RICA^2: Rubric-Informed, Calibrated Assessment of Actions

Abrar Majeedi, Viswanatha Reddy Gajjala, Satya Sai Srinath Namburi GNVV, and Yin Li(✉)

University of Wisconsin-Madison, Madison Wisconsin 53706, USA
{majeedi,vgajjala,sgnamburi,yin.li}@wisc.edu

Abstract. The ability to quantify how well an action is carried out, also known as action quality assessment (AQA), has attracted recent interest in the vision community. Unfortunately, prior methods often ignore the score rubric used by human experts and fall short of quantifying the uncertainty of the model prediction. To bridge the gap, we present RICA^2—a deep probabilistic model that integrates score rubric and accounts for prediction uncertainty for AQA. Central to our method lies in stochastic embeddings of action steps, defined on a graph structure that encodes the score rubric. The embeddings spread probabilistic density in the latent space and allow our method to represent model uncertainty. The graph encodes the scoring criteria, based on which the quality scores can be decoded. We demonstrate that our method establishes new state of the art on public benchmarks, including FineDiving, MTL-AQA, and JIGSAWS, with superior performance in *score prediction* and *uncertainty calibration*. Our code is available at https://abrarmajeedi.github.io/rica2_aqa/.

Keywords: Action Quality Assessment · Video Understanding

1 Introduction

Action quality assessment (AQA), aiming at quantifying how well an action is carried out, has been widely studied across scientific disciplines due to its broad range of applications. AQA is key to sports science and analytics. The right way of performing actions maximizes an athlete's performance and minimizes injury risk. AQA is crucial to occupational safety and health. High-quality actions mitigate the physical stress and strain in the workspace. AQA is pivotal for physical therapies. The quality of actions reveals the progress in rehabilitation. AQA also plays a major role in surgical education. Proficient actions improve the outcome and reduce complications.

Observational methods for AQA have been well established for various tasks, *e.g.*, gymnastics [27], manual material handling [41], and surgery [19]. These

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-73036-8_9.

A. Leonardis et al. (Eds.): ECCV 2024, LNCS 15121, pp. 143–161, 2025.
https://doi.org/10.1007/978-3-031-73036-8_9

methods involve a human expert observing an action and decomposing it into a series of key steps. Each of these steps, or a subset of them, can be grouped into a factor and then evaluated using a Likert scale [17] following a pre-defined criterion. Ratings for individual factors, sometimes complemented with impression-based global ratings, are then summarized into a final quality score [19]. Multiple expert ratings are often considered to account for the variance in the scores. While these methods are commonly adopted, they require significant input from human experts and are thus costly and inefficient.

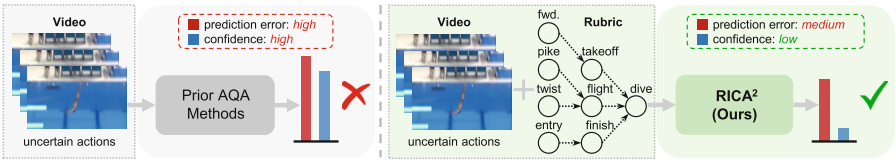


Fig. 1. RICA^2 integrates score rubric used by human experts and accounts for prediction uncertainty, resulting in *accurate* predictions and *calibrated* uncertainty estimates.

There is a burgeoning interest in the vision community to develop video-based AQA [25,26,35,43]. While current solutions have made steady progress across benchmarks [10,24,44], their decision-making processes differ largely from prior observational methods. Almost all prior solutions learn deep models to directly map input videos to scores. Many of them employ an exemplar-based approach, in which a model predicts relative scores by referencing exemplar videos with similar actions and known scores [2,44,47]. Few of them have considered the structure of the actions or their scoring criteria used by observational methods.

Further, existing AQA methods face a key challenge in the accurate quantification of model uncertainty *i.e.*, the uncertainty of model prediction that is calibrate to the expected error [12]. Knowing this uncertainty is particularly helpful for AQA, *e.g.*, when assessing the quality of high-stakes competitions or surgical procedures. With proper calibration, videos that have uncertain predictions can be passed to human experts for a thorough evaluation. Several recent works have started to consider the variance among scores from multiple human experts [35,48,50,51]. Unfortunately, they still fall short of considering prediction uncertainty, leaving this challenge largely unaddressed.

To bridge the gap, we develop a deep probabilistic model for AQA by integrating score rubrics and modeling the uncertainty of the prediction (see Fig. 1). Central to our method lies in the stochastic embedding of action steps, defined on a graph structure that encodes the score rubric. The embeddings spread probabilistic density in latent space and allow our method to represent model uncertainty. The graph encodes the scoring criteria, based on which the quality scores can be decoded. We also present a training scheme and describe an approach to estimate uncertainty. Putting things together, our method, dubbed

RICA^2 (Rubric-informed, Calibrated Assessment of Actions), yields accurate action scores with additional uncertainty estimates.

We evaluate RICA^2 on several public AQA datasets, covering sports and surgical videos. Particularly, RICA^2 establishes new state of the art on FineDiving [44], MTL-AQA [24] and JIGSAWS [10]. On FineDiving [44] – the largest and most challenging AQA benchmark, RICA^2 outperforms latest methods in prediction accuracy (a boost of +0.94% in Spearman's Rank Correlation Coefficient ($SRCC$)) and demonstrates significantly improved uncertainty calibration (a gain of +0.178 in Kendall Tau [13]). Similarly, on MTL-AQA [24], the most commonly used dataset for AQA, RICA^2 attains state-of-the-art $SRCC$, and again largely improved calibration (a gain of +0.444 in Kendall Tau). On JIGSAWS [10], RICA^2 beats the previous best results by a relative margin of +3.37% in $SRCC$. Further, we present extensive experiments to evaluate the key design of RICA^2.

Our main **contributions** are summarized into three folds.

- We propose *RICA*^{2}, a novel deep probabilistic method that incorporates scoring rubrics and uncertainty modeling for AQA, resulting in accurate scores and calibrated uncertainty estimates.
- Our technical innovations lie in (a) a graph neural network to model the scoring rubric in conjunction with stochastic embeddings on the graph to account for prediction uncertainty and (b) a training scheme under the variational information bottleneck framework.
- Our extensive set of experiments demonstrates that RICA^2 achieves state-of-the-art results in AQA, significantly outperforming prior methods in both prediction accuracy and calibration of uncertainty estimates.

2 Related Work

Action quality assessment (AQA). Early works in AQA [11,26] employed handcrafted features to estimate quality scores in videos. More recent methods developed various deep models, including convolutional [35,47], graph [23], recurrent [25,43], and Transformer [2,44] networks. AQA has also been widely considered in surgical education [18], rehabilitation [28], and ergonomics [5].

Recently, exemplar-based methods [2,44,47] have emerged as a promising solution for AQA due to their impressive performance across benchmarks. These methods predict the relative score of an input video by comparing it to selected exemplar videos with similar action steps and known scores. A limitation of this paradigm is the requirement of exemplar videos at inference time. This strategy largely deviates from existing observational methods used by human experts and leads to significantly higher computational costs. While RICA^2 also uses action steps in the input video, it further integrates the scoring rubric of these steps and offers a solution for no-reference AQA *i.e. without using exemplars.*

Several recent works have started to consider the modeling of score uncertainty in AQA [35,48,50,51]. For example, Tang et al. [35] proposed to model the final scores using a Gaussian distribution. They presented a model (MUSDL)

trained to predict the score distribution. This distribution learning idea was further extended in [48,50,51]. However, modeling the score distribution does not warrant the quantification of model uncertainty, as the output distributions might not be calibrated with prediction errors. While RICA^2 also predicts a Gaussian distribution for the scores, our key design is to consider stochastic embeddings to quantify prediction uncertainty, resulting in *calibrated uncertainty estimates*.

The most relevant work is IRIS [20]. IRIS incorporates score rubric into a convolutional network for AQA. This is done by segmenting key steps in the video and predicting sub-scores for individual steps. Similar to IRIS, RICA^2 also considers rubric in a deep model. However, RICA^2 adapts a graph network, treats sub-scores as latent embeddings, predicts the final score, and further quantifies prediction uncertainty. These differences allow RICA^2 to be trained on major public datasets with only final scores, and to output calibrated uncertainty estimates, both of which cannot be achieved by IRIS.

Modeling Uncertainty with Stochastic Embedding. Stochastic embedding, initially introduced in NLP [21,38], treats each embedding as a distribution. This approach has gained recent attention for modeling uncertainty in deep models. Oh et al. [22] considered probabilistic embeddings for metric learning and proposed to model uncertainty based on the stochasticity of embeddings. This idea was further adopted in many vision tasks, including face verification [32], age estimation [16], pose estimation [34], and cross-modal retrieval [6]. Another related line of work is the conditional variational autoencoder [33], where a probabilistic representation of the input is used for a prediction task. Our approach shares a similar idea of using stochastic embeddings to model uncertainty yet is specifically designed for AQA. Our method significantly extends prior idea to embed action steps on a graph structure, and to propagate these stochastic embeddings on the graph.

Graph Neural Networks (GNNs). GNNs [8,15,30] offer a powerful tool to leverage the relational inductive bias inherent in data [3,45]. This inductive bias is beneficial to aggregate a global representation from a group of local ones [29]. Recently, Zhou et al. [51] proposed a hierarchical graph convolutional network for AQA, in which a GNN was used for video representation learning. In contrast, we adapt graph networks to model score rubrics used by observational methods.

3 AQA with Score Rubric and Uncertainty Modeling

Our goal is to assess the quality of an action within an input video. Let X be the video with the action and Y as its quality score. Our method further considers the structure of the action and a scoring rubric based on the structure.

Action Steps. We assume that the action in X comprises a known, ordered set of key steps, denoted as $\mathbb{S} = (s_1, s_2, ..., s_k)$. Each s[1] represents a necessary sub-action for successfully executing the action. Further, s is associated with a text

[1] For the sake of brevity, we omit the subscript as long as there is no confusion.

description that elucidates the specifics of the corresponding step, *e.g.*, "a front-facing takeoff" for diving. This assumption is especially well suited for structured actions, such as diving or surgery, where the key steps are predetermined and follow a specific sequence. Note that the timing of the key steps is not presumed. Even if key steps are unavailable, they can be detected using action recognition methods [4,9] (see supplement Sec. C.4).

Scoring Rubric. We further assume a pre-specified scoring rubric based on the key steps—a common strategy in technical skill assessment [19,27,41]. Specifically, each action step s_k is independently scored, *i.e.*, $s_k \mapsto y_k$. Subsequently, a rule-based rubric is employed to aggregate individual scores $\{y_k\}$ and calculate a final quality score Y, *i.e.*, $\{y_k\} \mapsto Y$, in which steps might be grouped into intermediate stages (see an example in Fig. 2 (a-b)). This rubric follows a deterministic yet often non-injective mapping, *e.g.* a many-to-one mapping such as summation.

Method Overview. We now present RICA^2—a deep probabilistic model for AQA that leverages known action steps and incorporates the scoring rubric for modeling. Importantly, RICA^2 accounts for prediction uncertainty, *i.e.*, when the model prediction *can* and *cannot* be trusted. Figure 2 presents an overview of RICA^2. It consists of two main model components: (a) a graph neural network that integrates the key steps and scoring rubric (Sect. 3.1); and (b) stochastic embeddings defined on the graph to capture prediction uncertainty (Sect. 3.2), coupled with (c) a learning scheme under the variational information bottleneck framework (Sect. 3.3). In what follows, we delve into the details of RICA^2.

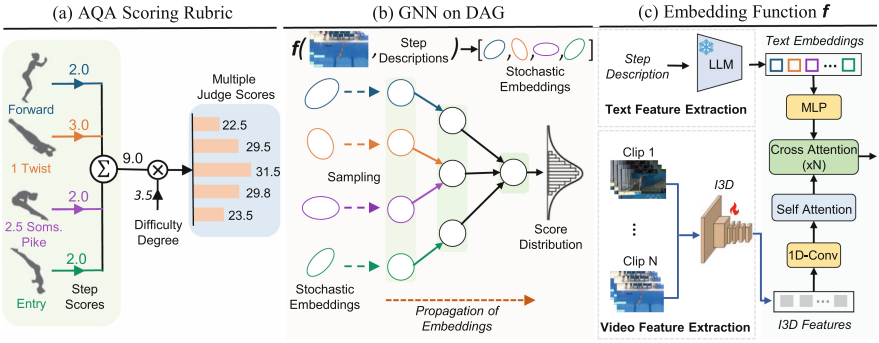


Fig. 2. Overview of RICA^2. Leveraging scoring rubrics (a), RICA^2 integrates a graph representation of action step and rubric with uncertainty modeling (b). Specifically, RICA^2 takes an input of the video and its key action steps, encodes the input into embeddings (c), refines the embeddings through a deep probabilistic model, and outputs an action score in tandem with its uncertainty estimate.

3.1 Integrating Actions Steps and Scoring Rubric with Graph

Steps and Rubric as Graph. We encode action steps and the corresponding scoring rubric with a directed acyclic graph (DAG). This DAG is denoted as $\mathcal{G} = (\mathbb{V}, \mathbb{E})$ with $\mathbb{V}$ as the set of nodes and $\mathbb{E}$ as the set of directed edges. $\mathbb{V}$ consists of three types of nodes: (1) the leaf nodes, denoted as V^s, correspond to individual action steps performed in the input video X; (2) the intermediate nodes capturing possible intermediate stages in the scoring criteria; and (3) a designated root node V^r representing the final score of the action. Further, the edges $\mathbb{E}$ indicate the scoring rubric, connecting steps (leaf nodes) to stages (intermediate nodes), and stages (intermediate nodes) to the final score (root node). We note that $\mathcal{G}$ varies for every input video X (assuming a single action), as different steps might be performed. Figure 2 (a-b) show the example in diving where the key steps and scoring rubric are encoded using our DAG. Additional examples can be found in our supplement Fig. B.

Learning for Quality Assessment. Our approach involves a two-step process for quality assessment. First, we employ an embedding function f, designed to map individual steps into a latent space representing action quality. Secondly, we leverage the key step embeddings $\{Z^s\}$ along with the score rubric encoded in $\mathcal{G}$ to learn a scoring function h. These functions f and h are defined as follows:

$$f : X, \mathcal{G} \mapsto \{Z^s\}; \quad h : \{Z^s\}, \mathcal{G} \mapsto Y, \tag{1}$$

where $\{Z^s\}$ are the embeddings for the set of steps $\mathbb{S}$ in X, corresponding to the leaf nodes $\{V^s\}$ on the DAG.

3.2 Modeling Score Uncertainty with Stochastic Embeddings

To model the prediction uncertainty, we adopt *stochastic step embeddings* defined on the leaf nodes, such that $Z^s \in \mathbb{R}^D \sim p(Z^s|X, \{V^s\})$. Unlike deterministic embeddings, where Z^s would be a fixed vector, stochastic embedding characterizes the distribution of Z^s, allowing for uncertainty control. Specifically, we model $p(Z^s|X, V^s)$ as a Gaussian distribution in $\mathbb{R}^D$ with mean μ^s and diagonal covariance Σ^s. The embedding function f is thus tasked to predict the mean and covariance for the key steps $\mathbb{S}$, *i.e.*, $\{\mu^s, \Sigma^s\} = f(X, \{V^s\})$.

Propagating Stochastic Embeddings on the Graph. Our scoring function h takes the stochastic embeddings Z^s for leaf nodes in $\mathcal{G}$ (provided by f), further computes the embeddings for all nodes in $\mathcal{G}$, and finally decodes a quality score Y from the embedding Z^r of the root node V^r. To this end, we propose an extension of graph neural networks (GNNs), in which stochastic embeddings Z^s are propagated from leaf nodes V^s to the root node V^r based on the graph structured informed by the scoring rubric of a particular task. Key to this GNN lies in a lightweight MLP that operates on each node, taking as input the embeddings of its direct predecessors, and generating a new embedding that is further

propagated to its successors. This scoring function h is thus given by

$$\underbrace{Z^s \sim \mathcal{N}\left(\mu^s, \Sigma^s\right),\ \forall s \in\ \mathbb{S}}_{\text{Sampling from leaf nodes}};\quad \underbrace{Z^{\neg s} = G\left(\Sigma_{V^j \in \mathcal{P}(V^{\neg s})} Z^j\right)}_{\text{Propagating on the DAG}};\quad \underbrace{\hat{Y} = \mathcal{S}\left(Z^r\right)}_{\text{Deocoding the score}} \tag{2}$$

where $V^{\neg s}$ denotes a non-leaf node with its embedding $Z^{\neg s}$. $\mathcal{P}(V^{\neg s})$ is the set of predecessors of $V^{\neg s}$, $G(\cdot)$ is the MLP aggregating features from predecessors, and $\mathcal{S}(\cdot)$ is another MLP decoding the final score $\hat{Y}$ from the root node V^r.

It is important to note that each leaf embedding Z^s is stochastic, characterized by a Gaussian distribution ($p(Z^s|X, \mathcal{G}) = \mathcal{N}(\mu^s, \Sigma^s)$), with parameters predicted by f. The non-leaf embeddings are however deterministic given samples from leaf distributions. This design is motivated by our assumption of the scoring rubric, where uncertainty lies only in assessing action steps and identical individual scores will yield the same final score.

3.3 Learning with Variational Information Bottleneck

With stochastic embeddings, training of RICA^2 is a challenge. We design a training scheme under the variational information bottleneck framework.

Variational Information Bottleneck (VIB). To train our model $p(Y|X, \mathcal{G})$ with stochastic step embeddings $\{Z^s\}$, we adopt the information bottleneck principle [36], leading to the maximization of the following objective

$$I(\{Z^s\}; Y|\mathcal{G})\ -\ \beta I(\{Z^s\}; X|\mathcal{G}), \tag{3}$$

where I is the conditional mutual information, and $\beta > 0$ controls the tradeoff between the sufficiency of using step embeddings $\{Z^s\}$ for predicting Y given $\mathcal{G}$, and the size of the embeddings $\{Z^s\}$ derived from X and $\mathcal{G}$.

While mutual information is computationally intractable for high dimensional $\{Z^s\}$, a common solution [1] is to assume Markov property ($p(Z|X, Y, \mathcal{G}) = p(Z|X, \mathcal{G})$) and conditional independence ($p(\{Z^s\}|X, \mathcal{G}) = \prod_s p(Z^s|X, \mathcal{G})$), followed by the variational approximation for a tractable lower bound

$$-\mathcal{L}_{\text{VIB}} = \mathbb{E}_{Z^s \sim p(Z^s|X,\mathcal{G}), \forall s \in \mathbb{S}} \left[\log p(Y|\{Z^s\}, \mathcal{G})\right] - \beta \Sigma_{s \in \mathbb{S}}\, \text{KL}\left(p(Z^s|X, \mathcal{G})||p(Z^s|\mathcal{G})\right), \tag{4}$$

where $p(Y|\{Z^s\}, \mathcal{G})$ is modeled by the scoring function h, KL denotes the Kullback–Leibler divergence, and $p(Z^s|\mathcal{G})$ is an approximate marginal prior.

VIB Loss. The first term in Eq. (4) defines the log-likelihood of the score given the input. By assuming that output scores follow a Gaussian with a fixed variance σ, this term can be reduced to a mean squared error (MSE) loss

$$\mathcal{L}_{MSE} = \frac{1}{N} \Sigma_i^N (\hat{Y}_i - Y_i)^2 / \sigma^2, \tag{5}$$

where Y_i is the predicted score for a video indexed by i, $\hat{Y}_i$ is the corresponding ground-truth score, and N is the total number of videos in the training set.

The second term in Eq. (4) regularizes the latent space and encodes prediction uncertainty. By assuming a marginal prior of $\mathcal{N}(0, I)$ for $p(Z^s|\mathcal{G})$, we have

$$\begin{aligned}\mathcal{L}_{KL} &= \Sigma_{s\in\mathbb{S}}\ KL\left(\mathcal{N}(\mu^s(x), \Sigma^s(x)||\mathcal{N}(0, I)\right) \\ &= \frac{1}{2}\Sigma_{s\in\mathbb{S}}\ \Sigma_{j=1}^{D}\left((\mu_j^s)^2 + (\sigma_j^s)^2 - \log(\sigma_j^s)^2 - 1\right),\end{aligned} \tag{6}$$

where μ_j^s and σ_j^s, respectively, are the j-th dimension of the mean ($\mu^s(x)$) and variance (square root of the diagonal of $\Sigma^s(x)$), for the step s.

The VIB loss ($\mathcal{L}_{VIB}$) is thus given by

$$\mathcal{L}_{VIB} = \mathcal{L}_{MSE} + \beta\mathcal{L}_{KL}. \tag{7}$$

$\mathcal{L}_{VIB}$ consists of (a) the MSE loss $\mathcal{L}_{MSE}$ from the negative log-likelihood of the predicted scores, aiming at minimizing prediction errors; and (b) the KL divergence $\mathcal{L}_{KL}$ between the predicted Gaussian and the prior, regularizing the stochastic embeddings. Further, the coefficient β balances between two loss terms.

During training, samples are drawn to compute the loss function. The output is matched to the Gaussian distribution with its mean equal to the average of the judge scores. The reparameterization trick [14] is used to allow the backpropagation of gradients through the sampling process.

Estimating Uncertainty. The diagonal covariance $\Sigma^s(X)$ models the uncertainty of the predicted quality score of a step s for an input video X. A larger value in its diagonal represents a wider distribution of scores and, hence, a lower confidence in the prediction. Following [16,22] we generate uncertainty scores by summing up the harmonic means of the predicted variances for individual steps

$$\text{uncertainty}(Y) = \Sigma_{s\in\mathbb{S}} D/\Sigma_{j=1}^{D}(\sigma_j^s)^{-1}, \tag{8}$$

where D is the dimensionality of the stochastic embeddings. Again, σ_j^s is the j-th dimension of the predicted variance.

Stochastic vs. Deterministic Modeling. An interesting variant of RICA^2 is to disable its stochastic component. Conceptually, this is equal to considering step embeddings Z^r as vectors and removing the KL loss $\mathcal{L}_{KL}$. We refer to this deterministic version of our model as RICA^2†. Without stochastic embeddings, RICA^2† is unable to estimate prediction uncertainty, yet often yields slightly lower prediction errors. This trade-off is also observed in prior works [6,16,32]. We include this variant of our model in the experiments.

3.4 Model Instantiation and Implementation

Video and Step Representation. For an input video X, we adapt a pre-trained video backbone(*e.g.*, I3D [4]) to extract its clip-level features, which are further pooled to produce video features $(x_1, x_2, ..., x_T)$ with fixed-length T. To represent action steps, we make use of a pre-trained language model [7] (Flan-T5)

to extract text features from their step descriptions, resulting in an ordered set of text embeddings $(s_1, s_2, ..., s_K)$ for K steps. Note that the language model is not part of RICA^2. It is used solely to extract embeddings for text descriptions of the action steps (see supplement Tables I-L).

Embedding Function f. Our embedding function f is realized using a Transformer model [37] (see Fig. 2(c)). f first processes video features $(x_1, x_2, ..., x_T)$ with a self-attention block and text embeddings $(s_1, s_2, ..., s_K)$ using a MLP. It further makes use of cross-attention blocks (2x) to fuse video and text features, where video features are used to compute keys and values, and text embeddings of steps are projected into queries. Further, f decodes stochastic embeddings of individual steps by predicting a mean vector $\mu^s \in \mathbb{R}^D$ and a diagonal covariance vector $\Sigma^s \in \mathbb{R}^D$ for each step s.

Scoring Function h. With Gaussian distributions for all steps specified by $\{\mu^s, \Sigma^s\}$, we encode the steps and score rubric into a video-specific DAG $\mathcal{G}$, and realize h as a GNN defined on $\mathcal{G}$ following Eq. (2). h is parameterized by its aggregation function G, which is shared among nodes of the same type. G is implemented using an averaging operation followed by a MLP (2 layers). Finally, h decodes the final score at the root note V^r of $\mathcal{G}$.

Training with Auxiliary Losses. While the VIB loss (Eq. (7)) is sufficient for training, it falls short of considering the temporal ordering of steps. This is because of the conditional independence assumption needed for the derivation of VIB, *i.e.*, $p(\{Z^s\}|X, \mathcal{G}) = \prod_s p(Z^s|X, \mathcal{G})$, where the ordering of $\{Z^s\}$ is discarded. To bridge the gap, we incorporate an auxiliary loss term $\mathcal{L}_{Aux}$ inspired by [2]. Specifically, we re-purpose the last cross-attention map ($\mathbb{R}^{K\times T}$) from f as a *step detector*. This is done by computing a temporally-weighted center across the attention of each action step to every video time step (*i.e.*, column-wise). We then enforce that (a) this center is co-located with the peak of the attention along video time steps using a sparsity loss [2]; and (b) all centers follow the temporal ordering of corresponding action steps using a ranking loss [2]. These two terms are summed up as the auxiliary loss, and further added to the VIB loss with a small weight (0.1). In our ablation, we empirically verify that adding the auxiliary loss leads to a minor performance boost.

Inference with Sampling. At the inference time, we enhance robustness by sampling 20 times and averaging their predictions to compute the final score.

4 Experiments and Results

Datasets. Our evaluations are primarily reported on three publicly available benchmark datasets, namely FineDiving [44], MTL-AQA [24], and JIGSAWS [10] in the main paper. In supplement Sec. B, we also include results on the Cataract-101 [31] with cataract surgery videos.

Evaluation Metrics. For all our experiments, we consider metrics on both the *accuracy* of the prediction and the *calibration* of the uncertainty estimates.

• For *accuracy*, we use two widely adopted metrics for AQA [2,35,44], namely Spearman's rank correlation ($SRCC$) and relative L2 distance ($R\ell_2$). $SRCC$ measures how well the predicted scores are ranked w.r.t. the ground truth, while $R\ell_2$ summarizes the prediction errors. A model with more accurate predictions will have higher $SRCC$ and lower $R\ell_2$.

• For *calibration*, we report the uncertainty versus error curve following [16,22]. To plot this curve, test samples are sorted by increasing uncertainty and divided into 10 equal-sized bins. The mean absolute error (MAE) is then computed for items in each bin. We also follow [16,22] in employing Kendall's tau (τ) [13], a numerical measure ranging from -1 to 1 to quantify the correlation between the uncertainties and the prediction errors. A higher τ indicates better calibration, signifying that a model's uncertainty better aligns with prediction errors.

Baselines. RICA^2 is benchmarked against a set of strong baselines, including exemplar-free methods such as DAE [48], USDL and MUSDL [35], and exemplar-based ones such as CoRE [47], TPT [2] and TSA [44]. We further include the deterministic version of our model RICA^2†, which trades the ability of uncertainty estimation for a minor boost in accuracy. Several baselines adopt a direct regression approach, without providing a confidence or uncertainty measure for predictions. USDL [35] and TPT [2] implicitly offer a confidence value. In these works, the probability of the predicted score bin serves as a proxy for uncertainty, computed as (1.0−confidence). DAE [48] outputs the standard deviation of the score distribution, which represents uncertainty.

We seek to ensure a fair comparison yet recognize that methods in our benchmark may consider different settings and/or various types of input. Most prior exemplar-free methods only consider a video as input. While RICA^2 does not utilize exemplars, it takes additional input of step information, *i.e.*, step presence and their temporal ordering. On the other hand, previous exemplar-based methods also require the step information as used by RICA^2, in addition to an input video and an exemplar database. Notably, step information is used to select exemplars, leading to improved results. For example, for diving videos, CoRE [47], TPT [2] and TSA [44] use the diving number (DN) encoding steps and their ordering. Further, TSA [44] also requires the timing of individual steps during training. While it is infeasible to standardize the settings of all methods, we compare to the best reported results in our experiments.

4.1 Results on FineDiving

Dataset. FineDiving [44] is the largest public dataset for AQA, with 3000 video samples capturing various diving actions. The dataset covers 52 different action types, 29 sub-action types, and 23 difficulty degree types, providing a rich and diverse set of examples for AQA. While this dataset contains temporal annotations for the steps, which can be used to improve the performance of AQA as demonstrated in [44], we do not use these annotations for RICA^2.

Experiment Setup. We adhere to the experimental setup of the most recent baseline [44] using their train-test split, with 2251 videos for training and 749

videos for testing. We follow the input video settings used in [44] for RICA^2 and the baselines. Specifically, for each video, we uniformly sample 96 frames, which are segmented into 9 overlapping clips, each containing 16 consecutive frames. We refer to supplement Sec. A.1 for further implementation details.

Results. Table 1a presents our results on FineDiving. Both our stochastic and deterministic versions (RICA^2 and RICA^2†) outperform the state-of-the-art TPT [2], an exemplar-based method. RICA^2 shows a relative margin of 0.7% / 1.4% on $SRCC$ / $R\ell_2$, and RICA^2† has a relative margin of 0.9% / 9.6% on $SRCC$ / $R\ell_2$. This improvement is more pronounced when compared with the exemplar-free methods (MUSDL, DAE-MT) showcasing a significant relative gain of 4.9%, 1.5% on $SRCC$ and 29.8%, 21.6% on $R\ell_2$. While the deterministic RICA^2† has slightly higher accuracy, our stochastic RICA^2 demonstrates superior calibration of its uncertainty estimate ($\tau_{\text{RICA}^2} = 0.64$ vs. $\tau_{\text{TPT}} = -0.56$). Figure 3a further shows uncertainty calibration results. Uncertainty estimates from RICA^2 have a clear upward trend, indicating a higher calibration level. While MUSDL [35] also exhibits a reasonable level of calibration ($\tau_{\text{MUSDL}} = 0.47$ vs. $\tau_{\text{RICA}^2} = 0.64$), the errors are significantly higher than RICA^2 across all uncertainty levels.

Table 1. Main results on (a) FineDiving and (b) MTL-AQA. Prediction accuracy ($SRCC$ and $R\ell_2$) and uncertainty calibration (τ) metrics are reported. We compare our method with exemplar-based and exemplar-free baselines.

(a) Results on FineDiving

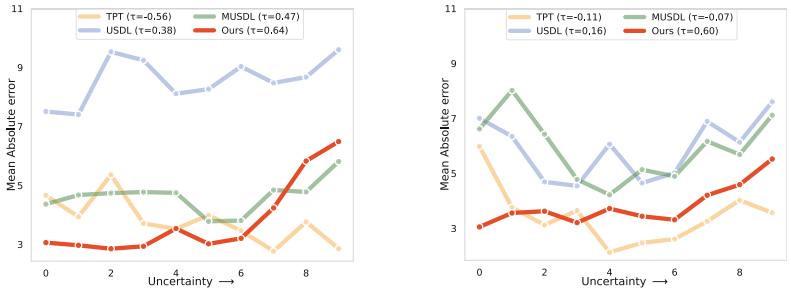
		Metrics		
		$SRCC(\uparrow)$	$R\ell_2(\downarrow)$	$\tau(\uparrow)$
Exemplar based	CoRe [47]	0.9061	0.3615	-
	TSA [44]	0.9203	0.3420	-
	TPT [2]	0.9333	0.2877	-0.5556
Exemplar free	USDL [35,44]	0.8913	0.3822	0.3778
	MUSDL [35,44]	0.8978	0.3704	0.4667
	DAE[48]	0.8820	0.4919	-0.1999
	DAE-MT [48]	0.9285	0.3320	-0.4667
	RICA^2 (Ours)	0.9402	0.2838	**0.6444**
	RICA^2† (Ours)	**0.9421**	**0.2600**	-

(b) Results on MTL-AQA

		Metrics		
		$SRCC(\uparrow)$	$R\ell_2(\downarrow)$	$\tau(\uparrow)$
Exemplar based	TSA-Net [40]	0.9422	-	-
	CoRe [47]	0.9512	0.2600	-
	DAE-CoRe [48]	0.9589	-	-
	TPT [2]	0.9607	0.2378	-0.1111
Exemplar free	C3D-AVG-MTL [24]	0.9044	-	-
	USDL [35]	0.9231	0.4680	0.1556
	MUSDL [35]	0.9273	0.4510	-0.0667
	DAE [48]	0.9231	-	-
	DAE-MT [48]	0.9490	0.2738	-0.4222
	RICA^2 (Ours)	0.9594	0.2580	**0.6000**
	RICA^2† (Ours)	**0.9620**	**0.2280**	-

4.2 Results on MTL-AQA

Dataset. MTL-AQA [24] is one of the most commonly used datasets for AQA. It consists of 1412 samples collected from 16 events with diverse views. The dataset has a rich set of annotations, including the steps performed during the dive, the difficulty score associated with the dive, and the individual judge scores.



(a) Uncertainty calibration on FineDiving **(b)** Uncertainty calibration on MTL-AQA

Fig. 3. Uncertainty vs. prediction error (MAE) on (a) Finediving and (b) MTL-AQA. Results are reported on the test splits, with the X-axis as the uncertainty bin index (uncertainty increases from left to right) and the Y-axis as the MAE in the bin. In comparison to baselines, RICA^2 has improved calibration with lower prediction errors.

Experiment Setup. We follow the evaluation protocol of [2,43,47], dividing the dataset into the standard train set of 1059 videos and a test set of 353 videos. Further, we use the same input video settings as TPT [2] in our experiments to ensure a fair comparison. Specifically, for each video, we uniformly sample 103 frames segmented into 20 overlapping clips, each containing 8 continuous frames. Please refer to supplement Sec. A.2 for more details.

Results. Table 1b summarizes our results on MTL-AQA. Similar to FineDiving, RICA^2 shows state-of-the-art results on MTL-AQA across all evaluation metrics. Specifically, our RICA^2† outperforms the best exemplar-free model (DAE-MT) by a relative margin of 1.4% / 16.7% on $SRCC$ / $R\ell_2$. When compared with the competitive exemplar-based TPT, our has slightly better $SRCC$ ($SRCC_{TPT} = 0.9607$ vs $SRCC_{\text{RICA}^{2\dagger}} = 0.9620$) and $R\ell_2$ (+4.1% relative margin). Again, compared to previous methods, our stochastic model shows improved calibration ($\tau_{\text{RICA}^2} = 0.60$ vs. $\tau_{USDL} = 0.16$ vs. $\tau_{TPT} = -0.11$) as shown in Fig. 3b.

4.3 Results on JIGSAWS

Dataset. In addition to diving videos, we also evaluate RICA^2 on JIGSAWS [10]—a robotic surgical video dataset. The dataset includes three tasks: "Suturing (S)," with 39 recordings, "Needle Passing (NP)," with 26 recordings and "Knot Tying (KT)" with 36 recordings. JIGSAWS is widely used for action quality assessment, despite its small scale.

Experiment Setup. Due to the limited number of samples in the dataset (as few as 7 videos in the test set), cross-validation is often considered for evaluation on JIGSAWS. To ensure a fair comparison, we follow the commonly adopted splits from [35], and the input video setting from [2]. Specifically, for each video, we uniformly sample 160 frames which are segmented into 20 non-overlapping

Table 2. Results on JIGSAWS [10] dataset. Only prediction accuracy ($SRCC$) is considered due to the limited sample size. RICA^2 outperforms all prior approaches.

		Task			
		S	NP	KT	Avg
Exemplar based	CoRe [47]	0.84	0.86	0.86	0.85
	TPT [2]	0.88	0.88	**0.91**	0.89
Exemplar free	ST-GCN [46]	0.31	0.39	0.58	0.43
	TSN [39]	0.34	0.23	0.72	0.46
	JRG [23]	0.36	0.54	0.75	0.57
	USDL [35]	0.64	0.63	0.61	0.63
	MUSDL [35]	0.71	0.69	0.71	0.70
	DAE [48]	0.73	0.72	0.72	0.72
	DAE-MT [48]	0.78	0.74	0.74	0.76
	RICA^2† (Ours)	0.88	0.93	0.88	0.90
	RICA^2 (Ours)	**0.92**	**0.94**	**0.90**	**0.92**

clips. We opt to not include score calibration curves due to the limited sample size of the test sets. Additionally, the key steps in JIGSAWS are general motions (*e.g.* reaching for the needle, orienting the needle, etc.) and thus cannot be localized to any specific section of the video. Thus, we do not use the auxiliary losses for this experiment. More details are described in supplement Sec. A.3.

Results. Table 2 summarizes our results on JIGSAWS. Similar to previous datasets, our models exhibit notable advancements over the previous state-of-the-art model TPT [35], showcasing substantial improvements of 1.1% (RICA^2) to 3.4% (RICA^2†) in terms of average $SRCC$ relative to the exemplar-based state-of-the-art TPT [2]. When compared to the exemplar-free methods, our approach demonstrates an impressive 18.4% (RICA^2) to 21.0% (RICA^2†) relative gain in average $SRCC$ compared to the latest method DAE-MT [48].

4.4 Ablation Studies

To understand our model design choices, we conduct ablation studies on the MTL-AQA [24] dataset. Additional ablations are in supplement Sec. C.2.

Experiment Setup. To simplify our experiments, we opt for running our ablations using fixed I3D features. This allows us to precisely evaluate the contribution of different components of our model. Specifically, we choose I3D weights from an intermediate checkpoint of our trained model and extract features for all videos with the frozen backbone.

Base model. Our ablation constructs a base model using *randomly initialized* step embeddings, an averaging of these embeddings after cross attention with the video features, followed by an MLP for scoring. This base model is trained using only the MSE loss. We then gradually add modules from RICA^2 and study

their effects. Table 3 presents our results using the same features and training epochs, with our base model in row 1.

Table 3. Ablation studies of model components on MTL-AQA dataset. * indicates that the text embeddings were frozen during training.

Step Rep.	DAG (Rubric)	$\mathcal{L}_{KL}$	$\mathcal{L}_{Aux}$	Metrics			
				$SRCC(\uparrow)$	$R\ell_2(\downarrow)$	$\tau(\uparrow)$	Avg. Rank (↓)
Random	×	×	×	0.9426	0.3882	-	5.50
Text	×	×	×	0.9431	0.3509	-	4.50
Text	✓	×	×	0.9430	0.3336	-	4.50
Text*	✓	×	×	0.9437	0.3335	-	3.50
Text*	✓	✓	×	0.9448	0.3329	0.4222	1.83
Text*	✓	✓	✓	**0.9460**	**0.3303**	**0.4222**	**1.17**

Text Embeddings as Step Representations. We first replace randomly initialized step representations with the text embeddings of step descriptions. This leads to a major boost in $R\ell_2$ (Table 3 row 1 vs. row 2), by leveraging knowledge encoded in the LLM [7]. Further, we find that freezing the text embeddings leads to comparable results and faster convergence (Table 3 row 3 vs. row 4).

Does the Scoring Rubric Help? We also investigate the effects of encoding steps and rubric as a DAG—a key design of our model. Adding the DAG results in a noteworthy boost in $R\ell_2$ (Table 3 row 2 vs. row 3). This improvement can be ascribed to the DAG's proficiency in dissecting the action quality across steps.

Effects of Loss Functions. We now study the loss terms. Our loss function has three terms (a) the MSE loss ($\mathcal{L}_{MSE}$) to minimize prediction error, (b) the KL loss ($\mathcal{L}_{KL}$) to regularize the stochastic embeddings, and (c) the auxiliary loss ($\mathcal{L}_{Aux}$) to ensure temporal ordering of steps. Adding the KL loss $\mathcal{L}_{KL}$ yields similar results in $SRCC$ and $R\ell_2$, yet enables calibrated uncertainty estimation. Further attaching the auxiliary loss $\mathcal{L}_{Aux}$ leads to improvement in both $SRCC$ and $R\ell_2$, while maintaining the calibration performance.

Evaluating the Cross-Attention Maps. To gain insight into RICA^2, we now examine the cross-attention maps between the step representations and video features in our learned embedding function f. Figure 4 visualizes the attention map on two test videos on FineDiving. These maps reveal that a step representation is likely to attend to video features during which the step occurs, indicating that RICA^2 learns to encode the temporal location of individual steps.

We further evaluate this *localization* ability following the Pointing game protocol [49], widely considered in weakly supervised / unsupervised localization tasks [42,52]. Pointing Game compares a generated heatmap with an annotated time interval and counts the chance of the heatmap's peak falling into the specified interval. Our evaluation focuses on the FineDiving dataset since it is the

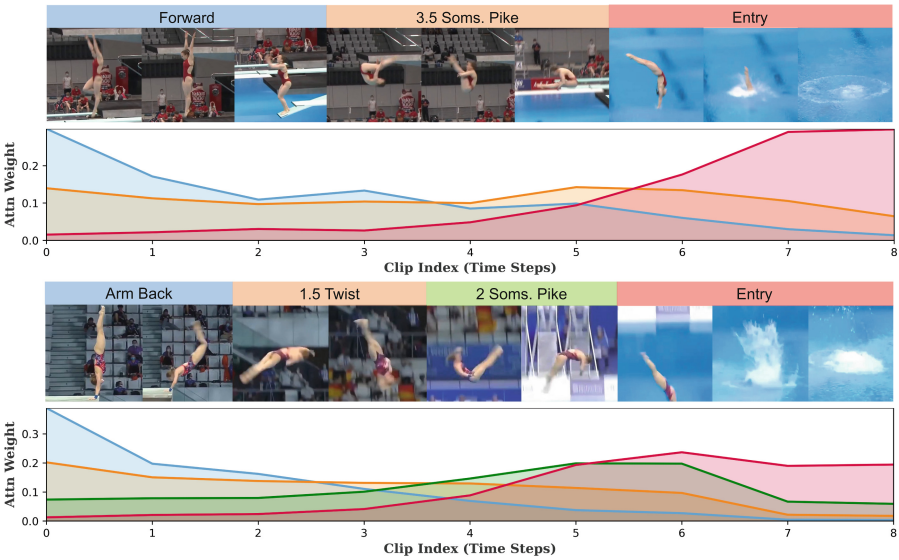


Fig. 4. Visualization of the cross-attention maps. Y-axis : attention value; Y-axis: clip indices (time). Each curve shows an attention map from a step representation to the temporal video features. The frames shown above are aligned with the timing of the corresponding attention plot. Curves and steps are colored accordingly.

only dataset providing annotated time intervals for individual steps. When evaluated on the full test set, attention maps from our model attain an accuracy of 61.4% in the Pointing game protocol, significantly outperforming the chance level accuracy of 30.7% (given each video has 3.26 steps on average). Note that we did not use any annotated segmentation data for training.

5 Conclusion and Discussion

In this paper, we present a deep probabilistic model for action quality assessment in videos. Our key innovation is to integrate score rubrics and to model prediction uncertainty. Specifically, we propose to adapt stochastic embeddings to quantify the uncertainty of individual steps, and to decode action scores using a variant of graph neural network operating on a DAG encoding the score rubric. Our method offers an exemplar-free approach for AQA, achieves new state-of-the-art results in terms of prediction accuracy on public benchmarks, and demonstrates superior calibration of the output uncertainty estimates. We believe that our work provides a solid step towards AQA. We hope that our method and findings can shed light on the challenging problem of trustworthy video recognition.

Acknowledgement:. This work was supported by the UW Madison Office of the Vice Chancellor for Research with funding from the Wisconsin Alumni Research Foundation, by National Science Foundation under Grant No. CNS 2333491, and by the Army Research Lab under contract number W911NF-2020221.

References

1. Alemi, A.A., Fischer, I., Dillon, J.V., Murphy, K.: Deep variational information bottleneck. In: International Conference on Learning Representations (2016)
2. Bai, Y., et al.: Action quality assessment with temporal parsing transformer. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) ECCV 2022. LNCS, pp. 422–438. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-19772-7_25
3. Battaglia, P.W., et al.: Relational inductive biases, deep learning, and graph networks. arXiv preprint arXiv:1806.01261 (2018)
4. Carreira, J., Zisserman, A.: Quo Vadis, action recognition? A new model and the kinetics dataset. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6299–6308 (2017)
5. Chen, C.H., Hu, Y.H., Yen, T.Y., Radwin, R.G.: Automated video exposure assessment of repetitive hand activity level for a load transfer task. Hum. Factors **55**(2), 298–308 (2013)
6. Chun, S., Oh, S.J., De Rezende, R.S., Kalantidis, Y., Larlus, D.: Probabilistic embeddings for cross-modal retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8415–8424 (2021)
7. Chung, H.W., et al.: Scaling instruction-finetuned language models. J. Mach. Learn. Res. **25**(70), 1–53 (2024). http://jmlr.org/papers/v25/23-0870.html
8. Duvenaud, D.K., et al.: Convolutional networks on graphs for learning molecular fingerprints. Adv. Neural Inf. Process. Syst. **28** (2015)
9. Feichtenhofer, C., Fan, H., Malik, J., He, K.: SlowFast networks for video recognition. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 6202–6211 (2019). https://doi.org/10.1109/ICCV.2019.00630
10. Gao, Y., et al.: Jhu-isi gesture and skill assessment working set (JIGSAWS): a surgical activity dataset for human motion modeling. In: Modeling and Monitoring of Computer Assisted Interventions (M2CAI) – MICCAI Workshop (2014)
11. Gordon, A.S.: Automated video assessment of human performance. In: Proceedings of AI-ED, vol. 2 (1995)
12. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: International Conference on Machine Learning, pp. 1321–1330. PMLR (2017)
13. Kendall, M.G.: A new measure of rank correlation. Biometrika **30**(1/2), 81–93 (1938)
14. Kingma, D.P., Welling, M.: Auto-encoding variational Bayes. In: International Conference on Learning Representations (2014)
15. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. In: International Conference on Learning Representations (2017). https://openreview.net/forum?id=SJU4ayYgl
16. Li, W., Huang, X., Lu, J., Feng, J., Zhou, J.: Learning probabilistic ordinal embeddings for uncertainty-aware regression. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13896–13905 (2021)

17. Likert, R.: A Technique for the Measurement of Attitudes. Archives of Psychology (1932)
18. Liu, D., et al.: Towards unified surgical skill assessment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9522–9531 (2021)
19. Martin, J., Martin, J., et al.: Objective structured assessment of technical skill (OSATS) for surgical residents. Br. J. Surg. **84**(2), 273–278 (1997)
20. Matsuyama, H., Kawaguchi, N., Lim, B.Y.: IRIS: interpretable rubric-informed segmentation for action quality assessment. In: Proceedings of the 28th International Conference on Intelligent User Interfaces, pp. 368–378 (2023)
21. Neelakantan, A., Shankar, J., Passos, A., McCallum, A.: Efficient non-parametric estimation of multiple embeddings per word in vector space. In: Moschitti, A., Pang, B., Daelemans, W. (eds.) Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, a meeting of SIGDAT, a Special Interest Group of the ACL, pp. 1059–1069. ACL (2014)
22. Oh, S.J., Gallagher, A.C., Murphy, K.P., Schroff, F., Pan, J., Roth, J.: Modeling uncertainty with hedged instance embeddings. In: International Conference on Learning Representations (2019). https://openreview.net/forum?id=r1xQQhAqKX
23. Pan, J.H., Gao, J., Zheng, W.S.: Action assessment by joint relation graphs. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6331–6340 (2019)
24. Parmar, P., Morris, B.T.: What and how well you performed? A multitask learning approach to action quality assessment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 304–313 (2019)
25. Parmar, P., Tran Morris, B.: Learning to score Olympic events. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 20–28 (2017)
26. Pirsiavash, H., Vondrick, C., Torralba, A.: Assessing the quality of actions. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8694, pp. 556–571. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10599-4_36
27. Prassas, S., Kwon, Y.H., Sands, W.A.: Biomechanical research in artistic gymnastics: a review. Sports Biomech. **5**(2), 261–291 (2006)
28. Qiu, Y., Wang, J., Jin, Z., Chen, H., Zhang, M., Guo, L.: Pose-guided matching based on deep learning for assessing quality of action on rehabilitation training. Biomed. Sig. Process. Control **72**, 103323 (2022)
29. Santoro, A., et al.: A simple neural network module for relational reasoning. Adv. Neural Inf. Process. Syst. **30** (2017)
30. Scarselli, F., Gori, M., Tsoi, A.C., Hagenbuchner, M., Monfardini, G.: The graph neural network model. IEEE Trans. Neural Netw. **20**(1), 61–80 (2008)
31. Schoeffmann, K., Taschwer, M., Sarny, S., Münzer, B., Primus, M.J., Putzgruber, D.: Cataract-101: video dataset of 101 cataract surgeries. In: César, P., Zink, M., Murray, N. (eds.) Proceedings of the 9th ACM Multimedia Systems Conference, MMSys 2018, Amsterdam, The Netherlands, June 12-15, 2018, pp. 421–425. ACM (2018)
32. Shi, Y., Jain, A.K.: Probabilistic face embeddings. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6902–6911 (2019)
33. Sohn, K., Lee, H., Yan, X.: Learning structured output representation using deep conditional generative models. Adv. Neural Inf. Process. Syst. **28** (2015)

34. Sun, J.J., Zhao, J., Chen, L.C., Schroff, F., Adam, H., Liu, T.: View-invariant probabilistic embedding for human pose. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12350, pp. 53–70. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58558-7_4
35. Tang, Y., et al.: Uncertainty-aware score distribution learning for action quality assessment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9839–9848 (2020)
36. Tishby, N.: The information bottleneck method. In: Proceedings of the 37th Allerton Conference on Communication and Computation, 1999 (1999)
37. Vaswani, A., et al.: Attention is all you need. Adv. Neural Inf. Process. Syst. **30** (2017)
38. Vilnis, L., McCallum, A.: Word representations via Gaussian embedding. In: International Conference on Learning Representations (2015)
39. Wang, L., et al.: Temporal segment networks: towards good practices for deep action recognition. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) Computer Vision – ECCV 2016. ECCV 2016. LNCS, vol. 9912. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46484-8_2
40. Wang, S., Yang, D., Zhai, P., Chen, C., Zhang, L.: TSA-NET: tube self-attention network for action quality assessment. In: Proceedings of the 29th ACM International Conference on Multimedia, pp. 4902–4910 (2021)
41. Waters, T.R., Putz-Anderson, V., Garg, A.: Applications Manual for the Revised NIOSH Lifting Equation (1994)
42. Xiao, F., Sigal, L., Jae Lee, Y.: Weakly-supervised visual grounding of phrases with linguistic structures. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5945–5954 (2017)
43. Xu, C., Fu, Y., Zhang, B., Chen, Z., Jiang, Y.G., Xue, X.: Learning to score figure skating sport videos. IEEE Trans. Circuits Syst. Video Technol. **30**(12), 4578–4590 (2019)
44. Xu, J., Rao, Y., Yu, X., Chen, G., Zhou, J., Lu, J.: FineDiving: a fine-grained dataset for procedure-aware action quality assessment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2949–2958 (2022)
45. Xu, K., Li, J., Zhang, M., Du, S.S., ichi Kawarabayashi, K., Jegelka, S.: What can neural networks reason about? In: International Conference on Learning Representations (2020). https://openreview.net/forum?id=rJxbJeHFPS
46. Yan, S., Xiong, Y., Lin, D.: Spatial temporal graph convolutional networks for skeleton-based action recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32 (2018)
47. Yu, X., Rao, Y., Zhao, W., Lu, J., Zhou, J.: Group-aware contrastive regression for action quality assessment. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 7899–7908. IEEE Computer Society, Los Alamitos, CA, USA (2021)
48. Zhang, B., Chen, J., Xu, Y., Zhang, H., Yang, X., Geng, X.: Auto-encoding score distribution regression for action quality assessment. Neural Comput. Appl. **36**(2), 929–942 (2023)
49. Zhang, J., Bargal, S.A., Lin, Z., Brandt, J., Shen, X., Sclaroff, S.: Top-down neural attention by excitation backprop. Int. J. Comput. Vis. **126**(10), 1084–1102 (2018)
50. Zhou, C., Huang, Y.: Uncertainty-driven action quality assessment. arXiv preprint arXiv:2207.14513 (2022)

51. Zhou, K., Ma, Y., Shum, H.P.H., Liang, X.: Hierarchical graph convolutional networks for action quality assessment. IEEE Trans. Circ. Syst. Vid. Technol. **33**(12), 7749–7763 (2023)
52. Zhu, Y., Zhou, Y., Ye, Q., Qiu, Q., Jiao, J.: Soft proposal networks for weakly supervised object localization. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1841–1850 (2017)