):5258, 2020. 11
48. Yansong Tang, Jiwen Lu, and Jie Zhou. Comprehensive instructional video analysis: The coin dataset and performance evaluation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(9):3138–3153, 2020. 11
49. Tevin Moodley and Dustin van der Haar. Casa: Cricket action similarity assessment in video footage using deep metric learning. In *Southern African Conference for Artificial Intelligence Research*, pages 139–153. Springer, 2022. 11
50. Paritosh Parmar and Brendan Morris. Win-fail action recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 161–171, 2022. 11
51. Qing-Jun Xing, Yuan-Yuan Shen, Run Cao, Shou-Xin Zong, Shu-Xiang Zhao, and Yan-Fei Shen. Functional movement screen dataset collected with two azure kinect depth sensors. *Scientific Data*, 9(1):104, 2022. 11
52. Sheng-Lan Liu, Yu-Ning Ding, Si-Fan Zhang, Wen-Yue Chen, Ning Zhou, Hao Liu, and Gui-Hong Lao. Fine-grained action analysis: A multi-modality and multi-task dataset of figure skating. *arXiv preprint arXiv:2307.02730*, 2023. 11

53. Yansong Tang, Jinpeng Liu, Aoyang Liu, Bin Yang, Wenxun Dai, Yongming Rao, Jiwen Lu, Jie Zhou, and Xiu Li. Flag3d: A 3d fitness activity dataset with language instruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22106–22117, 2023. 11
54. Ming Fang, Xinning Du, Qi Liu, Yunpeng Zhou, Qiwen Liang, and Shuhua Liu. Which is the better teacher action? a new ranking model and dataset. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 7695–7699. IEEE, 2024. 11
55. Ziliang Gan, Lei Jin, Yi Cheng, Yu Cheng, Yinglei Teng, Zun Li, Yawen Li, Wenhan Yang, Zheng Zhu, and Junliang Xing. Skatingverse: A large-scale benchmark for comprehensive evaluation on human action understanding. *IET Computer Vision*, 2024. 11
56. Jianwei Li, Jun Xue, Rui Cao, Xiaoxia Du, Siyu Mo, Kehao Ran, and Zeyan Zhang. Finerehab: A multi-modality and multi-task dataset for rehabilitation analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3184–3193, 2024. 11, 28
57. Takasuke Nagai, Shoichiro Takeda, Satoshi Suzuki, and Hitoshi Seshimo. Mmw-aqa: Multimodal in-the-wild dataset for action quality assessment. *IEEE Access*, 2024. 11
58. Vinay Venkataraman, Ioannis Vlachos, and Pavan K Turaga. Dynamical regularity for action analysis. In *BMVC*, volume 67, pages 1–12, 2015. 11, 12, 13, 16, 17, 18, 19
59. Andrew S Gordon. Automated video assessment of human performance. In *Proceedings of AI-ED*, volume 2, pages 541–546, 1995. 11
60. Marko Jug, Janez Perš, Branko Dežman, and Stanislav Kovačič. Trajectory based assessment of coordinated human activity. In *International Conference on Computer Vision Systems*, pages 534–543. Springer, 2003. 11
61. Kamil Wnuk and Stefano Soatto. Analyzing diving: A dataset for judging action quality. In *Asian conference on computer vision*, pages 266–276. Springer, 2010. 11
62. Yao-Jen Chang, Shu-Fang Chen, and Jun-Da Huang. A kinect-based system for physical rehabilitation: A pilot study for young adults with motor disabilities. *Research in Developmental Disabilities*, 32(6):2566–2570, 2011. 11
63. Lingling Tao, Ehsan Elhamifar, Sanjeev Khudanpur, Gregory D Hager, and René Vidal. Sparse hidden markov models for surgical gesture classification and skill evaluation. In *Information Processing in Computer-Assisted Interventions*, pages 167–177. Springer, 2012. 11
64. Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4489–4497, 2015. 12, 13, 18, 19
65. Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, November 1997. 12
66. 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. 13, 15, 16, 17, 18, 21, 23, 27
67. Joao Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman. A short note about kinetics-600. *arXiv preprint arXiv:1808.01340*, 2018. 13, 27
68. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 13, 16, 17, 18
69. Zhaofan Qiu, Ting Yao, and Tao Mei. Learning spatio-temporal representation with pseudo-3d residual networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5533–5541, 2017. 13, 14
70. Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3202–3211, 2022. 13, 14
71. Hazel R Doughty. *Skill determination from long videos.* Thesis, 2021. 13
72. Xiang Xiang, Ye Tian, Austin Reiter, Gregory D Hager, and Trac D Tran. S3d: Stacking segmental p3d for action quality assessment. In *IEEE International Conference on Image Processing*, pages 928–932. IEEE, 2018. 13, 14
73. Li-Jia Dong, Hong-Bo Zhang, Qinghongya Shi, Qing Lei, Ji-Xiang Du, and Shangce Gao. Learning and fusing multiple hidden substages for action quality assessment. *Knowledge-Based Systems*, 229:107388, 2021. 13, 14
74. Qing Lei, Hongbo Zhang, and Jixiang Du. Temporal attention learning for action quality assessment in sports video. *Signal, Image and Video Processing*, 15(7):1575–1583, 2021. 14, 15
75. Angchi Xu, Ling-An Zeng, and Wei-Shi Zheng. Likert scoring with grade decoupling for long-term action assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3232–3241, 2022. 14
76. Abhay Iyer, Mohammad Alali, Hemanth Bodala, and Sunit Vaidya. Action quality assessment using transformers. *arXiv preprint arXiv:2207.12318*, 2022. 14
77. Yang Bai, Desen Zhou, Songyang Zhang, Jian Wang, Errui Ding, Yu Guan, Yang Long, and Jingdong Wang. Action quality assessment with temporal parsing transformer. In *European Conference on Computer Vision*, pages 422–438. Springer, 2022. 14
78. Yu Zhang, Wei Xiong, and Siya Mi. Learning time-aware features for action quality assessment. *Pattern Recognition Letters*, 158:104–110, 2022. 14, 15
79. Kumie Gedamu, Yanli Ji, Yang Yang, Jie Shao, and Heng Tao Shen. Fine-grained spatio-temporal parsing network for action quality assessment. *IEEE Transactions on Image Processing*, 32:6386–6400, 2023. 14
80. Pu-Xiang Lian and Zhi-Gang Shao. Improving action quality assessment with across-staged temporal reasoning on imbalanced data. *Applied Intelligence*, 53(24):30443–30454, 2023. 14, 15
81. Yanchao Liu, Xina Cheng, and Takeshi Ikenaga. A figure skating jumping dataset for replay-guided action quality assessment. In *Proceedings of the ACM International Conference on Multimedia*, pages 2437–2445, 2023. 14, 15
82. Hong-Bo Zhang, Li-Jia Dong, Qing Lei, Li-Jie Yang, and Ji-Xiang Du. Label-reconstruction-based pseudo-subscore learning for action quality assessment in sporting events. *Applied Intelligence*, 53(9):10053–10067, 2023. 13, 14
83. Kanglei Zhou, Yue Ma, Hubert PH Shum, and Xiaohui Liang. Hierarchical graph convolutional networks for action quality assessment. *IEEE Transactions on Circuits

and Systems for Video Technology, 33(12):7749–7763, 2023. 14, 15

84. Qi An, Mengshi Qi, and Huadong Ma. Multi-stage contrastive regression for action quality assessment. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4110–4114. IEEE, 2024. 14, 15, 26
85. Feng Huang and Jianjun Li. Assessing action quality with semantic-sequence performance regression and densely distributed sample weighting. *Applied Intelligence*, 54(4):3245–3259, 2024. 14, 15
86. Xiao Ke, Huangbiao Xu, Xiaofeng Lin, and Wenzhong Guo. Two-path target-aware contrastive regression for action quality assessment. *Information Sciences*, 664:120347, 2024. 14, 15, 26
87. 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. 14, 15
88. Zhuang Luo, Yang Xiao, Feng Yang, Joey Tianyi Zhou, and Zhiwen Fang. Rhythmer: Ranking-based skill assessment with rhythm-aware transformer. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024. 14, 15
89. Colin Lea, Michael D Flynn, Rene Vidal, Austin Reiter, and Gregory D Hager. Temporal convolutional networks for action segmentation and detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 156–165, 2017. 13
90. A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 14
91. Hang Fang, Wengang Zhou, and Houqiang Li. End-to-end action quality assessment with action parsing transformer. In *IEEE International Conference on Visual Communications and Image Processing*, pages 1–5. IEEE, 2023. 14
92. Yang Bai. *Towards Interaction-level Video Action Understanding*. Thesis, 2023. 14
93. Chaoyu Han, Fangyao Shen, Lina Chen, Xiaoyi Lian, Hongjie Gou, and Hong Gao. Mla-lstm: A local and global location attention lstm learning model for scoring figure skating. *Systems*, 11(1):21, 2023. 15
94. Xiuchun Lin, Yichao Liu, Chen Feng, Zhide Chen, Xu Yang, and Hui Cui. Automatic evaluation method for functional movement screening based on multi-scale lightweight 3d convolution and an encoder–decoder. *Electronics*, 13(10):1813, 2024. 15
95. Ning Hao, Sihan Ruan, Yiheng Song, Jiashun Chen, and Longgang Tian. The establishment of a precise intelligent evaluation system for sports events: Diving. *Heliyon*, 9(11), 2023. 15
96. Youssef Mourchid and Rim Slama. D-stgcnt: A dense spatio-temporal graph conv-gru network based on transformer for assessment of patient physical rehabilitation. *Computers in Biology and Medicine*, 165:107420, 2023. 15
97. Lina Chen, Junbo Zhang, Weijie Wu, Chaoyu Han, and Hong Gao. Long video scoring method fusing high-precision pose and spatio-temporal attention modules. In *Asia-Pacific Web and Web-Age Information Management Joint International Conference on Web and Big Data*, pages 466–475. Springer, 2024. 15
98. Wei Wang, Hongyu Wang, Yingguang Hao, and Qiong Wang. Action quality assessment with multi-scale temporal attention mechanism. In *International Conference on Advanced Algorithms and Control Engineering*, pages 247–251. IEEE, 2024. 15
99. Mégane Millan and Catherine Achard. Fine-tuning siamese networks to assess sport gestures quality. In *International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, pages 57–65, 2020. 15
100. Jicheng Li, Anjana Bhat, and Roghayeh Barmaki. Improving the movement synchrony estimation with action quality assessment in children play therapy. In *Proceedings of the International Conference on Multimodal Interaction*, pages 397–406, 2021. 15
101. 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. 15
102. 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. 15
103. Hou-Chin Liu and Chung-Ta King. A low-cost virtual coach for diagnosis and guidance in baseball/softball batting training. In *International Conference on Systems*, 2019. 15
104. Jiahao Wang, Zhengyin Du, Annan Li, and Yunhong Wang. Assessing action quality via attentive spatio-temporal convolutional networks. In *Chinese Conference on Pattern Recognition and Computer Vision*, pages 3–16. Springer, 2020. 15
105. Tianyu Wang, Yijie Wang, and Mian Li. Towards accurate and interpretable surgical skill assessment: A video-based method incorporating recognized surgical gestures and skill levels. In *Medical Image Computing and Computer Assisted Intervention*, pages 668–678. Springer, 2020. 15, 16
106. Dave Epstein and Carl Vondrick. Learning goals from failure. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11194–11204, 2021. 15
107. Paritosh Parmar and Brendan Morris. Hallucinet-ing spatiotemporal representations using a 2d-cnn. *Signals*, 2(3):604–618, 2021. 15, 28
108. Behnoosh Parsa and Ashis G Banerjee. A multi-task learning approach for human activity segmentation and ergonomics risk assessment. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2352–2362, 2021. 15, 28
109. Behnoosh Parsa. *Deep Learning Methods for Video-Based Human Activity Recognition in Industrial Settings*. Thesis, 2020. 15
110. Ali Abedi and Shehroz S Khan. Improving state-of-the-art in detecting student engagement with resnet and tcn hybrid network. In *Conference on Robots and Vision*, pages 151–157. IEEE, 2021. 15
111. Qiaoge Li, Zhenghang Cui, Itaru Kitahara, and Ryusuke Sagawa. Precise gymnastic scoring from tv playback. In *IEEE Global Conference on Consumer Electronics*, pages 412–415. IEEE, 2022. 15
112. Xinyu Wang, Jianwei Li, and Haiqing Hu. Skeleton-based action quality assessment via partially connected lstm with triplet losses. In *Chinese Conference on Pattern Recognition and Computer Vision*, pages 220–232. Springer, 2022. 15
113. Langxi Liu, Pengjun Zhai, Dulei Zheng, and Yu Fang. Multi-stage action quality assessment method. In *Proceedings of the International Conference on Control,*

Robotics and Intelligent System, pages 116–122, 2023. 15
114. Honghao Gao, Si Yu, Muddesar Iqbal, and Mohsen Guizani. Resfnn: Residual structure-based feedforward neural network for action quality assessment in sports consumer electronics. *IEEE Transactions on Consumer Electronics*, 2024. 15
115. Shafkat Farabi, Hasibul Himel, Fakhruddin Gazzali, Md Bakhtiar Hasan, Md Hasanul Kabir, and Moshiur Farazi. Improving action quality assessment using weighted aggregation. In *Iberian Conference on Pattern Recognition and Image Analysis*, pages 576–587. Springer, 2022. 15
116. Chenglong Li, Qiwen Zhu, Tubiao Liu, Jin Tang, and Yu Su. Hand hygiene assessment via joint step segmentation and key action scorer. *arXiv preprint arXiv:2209.12221*, 2022. 15
117. Xinpeng Ding, Xiaowei Xu, and Xiaomeng Li. Sedskill: Surgical events driven method for skill assessment from thoracoscopic surgical videos. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 35–45. Springer, 2023. 15
118. Pramod Murthy, Bertram Taetz, Arpit Lekhra, and Didier Stricker. Divenet: Dive action localization and physical pose parameter extraction for high performance training. *IEEE Access*, 11:37749–37767, 2023. 15
119. Zeying Li, Hongtao Chen, Jing Cai, and Yanbing Xue. Segmentation and quality assessment of continuous fitness movements based on vision. In *International Conference on Intelligent Computing*, pages 96–107. Springer, 2024. 15
120. Yazan Abu Farha and Jurgen Gall. Ms-tcn: Multi-stage temporal convolutional network for action segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3575–3584, 2019. 15
121. 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. 16, 17
122. Takasuke Nagai, Shoichiro Takeda, Masaaki Matsumura, Shinya Shimizu, and Susumu Yamamoto. Action quality assessment with ignoring scene context. In *IEEE International Conference on Image Processing*, pages 1189–1193. IEEE, 2021. 16, 17
123. 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. 16, 17
124. Mingzhe Li, Hong-Bo Zhang, Qing Lei, Zongwen Fan, Jinghua Liu, and Ji-Xiang Du. Pairwise contrastive learning network for action quality assessment. In *European Conference on Computer Vision*, pages 457–473. Springer, 2022. 16, 17, 26
125. Keyi Huang, Yi Tian, Chen Yu, and Yaping Huang. Dual-referenced assistive network for action quality assessment. *Neurocomputing*, 614:128786, 2025. 16, 17
126. Faegheh Sardari, Adeline Paiement, and Majid Mirmehdi. View-invariant pose analysis for human movement assessment from rgb data. In *Image Analysis and Processing*, pages 237–248. Springer, 2019. 16
127. Yuhang Qiu, Jiping Wang, Zhe Jin, Honghui Chen, Mingliang Zhang, and Liquan Guo. Pose-guided matching based on deep learning for assessing quality of action on rehabilitation training. *Biomedical Signal Processing and Control*, 72:103323, 2022. 16
128. Yun Zhu, Shuang Liang, Peng Li, and Xiaojun Wu. Robust human pose quality assessment using optimal subpattern assignment. In *International Conference on Control, Automation and Information Sciences*, pages 419–423. IEEE, 2022. 16
129. Seiji Hirosawa, Takaaki Kato, Takayoshi Yamashita, and Yoshimitsu Aoki. Action quality assessment model using specialists' gaze location and kinematics data—focusing on evaluating figure skating jumps. *Sensors*, 23(22):9282, 2023. 16
130. Seiji Hirosawa, Takayoshi Yamashita, and Yoshimitsu Aoki. Expert's gaze-based prediction model for assessing the quality of figure skating jumps. In *International Symposium on Computer Science in Sport*, pages 42–52. Springer, 2023. 16
131. Seiji Hirosawa. *Computer Vision-based Action Quality Assessment Incorporating Human Expert's Knowledge: Focusing on the Figure Skating Jump*. Thesis, 2024. 16
132. Qing Lei, Hong-Bo Zhang, Ji-Xiang Du, Tsung-Chih Hsiao, and Chih-Cheng Chen. Learning effective skeletal representations on rgb video for fine-grained human action quality assessment. *Electronics*, 9(4):568, 2020. 16
133. Mahdiar Nekoui, Fidel Omar Tito Cruz, and Li Cheng. Falcons: Fast learner-grader for contorted poses in sports. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 900–901, 2020. 16
134. Shun Fan, Yuantai Wei, Jingfei Xia, and Feng Zheng. Hightlight video detection in figure skating. In *Chinese Conference on Pattern Recognition and Computer Vision*, pages 651–664. Springer, 2022. 16
135. Zhitao Zhang, Zhengyou Wang, Shanna Zhuang, and Jiahui Wang. Toward action recognition and assessment using sfagcn and combinative regression model of spatiotemporal features. *Applied Intelligence*, 53(1):757–768, 2023. 16
136. 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. 16
137. Renguang Chen, Guolong Zheng, Xu Yang, Zhide Chen, Jiwu Shu, Wencheng Yang, Kexin Zhu, and Chen Feng. Unlabeled action quality assessment based on multi-dimensional adaptive constrained dynamic time warping. *arXiv preprint arXiv:2410.14161*, 2024. 16
138. Ulises Gallardo, Fernando Caro, Eluney Hernández, Ricardo Espinosa, and Gilberto Ochoa-Ruiz. Gymetricpose: A light-weight angle-based graph adaptation for action quality assessment. In *IEEE International Symposium on Computer-Based Medical Systems*, pages 43–50. IEEE, 2024. 16
139. Hongli Wen and Yang Xu. Learning to score sign language with two-stage method. *arXiv preprint arXiv:2404.10383*, 2024. 16
140. Gang Wu, Tan Li, Yuqi Zhou, Jin Guo, Jingyu Zhu, Nanjiang Chen, Weining Song, Yalan Xing, Xianghui Meng, and Yanwen Lin. Research on worker action recognition and evaluation in intelligent manufacturing training based on industrial metaverse. In *Asia Simulation Conference*, pages 365–380. Springer, 2024. 16
141. Chengju Zhou, Jiayu Zeng, Lina Qiu, Shuxi Wang, Pingzhi Liu, and Jiahui Pan. An attention-based adaptive spatial–temporal graph convolutional network for

long-video ergonomic risk assessment. *Engineering Applications of Artificial Intelligence*, 131:107780, 2024. 16
142. Renato Manuel Lemos Baptista. *Human motion analysis using 3D skeleton representation in the context of real-world applications: from home-based rehabilitation to sensing in the wild*. Thesis, 2021. 16
143. Mahdiar Nekoui. *Intelligent Video-based Quality Assessment of Human Activities*. Thesis, 2022. 16
144. Yaparla Ganesh, Allaparthi Sri Teja, Sai Krishna Munnangi, and Garimella Rama Murthy. A novel framework for fine grained action recognition in soccer. In *Advances in Computational Intelligence*, pages 137–150. Springer, 2019. 16
145. 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. 16
146. Jianwei Li, Qingrui Hu, Tianxiao Guo, Siqi Wang, and Yanfei Shen. What and how well you exercised? an efficient analysis framework for fitness actions. *Journal of Visual Communication and Image Representation*, 80:103304, 2021. 16
147. David Freire-Obregón, Javier Lorenzo-Navarro, Oliverio J Santana, Daniel Hernández-Sosa, and Modesto Castrillón-Santana. Towards cumulative race time regression in sports: I3d convnet transfer learning in ultra-distance running events. In *International Conference on Pattern Recognition*, pages 805–811. IEEE, 2022. 16
148. Yan Wang, JinWei Wang, and Xue Bai. Three-stream fusion networks for student engagement recognition based on timesformer. In *International Conference on Artificial Intelligence and Computer Engineering*, volume 12610, pages 758–764. SPIE, 2023. 16
149. Wenting Zhao, Shigang Wang, Yan Zhao, Jian Wei, and Tianshu Li. Knowledge and data co-driven intelligent assessment of chinese zither fingerings. *Displays*, 78:102442, 2023. 16
150. Dimitrios Anastasiou, Yueming Jin, Danail Stoyanov, and Evangelos Mazomenos. Keep your eye on the best: contrastive regression transformer for skill assessment in robotic surgery. *IEEE Robotics and Automation Letters*, 8(3):1755–1762, 2023. 16
151. David Freire-Obregón, Javier Lorenzo-Navarro, Oliverio J Santana, Daniel Hernández-Sosa, and Modesto Castrillón-Santana. An x3d neural network analysis for runner's performance assessment in a wild sporting environment. In *International Conference on Machine Vision and Applications*, pages 1–5. IEEE, 2023. 16
152. Haoyang Zhou, Teng Hou, and Jitao Li. Prior knowledge-guided hierarchical action quality assessment with 3d convolution and attention mechanism. In *Journal of Physics: Conference Series*, volume 2632, page 012027. IOP Publishing, 2023. 16
153. Satoshi Kondo. Zeal: Surgical skill assessment with zero-shot tool inference using unified foundation model. *arXiv preprint arXiv:2407.02738*, 2024. 16
154. Yuxin Huang, Yiwei Yuan, Xiangyu Zeng, Ling Xie, Yiyu Fu, Guanghui Yue, and Baoquan Zhao. Full-reference motion quality assessment based on efficient monocular parametric 3d human body reconstruction. In *IEEE International Conference on Multimedia and Expo*, pages 1–6. IEEE, 2024. 16, 18
155. John W Emerson, Miki Seltzer, and David Lin. Assessing judging bias: An example from the 2000 olympic games. *The American Statistician*, 63(2):124–131, 2009. 17
156. Charles J Ansorge and John K Scheer. International bias detected in judging gymnastic competition at the 1984 olympic games. *Research Quarterly for Exercise and Sport*, 59(2):103–107, 1988. 17
157. Hillary N Morgan and Kurt W Rotthoff. The harder the task, the higher the score: Findings of a difficulty bias. *Economic Inquiry*, 52(3):1014–1026, 2014. 17
158. 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. 17, 18, 19
159. Caixia Zhou, Yaping Huang, and Haibin Ling. Uncertainty-driven action quality assessment. *arXiv preprint arXiv:2207.14513*, 2022. 17, 18, 19
160. Anandi Karunaratne, Chamin Jayasooriya, Sampath Deegalla, and Rajitha Navarathna. Objectively measure player performance on olympic weightlifting. In *International Conference on Information and Automation for Sustainability*, pages 410–415. IEEE, 2021. 17
161. Ming-Zhe Li, Hong-Bo Zhang, Li-Jia Dong, Qing Lei, and Ji-Xiang Du. Gaussian guided frame sequence encoder network for action quality assessment. *Complex & Intelligent Systems*, 9(2):1963–1974, 2023. 17, 18, 19
162. Abrar Majeedi, Viswanatha Reddy Gajjala, Satya Sai Srinath Namburi GNVV, and Yin Li. Rica2: Rubric-informed, calibrated assessment of actions. *arXiv preprint arXiv:2408.02138*, 2024. 17, 18, 19
163. Boyu Zhang, Jiayuan Chen, Yinfei Xu, Hui Zhang, Xu Yang, and Xin Geng. Auto-encoding score distribution regression for action quality assessment. *Neural Computing and Applications*, 36(2):929–942, 2024. 17, 18, 19
164. 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, 2021. 17, 18, 19
165. WenHao Sun, YanXiang Hu, Bo Zhang, XinRan Chen, CaiXia Hao, and YaRu Gao. A novel blind action quality assessment based on multi-headed gru network and attention mechanism. In *International Conference on Artificial Intelligence, Automation, and High-Performance Computing*, volume 12717, pages 835–843. SPIE, 2023. 18
166. Chung-In Joung, Seunghwan Byun, and Seungjun Baek. Contrastive learning for action assessment using graph convolutional networks with augmented virtual joints. *IEEE Access*, 2023. 18
167. 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. 18
168. Hiteshi Jain. *Assessment of Human Actions in Videos*. Thesis, 2020. 18
169. Zexing Du, Di He, Xue Wang, and Qing Wang. Learning semantics-guided representations for scoring figure skating. *IEEE Transactions on Multimedia*, 2023. 20
170. Qing Lei, Huiying Li, Hongbo Zhang, Jixiang Du, and Shangce Gao. Multi-skeleton structures graph convolutional network for action quality assessment in long videos. *Applied Intelligence*, 53(19):21692–21705, 2023. 20, 21

171. Jingfei Xia, Mingchen Zhuge, Tiantian Geng, Shun Fan, Yuantai Wei, Zhenyu He, and Feng Zheng. Skating-mixer: Long-term sport audio-visual modeling with mlps. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 2901–2909, 2023. 20, 21
172. Kaili Zheng, Ji Wu, Jialin Zhang, and Chenyi Guo. A skeleton-based rehabilitation exercise assessment system with rotation invariance. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:2612–2621, 2023. 20, 21
173. Yuning Ding, Sifan Zhang, Liu Shenglan, Jinrong Zhang, Wenyue Chen, Duan Haifei, Bingcheng Dong, and Tao Sun. 2m-af: A strong multi-modality framework for human action quality assessment with self-supervised representation learning. In *Proceedings of the ACM International Conference on Multimedia*, pages 1564–1572, 2024. 20, 21
174. Kumie Gedamu, Yanli Ji, Yang Yang, Jie Shao, and Heng Tao Shen. Visual-semantic alignment temporal parsing for action quality assessment. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024. 20
175. Tian He, Yang Chen, Ling Wang, and Hong Cheng. An expert-knowledge-based graph convolutional network for skeleton-based physical rehabilitation exercises assessment. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2024. 20, 21
176. Sania Zahan, Ghulam Mubashar Hassan, and Ajmal Mian. Learning sparse temporal video mapping for action quality assessment in floor gymnastics. *IEEE Transactions on Instrumentation and Measurement*, 2024. 20, 21
177. Shiyi Zhang, Sule Bai, Guangyi Chen, Lei Chen, Jiwen Lu, Junle Wang, and Yansong Tang. Narrative action evaluation with prompt-guided multimodal interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18430–18439, 2024. 20
178. Paritosh Parmar. *On Action Quality Assessment*. Thesis, 2019. 19
179. Huangbiao Xu, Xiao Ke, Yuezhou Li, Rui Xu, Huanqi Wu, Xiaofeng Lin, and Wenzhong Guo. Vision-language action knowledge learning for semantic-aware action quality assessment. In *European Conference on Computer Vision*, pages 423–440. Springer, 2024. 20
180. Huiying Li, Qing Lei, Hongbo Zhang, Jixiang Du, and Shangce Gao. Skeleton-based deep pose feature learning for action quality assessment on figure skating videos. *Journal of Visual Communication and Image Representation*, 89:103625, 2022. 21
181. Jianwei Li, Haiqing Hu, Qingjun Xing, Xinyu Wang, Jinyang Li, and Yanfei Shen. Tai chi action quality assessment and visual analysis with a consumer rgb-d camera. In *IEEE International Workshop on Multimedia Signal Processing*, pages 1–6. IEEE, 2022. 21
182. Wen-Bo Cui, Wen-Ai Song, Zi-Tong Pei, Yi Lei, Qing Wang, Yan-Jie Chen, and Ji-Jiang Yang. Study on assessment methods of developmental coordination disorder in children. In *IEEE Annual Computers, Software, and Applications Conference*, pages 1507–1512. IEEE, 2023. 21
183. Chyan Zheng Siow, Wei Hong Chin, and Naoyuki Kubota. Evaluating simple exercises with a fuzzy system based on human skeleton poses. In *IEEE International Conference on Fuzzy Systems*, pages 1–6. IEEE, 2023. 21
184. Long Yao, Qing Lei, Hongbo Zhang, Jixiang Du, and Shangce Gao. A contrastive learning network for performance metric and assessment of physical rehabilitation exercises. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2023. 21
185. Dinghuang Zhang, Dalin Zhou, and Honghai Liu. Action quality assessment for asd behaviour evaluation. In *International Conference on Machine Learning and Cybernetics*, pages 483–488. IEEE, 2023. 21
186. Yuan-Yuan Shen, Qing-Jun Xing, and Yan-Fei Shen. Markerless vision-based functional movement screening movements evaluation with deep neural networks. *Iscience*, 27(1), 2024. 21
187. Siyuan Xu, Peilin Chen, Yue Liu, Meng Wang, Shiqi Wang, and Sam Kwong. Reveal fluidity behind frames: A multi-modality framework for action quality assessment. In *IEEE International Workshop on Multimedia Signal Processing*, pages 1–6. IEEE, 2024. 21
188. Nathan Louis. *Improving Articulated Pose Tracking and Contact Force Estimation for Qualitative Assessment of Human Actions*. Thesis, 2024. 21
189. Dinghuang Zhang. *Hand-Eye Behaviour Analytics for Children with Autism Spectrum Disorder*. Thesis, 2023. 21
190. Sania Zahan. *Human Motion Analysis*. Thesis, 2024. 21
191. Doyoung Kim, Taewan Kim, Inwoong Lee, and Sanghoon Lee. Kinematic diversity and rhythmic alignment in choreographic quality transformers for dance quality assessment. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024. 21
192. Jia-Hui Pan, Jibin Gao, and Wei-Shi Zheng. Adaptive action assessment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):8779–8795, 2022. 22
193. Shao-Jie Zhang, Jia-Hui Pan, Jibin Gao, and Wei-Shi Zheng. Adaptive stage-aware assessment skill transfer for skill determination. *IEEE Transactions on Multimedia*, 2023. 22
194. 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 *International Joint Conference on Artificial Intelligence*, 2024. 22
195. 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. 22, 23, 24
196. Yuan-Ming Li, Ling-An Zeng, Jing-Ke Meng, and Wei-Shi Zheng. Continual action assessment via task-consistent score-discriminative feature distribution modeling. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024. 22, 23, 24
197. 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 *European Conference on Computer Vision*, pages 375–392. Springer, Cham, 2024. 22, 23, 24
198. Amirhossein Dadashzadeh. *Learning Strategies for Parkinson's Disease Severity Assessment*. Thesis, 2024. 23
199. Chanjin Seo, Masato Sabanai, Yuta Goto, Koji Tagami, Hiroyuki Ogata, Kazuyuki Kanosue, and Jun Ohya. Extracting and interpreting unknown factors with classifier for foot strike types in running. In *International Conference on Pattern Recognition*, pages 3217–3224. IEEE, 2021. 23

200. Tianyu Wang, Minhao Jin, and Mian Li. Towards accurate and interpretable surgical skill assessment: a video-based method for skill score prediction and guiding feedback generation. *International Journal of Computer Assisted Radiology and Surgery*, 16(9):1595–1605, 2021. 23
201. Seiji Hirosawa, Michiko Watanabe, and Yoshimitsu Aoki. Determinant analysis and developing evaluation indicators of grade of execution score of double axel jump in figure skating. *Journal of Sports Sciences*, 40(4):470–481, 2022. 23
202. Hitoshi Matsuyama, Nobuo Kawaguchi, and Brian Y Lim. Iris: Interpretable rubric-informed segmentation for action quality assessment. In *Proceedings of the International Conference on Intelligent User Interfaces*, pages 368–378, 2023. 23, 24
203. Xu Dong, Xinran Liu, Wanqing Li, Anthony Adeyemi-Ejeye, and Andrew Gilbert. Interpretable long-term action quality assessment. *arXiv preprint arXiv:2408.11687*, 2024. 23, 24
204. 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. 24, 25
205. Jiahao Wang, Yunhong Wang, Nina Weng, Tianrui Chai, Annan Li, Faxi Zhang, and Sansi Yu. Will you ever become popular? learning to predict virality of dance clips. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 18(2):1–24, 2022. 24, 25
206. Hiteshi Jain and Gaurav Harit. An unsupervised sequence-to-sequence autoencoder based human action scoring model. In *IEEE Global Conference on Signal and Information Processing*, pages 1–5. IEEE, 2019. 25
207. Chen Du, Sarah Graham, Colin Depp, and Truong Nguyen. Assessing physical rehabilitation exercises using graph convolutional network with self-supervised regularization. In *Annual International Conference of the IEEE Engineering in Medicine & Biology Society*, pages 281–285. IEEE, 2021. 25
208. Konstantinos Roditakis, Alexandros Makris, and Antonis Argyros. Towards improved and interpretable action quality assessment with self-supervised alignment. In *Proceedings of the PErvasive Technologies Related to Assistive Environments Conference*, pages 507–513, 2021. 25, 26
209. Shao-Jie Zhang, Jia-Hui Pan, Jibin Gao, and Wei-Shi Zheng. Semi-supervised action quality assessment with self-supervised segment feature recovery. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(9):6017–6028, 2022. 25, 26
210. Kyoungoh Lee, Yeseung Park, Jungwoo Huh, Jiwoo Kang, and Sanghoon Lee. Self-updatable database system based on human motion assessment framework. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(10):7160–7176, 2022. 25
211. Yun Zhong, Fan Zhang, and Yiannis Demiris. Contrastive self-supervised learning for automated multimodal dance performance assessment. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1–5. IEEE, 2023. 25
212. Kumie Gedamu, Yanli Ji, Yang Yang, Jie Shao, and Heng Tao Shen. Self-supervised subaction parsing network for semi-supervised action quality assessment. *IEEE Transactions on Image Processing*, 2024. 25, 26
213. Wulian Yun, Mengshi Qi, Fei Peng, and Huadong Ma. Semi-supervised teacher-reference-student architecture for action quality assessment. In *European Conference on Computer Vision*, pages 161–178. Springer, Cham, 2024. 25, 26
214. 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*, volume 38, pages 10270–10278, 2024. 25
215. Linfeng Dong, Wei Wang, Yu Qiao, and Xiao Sun. Lucidaction: A hierarchical and multi-model dataset for comprehensive action quality assessment. In *Advances in Neural Information Processing Systems*, 2024.
216. Jiang Liu, Huasheng Wang, Katarzyna Stawarz, Shiyin Li, Yao Fu, and Hantao Liu. Vision-based human action quality assessment: A systematic review. *Expert Systems with Applications*, page 125642, 2024. 27
217. Kanglei Zhou, Ruizhi Cai, Liyuan Wang, Hubert PH Shum, and Xiaohui Liang. A comprehensive survey of action quality assessment: Method and benchmark. *arXiv preprint arXiv:2412.11149*, 2024. 27
218. Genesis Authors. Genesis: A universal and generative physics engine for robotics and beyond, December 2024. 27
219. Wenguan Wang, Yi Yang, and Fei Wu. Towards data-and knowledge-driven ai: a survey on neuro-symbolic computing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 28
220. Ruocheng Guo, Lu Cheng, Jundong Li, P Richard Hahn, and Huan Liu. A survey of learning causality with data: Problems and methods. *ACM Computing Surveys*, 53(4):1–37, 2020. 28
221. Daniel C Castro, Ian Walker, and Ben Glocker. Causality matters in medical imaging. *Nature Communications*, 11(1):3673, 2020. 28
222. Paritosh Parmar, Eric Peh, and Basura Fernando. Learning to Visually Connect Actions and their Effects. *arXiv preprint arXiv:2401.10805*, 2024. 25, 26
223. Kristen Grauman, et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19383–19400, 2024. 9