A Comprehensive Survey of Action Quality Assessment: Method and Benchmark

Kanglei Zhou[a,b], Ruizhi Cai[b], Liyuan Wang[a], Hubert P. H. Shum[c], Xiaohui Liang[b,d,*]

[a]*Department of Psychological and Cognitive Sciences, Tsinghua University, No. 30 Shuangqing Road, Haidian District, Beijing 100084, China*
[b]*State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, No. 37 Xueyuan Road, Haidian District, Beijing 100191, China*
[c]*Department of Computer Science, Durham University, Stockton Rd, Durham DH1 3LE, United Kingdom*
[d]*Zhongguancun Laboratory, Beijing, China*

Abstract

Action Quality Assessment (AQA) aims to automatically evaluate how well human actions are performed and has been widely applied in sports analysis, skill assessment, and healthcare. However, AQA studies are often developed under heterogeneous datasets and evaluation settings, making systematic comparison across methods difficult. To address these challenges, we present a comprehensive survey of recent advances in AQA. In particular, we propose a modality-driven hierarchical taxonomy that organizes existing methods into video-based, skeleton-based, and multi-modal approaches, and analyze the methodological evolution of representative models. We further establish a unified benchmark for representative video-based AQA methods by integrating diverse datasets and standardized evaluation protocols, enabling consistent comparison in terms of both accuracy and computational efficiency. Finally, we analyze emerging research trends, identify key challenges in current AQA research, and outline future directions ranging from near-term methodological advances to longer-term opportunities enabled by emerging AI paradigms. The project webpage is available at https://ZhouKanglei.github.io/AQA-Survey.

Keywords: Action Quality Assessment, Sports Scoring, Skill Assessment, Exercise Assessment

*Corresponding author.
Email addresses: zhoukanglei@tsinghua.edu.cn (Kanglei Zhou), craaaaazy@buaa.edu.cn (Ruizhi Cai), wly19@tsinghua.org.cn (Liyuan Wang), hubert.shum@durham.ac.uk (Hubert P. H. Shum), liang_xiaohui@buaa.edu.cn (Xiaohui Liang)
Preprint submitted to Pattern Recognition

1. Introduction

Action Quality Assessment (AQA) [1, 2] aims to automatically assess how well an action is performed. Unlike action recognition, which focuses on identifying different action categories, AQA evaluates performance differences within the same action category. Consequently, accurately assessing action quality requires sensitivity to subtle variations in motion dynamics and domain-specific evaluation criteria, making it significantly more challenging than traditional recognition tasks. By providing an objective alternative to subjective judgments, AQA has been widely applied in domains such as sports analysis [3, 4], skill assessment [5, 6], and healthcare [7, 8]. More broadly, AQA can support intelligent perception systems by enabling quantitative evaluation of action performance, facilitating applications such as skill coaching and human–robot collaboration in embodied AI environments [9, 10]. Since AQA spans different domains with varying terminology, this survey uses *AQA* as the primary term, encompassing related concepts such as action scoring, skill assessment, and performance evaluation.

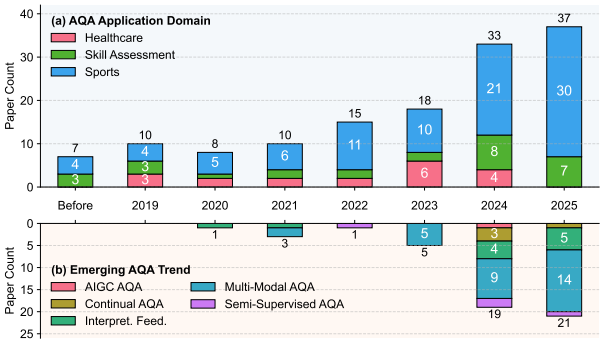

Figure 1: Annual statistics of methodological AQA papers in major CV and ML venues included in this survey. Papers are categorized by (a) application domains and (b) research directions, which have not been systematically summarized in prior AQA surveys.

Recent advancements in deep learning [11, 12] have driven substantial progress in AQA (see Fig. 1(a)), enabling increasingly powerful methodologies and expanding application scenarios (see Fig. 1(b)). Although several surveys [1, 2, 13, 14] have summarized AQA research (see Tab. 1), they remain limited in scope and methodological analysis. Early surveys primarily adopt domain-driven taxonomies [13, 14], which fail to capture shared methodological foundations across tasks, while trend-oriented reviews [2] summarize research developments but do not explicitly unify overlapping techniques across paradigms. In addition, none of the existing surveys provides a unified benchmark for systematic comparison of AQA methods. Furthermore, recent sur-

Table 1: Comparison of existing AQA surveys. , , and indicate sports, medical care, and skill assessment icons. † and ‡ denote two-level taxonomy standards. , , , , , , and represent video, skeleton, text, audio, gaze, and flow icons. and denote the dataset and metric icons.

Survey	Year	Domains	Taxonomy	Modality	Benchmark	Core Contributions
Lei et al. [13]	2019		† Feature type, ‡ Domain		None	An early overview of motion detection and data preprocessing techniques relevant to AQA.
Wang et al. [14]	2021		† Domain, ‡ Feature type		None	A domain-oriented survey emphasizing categorization by application fields and publishing venues.
Liu et al. [1]	2024		† Input mode		None	A descriptive review of representative AQA studies, summarizing methods and application scenarios.
Yin et al. [2]	2025		† Research trends		None	A trend-oriented survey, highlighting research challenges with a high-level discussion of evaluation practices.
Ours	2024		† Input mode, ‡ Mode specialty		6 with 7 Unified setting	A modality-driven hierarchical taxonomy, a unified benchmark, specific AQA tasks, challenges, and prospects.

veys [1, 2] mainly summarize existing studies but provide limited synthesis of emerging research trends and little discussion of future directions. These limitations reveal a fundamental gap in the current literature: existing surveys mainly summarize AQA studies from domain or trend perspectives and lack a unified methodological framework that organizes modeling paradigms and supports consistent evaluation across methods.

As a result, AQA research remains fragmented across datasets, application domains, and modeling paradigms (see Fig. 1), making it difficult to obtain a systematic understanding of the field. Such fragmentation manifests at multiple levels: heterogeneous datasets and evaluation protocols hinder reproducibility; methods are often validated in narrowly defined settings, limiting generalization; and the diversity of application domains obscures shared modeling principles across tasks. This fragmentation highlights the need for a more systematic and unified perspective on AQA research.

To bridge this gap, we present a systematic review of recent advances in AQA. **First**, we argue that the type of input data, rather than the application domain, is the primary determinant of AQA model design. Based on this premise, we propose a modality-driven hierarchical taxonomy. Specifically, we categorize approaches into video-based, skeleton-based, and multi-modal methods, further distinguishing them by modality-specific attributes. **Second**, we establish a unified AQA benchmark, to the best of our knowledge, integrating six widely used datasets and seven evaluation metrics. By standardizing experimental settings, we facilitate consistent comparisons across diverse methods, evaluating both accuracy and the often-overlooked computation overhead. **Third**, beyond summarizing existing work, we analyze emerging trends and outline both near-term and longer-term directions for AQA. Near-term directions focus on improving reliability and interpretability, while longer-term opportunities arise from emerging paradigms such as generative models and embodied AI.

This survey targets researchers developing AQA systems in domains such as sports analytics and healthcare, as well as related areas including human–computer interaction, where objective action evaluation supports decision-making. The remainder of this paper is organized as follows. Sec. 2 describes the review methodology, Sec. 3 introduces the foundational framework, Sec. 4 analyzes representative AQA methods, Sec. 5 reviews task-specific applications, Sec. 6 summarizes representative datasets and presents the benchmark, Sec. 7 discusses emerging trends, remaining challenges, and future research directions, and Sec. 8 concludes the whole paper.

2. Review Methodology

To ensure consistent and fair comparison across diverse AQA application domains, we follow the PRISMA framework (see Fig. 2) to improve transparency, reproducibility, and reliability in selecting representative studies for this survey.

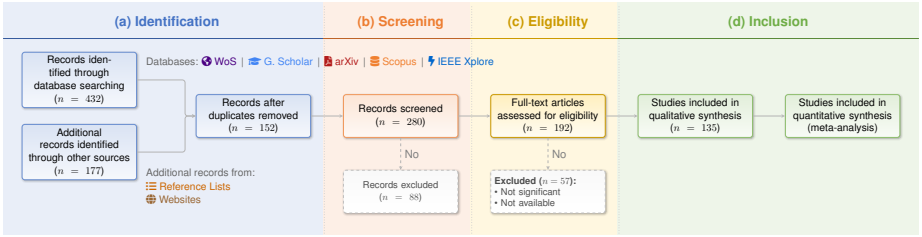

Figure 2: PRISMA flow diagram: (a) identification, (b) screening, (c) eligibility, and (d) inclusion.

Identification Strategy. We searched IEEE Xplore, Web of Science, Scopus, Google Scholar, and arXiv using keywords such as *action quality assessment*, *skill assessment*, and *action scoring*. Publications from 2014 to present were considered, spanning from early handcrafted to recent deep learning-based approaches. We further conducted backward citation analysis on representative papers to improve coverage.

Inclusion and Exclusion Criteria. Studies were included if they: (1) addressed AQA in terms of score, grade, or rank prediction; (2) proposed learning-based methods, benchmark datasets, or evaluation frameworks directly related to AQA; and (3) were peer-reviewed papers published in leading journals or conferences in computer vision and pattern recognition (e.g., TPAMI, IJCV, PR, CVPR, ICCV, ECCV), or widely recognized arXiv preprints with distinct technical contributions. To ensure quality, we

consider four criteria: (1) methodological clarity, (2) experimental rigor, (3) reproducibility, and (4) benchmark relevance. Benchmark relevance and methodological clarity are treated as essential inclusion criteria, and studies failing to meet them are excluded. Experimental rigor and reproducducibility are used for AQA and prioritization when multiple studies address similar problems. We further exclude non-English, tutorial, abstract-only, duplicate studies, and works without explicit quality prediction.

Screening and Selection Process. After deduplication, titles and abstracts were first screened to remove irrelevant records, followed by full-text assessment based on the above criteria. Papers with insufficient technical details or unclear experimental validation were excluded during full-text review. Disagreements were resolved through discussion among the authors. This resulted in 135 methodological AQA papers included in our final survey (see Fig. 2).

Data Extraction and Categorization. From each study, we extracted input modalities, model architectures, supervision types, datasets, and application domains. These attributes support our taxonomy and analysis across AQA methods.

Reproducibility and Resources. To promote transparency, we maintain a public webpage with curated papers and metadata, which will be continuously updated.

3. Fundamentals of Common AQA Setup

Although AQA spans diverse domains, most methods share a common framework. To enable comprehensive analysis, this section formulates a unified framework for multi-modal inputs and introduces typical scenarios and evaluation metrics.

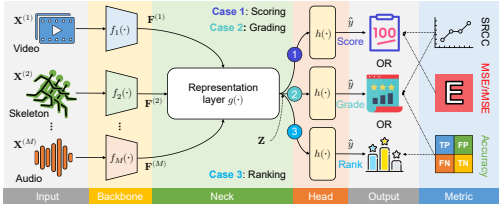

Figure 3: Three-stage AQA framework: (1) a backbone for modality-aware feature extraction; (2) a neck for representation learning that derives compact embeddings; and (3) a regression or classification head that outputs continuous scores (Case 1), discrete grades (Case 2), or ranks (Case 3).

3.1. Problem Formulation

Fig. 3 illustrates the unified framework for AQA, focusing on evaluating the quality of actions through diverse input modalities, including video, skeleton data, and

Table 2: Multidimensional comparison of different AQA scenarios.

Scenarios	Outputs	Precision of Feedback	Complexity of Annotation	Application Suitability
Continuous Scoring	$\hat{y} \in [a, b]$	★★★: Delivers highly exact feedback under certain criteria	★★★: Requires precise annotations often from domain experts	Sports (e.g., gymnastics [15], diving [16], figure skating [17])
Discrete Grading	$\hat{y} \in \{1, 2, \cdots\}$	★★☆: Offers output that is easier to interpret but less granular	★★☆: Demands annotations with moderate granularity	Skill assessment (e.g., surgery [18], physical exercise [7])
Discrete Ranking	$\hat{y} \in \{-1, 0, 1\}$	★☆☆: Provides general feedback with limited detail	★☆☆: Requires minimal effort	Comparative tasks (e.g., cooking [19], basketball [20], teaching [21])

sensor inputs. Let $\mathcal{X} = \left\{\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \cdots, \mathbf{X}^{(M)}\right\}$ represent a set of $M \geq 1$ input modalities, where each $\mathbf{X}^{(m)} = \left[\mathbf{x}_1^{(m)}; \mathbf{x}_2^{(m)}; \cdots; \mathbf{x}_T^{(m)}\right]$ corresponds to a sequence of observations for modality m over time $t \in \{0, 1, \cdots, T\}$. The AQA process in a neural network architecture can be divided into three primary components: feature extraction (backbone), representation learning layer (neck), and score prediction (head). Each input modality $\mathbf{X}^{(m)}$ is processed through a feature extractor $f_m(\cdot)$, generating modality-specific features $\mathbf{F}^{(m)} = f_m\left(\mathbf{X}^{(m)}\right)$. Then the extracted features are passed through the neck $g(\cdot)$, which fuses and transforms the features into an overall representation $\mathbf{Z} = g\left(\mathbf{F}^{(1)}, \mathbf{F}^{(2)}, \cdots, \mathbf{F}^{(M)}\right)$. Finally, the learned representation $\mathbf{Z}$ is passed through a quality prediction head $h(\cdot)$ to estimate the predicted quality score

$$\hat{y} = h\left(g\left(f_1\left(\mathbf{X}^{(1)}\right), f_2\left(\mathbf{X}^{(2)}\right), \cdots, f_M\left(\mathbf{X}^{(M)}\right)\right)\right). \tag{1}$$

Our formulation retains generality and can handle both unimodal and multi-modal AQA methods. For unimodal methods ($M = 1$), the focus is on a single input type, such as video (see Sec. 4.2) or skeletal data (see Sec. 4.3), each of which provides full human-centric cues crucial for AQA. For multi-modal methods ($M > 1$), auxiliary modalities aid in comprehensive analysis (see Sec. 4.4).

3.2. Typical Scenarios

In AQA, the form of the output varies depending on the requirements of different applications. Broadly, AQA methods typically produce continuous or discrete outputs.

3.2.1. Continuous Score. Continuous scores are commonly used in scenarios that require precise and detailed evaluation, such as in Olympic sports AQA [3, 4, 15–17]. The score $\hat{y}$ in Eq. (1) typically ranges within a specific interval, $[a, b]$ (e.g., $[0, 100]$), reflecting the exact quality of the action. The primary objective is to predict a continuous numerical value that accurately quantifies action quality, providing precise feedback essential for competitive evaluation.

3.2.2. Discrete Grade or Rank. Discrete outputs, such as grade and rank, simplify AQA by mapping performances to levels or relative orders, which is helpful when precise scores are unnecessary. **Grading** assigns each performance to a discrete level (e.g., "excellent", "good", "fair", "poor") [22, 23], providing interpretable and coarse evaluation suited for applications like skill assessment. **Ranking** compares performances to establish a relative order, useful for pairwise or group comparisons [19, 21, 24] where absolute scores are less critical. Given paired samples (i, j), the model outputs:

$$\hat{y}_{(i,j)} = \begin{cases} 1, & i \text{ outperforms } j \\ -1, & i \text{ underperforms } j \\ 0, & \text{no preference} \end{cases}. \tag{2}$$

3.2.3. Implications for Modeling. As summarized in Tab. 2, different AQA scenarios impose different supervision and modeling needs. *Continuous scoring* treats AQA as fine-grained regression for high-stakes settings but requires precise and robust labels; *discrete grading* favors efficiency and interpretability when coarse feedback suffices; *ranking* relies on relative comparisons, reducing annotation cost and emphasizing ordinal consistency. These choices reflect trade-offs between precision, cost, and robustness, motivating the joint use of correlation- and precision-based metrics in Sec. 3.3.

3.3. Evaluation Metrics

We mainly discuss metrics for evaluating prediction performance in AQA, while computational efficiency metrics are considered later in the unified benchmark.

3.3.1. Correlation Metrics. The correlation metric in AQA is the Spearman Rank Correlation Coefficient (SRCC). This metric assesses the strength and direction of the relationship between the predicted scores $\hat{y}_i$ $(i \in \{1, 2, \cdots, N\})$ and the true scores y_i based on their ranks $\hat{r}_i$ and r_i. SRCC can be defined as:

$$\text{SRCC} = \frac{\sum_{i=1}^{N} \left(r_i - \bar{r}\right)\left(\hat{r}_i - \bar{\hat{r}}\right)}{\sqrt{\sum_{i=1}^{N} \left(r_i - \bar{r}\right)^2} \sqrt{\sum_{i=1}^{N} \left(\hat{r}_i - \bar{\hat{r}}\right)^2}}, \tag{3}$$

where $\bar{r}$ and $\bar{\hat{r}}$ are the mean ranks of true and predicted scores, respectively. SRCC is particularly useful in scenarios where the precise numerical score is less important than the relative ranking of actions. This makes it a robust metric for evaluating ordinal data, as it emphasizes the consistency of rankings.

3.3.2. Precision Metrics. Two common types of precision metrics assess the accuracy of predicted scores, tailored to either continuous or discrete outputs (see Sec. 3.2).

Score error quantifies the difference between the predicted and actual scores, allowing for an assessment of the model's accuracy in predicting continuous quality scores. Common measures include Mean Absolute Error (MAE) and Mean Squared Error (MSE), which provide insights into the magnitude of prediction errors. Recently, relative Mean Squared Error (rMSE) [25, 26] has gained popularity as it avoids the impact of differing score scales across actions in different categories. rMSE is:

$$\text{rMSE} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{y_i - \hat{y}_i}{y_{\max} - y_{\min}}\right)^2 \times 100, \tag{4}$$

where $y_{\max}$ and $y_{\min}$ denote the maximum and minimum scores, respectively. Note that our definition differs slightly from the relative ℓ_2 in [25]. Normalization scales the score to a decimal, and squaring reduces it further. To address this, we multiply by 100 to rescale it for a precise comparison.

Accuracy measures the proportion of correctly predicted grades or ranked pairs relative to the total predictions made for applications where actions are categorized into discrete grades or ranks, which is:

$$\text{Accuracy} = \frac{\sum_{i=1}^{N} \mathbb{1}\left(\hat{y}_i = y_i\right)}{N} \times 100\%, \tag{5}$$

where $\mathbb{1}(\cdot)$ denotes the indicator function, which equals 1 if $\hat{y}_i = y_i$ and 0 otherwise.

3.3.3. Task-Specific Metrics. Beyond general AQA, task-specific metrics are often required to reflect application objectives. For AQA with temporal segmentation [27–29], Intersection over Union (IoU) evaluates segment-level localization quality, which is crucial for fine-grained feedback but less indicative of overall scoring accuracy. For continual AQA [30, 31], average forgetting measures knowledge retention over time,

highlighting the trade-off between stability and adaptability to new actions.

3.3.4. Metric Trade-offs across Scenarios. Different metrics capture complementary aspects and are suited to different scenarios. For ranking-based AQA, correlation metrics (e.g., SRCC) are preferred as they emphasize relative ordering but may tolerate large absolute errors. For scoring or grading with continuous or discrete outputs, precision metrics (e.g., rMSE, accuracy) are essential to measure numerical or categorical correctness, though they may not fully capture ranking consistency. Task-specific metrics further tailor evaluation to specialized goals but reduce cross-task comparability. Thus, many AQA settings jointly adopt correlation- and precision-based metrics.

4. A Modality-Driven Taxonomy of AQA Methods

A clear taxonomy is essential to address fragmentation in AQA methods. We organize AQA approaches by input modality rather than domains, as it directly determines network design and modeling strategy (see Sec. 4.1). Accordingly, methods are grouped into video-based, skeleton-based, and multi-modal categories (see Secs. 4.2 to 4.4), and we discuss how modality-specific trade-offs between data acquisition cost, representation capability, and robustness guide AQA system design (see Sec. 4.5).

4.1. Modality-Driven Taxonomy: Rationale and Overview

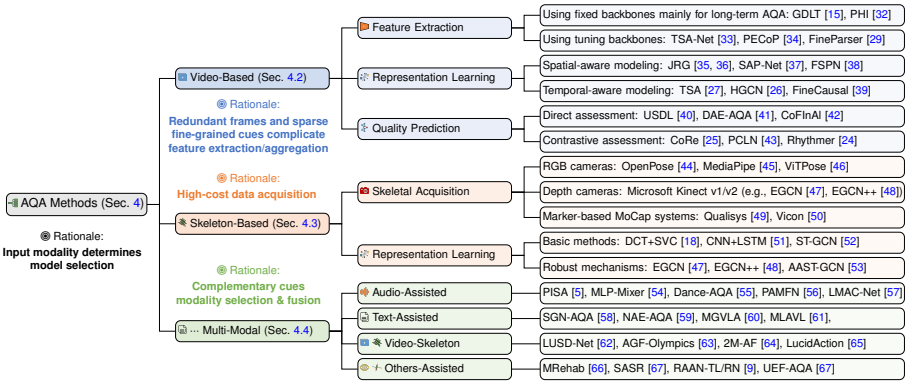

Figure 4: Our hierarchical taxonomy organized by input modalities and key method design rationales.

As illustrated in Fig. 4, we organize AQA methods using a modality-driven taxonomy because the input modality largely determines the available motion cues, data rep-

resentation, and modeling pipeline. At the **first** level, methods are divided into video-based, skeleton-based, and multi-modal branches. Video-based and skeleton-based AQA form two primary paradigms, as they capture motion through visual appearance and structured joint trajectories, respectively. At the **second** level, each branch is further refined according to its core challenges. Video-based AQA focuses on extracting subtle cues from long and redundant visual sequences, skeleton-based AQA emphasizes robust modeling under noisy pose data and costly acquisition, while multi-modal AQA centers on effectively fusing complementary modalities. In this survey, we distinguish paradigm-level advances from incremental improvements. The former introduces new modeling formulations (e.g., modality shifts or supervision paradigms), while the latter mainly refines existing pipelines through architectural modifications.

4.2. Video-Based Approach

Video-based AQA methods exploit rich RGB visual information and the wide availability of video data to capture fine-grained motion and contextual cues, making them the dominant paradigm in gymnastics [68, 69], diving [16, 70], etc. However, assessing action quality from raw videos faces intrinsic challenges, including long and redundant temporal sequences versus limited computational resources [4, 15], sparse and subtle quality cues easily overwhelmed by background variations [71, 72], coarse features transferred from action recognition with domain shift [32, 42], and limited labeled data due to high annotation cost [30, 73]. These challenges arise at different stages of the processing pipeline and motivate distinct design choices. Accordingly, we organize video-based AQA methods by *how they address these issues across three stages*: feature extraction, video-level representation learning, and quality prediction.

4.2.1. Feature Extraction. As illustrated in Fig. 5, early AQA methods [6, 74] relied on handcrafted features, which were adequate for small-scale datasets but limited in capturing rich spatiotemporal cues. With the emergence of deep learning, more expressive representations became possible. However, because AQA datasets remain relatively small due to costly annotations, training deep models from scratch is often impractical for long and highly redundant video data. Consequently, most works adopt backbone networks [75–79] pre-trained on large-scale image or action recogni-

tion datasets [75, 76, 80, 81]. Yet these models are not optimized for the fine-grained requirements of AQA, and the high computational cost of long videos further limits extensive fine-tuning. As a result, recent approaches either adapt the backbone to the AQA domain or refine representations at later stages. In the following, we review commonly used backbones and representative adaptation strategies.

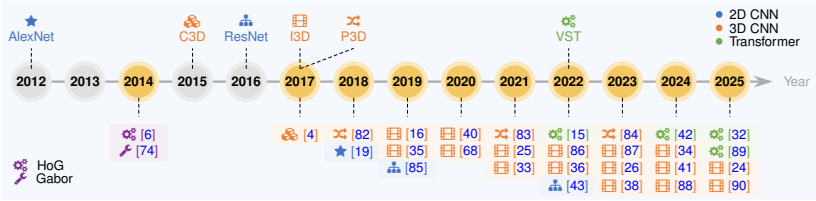

Figure 5: Timeline of backbone architectures (top) and representative video-based AQA papers (bottom).

Representative Backbones in AQA. As summarized in Tab. 3 and Fig. 5, commonly used backbones in video-based AQA can be broadly grouped into three categories: 2D CNNs, 3D CNNs, and transformer-based models, reflecting different inductive biases for modeling action dynamics. Early studies typically employed lightweight 2D CNNs, such as AlexNet [75] (e.g., [19]) and ResNet [91] (e.g., [43, 85]), whose strong spatial inductive bias enables efficient appearance modeling but lacks explicit temporal reasoning. To better capture motion dynamics, subsequent works increasingly adopted 3D CNNs, including C3D [76] (e.g., [4, 70]), P3D [77] (e.g., [82]), and especially I3D [78] (e.g., [26, 29, 35, 86]), which introduce a local spatiotemporal inductive bias through 3D convolutions and have become the dominant choice for short- to medium-length AQA tasks. More recently, transformer-based backbones such as VST [79], adopted in [15, 42, 58, 92], relax these locality assumptions and rely on global attention to model long-range temporal dependencies, making them particularly suitable for long-term AQA, albeit at a substantially higher computational cost.

Efficiency, Transferability, and Practical Choices. The different inductive biases of 2D CNNs, 3D CNNs, and transformers lead to distinct trade-offs in efficiency and temporal modeling. Compared with 2D CNNs, 3D CNNs incur substantially higher computational and memory costs, while transformer-based models further increase this burden due to global attention. To make training tractable, many studies adopt temporal down-sampling strategies, such as key clip selection [23] or uniform clip partitioning,

Table 3: Representative backbone architectures for video-based AQA and their characteristics.

Backbone	Type	Typical Works	Pre-Trained Data	Spatial-Temporal Modeling	Efficiency	Generalization
AlexNet [75]	2D CNN	[19]	ImageNet [75]	★☆☆☆☆: Strong spatial appearance; no explicit temporal modeling.	★★★★★ Lightweight and computationally efficient	★★★☆☆ Limited cross-domain generalization
ResNet [91]	2D CNN	[43, 85]	ImageNet [75]	★★☆☆☆ Strong spatial features; temporal modeling requires extra modules	★★★★☆ Efficient with deeper representations	★★★★☆ Good transferability with fine-tuning
C3D [76]	3D CNN	[4, 70]	Sports1M [76]	★★★★☆ Joint spatiotemporal modeling via 3D convolutions	★★☆☆☆ High computation and memory cost	★★★☆☆ Moderate transferability
P3D [77]	3D CNN	[82–84]	Kinetics-400 [78]	★★★☆☆ Factorized spatial and temporal convolutions	★★★★☆ More efficient than full 3D CNNs	★★★☆☆ Moderate transferability
I3D [78]	3D CNN	[26, 29, 35, 86]	Kinetics-400 [78]	★★★★★ Strong spatiotemporal modeling via inflated 2D filters	★★☆☆☆ Computationally heavy due to 3D operations	★★★★☆ Strong transferability with fine-tuning
VST [79]	Transformer	[15, 42, 58, 92]	Kinetics-600 [81]	★★★★★ Strong long-range temporal modeling with global attention	★☆☆☆☆ Expensive due to self-attention	★★★★☆ Good generalization in complex scenarios

although this may discard subtle yet critical cues and weaken long-range temporal coherence. Meanwhile, domain shift remains a fundamental issue. Backbones pre-trained for action recognition often encode coarse semantics that are insufficient for the fine-grained requirements of AQA [32, 34, 42]. In practice, a clear trade-off emerges: 2D CNNs favor efficiency, 3D CNNs balance spatiotemporal modeling and cost for moderate-length actions, while transformer-based backbones are more suitable for long and complex sequences that require long-range temporal reasoning.

Backbone Tuning Strategies. To mitigate the domain gap between pre-trained and AQA tasks, full fine-tuning is the most straightforward approach. However, it is prone to overfitting on small-scale AQA datasets [31]. Consequently, several works explore constrained backbone adaptation strategies that improve task relevance while limiting parameter updates. For example, TSA-Net [33] injects human-centric masks into I3D to emphasize critical regions with minimal overhead (see Fig. 7(a)), and this idea is further extended by FineParser [29] and Uni-FineParser [90]. In contrast, PECoP [34] introduces lightweight 3D components together with self-supervised objectives to learn in-domain spatiotemporal features, and then freezes the adapted backbone to reduce overfitting risk in downstream AQA tasks. These strategies offer a practical compromise between transferability and efficiency. For *long-term AQA* [15, 42, 58, 92], where minute-level videos are processed, directly fine-tuning the backbone is often impractical due to prohibitive computational cost. Such scenarios are therefore more commonly addressed by feature adaptation at later stages (see Sec. 4.2.2).

4.2.2. Video-Based Representation Learning. Although powerful backbones provide strong clip-level features, feature extraction (see Sec. 4.2.1) in AQA still faces two fundamental limitations: (1) to control computational cost, long videos are usually

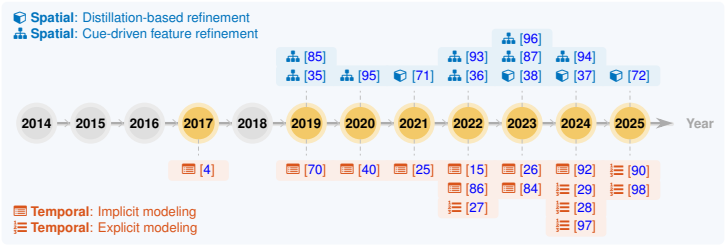

Figure 6: Timeline of representative video-based AQA methods in representation learning, grouped into spatial-aware (above the axis) and temporal-aware (below the axis) strategies over time.

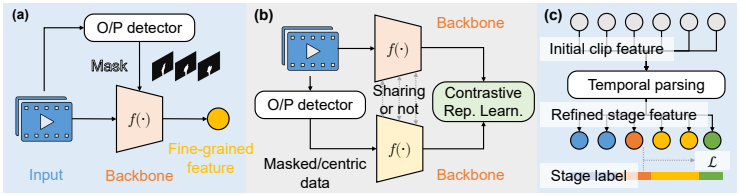

Figure 7: Three representative fine-grained reasoning paradigms in AQA. (a) Spatial reasoning injects object- or pose-centric masks into backbones to emphasize critical regions. (b) Siamese reasoning contrasts raw and masked inputs to highlight object-centric cues, sometimes using unshared branches for richer representations. (c) Temporal reasoning models procedural structure by parsing subactions, either implicitly or explicitly.

processed as short clips, which breaks temporal continuity and weakens long-range dependency modeling; and (2) features transferred from action recognition are often coarse and domain-shifted, and fine-tuning them is impractical for *long-term AQA* with minute-level videos due to prohibitive cost. Video-level representation learning is therefore introduced as an intermediate stage between feature extraction and quality prediction, aiming to refine clip-level features into more quality-aware representations and aggregate them into a coherent video-level embedding that preserves global temporal context. By explicitly modeling finer spatial and temporal cues and integrating information across clips, this stage mitigates both coarse semantics and fragmented temporal modeling. Most existing methods address both aspects to some extent, but differ in their primary emphasis. As summarized in Fig. 6, representative works are reviewed below and categorized into *spatial-aware* approaches [37, 38, 71, 93], which enhance human- and object-centric cues, and *temporal-aware* approaches [70, 86, 99, 100], which focus on procedural structure and long-range dependencies.

Spatial-Aware Modeling. *Spatial-aware* modeling aims to emphasize human- and object-centric regions while suppressing background distractions, since subtle quality differences in AQA are often reflected in localized body parts, tools, or interactions rather than global appearance. Instead of explicitly modifying the backbone with masks

(e.g., Fig. 7(a)), most works implicitly refine spatial representations during representation learning to improve flexibility (see Fig. 7(b)). These works mainly fall into two representative directions.

One direction is **distillation-based refinement**, which guides the model to focus on actor-centric regions through strategies such as human-masked supervision and scene-adversarial objectives [71], teacher–student consistency with pseudo actor-centric labels [37], background swapping augmentation [72], and causality-aware foreground–background modeling [39]. These strategies suppress background bias and transfer fine-grained spatial knowledge to the student network, enabling more discriminative representations for accurate AQA.

Another line of work follows **cue-driven feature refinement**, where spatial representations are enriched by integrating auxiliary cues, such as frame-wise spatial attention to highlight critical regions [85], semantic component modeling in surgical scenes [93], body-part graph reasoning based on pose estimation [35, 36], and interaction relationship [87, 95] or group structure modeling [96] for multi-agent actions. These designs introduce richer semantics into spatial modeling, leading to more interpretable representations of critical regions, while often relying on the quality of auxiliary cues. Cue-driven designs enrich spatial semantics but depend on external cue quality and add complexity, which may hurt robustness. In contrast, distillation-based refinement injects spatial priors implicitly, offering a lightweight and flexible alternative. This reveals a key trade-off between explicit semantic structure and implicit, data-driven refinement, motivating designs that balance expressiveness and efficiency.

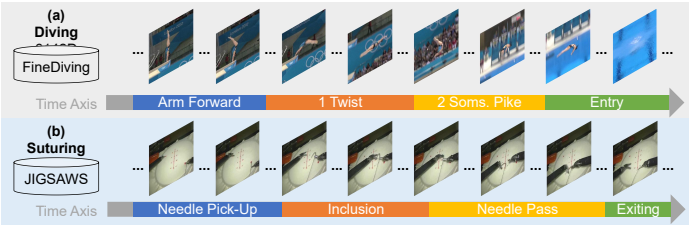

Figure 8: The procedural nature of actions in fine temporal modeling for AQA. (a) illustrates a diving example from the FineDiving dataset [27], while (b) depicts a suturing example from the JIGSAWS dataset [101].

Temporal-Aware Modeling. *Temporal-aware* modeling aims to recover the procedural structure of actions (see Fig. 8) from fragmented clip features and to capture

long-range dependencies that are critical for AQA. In many AQA settings, an action naturally consists of a sequence of subactions (e.g., approach, takeoff, flight, and entry in Fig. 8(a)), and errors in any stage may substantially affect the final score [26, 27, 29, 83]. However, due to efficiency constraints, videos are usually processed as disjoint clips, and naive aggregation (e.g., average pooling [4]) often leads to vague temporal modeling and loss of contextual cues. To address this, existing works mainly fall into two representative directions.

One line of work follows **implicit temporal modeling**, which learns temporal dependencies directly from data using sequence models. Early studies employed LSTMs or TCNs to aggregate clip features [18, 70, 102], while more recent methods adopt Transformers to better capture long-range interactions [15, 86, 99, 100]. Typical designs include grade-aware decoding via cross-attention [15], temporal parsing with learnable queries [86], hierarchical graph reasoning over clips [26, 103], multi-stage temporal parsing for subaction decomposition [83, 104, 105], causality-aware modeling [39, 39, 89], and event pattern analysis [22, 106]. These approaches are flexible and do not require extra annotations, enabling broad applicability across tasks, but often suffer from limited interpretability, higher computational cost, and the risk of shortcut learning or overfitting on small datasets [92, 107].

In contrast, **explicit temporal modeling** injects human knowledge of action procedures through temporal annotations. Representative works such as TSA and FineParser [27–29, 90] leverage stage-level supervision to guide feature aggregation and score prediction, leading to more interpretable and fine-grained assessment of each subaction. Weakly supervised variants further reduce annotation cost by aligning procedures across videos sharing similar structures [97, 98]. While these methods offer stronger interpretability and stage-level feedback, they rely heavily on costly temporal annotations, are less scalable, and often generalize poorly to unseen actions. Overall, temporal-aware modeling reveals a clear trade-off: implicit approaches favor flexibility and scalability but often lack transparency, whereas explicit designs provide structured and interpretable reasoning at the cost of increased annotation demands and limited generalization. These observations suggest that hybrid strategies, which combine data-driven temporal learning with lightweight procedural priors, may offer a promising direction

for improving the robustness and practicality of AQA.

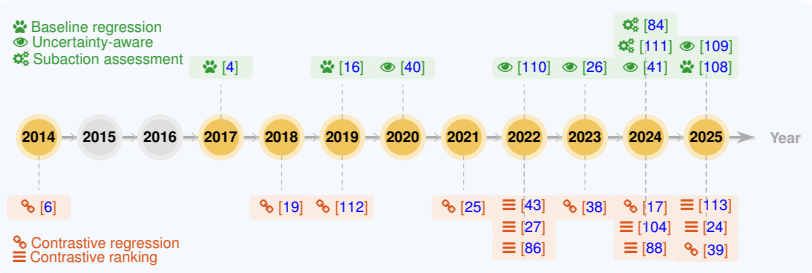

Figure 9: Timeline of representative quality prediction methods in AQA, grouped into direct assessment (top) and contrastive assessment (bottom) paradigms.

4.2.3. Quality Prediction. Quality prediction is the final stage of AQA, which maps video-level representations to assessment outputs and ultimately determines how action quality is quantified. From a modeling perspective, this stage differs mainly in whether quality is inferred from an individual action in isolation or from its relation to other actions. As visualized in Fig. 9, existing methods can therefore be broadly categorized into *direct* and *contrastive* assessments.

Direct Assessment. *Direct* assessment methods [16, 26, 30, 41, 42, 108, 114] formulate AQA as regression or classification, predicting absolute scores or grades by minimizing errors with respect to ground truth. They are simple to implement and computationally efficient, but often show limited sensitivity and generalization when quality differences are subtle or action contexts vary. To alleviate these issues, two representative strategies have been explored. **Uncertainty modeling** [26, 40, 41, 62] explicitly accounts for ambiguity in human judgments by predicting score distributions rather than deterministic values. USDL and MUSDL [40] soften labels with Gaussian distributions for single- and multi-judge settings, while DAE [41] employs VAE-based modeling to regress scores from latent distributions. These approaches improve robustness to noisy annotations, but often rely on assumptions about label variance and may generalize poorly with limited data. **Subaction assessment** [84, 94, 111, 115] decomposes an action into meaningful stages and predicts sub-scores that are later aggregated into a final score. This enables fine-grained diagnosis and improves generalization by leveraging the internal structure of actions.

Contrastive Assessment. *Contrastive assessment* methods [19, 25, 27, 38, 86,

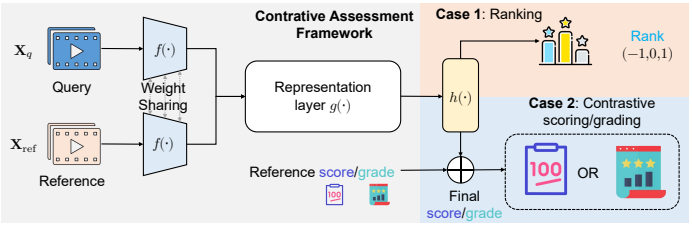

Figure 10: Pairwise contrastive framework for AQA. The model compares a target action with a reference action to learn relative quality cues. Case 1 uses relative ranking (see Eq. (2)) without exact labels, whereas Case 2 performs contrastive scoring or grading with labeled references.

116] learn action quality by modeling *relative differences* between actions rather than relying solely on absolute scores. Typically implemented with Siamese architectures (see Fig. 10), these methods compare a target video against reference examples, making them particularly sensitive to fine-grained performance variations. Originally introduced in ranking-based formulations [6, 19, 112], contrastive learning has since been extended to both ranking and scoring settings. **Ranking-oriented methods** [6, 19–21, 24] avoid absolute annotations and instead predict relative orderings between actions. This formulation substantially reduces annotation cost and is especially attractive when expert scores are expensive, subjective, or unreliable. Such approaches have been successfully applied across domains, including surgical skill assessment [22, 117] and teaching quality evaluation [21, 108]. **Contrastive regression methods** incorporate pairwise contrastive learning into score prediction [25, 27–29, 37, 38, 86, 116]. Representative designs include group-aware contrastive regression [25], hierarchical contrastive parsing of subactions [86], and multi-view or replay-based contrastive frameworks [17, 88]. By explicitly enforcing relative consistency alongside regression objectives, these methods significantly improve sensitivity to subtle quality differences.

Direct and Contrastive Trade-offs. As summarized in Tab. 4, direct and contrastive assessments differ in several ways. Direct methods efficiently regress absolute scores but rely on precise annotations and are less sensitive to subtle differences. In contrast, contrastive methods learn from relative comparisons, improving fine-grained discrimination and reducing annotation burden, albeit at higher computational cost. In practice, direct assessment suits scenarios requiring absolute scoring and real-time inference, while contrastive formulations favor fine-grained differentiation under limited labels. This trade-off has motivated hybrid strategies that use relative comparisons

Table 4: Comparison of direct and contrastive assessment methods.

Assessment	Typical Works	Easy to Implement	Sensitivity to Variations	Required Data	Generalization	Computation
Direct	[40, 41, 70, 94]	★★★★★: Simple	★★★☆☆: Hard to capture details	★★☆☆☆: Less for train	★★★☆☆: Moderate	★★★★★: Efficient
Contrastive	[19, 25, 27, 37, 86, 112]	★★☆☆☆: Complex pairwise design	★★★★★: Robust for subtle differences	★★★★★: More for pairs	★★★★☆: Better for relation focus	★★★☆☆: High for reference compute

Table 5: Statistics of popular skeleton acquisition techniques in AQA.

Data Acquisition	Input	Output	Typical Works
OpenPose [44]	Image	2D Joints	[120, 122]
MediaPipe [45]	Image	2D Joints	[123, 124]
ViTPose [46]	Image	2D Joints	[125]
CPM [126]	Image	2D Joints	[127]
MMPose [128]	Image	3D Joints	[129]
LTHP [130]	Multi-view Images	3D Joints	[131]
MotioAGF. [132]	2D Joints	3D Joints	[125]
PosePrior [133]	2D Joints	3D Joints	[127]
VideoPose3D [134]	Video	3D Joints	[53, 135, 136]
MubyNet [137]	Image	3D Pose & Shape	[138]
Depth Sensors (e.g., Kinect)	Depth	3D Joints	[7, 47–51, 139–143]
Marker-Based (MoCap)	Human Body	3D joints	[49, 50]

during training and absolute score prediction at inference, combining efficient global calibration with fine-grained supervision [108, 118].

4.3. Skeleton-Based Approach

Skeleton-based AQA methods [49, 119–121] assess action quality by modeling human joint trajectories in 2D or 3D space, explicitly focusing on body kinematics rather than visual appearance. Compared with video-based AQA, skeleton-based approaches inherently suppress background variations, making them well-suited for human-centric scenarios with limited object interaction, such as rehabilitation. However, this abstraction introduces several practical challenges [33, 74]. First, reliable pose acquisition is difficult in many real-world settings due to occlusions, viewpoint changes, and limited sensing conditions. Second, skeletal representations are inherently sparse and may miss subtle motion cues required for fine-grained quality assessment. Third, skeleton-only representations lack environmental context, which is often important for evaluating action quality when interactions with objects or scene elements are involved. Since quality prediction largely overlaps with video-based AQA, we focus this section on skeletal acquisition and representation learning, which most clearly reflect the unique challenges of skeleton-based AQA. Given the relatively limited number of skeleton-based studies, this section emphasizes their distinctive modeling challenges rather than providing an exhaustive methodological breakdown.

4.3.1. Skeletal Acquisition Methods. High-quality skeletal data significantly affects model performance. Acquisition methods include RGB cameras, depth sensors, and optical motion capture systems, each with distinct trade-offs (Tab. 5). **RGB camera-based methods** [53, 122, 124, 125, 138] estimate 2D/3D poses from video. They are low-cost and accessible, but face challenges such as occlusion, depth ambiguity, and pose inaccuracies. Improvements like 2D-to-3D lifting [125], multi-view reconstruction [138], and context integration have expanded their use in sports, healthcare, and fitness. **Depth sensors** (e.g., Kinect) [50, 51, 142] directly capture 3D joints with high accuracy, making them suitable for medical applications like physiotherapy [47]. However, they are more costly and sensitive to the environment. Combining RGB and depth (RGB-D) leverages both modalities for enhanced pose estimation, particularly in rehabilitation. **Optical motion capture (MoCap) systems** (e.g., Vicon, Qualisys) [49, 50] provide the highest precision for skeletal tracking, but require complex, expensive setups and are mainly used in controlled labs. RGB cameras are affordable and flexible, but less precise. Depth sensors improve 3D accuracy but have environmental and cost constraints. MoCap offers unmatched accuracy at the expense of flexibility and cost. The choice depends on the balance between accuracy and practicality.

4.3.2. Skeletal Representation Learning Methods. Unlike video-based representation learning, which focuses on pixel-based features from raw images or frames, skeletal representation learning directly works with abstract body joint coordinates and their dynamic interactions over time. The complexity of joint relationships makes skeletal representation learning distinct, as it emphasizes the spatial and temporal dependencies between body parts. This section first reviews typical feature representation methods and then introduces robust mechanisms for enhanced AQA.

Basic Learning Methods.. Skeletal representation learning in AQA has evolved from handcrafted descriptors to deep neural models. Early approaches relied on manually designed features combined with traditional classifiers, such as DCT-based descriptors or self-similarity features with SVMs [122], but these methods struggled to capture complex kinematic patterns and generalized poorly. With the advent of deep learning, CNNs, LSTMs, GCNs, and Transformers have been explored, among

which graph convolutional networks (GCNs) have become dominant due to their natural compatibility with skeletal data [50, 51, 119]. By representing joints as nodes and spatial–temporal connections as edges, GCNs provide a principled way to model structured body-part relations [139, 140]. The introduction of ST-GCN further enabled joint spatial–temporal modeling by integrating temporal convolution [52], laying the foundation for modern skeleton-based AQA. Subsequent works build upon this paradigm with enhanced strategies to improve assessment robustness [47, 48, 141, 142].

Robust Learning Mechanisms.. Beyond backbone architectures, recent advances focus on improving robustness and discriminability under noisy, sparse, and variable-length skeletal data. We summarize four representative mechanisms. *Hierarchical body-part modeling* exploits structured priors by decomposing the skeleton into body parts or motion hierarchies [49, 51, 120, 140]. By modeling joints, limbs, and body parts at different granularities, these methods enhance sensitivity to localized motion quality while preserving global coordination. *Multi-input and ensemble modeling* improves robustness by integrating complementary skeletal cues. EGCN and its extensions [47, 48] combine positional and orientational features through data- and model-level fusion, enabling richer motion representations. Related works further incorporate multi-task learning, jointly optimizing quality assessment with auxiliary objectives such as abnormality detection or reference comparison [50, 139]. *Contrastive and relational learning* introduces relative reasoning into skeletal representation learning. By comparing test actions with reference or standard motions, contrastive objectives enhance discrimination between subtle quality differences [141], complementing absolute regression-based learning. *Self-supervised learning* addresses annotation scarcity and domain variability by encouraging consistent motion representations without explicit labels [142, 144]. These approaches improve generalization by learning intrinsic motion patterns before or alongside supervised quality prediction.

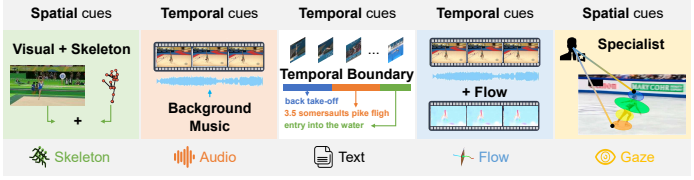

Figure 11: Illustration of typical multi-modal AQA methods.

4.4. Multi-Modal Approach

Multi-modal AQA methods improve robustness and assessment accuracy by integrating complementary information from multiple data sources, alleviating the intrinsic limitations of unimodal approaches. For example, RGB-based methods are sensitive to lighting and viewpoint changes, while pose or sensor signals may be noisy or incomplete. By fusing modalities such as audio, text, skeletons, optical flow, and wearable sensors, multi-modal AQA provides a more comprehensive understanding of action quality (see Fig. 11). Such approaches are particularly effective in domains requiring fine-grained and reliable assessment, including sports analysis [54–56, 58–60, 145] and healthcare [66, 67, 146–148]. We organize existing methods by auxiliary modality and discuss how each modality contributes to AQA.

Audio-Assisted Methods. Audio provides complementary temporal and rhythmic cues that are difficult to infer from visual motion alone, particularly in activities in which action quality is closely tied to timing, pace, or synchronization with sound. Early audio-assisted AQA studies therefore focused on rhythm-dominated domains such as figure skating [54], dance [55], piano performance [5], and long-term assessment [56, 57]. Skating-Mixer [54] jointly models audio and visual streams with memory recurrent units to capture long-term temporal dependencies in skating routines, while contrastive audio–visual learning was introduced to align dance movements with musical structure and beat dynamics [55]. Subsequent works extended audio–visual fusion to richer multi-modal settings. PAMFN [56] progressively aggregates RGB, optical flow, and audio features through modality-specific and shared branches to enhance fine-grained quality estimation. To handle long-term sequences, LMAC-Net [57] employs attention-driven fusion to align RGB, optical flow, and audio, improving temporal coherence over extended actions. More recent methods further incorporate semantic guidance: MLAVL [61] injects language prompts to align audio–visual grading with textual criteria, while MCMoE [149] uses mixture-of-experts to complete missing audio or visual modalities. These developments demonstrate that audio cues consistently enhance rhythm awareness, temporal alignment, and robustness in AQA.

Text-Assisted Methods. Text-assisted AQA incorporates criteria instructions and

language feedback into visual assessment, enabling more context-aware parsing and interpretable evaluation. Recent works increasingly explore vision-language interaction and prompt-based assessment, reflecting a broader trend of integrating language supervision into visual reasoning. One line of work uses textual information to enhance visual representation learning and improve assessment performance. VLAKL [60] embeds domain-specific terminology into multi-level representations to address coarse action segmentation, while SGN [58] adopts a teacher-student framework to transfer semantic knowledge from text to visual features. Another line of work focuses on generating language-based feedback to improve interpretability. NAE-AQA [59] reformulates score regression as video-text matching using prompt-guided transformers, producing textual descriptions rather than numerical scores to improve readability. ExpertAF [150] generates expert-like free-form feedback from RGB video and 3D pose conditioned on text prompts. CROSSTRAINER [151] mines transferable skill attributes from video-language pairs. EFA [152] introduces chain-of-thought reasoning for video-language form assessment, while QGVL [153] leverages quality-aware prompts to guide scoring. These studies highlight the growing role of language in injecting high-level semantics and interpretability into AQA systems.

Video-Skeleton Methods. Video-skeleton methods combine visual appearance with explicit human kinematics, capturing both contextual cues and precise motion dynamics. Such fusion is particularly beneficial when appearance alone is ambiguous or when fine-grained body control is critical. Early two-stream designs integrate RGB and pose features to improve robustness [154]. Subsequent works focus on long-term and complex actions. Zahan et al. [63] introduce the AGF-Olympics dataset and a discriminative attention module for long sports routines. 2M-AF [64] fuses RGB and skeleton streams via preference-based feature aggregation. LucidAction [65] provides multi-view RGB with 2D and 3D poses and adapts multiple AQA architectures. Recent advances further refine pose-guided fusion. MELO [118] introduces scene-aware contrastive fusion for multi-person routines. PGT [155] injects 2D and 3D pose priors into transformers. HP-MCoRe [105] adopts hierarchical pose-guided contrastive regression. DanceFix [156] demonstrates the effectiveness of RGB-pose fusion for skiing and group dance assessment. These methods show that video-skeleton integration

substantially improves fine-grained action understanding.

Other Modality-Assisted Methods. Beyond audio, text, and skeletons, several works incorporate additional sensing modalities to enhance AQA. Wearable and environmental sensors provide complementary kinematic and physiological signals. MMW-AQA [146] combines video, IMU, and GPS for windsurfing assessment. Gaze information is fused with kinematics to analyze figure skating jumps [147]. Optical flow is integrated with RGB using cross-attention for fine-grained evaluation [67]. UJ-AQA-CricketVision [157] trains a dual-stream autoencoder over RGB and keypoints. FLEX [148] further expands modality coverage by providing RGB, 3D pose, sEMG, and physiological signals for fitness benchmarking. These studies illustrate the potential of sensor-rich settings for precise and robust AQA.

4.5. Modality Selection and Trade-offs

As summarized in Tab. 6, the choice of modality in AQA reflects a fundamental trade-off between information richness, annotation cost, robustness, and computational efficiency. Video-based methods provide the most comprehensive visual cues and remain the dominant choice for complex actions, but they are sensitive to background variations and incur high computation and annotation cost. Skeleton-based approaches abstract actions into joint trajectories, offering efficiency and interpretability for human-centric scenarios, yet they are vulnerable to pose estimation errors and often lack contextual cues needed for fine-grained assessment. Multi-modal methods further enhance robustness by integrating complementary signals, but their reliance on synchronized data acquisition, modality alignment, and full-modality availability at inference significantly limits scalability and real-world deployment.

Table 6: Comparison of video-based, skeleton-based, and multi-modal AQA approaches from the perspective of supervision, robustness, efficiency, and application scenarios.

Modality	Information Richness	Annotation Cost	Robustness	Computational Cost	Interpretability	Suitable Scenarios
Video-based (RGB)	High	High	Medium	High	Low	Sports judging, in-the-wild videos
Skeleton-based	Medium	Medium	Medium–Low	Low	High	Rehabilitation, technique analysis
Multi-modal	Very High	Very High	High	Very High	Medium–High	Medical assessment, expert analysis

In practice, modality selection should be guided by application constraints rather than performance alone. Skeleton-based methods are preferable when precise kinematic analysis and efficiency are prioritized, video-based approaches suit scenarios where visual context is essential, and multi-modal designs are most appropriate when

annotation cost and inference complexity are acceptable in exchange for maximum robustness. An emerging direction is to design modality-flexible frameworks that can selectively activate auxiliary signals, balancing performance and practicality.

5. Task-Specific Applications

Despite strong benchmark performance, AQA still faces real-world challenges in label scarcity, continual adaptation, and interpretability. To address these challenges, we identify several underexplored research directions for practical AQA deployment.

Semi-Supervised AQA. *Labeled data scarcity* remains a fundamental bottleneck for AQA, as high-quality annotations often require expert knowledge and are costly to obtain. Semi-supervised AQA methods [37, 73, 158, 159] aim to alleviate this issue by exploiting large amounts of unlabeled data together with limited labeled samples, using techniques such as self-training, consistency regularization, and pseudo-labeling. Early work such as S^4AQA [73] explored masked segment recovery to learn robust representations from unlabeled videos, while TRS-AQA [158] adopted a teacher–student paradigm to generate reliable pseudo labels. These studies indicate that semi-supervised learning is a promising direction for scaling AQA to new domains, though ensuring pseudo-label reliability remains an open challenge.

Continual AQA. *Non-stationary variations* in action execution, recording conditions, and user populations pose significant challenges to the robustness of AQA systems. Continual AQA [30, 114] investigates how models can incrementally adapt to new data distributions without catastrophic forgetting, drawing on advances in continual learning [11]. Representative methods include PECoP [34], which mitigates domain shift during incremental training, methods that explicitly address the introduction of new action categories [114], and MAGR [30, 31], which leverages manifold projection and graph regularization to stabilize learning while preserving privacy. While these methods show encouraging progress, balancing adaptability, memory efficiency, and long-term stability remains an open research problem for practical AQA systems.

Interpretable AQA. *Limited interpretability* is also a major barrier to the real-world adoption of AQA, as numerical scores alone provide little actionable guidance

Table 7: Overview of popular AQA datasets, including modality, domain, number of classes, sample size, average frames, annotations, and access URLs. Datasets span sports, skill assessment, and healthcare.

Dataset	Year	Modality	Domain	# Class	# Samples	# Average Frames	Annotations	URL
MIT Olympic [74]	2014	Video 2D Skeleton	Sports	2	309	Dive: 150 Figure Skate: 4200	Score	
UNLV Olympic [4]	2017	Video	Sports	3	717	Dive: 150 Figure Skate: 4200 Vault: 75	Score	
AQA-7 [70]	2019	Video	Sports	7	1189	Dive: 97 105 156 Vault: 87 Big Air: 122 132 Trampoline: 634	Score	
MTL-AQA [16]	2019	Video	Sports	16	1412	96	Score	
Fis-V [3]	2019	Video	Sports	1	500	4300	TES PCS	
RG [68]	2020	Video	Sports	1	250	2375	Difficulty Execution Total	
FineDiving [27]	2022	Video	Sports	52	3000	105	Step Score	
TaiChi-24 [154]	2022	RGB-D 3D Skeleton	Sports	24	1408	72-695	Score	–
FS1000 [54]	2023	Video Audio	Sports	7	1604	5000	TES PCS Detailed PCS	
FineFS [62]	2023	Video 2D/3D Skeleton	Sports	4	1167	5000	Detailed Score Subaction Class Segmentation	
LOGO [96]	2023	Video	Sports	12	200	5100	Action Class Formation Score	
LucidAction [65]	2024	Video 2D/3D Skeleton	Sports	259	6702	103	Score Penalty	–
FineDiving-HM [29]	2024	Video Mask	Sports	52	3000	105	Mask + Score	
Trampoline-AQA [69]	2025	Video Optical Flow	Sports	1	206	≈902 (30.05 s @30fps)	Score Label Competition Label Meta Label	
UJ-AQA-CricketVision [157]	2025	Video 2D Skeleton	Sports	1	8500	28	Phase Scores	
FineDiving-Pose [105]	2025	Video 2D/3D Skeleton	Sports	1	3,000	96	Pose + Score	
BDJ [109]	2025	Video	Sports	1	920	400	Action Specification Score Execution Level Score Total Score	
FLEX [148]	2025	Video 3D Pose sEMG Physiological	Sports	20	7500	234	Skill Score Knowledge Graph	
CoT-AFA [152]	2025	Video Text	Sports	141	3,392	108	Form + CoT Labels	
JIGSAWS [101]	2014	Video	Skill Assessment	3	103	–	Surgemes Class Rating	
EPIC-Skills [19]	2018	Video	Skill Assessment	7	216	–	Relative Rank	
BEST [112]	2019	Video	Skill Assessment	5	500	6400	Relative Rank	
PISA [5]	2021	Video Audio	Skill Assessment	1	992	160	Skill Level Difficulty	
TAQR [21]	2024	Video	Skill Assessment	4	300	488	Relative Rank	
EgoExo-Fitness [161]	2024	Video (Ego/Exo)	Skill Assessment	12	6,131	600 10-30 s @ 30fps	Score Boundaries Comment	
EgoExoLearn [9]	2024	Video	Skill Assessment	8	3304	250	Relative Rank	
BASKET [117]	2025	Video	Skill Assessment	21	66,000	16,200 (~9 min @ 30fps)	Score	
TAQA [162]	2026	Video Text	Skill Assessment	4	3738	348	Score Text Prompt	
UI-PRMD [163]	2018	3D Skeleton Joint Pos. & Ori.	Healthcare	10	1326	–	Binary Class	
KIMORE [7]	2019	Video, 3D Skeleton Joint Pos. & Ori.	Healthcare	5	1560	–	Score	
EHE [50]	2021	3D Skeleton Joint Pos. & Ori.	Healthcare	6	869	–	Binary Class	
FineRehab [143]	2024	Video, 3D Skeleton Joint Pos. & Ori.	Healthcare	16	4215	–	Score	
GAIA [164]	2024	Video	AIGC	510	9180	70	Subject Completeness Interaction	
Human-AGVQA [165]	2025	Video Text	AIGC	44	6000	32	Quality Ratings	

to users. Interpretable AQA methods aim to augment predictions with understandable feedback. Narrative feedback approaches [59, 94, 138, 150, 160] usually generate descriptive explanations, error diagnoses, or coaching-style comments through vision-language foundation models. Systems such as AIFit [138], NSAQA [94], NAE-AQA [59], ExpertAF [150], and TechCoach [160] illustrate different levels of interpretability, ranging from keypoint-level analysis to free-form textual feedback. Complementary visual feedback mechanisms [8, 138, 150], including motion overlays and side-by-side comparisons, further help users understand performance differences. Despite this progress, generating informative explanations remains an open challenge.

6. Dataset and Benchmark

We review representative datasets and observe that video-based AQA is the most common setting. Thus, we use it as a case study to build a comprehensive benchmark.

6.1. Datasets

As shown in Tab. 7, existing AQA datasets exhibit three evolving trends.

First, there is a growing emphasis on **modality diversity and fine-grained annotation**, moving beyond single-stream RGB videos toward richer supervision signals. Early datasets such as MIT Olympic [74], UNLV Olympic [4], and AQA-7 [70] mainly provide video-level scores, whereas more recent benchmarks introduce multi-modal inputs and structured labels, including step-level annotations in FineDiving [27], detailed subaction scores and segmentation in FineFS [62], and pose-enhanced supervision in FineDiving-Pose [105]. Healthcare-oriented datasets such as FineRehab [143] and KIMORE [7] further incorporate 3D skeletal signals to support precise AQA.

Second, datasets are rapidly expanding in **scale and temporal coverage**, enabling more robust training and evaluation of data-hungry models. While earlier benchmarks are limited to a few hundred samples, large-scale datasets such as FS1000 [54], BASKET [117], and GAIA [164] provide thousands to tens of thousands of videos, often with long temporal durations. Recent long-sequence datasets, including LOGO [96] and Trampoline-AQA [69], explicitly target minute-level actions, posing new challenges for temporal modeling and efficiency.

Third, AQA datasets are expanding into **broader application domains**, reflecting the practical diversification of AQA research. AQA datasets have steadily expanded into more diverse application domains: early extensions focused on skill assessment and healthcare (e.g., EPIC-Skills [19], UI-PRMD [163], EHE [50]), while more recent releases further broaden the scope to teaching assessment (e.g., TAQR [21], TAQA [162]) and to emerging or fine-grained settings such as AI-generated action evaluation (GAIA [164], Human-AGVQA [165]) and traditional activities with specialized criteria (e.g., BDJ [109]).

6.2. A Unified AQA Benchmark

Existing AQA studies are often evaluated on disparate datasets with inconsistent protocols, hindering fair comparisons and obscuring real progress across methods. To mitigate this issue, we construct a unified benchmark centered on *video-based AQA*, which remains the most mature, widely adopted, and reproducible research setting in the literature. Skeleton-based and multi-modal approaches are not included at this stage, primarily due to the limited availability of standardized datasets and open-source implementations that support controlled benchmarking.

6.2.1. Experimental Setting. Our benchmark covers six widely used datasets, seven state-of-the-art baseline methods, and seven evaluation metrics.

Benchmark Datasets. We select datasets according to three criteria: (1) public availability with clearly defined train/test splits, (2) coverage of both short-term and long-term actions, and (3) compatibility with standard video-based pipelines without requiring additional sensors or annotations. Accordingly, we include short-term benchmarks such as MTL-AQA [16], AQA-7 [70], and FineDiving [27], as well as long-term datasets including Fis-V [3], RG [68], and LOGO [96].

Benchmark Methods. We include representative baselines selected based on four criteria: (1) publicly available implementations to ensure reproducibility; (2) competitive performance reported on standard AQA benchmarks; (3) methodological diversity, covering major paradigms in Sec. 4; and (4) compatibility with unified evaluation settings (see Sec. 3.3) for fair comparison. Specifically, USDL [40] and DAE [41] represent uncertainty-aware direct regression, CoRe [25] and T^2CR [88] represent contrastive regression and multi-view reasoning, GDLT [15] and CoFInAl [42] represent transformer-based methods, and HGCN [26] represents structured graph reasoning.

Evaluation Metrics. We report SRCC for ranking, MSE, and rMSE for accuracy, as well as computational metrics such as training time, model size, inference speed, and computational complexity, providing a balanced view of accuracy and efficiency.

Implementation Details. All benchmark results are **reproduced** using publicly available implementations under a unified evaluation protocol. Unless otherwise specified, experiments are conducted on a single RTX 3090 GPU with an I3D backbone

Table 8: Results of our video-based AQA benchmark. The training time unit is in hours. The average SRCC is calculated using Fisher-z transformation to ensure comparability across datasets. The best results are presented in **bold**, while the second-best results are underlined.

Method	Publisher	MTL-AQA [16]				AQA-7 $^{\text{Average}}$ [70]				FineDiving [27]				RG $^{\text{Average}}$ [68]				Fis-V $^{\text{Average}}$ [3]				LOGO [96]			
		SRCC	MSE	rMSE	Time	SRCC	MSE	rMSE	Time	SRCC	MSE	rMSE	Time	SRCC	MSE	rMSE	Time	SRCC	MSE	rMSE	Time	SRCC	MSE	rMSE	Time
MUSDL [40]	CVPR'20	0.9350	39.7753	0.3642	11.32	0.8292	268.1272	3.0767	1.57	0.8812	53.6996	0.4927	14.08	0.4897	12.0564	4.3557	**0.01**	0.3514	478.8537	41.1493	0.07	0.7044	368.2800	3.6828	**0.02**
		0.0052	5.9591	0.0545	0.03	0.0126	10.1333	0.1133	0.04	0.0017	1.4217	0.0130	0.03	0.0437	0.9665	0.3443	0.00	0.0112	0.3705	0.0592	0.00	0.0199	1.2255	0.1291	0.01
CoRe [25]	ICCV'21	0.9519	32.9234	0.3015	46.81	0.7838	256.4000	2.8641	4.44	0.9413	24.4000	0.2440	26.62	0.7348	**6.0166**	**2.2178**	0.12	0.7253	183.9800	1.8398	0.20	0.5852	45.3446	4.7783	0.10
		0.0311	18.6081	0.1704	0.90	0.0287	22.9929	0.2513	0.02	0.0647	25.7334	0.2361	1.72	0.0202	0.4138	0.1466	0.01	0.0082	0.5675	0.0751	0.01	0.0229	2.6094	0.2750	0.01
GDLT [15]	CVPR'22	0.9395	43.5769	0.3990	**10.84**	0.8057	238.9922	**2.7120**	1.73	0.9342	27.9100	0.2791	8.84	**0.7584**	6.2177	2.2767	0.04	0.7203	190.2050	1.9021	0.04	**0.7453**	401.5500	4.0155	0.04
		0.0019	1.3744	0.0126	0.05	0.0199	28.5499	0.3179	0.01	0.0038	1.7906	0.0164	0.08	0.0163	0.4778	0.1721	0.00	0.0182	0.9873	0.1302	0.00	0.0994	7.6091	0.8018	0.00
HGCN [26]	TCSVT'23	0.9522	30.7432	0.2815	14.79	**0.8451**	**235.1635**	2.7642	**1.53**	0.9383	24.7700	0.2477	9.14	0.7121	6.8748	2.5240	0.03	0.7270	227.8600	2.2786	**0.03**	0.7082	369.0600	3.6906	0.04
		0.0023	2.6689	0.0244	0.06	0.0121	8.9000	0.0942	0.04	0.0018	4.1503	0.0381	0.06	0.0392	1.2362	0.4725	0.00	0.0140	2.7601	0.2424	0.00	0.0492	9.2508	0.9748	0.01
DAE [41]	NCAA'24	0.9497	31.3355	0.2869	11.11	0.7916	307.4535	3.4763	1.99	0.9350	26.5100	0.2651	**8.58**	0.7412	7.1299	2.7019	0.03	**0.7447**	**181.1600**	**1.8116**	0.03	0.6701	356.6300	**3.5663**	0.02
		0.0009	2.9995	0.0275	0.02	0.0136	32.3747	0.4067	0.01	0.0009	1.3459	0.0123	0.06	0.0032	0.5730	0.2287	0.00	0.0058	2.1296	0.2655	0.00	0.0196	7.5021	0.7906	0.01
T^2CR [88]	INFS'24	**0.9529**	**29.8640**	**0.2735**	49.62	0.7910	303.2644	3.4754	4.98	**0.9424**	**23.5100**	**0.2351**	30.56	0.6581	7.0013	2.5215	0.28	0.6898	236.1867	2.3619	1.09	0.5199	43.3161	4.5645	0.26
		0.0012	0.1800	0.0020	0.52	0.0141	11.3500	0.1300	0.07	0.0012	0.1612	0.0015	0.03	0.0172	0.4346	0.1602	0.02	0.0131	0.7144	0.0834	0.06	0.0767	3.8466	0.4053	0.04
CoFInAl [42]	IJCAI'24	0.9461	37.7907	0.3461	14.80	0.8195	249.9134	2.7769	1.57	0.9317	36.4681	0.2887	8.64	0.7534	10.8178	4.0537	0.08	0.6974	401.9300	4.0193	0.05	0.5972	**38.6218**	4.0698	0.05
		0.0005	1.0202	0.0094	0.05	0.0127	9.2500	0.1000	0.04	0.0017	0.5107	0.0120	0.81	0.0129	1.0906	0.3848	0.01	0.0151	5.5130	0.5727	0.01	0.0777	1.4044	0.1480	0.00

Method	Publisher	Diving $^{\text{AQA-7}}$ [70]				Gym Vault $^{\text{AQA-7}}$ [70]				BigSki. $^{\text{AQA-7}}$ [70]				BigSnow. $^{\text{AQA-7}}$ [70]				Sync. 3m $^{\text{AQA-7}}$ [70]				Sync. 10m $^{\text{AQA-7}}$ [70]			
		SRCC	MSE	rMSE	Time	SRCC	MSE	rMSE	Time	SRCC	MSE	rMSE	Time	SRCC	MSE	rMSE	Time	SRCC	MSE	rMSE	Time	SRCC	MSE	rMSE	Time
MUSDL [40]	CVPR'20	0.8738	129.5963	1.2960	3.21	0.7300	133.8101	2.1927	1.44	0.5369	729.6780	8.4625	1.44	**0.7109**	**366.3390**	**3.6634**	1.64	0.9205	126.1327	1.5508	**0.85**	0.9416	**124.1512**	**1.3039**	0.86
		0.0120	6.5000	0.0650	0.04	0.0115	5.8000	0.0950	0.04	0.0180	25.0000	0.2800	0.05	0.0160	12.0000	0.1200	0.05	0.0100	6.0000	0.0700	0.02	0.0080	5.5000	0.0500	0.02
CoRe [25]	ICCV'21	0.8432	85.5388	0.8554	11.53	0.7520	**124.9170**	**2.0470**	3.41	0.6545	**365.8256**	**4.2427**	3.47	0.5328	516.1882	5.1619	4.17	0.9061	108.2117	1.3305	2.02	0.8452	337.7187	3.5469	2.03
		0.0145	14.6279	0.1463	0.01	0.0047	6.3346	0.1038	0.03	0.0263	20.1799	0.2340	0.02	0.0689	49.8321	0.4983	0.04	0.0147	17.8071	0.2189	0.02	0.0433	29.1758	0.3064	0.01
GDLT [15]	CVPR'22	0.8425	**82.4375**	**0.8244**	3.77	0.7665	159.1891	2.6086	1.50	0.6540	397.9806	4.6157	1.56	0.5805	526.6038	5.2661	1.79	0.9078	64.1058	0.7882	0.85	0.9049	207.0555	2.1746	0.86
		0.0104	4.2804	0.0428	0.01	0.0137	14.8464	0.2433	0.01	0.0319	52.1782	0.6052	0.01	0.0433	88.8919	0.8889	0.01	0.0095	3.2000	0.0400	0.01	0.0110	8.5000	0.0900	0.02
HGCN [26]	TCSVT'23	**0.8871**	101.4813	1.0148	**3.16**	**0.7725**	220.5605	3.6143	**1.36**	**0.6701**	483.7507	5.6104	**1.40**	0.6487	369.5956	3.6960	**1.59**	0.9174	96.7002	1.1890	0.86	**0.9507**	140.0894	1.4713	**0.85**
		0.0110	3.2000	0.0250	0.05	0.0120	8.9000	0.1200	0.05	0.0150	18.0000	0.1800	0.05	0.0180	13.0000	0.1300	0.05	0.0090	4.5000	0.0500	0.02	0.0075	5.8000	0.0600	0.02
DAE [41]	NCAA'24	0.8468	99.1924	0.9919	3.27	0.7526	158.1046	2.5908	1.43	0.5439	558.6434	6.4789	1.55	0.4824	589.9491	5.8995	3.23	**0.9329**	160.5100	1.9735	1.27	0.8883	278.3217	2.9231	1.19
		0.0038	13.7533	0.1376	0.02	0.0152	45.3578	0.7433	0.01	0.0139	45.1156	0.5232	0.02	0.0341	17.2916	0.1729	0.01	0.0060	55.4437	0.6817	0.01	0.0086	17.2860	0.1815	0.01
T^2CR [88]	INFS'24	0.8334	96.7862	0.9679	9.13	0.7585	187.0455	3.0651	4.93	0.5887	482.7082	5.5983	5.03	0.4790	546.7196	5.4672	5.55	0.9087	242.4116	2.9805	2.58	0.9109	265.3465	2.7868	2.71
		0.0140	4.1000	0.0300	0.10	0.0135	7.5000	0.1400	0.07	0.0160	17.5000	0.1900	0.08	0.0200	19.0000	0.2000	0.07	0.0105	9.5000	0.1100	0.05	0.0105	10.5000	0.1100	0.05
CoFInAl [42]	IJCAI'24	0.8652	150.5001	1.5050	3.17	0.7685	126.4775	2.0726	1.43	0.6119	394.2921	4.5729	1.45	0.5990	539.6546	5.3965	1.64	0.9073	**46.2191**	**0.5683**	0.85	0.9334	243.4674	2.5570	0.87
		0.0130	7.1000	0.0550	0.06	0.0125	5.2000	0.0850	0.04	0.0170	14.0000	0.1600	0.05	0.0150	18.0000	0.1800	0.05	0.0100	2.2000	0.0300	0.02	0.0090	9.0000	0.0900	0.02

Method	Publisher	Ball $^{\text{RG}}$ [68]				Clubs $^{\text{RG}}$ [68]				Hoop $^{\text{RG}}$ [68]				Ribbon $^{\text{RG}}$ [68]				TES $^{\text{Fis-V}}$ [3]				PCS $^{\text{Fis-V}}$ [3]			
		SRCC	MSE	rMSE	Time	SRCC	MSE	rMSE	Time	SRCC	MSE	rMSE	Time	SRCC	MSE	rMSE	Time	SRCC	MSE	rMSE	Time	SRCC	MSE	rMSE	Time
MUSDL [40]	CVPR'20	0.4897	435.5650	4.3557	**0.01**	0.4746	439.7667	4.3977	**0.01**	0.4911	452.2267	4.5223	**0.01**	0.6128	466.8100	4.6681	**0.01**	0.0000	7823.4400	78.2344	**0.02**	0.6506	407.5300	4.0753	**0.02**
		0.0503	1.1513	0.3075	0.00	0.0481	0.5349	0.2509	0.00	0.0230	1.4687	0.5528	0.00	0.0535	0.7110	0.2660	0.00	0.0000	0.0000	0.0000	0.00	0.0225	0.7409	0.1184	0.00
CoRe [25]	ICCV'21	0.7510	160.1400	1.6014	0.06	**0.7592**	232.4600	**2.3246**	0.06	0.7554	211.4200	2.1148	0.12	0.7439	246.2800	2.4628	0.06	0.6578	186.3000	1.8630	0.20	0.7812	181.6600	1.8166	0.20
		0.0323	0.7252	0.1937	0.02	0.0260	0.4628	0.2171	0.00	0.0049	0.2195	0.0827	0.00	0.0174	0.2478	0.0927	0.00	0.0099	0.4128	0.0347	0.01	0.0065	0.7223	0.1154	0.02
GDLT [15]	CVPR'22	**0.7759**	197.7200	1.9772	0.02	0.7365	263.7000	2.6370	0.02	0.7616	179.1000	1.7910	0.04	**0.8007**	202.0100	**2.0201**	0.02	0.6304	237.0700	2.3707	0.04	0.7911	143.3400	**1.4334**	0.04
		0.0255	0.5975	0.1595	0.00	0.0085	0.3780	0.1774	0.00	0.0278	0.6003	0.2259	0.00	0.0034	0.3356	0.1255	0.00	0.0202	0.7249	0.0608	0.00	0.0163	1.2498	0.1996	0.00
HGCN [26]	TCSVT'23	0.7007	197.3500	1.9735	0.02	0.7585	246.6900	2.4669	0.02	**0.7701**	176.4000	**1.7640**	0.03	0.7603	215.9700	2.1597	0.02	0.6830	273.5100	2.7351	0.03	0.7657	182.2100	1.8221	0.03
		0.0135	0.3744	0.1000	0.00	0.0550	0.8109	0.3804	0.00	0.0352	1.3832	0.5206	0.00	0.0532	2.3762	0.8889	0.00	0.0179	5.2356	0.4394	0.00	0.0101	0.2847	0.0454	0.00
DAE [41]	NCAA'24	0.7425	143.1000	**1.4310**	0.02	0.7518	402.1000	4.0210	0.02	0.7431	183.4900	1.8349	0.03	0.7399	388.5100	3.8851	0.02	0.6748	198.8100	1.9881	0.03	0.8014	163.5100	1.6351	0.03
		0.0037	0.4228	0.1129	0.00	0.0040	1.0763	0.5049	0.00	0.0022	0.1391	0.0523	0.00	0.0030	0.6537	0.2445	0.00	0.0058	1.9698	0.1653	0.00	0.0058	2.2894	0.3657	0.00
T^2CR [88]	INFS'24	0.6483	185.0900	1.8509	0.28	0.7335	237.6300	2.3763	0.28	0.6810	284.4400	2.8444	0.28	0.7769	231.5000	2.3150	0.28	**0.6896**	184.9400	**1.8494**	1.14	0.7536	178.8300	1.7883	1.05
		0.0178	0.3949	0.1055	0.00	0.0065	0.3280	0.1539	0.05	0.0133	0.6959	0.2619	0.01	0.0314	0.3198	0.1196	0.02	0.0132	0.8095	0.0679	0.09	0.0130	0.6193	0.0989	0.03
CoFInAl [42]	IJCAI'24	0.7616	**5.9593**	1.5916	0.14	0.7443	**6.4109**	3.0076	0.12	0.7482	**24.4233**	9.1924	0.11	0.7593	**6.4777**	2.4232	0.06	0.6539	**37.2062**	3.1223	0.06	**0.8152**	**17.4047**	2.7803	0.12
		0.0220	1.4496	0.3872	0.00	0.0113	0.6054	0.2840	0.02	0.0144	2.0579	0.7746	0.00	0.0039	0.2496	0.0934	0.00	0.0102	8.1239	0.6818	0.00	0.0199	2.9021	0.4636	0.01

pre-trained on Kinetics-400, trained for 100 epochs using the Adam optimizer (learning rate 1×10^{-4}, weight decay 1×10^{-4}). For datasets without released configurations, we follow official settings from prior works; for RG and Fis-V, official VST features are directly adopted. To ensure fairness and reproducibility, no per-dataset or per-method hyperparameter tuning is performed. Results are averaged over three runs (seeds 1022, 1023, and 1024) on the official splits, following common AQA benchmarking practice. Task-specific tuning may further improve individual methods.

6.2.2. Assessment Performance. Tab. 8 compares seven AQA methods across six benchmark datasets. Results are based on standardized implementations with official code for fair comparison, and the reported standard deviations indicate generally stable performance across runs. Transformer-based methods, such as CoFInAl [42], perform competitively on several long-sequence datasets (e.g., RG, LOGO), while GDLT [15] achieves strong SRCC performance on RG and LOGO. However, USDL/MUSDL [40] shows larger variance on challenging datasets such as Fis-V, suggesting sensitivity to noise and complex temporal dependencies. Contrastive regression methods (CoRe [25], T^2CR [88]) perform competitively on large-scale datasets (MTL-AQA, FineDiving),

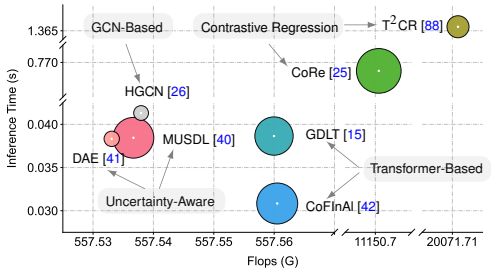

Figure 12: Computation comparison with selected baselines under the identical device settings on the MTL-AQA dataset. The x-axis represents the FLOPs, the y-axis indicates the inference time, and the bubble size corresponds to the number of parameters.

while their performance varies more across smaller datasets. Uncertainty-aware DAE [41] shows strong results on Fis-V, indicating the benefit of modeling prediction uncertainty. GCN-based HGCN [26] also performs competitively on AQA-7, suggesting the effectiveness of graph-based local and global feature modeling for robust performance.

6.2.3. Computational Performance. Tab. 8 and Fig. 12 compare training time, FLOPs, parameters, and inference speed across methods (statistics from MTL-AQA). Contrastive regression methods (CoRe [25], T^2CR [88]) are the most computationally intensive, with training times of 46 and 49 hours, respectively, and the longest inference times (up to 1.4s per sample), mainly due to exemplar computation and ensemble strategies. T^2CR's dual time-scale design further increases computation. In contrast, direct regression and transformer-based methods (e.g., CoFInAl [42]) are more efficient, with CoFInAl showing the fastest inference (0.03s). HGCN [26] and DAE [41] have the lowest parameter counts (around 12.5M), and DAE has the lowest FLOPs (557.53G), whereas T^2CR's complexity is over 30 times higher.

6.2.4. Experimental Observations. As shown in Tab. 8 and Fig. 12, our benchmark provides several key insights into the strengths and limitations of existing AQA methods. First, no single method consistently achieves the best performance across all datasets; methods that perform well on one dataset often fail to maintain the same advantage on others. Second, performance is strongly influenced by dataset characteristics, particularly differences in temporal length and domain structure. Third, higher predictive accuracy is often accompanied by increased computational cost, indicating a clear trade-off between performance and efficiency. These findings highlight fundamental limitations of existing AQA approaches and emphasize the need for models that are both computationally efficient and robust across diverse datasets.

7. Trends, Challenges, and Future Directions

As shown in Fig. 13, AQA is evolving beyond benchmark-driven accuracy toward more robust and adaptive systems for real-world deployment. This section synthesizes key methodological trends, remaining challenges, and future directions in AQA.

7.1. Current Trends

Recent AQA research is shifting from simple score regression toward structure-aware modeling. A prominent trend is the adoption of **fine-grained reasoning** [27, 29, 39] to support model-level interpretability, where internal representations are aligned with meaningful temporal segments or sub-actions. In parallel, increasing attention is paid to **output-level interpretability** [150, 160], moving beyond scalar scores toward diagnostic and actionable feedback. Driven by limited annotations, AQA is increasingly integrated with **semi/self-supervised** learning [37, 158] to better exploit sparse labels, **contrastive** objectives [25, 29] to enhance assessment robustness, and **multi-modal** data [60, 61] to leverage complementary cues. At the architectural level, transformers [15, 86] and **pre-trained models** [78, 79] are becoming common backbones, enabling long-range temporal modeling and improved data efficiency. Together, these trends indicate a transition toward more interpretable and generalizable AQA systems.

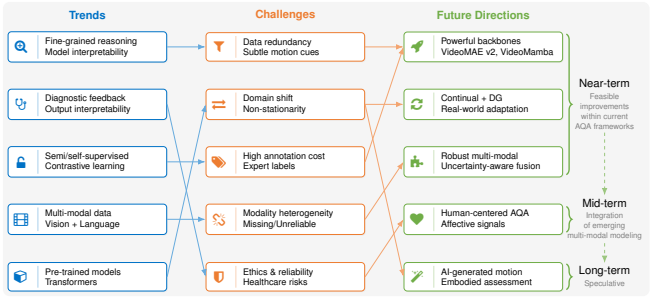

Figure 13: Evolutionary roadmap of AQA. The diagram connects current methodological **trends** (left) with emerging **challenges** (middle), and highlights potential **future directions** (right).

7.2. Under-explored Challenges

Despite recent advances, several challenges in AQA remain under-explored.

Data Redundancy versus Subtle Cues. High-dimensional video inputs contain considerable redundancy, while action quality depends on subtle motion variations. Balancing information suppression and cue amplification remains a core challenge.

Domain Shift and Non-stationarity. Gaps between pre-trained domains and AQA tasks, together with real-world non-stationary variations, significantly degrade model performance. Effectively adapting to such shifts remains largely unsolved.

High Annotation Cost. Fine-grained quality annotations, such as stage-level or temporal labels, require expert knowledge and are expensive to obtain. Learning reliable quality representations from coarse or sparse supervision remains challenging.

Modality Heterogeneity. Although multi-modal AQA improves performance, real-world data often involve missing or unreliable modalities. Designing robust models that gracefully degrade under incomplete multi-modal inputs remains under-explored.

Ethics and Reliability. In sensitive domains such as healthcare and rehabilitation, AQA systems must ensure reliability, fairness, and interpretability. Addressing ethical risks and enabling responsible deployment remain open challenges.

7.3. Future Directions

Building on the above challenges, we outline several promising directions for advancing AQA, organized roughly from near-term extensions of existing pipelines to longer-term opportunities enabled by emerging AI paradigms (see Fig. 13).

Integrating Powerful Foundation Models. Our analysis in Sec. 4 reveals that progress in AQA closely follows backbone evolution (see Fig. 5), and recent works [32, 42] show that stronger backbones consistently improve performance. Meanwhile, our benchmark analysis shows that current AQA datasets are limited by label scarcity (see Sec. 6), which constrains supervised learning and limits generalization. This motivates the adoption of foundation models that leverage large-scale pre-training to learn transferable representations, such as VideoMAE V2 [166] and VideoMamba [167] for spatiotemporal representation, MotionGPT [168] for skeletal modeling, and Video-LLaMA [169] for video-language understanding. While these models provide richer representations and cross-modal reasoning capabilities, effective adaptation to AQA-specific requirements remains an open challenge.

Continual Learning and Domain Generalization. Existing continual AQA studies [31, 114, 170] still largely rely on labeled data for model updates. Future research should jointly address continual adaptation and generalization to unseen domains, par-

ticularly in ego-view settings [161], where viewpoints and motion patterns evolve significantly over time, enabling robust AQA under non-stationary conditions. Moreover, these methods rely on rehearsal strategies that store previously seen data, which may raise privacy concerns and limit their applicability in real-world scenarios.

Robust Multi-modal Learning. Our analysis in Sec. 4 reveals that most multi-modal AQA methods assume complete and reliable modality inputs, limiting their applicability in real-world scenarios where modalities may be missing or noisy. This motivates the design of AQA models that are robust to incomplete or unreliable modalities [149, 170]. Future work should focus on flexible architectures with uncertainty-aware fusion and graceful degradation under partial multi-modal inputs.

Human-centered and Trustworthy AQA. As discussed in Sec. 5 and Sec. 6, AQA is often applied in sensitive domains such as healthcare [8], where reliability and interpretability are critical. However, current methods remain vulnerable to adversarial perturbations and lack sufficient interpretability for practical deployment. This motivates the development of trustworthy AQA systems that emphasize robustness, interpretability, and ethical considerations. Future work should also incorporate affective signals, such as emotion and engagement, to enable more human-centered assessment.

AI-generated and Embodied Action Assessment. The rapid progress of action generation foundation models [168, 171] has enabled realistic motion synthesis, making reliable evaluation increasingly important. In this context, AQA can play a critical role in assessing the realism, correctness, and execution quality of generated motions for applications such as e-commerce avatars and AI hosts [164]. Beyond motion synthesis, similar evaluation challenges arise in embodied learning and robotics, where agents must perform actions in physical or simulated environments [172, 173]. Related problems have also been studied in surgical skill assessment, where methods are used to evaluate how humans operate robotic or medical systems [87, 174, 175]. Future research should explore how AQA methods can be integrated with generative and embodied models to provide automatic and fine-grained evaluation, potentially forming a feedback loop that further improves motion generation and embodied action learning.

8. Conclusion

This survey reviews recent advances in AQA, proposes a modality-driven hierarchical taxonomy, and introduces a unified benchmark for evaluating both accuracy and computational efficiency. By organizing fragmented literature and inconsistent comparisons, we clarify the methodological evolution across modalities and applications. We also summarize key trends, open challenges, and future research directions. This survey has several limitations. Some emerging application-specific datasets are not discussed in depth due to space constraints. In addition, the benchmark focuses on video-based AQA, as skeleton-based and multi-modal methods still lack standardized datasets and reproducible pipelines. Thus, the evaluation emphasizes fair comparison under official settings rather than extensive task-specific tuning. We hope this survey serves as a useful reference and promotes more robust and unified AQA research.

Acknowledgements. This work was supported in part by NSFC under Grant 62272019.

References

[1] J. Liu, H. Wang, K. Stawarz et al., Vision-based human action quality assessment: A systematic review, ESWA 263 (2025).
[2] H. Yin, P. Parmar, D. Xu et al., A decade of action quality assessment: Largest systematic survey of trends, challenges, and future directions, IJCV (2026).
[3] C. Xu, Y. Fu, B. Zhang et al., Learning to score figure skating sport videos, TCSVT 30 (12) (2019) 4578–4590.
[4] P. Parmar, B. Tran Morris, Learning to score olympic events, in: CVPRW, 2017, pp. 20–28.
[5] P. Parmar, J. Reddy, B. Morris, Piano skills assessment, in: MMSP, 2021, pp. 1–5.
[6] Q. Zhang, B. Li, Relative hidden markov models for video-based evaluation of motion skills in surgical training, TPAMI 37 (6) (2014) 1206–1218.
[7] M. Capecci, M. G. Ceravolo, F. Ferracuti et al., The kimore dataset: Kinematic assessment of movement and clinical scores for remote monitoring of physical rehabilitation, TNSRE 27 (7) (2019) 1436–1448.
[8] K. Zhou, R. Cai, Y. Ma et al., A video-based augmented reality system for human-in-the-loop muscle strength assessment of juvenile dermatomyositis, TVCG 29 (5) (2023) 2456–2466.
[9] Y. Huang, G. Chen, J. Xu et al., Egoexolearn: A dataset for bridging asynchronous ego- and exo-centric view of procedural activities in real world, in: CVPR, 2024, pp. 22072–22086.
[10] C. H. Wu, K. Ashutosh, K. Grauman, Skillsight: Efficient first-person skill assessment with gaze, arXiv preprint arXiv:2511.19629 (2025).
[11] L. Wang, X. Zhang, H. Su et al., A comprehensive survey of continual learning: theory, method and application, TPAMI 46 (8) (2024) 5362–5383.

[12] Z. Chang, G. A. Koulieris, H. J. Chang et al., On the design fundamentals of diffusion models: A survey, PR (2025) 111934.
[13] Q. Lei, J.-X. Du, H.-B. Zhang et al., A survey of vision-based human action evaluation methods, Sensors 19 (19) (2019) 4129.
[14] S. Wang, D. Yang, P. Zhai et al., A survey of video-based action quality assessment, in: INSAI, IEEE, 2021, pp. 1–9.
[15] A. Xu, L.-A. Zeng, W.-S. Zheng, Likert scoring with grade decoupling for long-term action assessment, in: CVPR, 2022, pp. 3222–3231.
[16] P. Parmar, B. T. Morris, What and how well you performed? a multitask learning approach to action quality assessment, in: CVPR, 2019, pp. 304–313.
[17] Y. Liu, X. Cheng, T. Ikenaga, A figure skating jumping dataset for replay-guided action quality assessment, in: ACM MM, 2023, pp. 2437–2445.
[18] T. Wang, Y. Wang, M. Li, Towards accurate and interpretable surgical skill assessment: A video-based method incorporating recognized surgical gestures and skill levels, in: MICCAI, 2020, pp. 668–678.
[19] H. Doughty, D. Damen, W. Mayol-Cuevas, Who's better? who's best? pairwise deep ranking for skill determination, in: CVPR, 2018, pp. 6057–6066.
[20] G. Bertasius, H. Soo Park, S. X. Yu et al., Am i a baller? basketball performance assessment from first-person videos, in: ICCV, 2017, pp. 2177–2185.
[21] M. Fang, X. Du, Q. Liu et al., Which is the better teacher action? a new ranking model and dataset, in: ICASSP, 2024, pp. 7695–7699.
[22] X. Ding, X. Xu, X. Li, Sedskill: Surgical events driven method for skill assessment from thoracoscopic surgical videos, in: MICCAI, Vol. 14228, 2023, pp. 35–45.
[23] Y. Li, X. Chai, X. Chen, Scoringnet: Learning key fragment for action quality assessment with ranking loss in skilled sports, in: ACCV, Vol. 11366, 2019, pp. 149–164.
[24] Z. Luo, Y. Xiao, F. Yang et al., Rhythmer: Ranking-based skill assessment with rhythm-aware transformer, TCSVT 35 (1) (2025) 259–272.
[25] X. Yu, Y. Rao, W. Zhao et al., Group-aware contrastive regression for action quality assessment, in: ICCV, 2021, pp. 7899–7908.
[26] K. Zhou, Y. Ma, H. P. H. Shum et al., Hierarchical graph convolutional networks for action quality assessment, TCSVT 33 (12) (2023) 7749–7763.
[27] J. Xu, Y. Rao, X. Yu et al., Finediving: A fine-grained dataset for procedure-aware action quality assessment, in: CVPR, 2022, pp. 2939–2948.
[28] J. Xu, Y. Rao, J. Zhou et al., Procedure-aware action quality assessment: Datasets and performance evaluation, IJCV 132 (12) (2024) 6069–6090.
[29] J. Xu, S. Yin, G. Zhao et al., Fineparser: A fine-grained spatio-temporal action parser for human-centric action quality assessment, in: CVPR, 2024, pp. 14628–14637.
[30] K. Zhou, L. Wang, X. Zhang et al., Magr: Manifold-aligned graph regularization for continual action quality assessment, in: ECCV, Vol. 15069, 2024, pp. 375–392.
[31] K. Zhou, Q. Pan, X. Zhang et al., Continual action quality assessment via adaptive manifold-aligned graph regularization, arXiv preprint arXiv:2510.06842 (2025).
[32] K. Zhou, H. P. H. Shum, F. W. B. Li et al., Phi: Bridging domain shift in long-term action quality assessment via progressive hierarchical instruction, TIP 34 (2025) 3718–3732.
[33] S. Wang, D. Yang, P. Zhai et al., Tsa-net: Tube self-attention network for action quality assessment, in: ACM MM, 2021, pp. 4902–4910.
[34] A. Dadashzadeh, S. Duan, A. Whone et al., Pecop: Parameter efficient continual pretraining for action quality assessment, in: WACV, 2024, pp. 42–52.
[35] J.-H. Pan, J. Gao, W.-S. Zheng, Action assessment by joint relation graphs, in: ICCV, 2019, pp. 6340–6349.
[36] J. Pan, J. Gao, W.-S. Zheng, Adaptive action assessment, TPAMI 44 (12) (2022) 8779–

8795.
[37] K. Gedamu, Y. Ji, Y. Yang et al., Self-supervised sub-action parsing network for semi-supervised action quality assessment, TIP 33 (2024) 6057–6070.
[38] K. Gedamu, Y. Ji, Y. Yang et al., Fine-grained spatio-temporal parsing network for action quality assessment, TIP 32 (2023) 6386–6400.
[39] R. Han, K. Zhou, A. Atapour-Abarghouei et al., Finecausal: A causal-based framework for interpretable fine-grained action quality assessment, in: CVPRW, 2025, pp. 6008–6017.
[40] Y. Tang, Z. Ni, J. Zhou et al., Uncertainty-aware score distribution learning for action quality assessment, in: CVPR, 2020, pp. 9836–9845.
[41] B. Zhang, J. Chen, Y. Xu et al., Auto-encoding score distribution regression for action quality assessment, Neural Comput. Appl. 36 (2) (2024) 929–942.
[42] K. Zhou, J. Li, R. Cai et al., Cofinal: Enhancing action quality assessment with coarse-to-fine instruction alignment, in: IJCAI, 2024, pp. 1771–1779.
[43] M. Li, H.-B. Zhang, Q. Lei et al., Pairwise contrastive learning network for action quality assessment, in: ECCV, Vol. 13664, 2022, pp. 457–473.
[44] Z. Cao, T. Simon, S.-E. Wei et al., Realtime multi-person 2d pose estimation using part affinity fields, in: CVPR, 2017, pp. 7291–7299.
[45] C. Lugaresi, J. Tang, H. Nash et al., Mediapipe: A framework for building perception pipelines, arXiv preprint arXiv:1906.08172 (2019).
[46] Y. Xu, J. Zhang, Q. Zhang et al., Vitpose: Simple vision transformer baselines for human pose estimation, NeurIPS 35 (2022) 38571–38584.
[47] B. Yu, Y. Liu, X. Zhang et al., Egcn: An ensemble-based learning framework for exploring effective skeleton-based rehabilitation exercise assessment, in: IJCAI, 2022, pp. 3681–3687.
[48] X. B. Bruce, Y. Liu, K. C. C. Chan et al., Egcn++: A new fusion strategy for ensemble learning in skeleton-based rehabilitation exercise assessment, TPAMI 46 (9) (2024) 6471–6485.
[49] X. Wang, J. Wang, H. Hu, Skeleton-based action quality assessment via partially connected lstm with triplet losses, in: PRCV, Vol. 13536, 2022, pp. 220–232.
[50] X. Bruce, Y. Liu, K. C. Chan et al., Skeleton-based human action evaluation using graph convolutional network for monitoring alzheimer's progression, PR 119 (2021) 108095.
[51] Y. Liao, A. Vakanski, M. Xian, A deep learning framework for assessing physical rehabilitation exercises, TNSRE 28 (2) (2020) 468–477.
[52] S. Yan, Y. Xiong, D. Lin, Spatial temporal graph convolutional networks for skeleton-based action recognition, in: AAAI, Vol. 32, 2018.
[53] C. Zhou, J. Zeng, L. Qiu et al., An attention-based adaptive spatial–temporal graph convolutional network for long-video ergonomic risk assessment, EAAI 131 (2024) 107780.
[54] J. Xia, M. Zhuge, T. Geng et al., Skating-mixer: Long-term sport audio-visual modeling with mlps, in: AAAI, Vol. 37, 2023, pp. 2901–2909.
[55] Y. Zhong, F. Zhang, Y. Demiris, Contrastive self-supervised learning for automated multimodal dance performance assessment, in: ICASSP, 2023, pp. 1–5.
[56] L.-A. Zeng, W.-S. Zheng, Multimodal action quality assessment, TIP 33 (2024) 1600–1613.
[57] X. Wang, P.-J. Li, Y.-Y. Shen, Attention-driven multimodal alignment for long-term action quality assessment, Appl. Soft Comput. 183 (2025).
[58] Z. Du, D. He, X. Wang et al., Learning semantics-guided representations for scoring figure skating, TMM 26 (2023) 4987–4997.
[59] S. Zhang, S. Bai, G. Chen et al., Narrative action evaluation with prompt-guided multimodal interaction, in: CVPR, 2024, pp. 18430–18439.

[60] H. Xu, X. Ke, Y. Li et al., Vision-language action knowledge learning for semantic-aware action quality assessment, in: ECCV, Vol. 15100, 2024, pp. 423–440.
[61] H. Xu, X. Ke, H. Wu et al., Language-guided audio-visual learning for long-term sports assessment, in: CVPR, 2025, pp. 23967–23977.
[62] Y. Ji, L. Ye, H. Huang et al., Localization-assisted uncertainty score disentanglement network for action quality assessment, in: ACM MM, 2023, pp. 8590–8597.
[63] S. Zahan, G. M. Hassan, A. Mian, Learning sparse temporal video mapping for action quality assessment in floor gymnastics, IEEE Transactions on Instrumentation and Measurement (2024).
[64] Y. Ding, S. Zhang, S. Liu et al., 2m-af: A strong multi-modality framework for human action quality assessment with self-supervised representation learning, in: ACM MM, 2024, pp. 1564–1572.
[65] L. Dong, W. Wang, Y. Qiao et al., Lucidaction: A hierarchical and multi-model dataset for comprehensive action quality assessment, in: NeurIPS, 2024.
[66] M. A. A. H. Khan, H. Shahriar, Mrehab: Mutlimodal data acquisition and modeling framework for assessing stroke and cardiac rehabilitation exercises, in: COMPSAC, IEEE, 2022, pp. 452–453.
[67] D.-W. Kim, J. E. Park, M.-J. Kim et al., Automatic assessment of upper extremity function and mobile application for self-administered stroke rehabilitation, TNSRE 32 (2024) 652–661.
[68] L.-A. Zeng, F.-T. Hong, W.-S. Zheng et al., Hybrid dynamic-static context-aware attention network for action assessment in long videos, in: ACM MM, 2020, pp. 2526–2534.
[69] F. Lin, J. Huang, Z. Chen et al., Enhancing long-term action quality assessment: A dual-modality dataset and causal cross-modal framework for trampoline gymnastics, Sensors 25 (18) (2025) 5824.
[70] P. Parmar, B. T. Morris, Action quality assessment across multiple actions, in: WACV, 2019, pp. 1468–1476.
[71] T. Nagai, S. Takeda, M. Matsumura et al., Action quality assessment with ignoring scene context, in: ICIP, 2021, pp. 1189–1193.
[72] X. Zhang, H. Feng, M. S. Hossain et al., Scaled background swap: Video augmentation for action quality assessment with background debiasing, ACM TOMM 21 (8) (2025).
[73] S.-J. Zhang, J.-H. Pan, J. Gao et al., Semi-supervised action quality assessment with self-supervised segment feature recovery, TCSVT 32 (9) (2022) 6017–6028.
[74] H. Pirsiavash, C. Vondrick, A. Torralba, Assessing the quality of actions, in: ECCV, 2014, pp. 556–571.
[75] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, NeurIPS 25 (2012).
[76] D. Tran, L. Bourdev, R. Fergus et al., Learning spatiotemporal features with 3d convolutional networks, in: ICCV, 2015, pp. 4489–4497.
[77] Z. Qiu, T. Yao, T. Mei, Learning spatio-temporal representation with pseudo-3d residual networks, in: ICCV, 2017, pp. 5533–5541.
[78] J. Carreira, A. Zisserman, Quo vadis, action recognition? a new model and the kinetics dataset, in: CVPR, 2017, pp. 6299–6308.
[79] Z. Liu, J. Ning, Y. Cao et al., Video swin transformer, in: CVPR, 2022, pp. 3202–3211.
[80] W. Kay, J. Carreira, K. Simonyan et al., The kinetics human action video dataset, arXiv preprint arXiv:1705.06950 (2017).
[81] J. Carreira, E. Noland, A. Banki-Horvath et al., A short note about kinetics-600, arXiv preprint arXiv:1808.01340 (2018).
[82] X. Xiang, Y. Tian, A. Reiter et al., S3d: Stacking segmental p3d for action quality assessment, in: ICIP, 2018, pp. 928–932.

[83] L.-J. Dong, H.-B. Zhang, Q. Shi et al., Learning and fusing multiple hidden substages for action quality assessment, Knowl.-Based Syst. 229 (2021) 107388.
[84] H.-B. Zhang, L.-J. Dong, Q. Lei et al., Label-reconstruction-based pseudo-subscore learning for action quality assessment in sporting events, Appl. Intell. 53 (9) (2023) 10053–10067.
[85] Z. Li, Y. Huang, M. Cai et al., Manipulation-skill assessment from videos with spatial attention network, in: ICCV, 2019, pp. 4385–4395.
[86] Y. Bai, D. Zhou, S. Zhang et al., Action quality assessment with temporal parsing transformer, in: ECCV, Vol. 13664, 2022, pp. 422–438.
[87] J. Gao, J.-H. Pan, S.-J. Zhang et al., Automatic modelling for interactive action assessment, IJCV 131 (3) (2023) 659–679.
[88] X. Ke, H. Xu, X. Lin et al., Two-path target-aware contrastive regression for action quality assessment, Inf. Sci. 664 (2024) 120347.
[89] R. Han, K. Zhou, S. Chen et al., Caflow: Enhancing long-term action quality assessment with causal counterfactual flow, arXiv preprint arXiv:2511.21653 (2025).
[90] J. Xu, S. Yin, Y. Peng, Human-centric fine-grained action quality assessment, TPAMI 47 (8) (2025) 6242–6255.
[91] K. He, X. Zhang, S. Ren et al., Deep residual learning for image recognition, in: CVPR, 2016, pp. 770–778.
[92] X. Dong, X. Liu, W. Li et al., Interpretable long-term action quality assessment, in: BMVC, 2024.
[93] Z. Li, L. Gu, W. Wang et al., Surgical skill assessment via video semantic aggregation, in: MICCAI, 2022, pp. 410–420.
[94] L. Okamoto, P. Parmar, Hierarchical neurosymbolic approach for comprehensive and explainable action quality assessment, in: CVPRW, 2024, pp. 3204–3213.
[95] J. Gao, W.-S. Zheng, J.-H. Pan et al., An asymmetric modeling for action assessment, in: ECCV, Vol. 12375, 2020, pp. 222–238.
[96] S. Zhang, W. Dai, S. Wang et al., Logo: A long-form video dataset for group action quality assessment, in: CVPR, 2023, pp. 2405–2414.
[97] T. He, H. Liu, Y. Li et al., Collaborative weakly supervised video correlation learning for procedure-aware instructional video analysis, in: AAAI, Vol. 38, 2024, pp. 2112–2120.
[98] T. He, H. Liu, Z. Ni et al., Achieving procedure-aware instructional video correlation learning under weak supervision from a collaborative perspective, IJCV 133 (4) (2025) 2070–2095.
[99] H. Fang, W. Zhou, H. Li, End-to-end action quality assessment with action parsing transformer, in: VCIP, IEEE, 2023, pp. 1–5.
[100] Q. Lei, H. Zhang, J. Du, Temporal attention learning for action quality assessment in sports video, Signal, Image and Video Processing 15 (7) (2021) 1575–1583.
[101] Y. Gao, S. S. Vedula, C. E. Reiley et al., Jhu-isi gesture and skill assessment working set (jigsaws): A surgical activity dataset for human motion modeling, in: MICCAI Workshop, Vol. 3, 2014, p. 3.
[102] J. Wang, Z. Du, A. Li et al., Assessing action quality via attentive spatio-temporal convolutional networks, in: PRCV, Vol. 12306, 2020, pp. 3–16.
[103] J. Liu, H. Wang, W. Zhou et al., Adaptive spatiotemporal graph transformer network for action quality assessment, TCSVT 35 (7) (2025) 6628–6639.
[104] Q. An, M. Qi, H. Ma, Multi-stage contrastive regression for action quality assessment, in: ICASSP, 2024, pp. 4110–4114.
[105] M. Qi, H. Ye, J. Peng et al., Action quality assessment via hierarchical pose-guided multi-stage contrastive regression, TIP 34 (2025) 6461–6474.
[106] D. Liu, Q. Li, T. Jiang et al., Towards unified surgical skill assessment, in: CVPR, 2021,

pp. 9517–9526.
[107] X. Dong, X. Liu, W. Li et al., Uil-aqa: Uncertainty-aware clip-level interpretable action quality assessment, IJCV 134 (24) (2026).
[108] M. Fang, Y. Zhou, J. Ren et al., A teacher action quality assessment method based on label constraint strategy, in: ICASSP, 2025.
[109] C. Wang, X. Ding, X. Zhao et al., Adaptive frequency-aware network for action quality assessment 31 (5) (2025).
[110] C. Zhou, Y. Huang, H. Ling, Uncertainty-driven action quality assessment, arXiv preprint arXiv:2207.14513 (2022).
[111] H. Matsuyama, N. Kawaguchi, B. Y. Lim, Iris: Interpretable rubric-informed segmentation for action quality assessment, in: ICIUI, 2023, pp. 368–378.
[112] H. Doughty, W. Mayol-Cuevas, D. Damen, The pros and cons: Rank-aware temporal attention for skill determination in long videos, in: CVPR, 2019, pp. 7854–7863.
[113] H.-M. Qiu, H.-B. Zhang, Q. Lei et al., Learning referee evaluation and assessing action quality from coarse to fine in diving sport, Neurocomputing 648 (2025).
[114] Y.-M. Li, L.-A. Zeng, J.-K. Meng et al., Continual action assessment via task-consistent score-discriminative feature distribution modeling, TCSVT 34 (10) (2024) 9112–9124.
[115] P.-X. Lian, Z.-G. Shao, Improving action quality assessment with across-staged temporal reasoning on imbalanced data, Appl. Intell. 53 (24) (2023) 30443–30454.
[116] H. Jain, G. Harit, A. Sharma, Action quality assessment using siamese network-based deep metric learning, TCSVT 31 (6) (2021) 2260–2273.
[117] Y. Pan, C. Zhang, G. Bertasius, Basket: A large-scale video dataset for fine-grained skill estimation, in: CVPR, 2025, pp. 28952–28962.
[118] X. Ding, C. Wang, X. Zhao et al., Scene-aware contrastive regression for multi-person action quality assessment, Appl. Intell. 55 (16) (2025).
[119] A. Kanade, M. Sharma, M. Muniyandi, Attention-guided deep learning framework for movement quality assessment, in: ICASSP, 2023, pp. 1–5.
[120] Q. Lei, H. Li, H. Zhang et al., Multi-skeleton structures graph convolutional network for action quality assessment in long videos, Appl. Intell. 53 (19) (2023) 21692–21705.
[121] H. Li, Q. Lei, H. Zhang et al., Skeleton-based deep pose feature learning for action quality assessment on figure skating videos, JVCIR 89 (2022) 103625.
[122] Q. Lei, H.-B. Zhang, J.-X. Du et al., Learning effective skeletal representations on rgb video for fine-grained human action quality assessment, Electronics 9 (4) (2020) 568.
[123] A. Kryeem, S. Raz, D. Eluz et al., Personalized monitoring in home healthcare: An assistive system for post hip replacement rehabilitation, in: ICCV, 2023, pp. 1868–1877.
[124] A. Abedi, M. Malmirian, S. S. Khan, Cross-modal video to body-joints augmentation for rehabilitation exercise quality assessment, arXiv preprint arXiv:2306.09546 (2023).
[125] U. Gallardo, F. Caro, E. Hernández et al., Gymetricpose: A light-weight angle-based graph adaptation for action quality assessment, in: CBMS, IEEE, 2024, pp. 43–50.
[126] S.-E. Wei, V. Ramakrishna, T. Kanade et al., Convolutional pose machines, in: CVPR, 2016, pp. 4724–4732.
[127] T. Wang, M. Jin, J. Wang et al., Towards a data-driven method for rgb video-based hand action quality assessment in real time, in: Annual ACM Symposium on Applied Computing, 2020, pp. 2117–2120.
[128] MMPose Contributors, MMPose: Openmmlab pose estimation toolbox and benchmark, https://github.com/open-mmlab/mmpose, accessed: December 01, 2024 (2020).
[129] Z. Li, H. Chen, J. Cai et al., Segmentation and quality assessment of continuous fitness movements based on vision, in: International Conference on Intelligent Computing, Springer, 2024, pp. 96–107.

[130] K. Iskakov, E. Burkov, V. Lempitsky et al., Learnable triangulation of human pose, in: ICCV, 2019, pp. 7718–7727.
[131] K. Zheng, J. Wu, J. Zhang et al., A skeleton-based rehabilitation exercise assessment system with rotation invariance, TNSRE 31 (2023) 2612–2621.
[132] S. Mehraban, V. Adeli, B. Taati, Motionagformer: Enhancing 3d human pose estimation with a transformer-gcnformer network, in: WACV, 2024, pp. 6920–6930.
[133] C. Zimmermann, T. Brox, Learning to estimate 3d hand pose from single rgb images, in: ICCV, 2017, pp. 4903–4911.
[134] D. Pavllo, C. Feichtenhofer, D. Grangier et al., 3d human pose estimation in video with temporal convolutions and semi-supervised training, in: ICCV, 2019, pp. 7753–7762.
[135] C.-I. Joung, S. Byun, S. Baek, Contrastive learning for action assessment using graph convolutional networks with augmented virtual joints, IEEE Access (2023).
[136] B. Garg, A. Postlmayr, P. Cosman et al., Short: Deep learning approach to skeletal performance evaluation of physical therapy exercises, in: ACM/IEEE CHASE, 2023, pp. 168–172.
[137] A. Zanfir, E. Marinoiu, M. Zanfir et al., Deep network for the integrated 3d sensing of multiple people in natural images, NeurIPS 31 (2018).
[138] M. Fieraru, M. Zanfir, S. C. Pirlea et al., Aifit: Automatic 3d human-interpretable feedback models for fitness training, in: CVPR, 2021, pp. 9919–9928.
[139] C. Li, X. Ling, S. Xia, A graph convolutional siamese network for the assessment and recognition of physical rehabilitation exercises, in: ICANN, Springer, 2023, pp. 229–240.
[140] S. Deb, M. F. Islam, S. Rahman et al., Graph convolutional networks for assessment of physical rehabilitation exercises, TNSRE 30 (2022) 410–419.
[141] L. Yao, Q. Lei, H. Zhang et al., A contrastive learning network for performance metric and assessment of physical rehabilitation exercises, TNSRE 31 (2023) 3790–3802.
[142] C. Du, S. Graham, C. Depp et al., Assessing physical rehabilitation exercises using graph convolutional network with self-supervised regularization, in: EMBC, 2021, pp. 281–285.
[143] J. Li, J. Xue, R. Cao et al., Finerehab: A multi-modality and multi-task dataset for rehabilitation analysis, in: CVPRW, 2024, pp. 3184–3193.
[144] M. Nekoui, L. Cheng, Enhancing human motion assessment by self-supervised representation learning., in: BMVC, 2021, p. 322.
[145] K. Gedamu, Y. Ji, Y. Yang et al., Visual-semantic alignment temporal parsing for action quality assessment, TCSVT 35 (3) (2025) 2436–2449.
[146] T. Nagai, S. Takeda, S. Suzuki et al., Mmw-aqa: Multimodal in-the-wild dataset for action quality assessment, IEEE Access (2024).
[147] S. Hirosawa, T. Kato, T. Yamashita et al., Action quality assessment model using specialists' gaze location and kinematics data—focusing on evaluating figure skating jumps, Sensors 23 (22) (2023) 9282.
[148] H. Yin, L. Gu, P. Parmar et al., Flex: A large-scale multi-modal multi-action dataset for fitness action quality assessment, arXiv preprint arXiv:2506.03198 (2025).
[149] H. Xu, H. Wu, X. Ke et al., Mcmoe: Completing missing modalities with mixture of experts for incomplete multimodal action quality assessment, arXiv preprint arXiv:2511.17397 (2025).
[150] K. Ashutosh, T. Nagarajan, G. Pavlakos et al., Expertaf: Expert actionable feedback from video, in: CVPR, 2025, pp. 13582–13594.
[151] K. Ashutosh, K. Grauman, Learning skill-attributes for transferable assessment in video, in: NeurIPS, 2025.
[152] M. Qi, Y. Wu, X. Zhang et al., Explainable action form assessment by exploiting multimodal chain-of-thoughts reasoning, arXiv preprint arXiv:2512.15153 (2025).
[153] H. Xu, H. Wu, X. Ke et al., Quality-guided vision-language learning for long-term action

quality assessment, TMM 27 (2025) 7326–7339.
[154] J. Li, H. Hu, Q. Xing et al., Tai chi action quality assessment and visual analysis with a consumer rgb-d camera, in: MMSP, 2022.
[155] Y. Zhang, X. Li, W. Chai et al., Pose-guided transformer for fine-grained action quality assessment, TCSVT 35 (8) (2025) 7940–7952.
[156] H. Xu, X. Ke, H. Wu et al., Dancefix: An exploration in group dance neatness assessment through fixing abnormal challenges of human pose, in: AAAI, Vol. 39, 2025, pp. 8869–8877.
[157] T. Moodley, D. van der Haar, I3d-ae-lstm: A 2-stream autoencoder for action quality assessment using a newly created cricket batsman video dataset, in: WACV, 2025, pp. 5470–5478.
[158] W. Yun, M. Qi, F. Peng et al., Semi-supervised teacher-reference-student architecture for action quality assessment, in: ECCV, Vol. 15132, 2024, pp. 161–178.
[159] L. Ye, K. Gedamu, J. Shao, Decoupling representations with quantized vectors for semi-supervised action quality assessment, in: ICME, 2025, pp. 1–6.
[160] Y.-M. Li, A.-L. Wang, K.-Y. Lin et al., Techcoach: Towards technical-point-aware descriptive action coaching, arXiv preprint arXiv:2411.17130 (2024).
[161] Y.-M. Li, W.-J. Huang, A.-L. Wang et al., Egoexo-fitness: Towards egocentric and exocentric full-body action understanding, in: ECCV, 2024, pp. 363–382.
[162] M. Fang, Y. Zhou, Q. Liu et al., A contrastive video language multimodal method for teacher action quality assessment, ESWA 297 (C) (2026) 129501.
[163] A. Vakanski, H.-p. Jun, D. Paul et al., A data set of human body movements for physical rehabilitation exercises, Data 3 (1) (2018) 2.
[164] Z. Chen, W. Sun, Y. Tian et al., Gaia: Rethinking action quality assessment for ai-generated videos, NeurIPS 37 (2024) 40111–40144.
[165] Z. Zhang, W. Sun, X. Li et al., Human-activity agv quality assessment: A benchmark dataset and an objective evaluation metric, in: ACM MM, 2025, pp. 6771–6780.
[166] L. Wang, B. Huang, Z. Zhao et al., Videomae v2: Scaling video masked autoencoders with dual masking, in: CVPR, 2023, pp. 14549–14560.
[167] J. Park, H.-S. Kim, K. Ko et al., Videomamba: Spatio-temporal selective state space model, in: ECCV, Springer, 2024, pp. 1–18.
[168] B. Jiang, X. Chen, W. Liu et al., Motiongpt: Human motion as a foreign language, NeurIPS 36 (2023) 20067–20079.
[169] H. Zhang, X. Li, L. Bing, Video-llama: An instruction-tuned audio-visual language model for video understanding, 2023, pp. 543–553.
[170] K. Zhou, C. Li, Q. Pan et al., Brima: Bridged modality adaptation for multi-modal continual action quality assessment, in: CVPR, 2026.
[171] Y. Yuan, J. Song, U. Iqbal et al., Physdiff: Physics-guided human motion diffusion model, in: CVPR, 2023, pp. 16010–16021.
[172] G. Paolo, J. Gonzalez-Billandon, B. Kégl, Position: A call for embodied ai, in: ICML, 2024.
[173] Z. Feng, R. Xue, L. Yuan et al., Multi-agent embodied ai: Advances and future directions, arXiv preprint arXiv:2505.05108 (2025).
[174] J. L. Laughlin, L. R. Johnson, B. Ghanekar et al., Technical skills assessment in robotic surgery: A review of recent methods, Methodist DeBakey Cardiovascular Journal 21 (5) (2025) 49.
[175] M. W. Boal, D. Anastasiou, F. Tesfai et al., Evaluation of objective tools and artificial intelligence in robotic surgery technical skills assessment: a systematic review, British Journal of Surgery 111 (1) (2024) znad331.

Appendix A. Database Query Details

To ensure a comprehensive and reproducible literature collection process, we conducted a systematic search across five major academic databases, including IEEE Xplore, Web of Science, Scopus, Google Scholar, and arXiv. The search focused on publications from 2014 to 2025, covering the rapid development period of deep learning-based AQA methods.

The query strategy was designed to capture representative studies related to AQA, skill assessment, and score prediction tasks in both sports and medical domains. The core search terms included:

```
("action quality assessment" OR "action scoring" OR
"skill assessment" OR "performance evaluation" OR "score
prediction")
```

Since different databases employ different indexing mechanisms, database-specific query formulations were adopted as follows:

IEEE Xplore

```
("All Metadata":"action quality assessment" OR "All
Metadata":"action scoring" OR "All Metadata":"skill
assessment")
```

Web of Science

```
TS=("action quality assessment" OR "action scoring"
OR "skill assessment")
```

Scopus

```
TITLE-ABS-KEY("action quality assessment" OR "action
scoring" OR "skill assessment")
```

Google Scholar

Keyword combinations were manually queried using phrases such as *action quality assessment*, *skill assessment*, and *action scoring*, with relevance ranking and citation impact considered during screening.

arXiv

```
all:"action quality assessment" OR all:"skill assessment"
```

The initial search results were further refined through title screening, abstract review, and full-text examination according to the inclusion and exclusion criteria described in Section 2. Duplicate studies, non-peer-reviewed works, papers without quantitative evaluation, and studies unrelated to score prediction were excluded to ensure the final benchmark set was representative, reproducible, and technically rigorous.