RICA²: 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 RICA²—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 [34], manual material handling [49], and surgery [25]. These 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 [21] following a pre-defined criterion. Ratings for individual factors, sometimes complemented with impression-based



Fig. 1: RICA² integrates score rubric used by human experts and accounts for prediction uncertainty, resulting in *accurate* predictions and *calibrated* uncertainty estimates.

global ratings, are then summarized into a final quality score [25]. 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.

There is a burgeoning interest in the vision community to develop video-based AQA [32, 33, 42, 51]. While current solutions have made steady progress across benchmarks [11, 31, 52], 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, 52, 56]. 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 [13]. 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 [42,57,59,60]. 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 RICA² (Rubric-informed, Calibrated Assessment of Actions), yields accurate action scores with additional uncertainty estimates.

We evaluate RICA² on several public AQA datasets, covering sports and surgical videos. Particularly, RICA² establishes new state of the art on FineDiving [52], MTL-AQA [31] and JIGSAWS [11]. On FineDiving [52] – the largest and most challenging AQA benchmark, RICA² 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 [17]). Similarly, on MTL-AQA [31], the most commonly used dataset for AQA, RICA² attains state-of-the-art SRCC, and again largely improved calibration (a gain of +0.444 in Kendall Tau). On JIGSAWS [11], RICA² 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 RICA².

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 RICA² achieves stateof-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 [12,33] employed handcrafted features to estimate quality scores in videos. More recent methods developed various deep models, including convolutional [42,56], graph [30], recurrent [32,51], and Transformer [2,52] networks. AQA has also been widely considered in surgical education [22], rehabilitation [35], and ergonomics [5].

Recently, exemplar-based methods [2, 52, 56] 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 RICA² 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 [42, 57, 59, 60]. For example, Tang et al. [42] 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 [57,59,60]. 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 RICA² 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 [26]. IRIS incorporates score rubric into a convolutional network for AQA. This is done by segmenting key steps in the

A. Majeedi et al.

4

video and predicting sub-scores for individual steps. Similar to IRIS, RICA² also considers rubric in a deep model. However, RICA² adapts a graph network, treats sub-scores as latent embeddings, predicts the final score, and further quantifies prediction uncertainty. These differences allow RICA² 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 [27,45], treats each embedding as a distribution. This approach has gained recent attention for modeling uncertainty in deep models. Oh et al. [28] 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 [39], age estimation [20], pose estimation [41], and cross-modal retrieval [6]. Another related line of work is the conditional variational autoencoder [40], 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 [9,19,37] offer a powerful tool to leverage the relational inductive bias inherent in data [3,53]. This inductive bias is beneficial to aggregate a global representation from a group of local ones [36]. Recently, Zhou et al. [60] 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 subaction for successfully executing the action. Further, s is associated with a text 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 [4,10] (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 [25, 34, 49]. 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

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

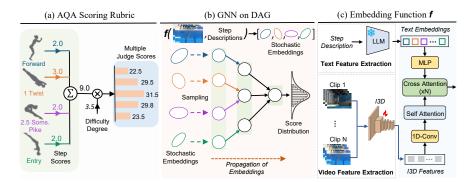


Fig. 2: Overview of RICA². Leveraging scoring rubrics (a), RICA² integrates a graph representation of action step and rubric with uncertainty modeling (b). Specifically, RICA² takes an input of the video and its key action steps, encodes the input into embeddings (c), refines the embeddings through a deep probabilistic model, and outputs an action score in tandem with its uncertainty estimate.

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 RICA²— a deep probabilistic model for AQA that leverages known action steps and incorporates the scoring rubric for modeling. Importantly, RICA² accounts for prediction uncertainty, *i.e.*, when the model prediction can and cannot be trusted. Fig. 2 presents an overview of RICA². It consists of two main model components: (a) a graph neural network that integrates the key steps and scoring rubric (Sec. 3.1); and (b) stochastic embeddings defined on the graph to capture prediction uncertainty (Sec. 3.2), coupled with (c) a learning scheme under the variational information bottleneck framework (Sec. 3.3). In what follows, we delve into the details of RICA².

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. Fig. 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:

$$f: X, \mathcal{G} \mapsto \{Z^s\}; \quad h: \{Z^s\}, \mathcal{G} \mapsto Y,$$
 (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 Y^r . To this end, we propose an extension of graph neural networks (GNNs), in which stochastic embeddings Z^s are propagated from leaf nodes Y^s to the root node Y^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 $S(\cdot)$ 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 RICA² 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 [43], leading to the maximization of the following objective

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

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

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

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

$$\mathcal{L}_{KL} = \Sigma_{s \in \mathbb{S}} KL\left(\mathcal{N}(\mu^s(x), \Sigma^s(x) || \mathcal{N}(0, I)\right)$$

$$= \frac{1}{2} \Sigma_{s \in \mathbb{S}} \Sigma_{j=1}^D \left((\mu_j^s)^2 + (\sigma_j^s)^2 - \log(\sigma_j^s)^2 - 1 \right),$$
(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 [18] 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 [20,28] we generate uncertainty scores by summing up the harmonic means of the predicted variances for individual steps

uncertainty(Y) =
$$\Sigma_{s \in \mathbb{S}} D / \Sigma_{j=1}^{D} (\sigma_{j}^{s})^{-1}$$
, (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 RICA² 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 RICA²†. Without stochastic embeddings, RICA²† is unable to estimate prediction uncertainty, yet often yields slightly lower prediction errors. This trade-off is also observed in prior works [6, 20, 39]. 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 [8] (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 RICA². 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 [44] (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 [52], MTL-AQA [31], and JIGSAWS [11] in the main paper. In supplement Sec. B, we also include results on the Cataract-101 [38] 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 [2,42,52], 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 [20, 28]. 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 [20, 28] in employing Kendall's tau (τ) [17], 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. RICA² is benchmarked against a set of strong baselines, including exemplar-free methods such as DAE [57], USDL and MUSDL [42], and exemplar-based ones such as CoRE [56], TPT [2] and TSA [52]. We further include the deterministic version of our model RICA²†, 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 [42] 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 [57] 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 RICA² 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 RICA², 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 [56], TPT [2] and TSA [52] use the diving number (DN) encoding steps and their ordering. Further, TSA [52] 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 [52] 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 [52], we do not use these annotations for RICA².

Experiment setup. We adhere to the experimental setup of the most recent baseline [52] using their train-test split, with 2251 videos for training and 749 videos for testing. We follow the input video settings used in [52] for RICA² 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. Tab. 1a presents our results on FineDiving. Both our stochastic and deterministic versions (RICA² and RICA²†) outperform the state-of-the-art TPT [2], an exemplar-based method. RICA² shows a relative margin of 0.7% / 1.4% on SRCC / $R\ell_2$, and RICA²† 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 RICA²† has slightly higher accuracy, our stochastic RICA² demonstrates superior calibration of its uncertainty estimate ($\tau_{RICA}^2 = 0.64$ vs. $\tau_{TPT} = -0.56$). Fig. 3a further shows uncertainty calibration results. Uncertainty estimates from RICA² have a clear upward trend, indicating a higher calibration level. While MUSDL [42] also exhibits a reasonable level of calibration ($\tau_{MUSDL} = 0.47$ vs. $\tau_{RICA}^2 = 0.64$), the errors are significantly higher than RICA² across all uncertainty levels.

4.2 Results on MTL-AQA

Dataset. MTL-AQA [31] 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.

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

		•			•							
	(a) Results o	n FineDi	ving			(b) Results on MTL-AQA						
		1	Metrics					1	Metrics			
		$\overline{SRCC(\uparrow)}$		$\tau(\uparrow)$				$SRCC(\uparrow)$	$R\ell_2(\downarrow)$	$\tau(\uparrow)$		
	CoRe [56]	0.9061	0.3615	- (1)			TSA-Net [48]	0.9422	-	-		
Exemplar	TSA [52]	0.9203	0.3420	_		Exemplar	CoRe [56]	0.9512	0.2600	-		
based	TPT [2]	0.9333		-0.5556	based		DAE-CoRe [57]	0.9589	-	-		
							TPT [2]	0.9607	0.2378	-0.1111		
	USDL [42,52]	0.8913		0.3778		•	C3D-AVG-MTL [31]					
	MUSDL [42, 52]	0.8978	0.3704				USDL [42]	0.9231		0.1556		
Exemplar	F 1	0.8820		-0.1999		Exemplar	MUSDL [42]	0.9273	0.4510	-0.0667		
free	DAE-MT [57]	0.9285	0.3320	-0.4667		free	DAE [57]	0.9231	0.0720	0.4000		
	RICA ² (Ours)	0.9402	0.2838	0.6444			DAE-MT [57] RICA ² (Ours)	0.9490 0.9594		-0.4222 0.6000		
	RICA ² † (Ours)	0.9421	0.2600	-			RICA (Ours)	0.9620				
Mean Absolute error 6	TPT (t=0.56 USDL (t=0.3	8) Ours	DL (τ=0.47) (τ=0.64)			Mean Absolute error 6	USDL (T=0.16)	MUSDL (τ=-0.07 Durs (τ=0.60)				
() :		ertainty → ⁶	8			0	Uncertainty -		8 Tr. 10			
(a)	Uncertainty cali	bration of	on Finel	Jiving	(b) Uncer	tainty calibratio	on on M'	TL-AQ	A		

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, RICA² has improved calibration with lower prediction errors.

Experiment setup. We follow the evaluation protocol of [2,51,56], 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. Tab. 1b summarizes our results on MTL-AQA. Similar to FineDiving, RICA² shows state-of-the-art results on MTL-AQA across all evaluation metrics. Specifically, our RICA²† 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 $SRCC_{RICA^2\dagger} = 0.9620$) and $R\ell_2$ (+4.1% relative margin). Again, compared to previous methods, our stochastic model shows improved calibration ($\tau_{RICA^2} = 0.60$ vs. $\tau_{USDL} = 0.16$ vs. $\tau_{TPT} = -0.11$) as shown in Fig. 3b.

Table 2: Results on JIGSAWS [11] dataset. Only prediction accuracy (SRCC) is considered due to the limited sample size. RICA² outperforms all prior approaches.

			Ta	sk	
		\mathbf{S}	NP	KT	Avg
Exemplar	CoRe [56]	0.84	0.86	0.86	0.85
based	TPT [2]	0.88	0.88	0.91	0.89
	ST-GCN [55]	0.31	0.39	0.58	0.43
	TSN [47]	0.34	0.23	0.72	0.46
	JRG [30]	0.36	0.54	0.75	0.57
Exemplar	USDL [42]	0.64	0.63	0.61	0.63
free	MUSDL [42]	0.71	0.69	0.71	0.70
	DAE [57]	0.73	0.72	0.72	0.72
	DAE-MT [57]	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

4.3 Results on JIGSAWS

Dataset. In addition to diving videos, we also evaluate RICA² on JIGSAWS [11] — 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 [42], and the input video setting from [2]. Specifically, for each video, we uniformly sample 160 frames which are segmented into 20 non-overlapping 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. Tab. 2 summarizes our results on JIGSAWS. Similar to previous datasets, our models exhibit notable advancements over the previous state-of-the-art model TPT [42], showcasing substantial improvements of 1.1% (RICA²) to 3.4% (RICA² \dagger) 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% (RICA²) to 21.0% (RICA² \dagger) relative gain in average SRCC compared to the latest method DAE-MT [57].

4.4 Ablation Studies

To understand our model design choices, we conduct ablation studies on the MTL-AQA [31] 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

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

Step	DAG	<i>C</i>	<i>C</i> .	Metrics								
Rep.	(Rubric)	\mathcal{L}_{KL}	\mathcal{L}_{Aux}	$\overline{SRCC}(\uparrow)$	$R\ell_2(\downarrow)$	$oldsymbol{ au}(\uparrow)$	Avg. Rank (\downarrow)					
Random	×	×	×	0.9426	0.3882	-	5.50					
Text	×	×	×	0.9431	0.3509	-	4.50					
Text	\checkmark	×	×	0.9430	0.3336	-	4.50					
Text*	\checkmark	×	×	0.9437	0.3335	-	3.50					
Text*	\checkmark	\checkmark	×	0.9448	0.3329	0.4222	1.83					
Text^*	\checkmark	\checkmark	\checkmark	0.9460	0.3303	0.4222	1.17					

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 RICA² and study their effects. Tab. 3 presents our results using the same features and training epochs, with our base model in row 1.

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 $R\ell_2$ (Tab. 3 row 1 vs. row 2), by leveraging knowledge encoded in the LLM [8]. Further, we find that freezing the text embeddings leads to comparable results and faster convergence (Tab. 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 $R\ell_2$ (Tab. 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 RICA², we now examine the cross-attention maps between the step representations and video features in our learned embedding function f. Fig. 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 RICA² learns to encode the temporal location of individual steps.

We further evaluate this *localization* ability following the Pointing game protocol [58], widely considered in weakly supervised / unsupervised localization tasks [50,61]. Pointing Game compares a generated heatmap with an annotated

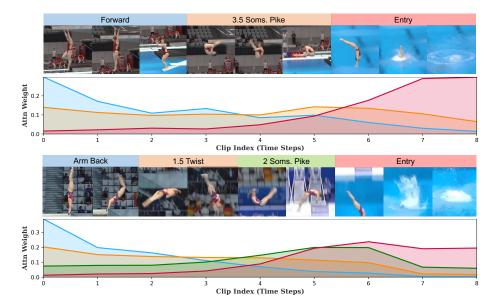


Fig. 4: Visualization of the cross-attention maps. Y-axis: attention value; Y-axis: clip indices (time). Each curve shows an attention map from a step representation to the temporal video features. The frames shown above are aligned with the timing of the corresponding attention plot. Curves and steps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

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 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.

Supplement

In this supplement, we describe (1) technical, implementation, and experiment details for individual datasets, as well as our loss function (Sec. A); (2) additional results for surgical skill assessment on Cataract-101 dataset (Sec. B); (3) additional ablations including the study of architectural design and loss coefficients, choice of text embeddings and additional cross-attention plots, and the consideration of action recognition methods (Sec. C); (4) details of the scoring rubric considered in the model, with examples from FINA diving manual (Sec. D); and (5) further discussion of our work (Sec. E). We hope this document complements our paper.

For sections, figures, and tables, we use numbers (e.g., Sec. 1) to refer to the main paper and capital letters (e.g., Sec. A) to refer to this supplement.

A Technical, Implementation, and Experiment Details

We describe implementation and experiment details for 4 datasets considered in our paper, including FineDiving [52], MTL-AQA [31], JIGSAWS [11], and Cataract-101 [38]. We further present technical details of our loss functions.

A.1 Details for FineDiving

We follow the implementation from TSA [52] to process the input videos. Specifically, we uniformly sample 96 frames from each video, segmented into 9 overlapping clips of 16 consecutive frames, with a stride of 10. The frames are resized to a resolution of 200×112 . During training, we employ a random crop of 112×112 , while a center crop of size 112×112 is performed during testing.

We use the I3D backbone [4] to extract video features, following prior works [2, 52]. We generate text descriptions of steps with the help of GPT-4 [29], as shown in Table J. These descriptions are embedded using Flan-T5 XXL [7, 54].

For training our model, we utilize AdamW [24] optimizer with a linear warmup for 5 epochs and a total of 350 epochs. The training batch size is set to 8, while the learning rates for the I3D backbone, transformer blocks, and the head (DAG) are set to 1×10^{-5} , 3×10^{-5} , and 5×10^{-4} , respectively. We also experimented with learning rate decay, yet did not find it helpful with our long training schedule.

A.2 Details for MTL-AQA

We follow TPT [2] for processing the input videos. Specifically, we uniformly sample 103 frames from each video, segmented into 20 overlapping clips of 8

consecutive frames with a stride of 5. The frames are resized to a resolution of 455×256 . During training, we employ a random crop of 224×224 , while a center crop of size 224×224 is performed during testing.

Again, we use the I3D backbone [4] to extract video features, generate text descriptions of individual steps with the help of GPT-4 [29] (see Table I), and further embed these descriptions using Flan-T5 XXL [7,54]

For training, we use the AdamW [24] optimizer with a linear warmup for 5 epochs and a total of 350 epochs. Other hyperparameters are kept the same as for FineDiving (batch size of 8 with learning rates for I3D backbone, transformer blocks, and the head (DAG) as 1×10^{-5} , 3×10^{-5} , and 5×10^{-4} , respectively).

A.3 Details for JIGSAWS

We adopt the video input configuration from TPT [2]. We uniformly sample 160 frames from each video, and these frames are split into 20 non-overlapping clips of 8 consecutive frames with a stride of 8. The frames are resized to a resolution of 455×288 . During training, we employ a random crop of size 224×224 , while during testing, a center crop of the same size (224×224) is performed.

We follow the same protocol to extract video features (using I3D) and step embeddings (using Flan-T5 XXL), as FineDiving and MTL-AQA experiments. Step descriptions generated with the help of GPT-4 are shown in Table K.

For training, we utilize AdamW [24] optimizer with a linear warmup for the initial 5 epochs and a total of 350 epochs. Due to the smaller size of the dataset, a small training batch size of 2 is employed. The learning rates for the I3D backbone, transformer blocks, and the head (DAG) are set to 1×10^{-5} , 3×10^{-5} , and 5×10^{-4} , respectively, maintaining consistency with the previous settings.

We note that in the JIGSAWS dataset, the steps are not performed in any specific order. Moreover, steps are repeated multiple times within each video. Consequently, we do not employ the auxiliary losses (\mathcal{L}_{Aux}) of ranking and sparsity for our experiments on JIGSAWS.

A.4 Cataract-101

Videos in this dataset record complex surgical procedures (cataract surgery) and are thus significantly longer than other datasets (10 minutes vs. a few seconds). Consequently, we consider a stronger video backbone — SlowFast-8 \times 8-R50 [10]. We split an input video into sliding windows (clips) of 64 frames with a temporal stride of 16 frames. For each clip, we resize all frames such that their shortest side is kept at 256. After a center-crop of size 224×224 , the clips are fed into the SlowFast model pre-trained on Kinetics. The extracted video features are further used as input to our model for training and inference. It is worth noting that we do not update the SlowFast model during training, and thus, these video features remain fixed.

We follow the same protocol to extract step embeddings (using Flan-T5 XXL), with step descriptions in Table L. We utilize AdamW [24] optimizer with a linear

warmup for 5 epochs and a total of 350 epochs for training. The batch size is 2 and the learning rate is 1×10^{-4} .

A.5 Details of Loss Functions

As discussed in Sec. 3.3 and 3.4, our loss function consists of the VIB loss (\mathcal{L}_{VIB}) and the auxiliary loss (\mathcal{L}_{Aux}). Further, \mathcal{L}_{VIB} comprises the MSE loss (\mathcal{L}_{MSE}) and the KL loss (\mathcal{L}_{KL}), and \mathcal{L}_{Aux} includes the ranking loss (\mathcal{L}_{rank}) and the sparsity loss ($\mathcal{L}_{sparsity}$). Our overall loss is thus given by

$$\mathcal{L} = \mathcal{L}_{MSE} + \beta \mathcal{L}_{KL} + \gamma \left(\mathcal{L}_{rank} + \mathcal{L}_{sparsity} \right), \tag{9}$$

where β and γ are the loss weights. β controls the bottleneck effect, and γ manages additional regularization (e.g., temporal ordering).

The auxiliary losses of ranking (\mathcal{L}_{rank}) and sparsity $(\mathcal{L}_{sparsity})$ are designed to enforce the ordering of the step representation. Specifically, we follow TPT [2] and consider the last cross-attention map from our embedding function f. Concretely, given S step representations as the queries, i.e. $Q \in \mathbb{R}^{S \times D}$, and video features defined over T time steps as the keys, i.e. $K \in \mathbb{R}^{T \times D}$, we denote their cross-attention map as $A \in \mathbb{R}^{S \times T}$. Considering the cross-attention map for each step (i.e. a row in A), a corresponding temporally-weighted center $(\bar{\alpha}_s)$, as a detector of the step's temporal location, is calculated as

$$\bar{\alpha}_s = \Sigma_{t=1}^T t \cdot A_{s,t},\tag{10}$$

where $\alpha_{s,t}$ is the similarity between a step representation and a video feature at time step t. Our auxiliary losses are defined on top of these centers.

The sparsity loss $\mathcal{L}_{sparsity}$ is defined to discourage the spread of query attention densities. $\mathcal{L}_{sparsity}$ is given by

$$L_{sparsity} = \sum_{s=1}^{S} \sum_{t=1}^{T} |t - \bar{\alpha}_s| \cdot \alpha_{s,t}$$
(11)

The ranking loss \mathcal{L}_{rank} is designed to encourage all centers to follow the pre-specified step ordering. \mathcal{L}_{rank} is written as

$$L_{rank} = \Sigma_{s=1}^{S-1} \max(0, \bar{\alpha}_s - \bar{\alpha}_{s+1} + m) + \max(0, 1 - \bar{\alpha}_1 + m) + \max(0, \bar{\alpha}_S - T + m)$$
(12)

B Results on Cataract-101

We now present our results on Cataract-101, a video dataset for surgical skill assessment. These results are omitted from the main paper due to lack of space.

Dataset. Cataract-101 [38] is a publicly available dataset of 101 cataract surgery videos, with each recorded procedure annotated with frame-level labels and the performing surgeon's corresponding skill score. The skill scores are discrete, labeled as either expert or novice.

Experiment setup. We train our model on 80 randomly picked videos and test on the remaining 21 videos. Due to the significantly longer duration of videos

Table A: Ablation on loss weight β . We report the accuracy and calibration metrics.

	Metrics									
$oldsymbol{eta}$	$\overline{SRCC(\uparrow)}$	$R\ell_2(\downarrow)$	$ au(\uparrow)$							
10^{-1}	0.9450	0.3614	0.5556							
10^{-2}	0.9456	0.3529	0.4667							
10^{-3}	0.9463	0.3478	0.3333							
10^{-4}	0.9451	0.3289	0.3333							
10^{-5}	0.9449	0.3393	0.4222							

(exceeding 10 minutes compared to a few seconds for diving videos), we perform our experiments with pre-extracted features. In this case, we choose a more advanced model for feature extraction. Specifically, we utilize SlowFast [10] model pre-trained on Kinetics [16] to extract features for the videos and train our model on the extracted features. Given the binary labels, we switch from the mean squared error (MSE) loss to binary cross entropy (BCE) loss, which is compatible with our maximum log-likelihood interpretation of the loss.

We report the average accuracy as our evaluation metric and compare our results to a baseline method TUSA [23] specifically designed for surgical skill assessment. TUSA is trained and evaluated using the same features as our method.

Results and discussion. Both TUSA and our method reach an impressive 100% accuracy. Randomizing train-test splits leads to similar perfect accuracy. Our analysis of this dataset shows the duration of the surgery is a good predictor of the surgeon's skills; expert surgeons often perform this routine surgery faster than novice surgeons.

C Additional Ablation Studies

We further discuss additional ablation studies. Similar to the ablations in the main paper (Sec. 4.4), we run the experiments on the MTL-AQA [31] dataset with the same pre-extracted I3D features unless otherwise specified.

C.1 Effects of Loss Coefficients

We first evaluate the effects of loss weights β and γ . To this end, we fix γ and vary β , which balances the objectives of minimizing prediction error and ensuring uncertainty estimation. Table A shows the results. Lower values of β often lead to minorly improved accuracy metrics, yet higher values of β allow better calibration. Through the experiments, we empirically observe that an optimal balance can be attained with a training schedule in which β is initially set to 10^{-5} and incrementally increased to a maximum value of $\beta = 0.005$ during the training process. All results for RICA² in our study are reported using this annealing scheme, which was also discussed in prior work [18].

Table B: Ablation on design choices. We vary the design of major components in RICA^2 and report the results on MTL-AQA. $\#\mathrm{Convs}$ denotes the number of 1D convolution blocks, $\#\mathrm{Enc}$ denotes number of encoder blocks and $\#\mathrm{Dec}$ denotes the number of decoder blocks

Embed	#Convs	#Enc	#Dog	DAG	c	<i>c</i> .	C		M	letrics	
\mathbf{Dim}	#Convs	#Enc	#Dec	(Rubric)	\mathcal{L}_{spar}	Lrank	LKL	$\overline{SRCC}(\uparrow)$	$R\ell_2(\downarrow)$	$oldsymbol{ au}(\uparrow)$	Avg. Rank (\downarrow)
256	2	0	0	×	×	×	×	0.9267	0.4213	-	6.67
512	2	2	0	×	×	×	×	0.9263	0.3843	-	6.33
512	2	2	0	×	×	×	×	0.9272	0.4090	-	6.00
512	2	2	2	×	×	×	×	0.9444	0.3707	-	4.33
512	2	2	2	\checkmark	×	×	×	0.9437	0.3335	-	4.00
512	2	2	2	\checkmark	\checkmark	×	×	0.9440	0.3562	-	4.33
512	2	2	2	✓	✓	\checkmark	×	0.9451	0.3458	-	3.33
512	2	2	2	\checkmark	✓	✓	\checkmark	0.9460	0.3303	0.4222	1.00

We further experiment with different values for γ , which denotes the weight for the auxiliary losses of ranking and sparsity. We empirically find our results are insensitive to different values and set $\gamma = 0.1$ for our experiments.

C.2 Design Choices

We vary the design of our model and examine their contributions to final performance. Specifically, our model extracts video features and step embeddings, further encodes individual features, uses cross-attention Transformer blocks to fuse them and decode step representation, and finally decodes a distribution of scores using a DAG representing the rubric. We study the design of encoding (convolution and self-attention) and decoding modules (cross-attention), as well as the choice of loss terms.

Table B shows the results. In addition to what is presented in the main paper, we highlight some key choices to boost performance in Table B: (a) using embedding dim of 512 over 256 and utilizing 2 decoder blocks over 1 decoder block; and (b) combining all loss terms. We can clearly notice our full model (shown in the last row of Table B) leads to the best performance in accuracy and score calibration. Specifically, this model combines two 1D-conv blocks, 2 self-attention blocks, integration of text queries, and 2 decoder blocks, followed by the DAG, and is optimized with \mathcal{L}_{KL} and the auxiliary losses.

C.3 Effects of Text Queries

To assess the influence of various language models, we leverage text embeddings derived from three different models: OpenClip [14], $E5_{large}$ [46], and Flan-T5 XXL [7,54]. The corresponding results are detailed in Table C. Our empirical findings indicate that embeddings derived from Flan-T5 XXL [7,54] result in the best performance on accuracy metrics (SRCC and $R\ell_2$).

Table C: Ablation on text embeddings. We experiment with different language models and report results on MTL-AQA. *Dim* denotes the text embedding dimensions.

Text Model	Dim			
Text Model	Dilli	$SRCC(\uparrow)$	$R\ell_2(\downarrow)$	$\tau(\uparrow)$
OpenClip [14]	512	0.9455	0.4339	0.2444
$E5_{large}$ [46]	1024	0.9431	0.3410	0.4222
Flan-T5 XXL [7]	4096	0.9460	0.3303	0.4222

C.4 What if Steps Are Not Available?

When step information is not available at inference time, action recognition methods offer a solution for identifying steps in the input video. To explore this scenario, we experimented on the FineDiving dataset [52], utilizing pre-extracted I3D features. For this task, we designed a simple action recognition model comprising a 2-layer MLP projection block, 2 Transformer blocks, and 2 MLPs. This model was trained on the training split to predict the dive number (i.e. the steps in a video). Subsequently, we evaluated the trained model on the test set, achieving an average step recognition accuracy of 82%. The generated step predictions on the test set were then incorporated into RICA², replacing the ground truth step presence information. Using the model-predicted step information resulted in an SRCC and $R\ell_2$ of 0.9379 and 0.2779 respectively while the ground truth SRCC and $R\ell_2$ are 0.9389 and 0.2750. These results demonstrate the feasibility of leveraging action recognition methods to provide step information for RICA² during inference.

C.5 Additional Visualization of Our Results

We present additional visualization of sample results in Fig. A. These samples are from the test set of FineDiving, and the visualization follows the same format as Fig. 4 of our main paper.

D Score Rubric and Our DAG Representation

A key component of RICA² is the integration of a scoring rubric in the form of a directed acyclic graph (DAG). As mentioned in Sec. 3, we assume that each video contains a set of steps, and each action step is independently scored. Subsequently, a rule-based rubric is employed to aggregate individual scores and calculate a final quality score, in which steps might be grouped into intermediate stages. Our DAG representation applies to a broad range of technical skill assessment tools.

We use the FINA diving manual [15] as an example to further illustrate our representation. For an individual dive, a sequence of action items is predetermined, with various combinations leading to different difficulty degrees (see examples in Table G). A panel of 7 to 11 judges will assess the dive, each assigning a score on a scale of 0-10. Judges will follow specified guidelines to evaluate the performance

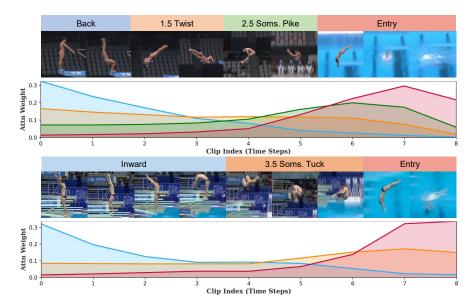


Fig. A: Visualization of the cross-attention maps on two test videos from Fine-Diving [52] with the Y-axis as attention value and the X-axis as clip indices (time). Each curve shows an attention map from a step representation to the temporal video features. The frames shown are aligned with the time axis of the corresponding attention plot. Curves and steps are colored accordingly

of the approach and takeoff (Table D), the flight (Table E), and the entry into the water (Table F) to arrive at a final score. Other considerations e.g. technique, execution and overall performance may be taken into account. The top two and the bottom two scores are discarded. The remaining scores are summed up and multiplied by the difficulty degree, yielding the final rating. Table H illustrates a completed diving sheet from a real diving event.

We note that RICA² does not employ the detailed exact rubric of FINA. Instead, our method follows its general structures by assuming key steps, the scoring of these steps, and the combination of individual scores — a central concept in technical skill assessment [25, 34, 49].

E Further Discussions

E.1 Benchmark Settings

We compare the setting of the baseline exemplar-based methods. Exemplar-based methods rely on exemplar videos and their corresponding scores during both the training and testing phases. Latest methods, such as TPT [2], TSA [52], and CoRE [56], utilize dive numbers to select the exemplar videos and their associated quality scores. Notably, these dive numbers uniquely encode the presence and sequence of steps being executed, as shown in Table 7 of FineDiving [52]. Moreover,

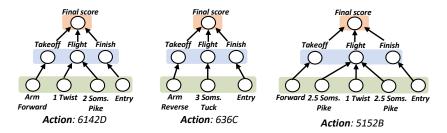


Fig. B: Sample DAGs for three diving actions in Finediving.

TSA incorporates not only the step presence but also precise start and end time stamps of each performed step.

In sharp contrast, our method takes an input of step presence and ordering during training and inference, without the need for exemplar videos, their scores or timestamp data. Different from prior methods, our method additionally considers text descriptions of steps, obtained by intuitively expanding step names into simple sentences as shown in Tables I, J, and K. While our method assumes a pre-specified scoring rubric, this rubric is incorporated into our model design, and not as part of the input.

E.2 Trustworthy Visual Recognition

Over the last decade, we have witnessed major advances in visual recognition with deep learning, leading to significantly improved results on public datasets. Despite the superior performance of these deep models, one remaining question is how users can trust their output, knowing that mistakes can be made by these models. This trustworthy visual recognition has multiple facets. We argue two of the key aspects are to provide credible confidence of the output and to consider a human-interpretable decision-making process. Our work in this paper takes a step towards these two critical aspects while addressing the challenging problem of video-based AQA. We presented a deep model for AQA that integrates the human score rubric and models the output confidence. Our method could facilitate high-stakes applications of AQA, including competitive sports and healthcare, where understanding and trusting the model's decisions are crucial, and low-confident samples can be readily passed to human experts.

E.3 Ethical Concerns

Our work focuses on the technical aspect of action quality assessment, and as such we do not anticipate major ethical concerns.

Table D: Guidelines for judging the entry into water [15].

Fault	Range of deduction	Comments			
Unbalance take-off	½ - 2 points				
Improper angle of take-off	½ - 2 points				
Armstand unbalance position	½ - 2 points				
Armstand no balance at all		Deduction 2 points			
Armstand no control, no vertical		Deduction 1 point			

Table E: Guidelines for judging the approach and take-off [15].

Fault	Range of deduction
Insufficient height	½ to 2 points
Dive is too close to the platform (but does not hit the platform)	½ to 2 points (according to opinion)
Dive is unsafely close to the platform with the head (but does not hit the platform)	2 maximum award
Dive hits the platform with feet or hands (does not affect the dive)	4 ½ maximum award
Dive hits the platform with the head	2 maximum award

Table F: Guidelines for judging the flight [15].

Fault	Range of deduction	Comments
Dive not vertical on entry	Judge's discretion	
Dive which is more than 5 degrees off vertical		It can't be classified as very good dive
Dive which is more than 35 degrees off vertical		Is classified as deficient dive or lower (2 ½ - 4 ½)
Dive twisted on the entry	Judge's discretion	
Dive twisted on the entry more or less than 5°		It can't be classified as very good or excellent dive
Dive twisted on the entry more or less than 15°		Is classified as deficient dive or more (2 ½ - 4 ½)
Dive twisted on entry more or less than 35°		Is classified as deficient dive or lower (2 ½ - 4 ½)
Dive twisted on entry more or less than 90°	Failed dive	
Arms are not in the correct position at the entry	½ - 2 points	
Arms are above the shoulders at the entry	4 ½ maximum	

Table G: How the difficulty degree for a dive is determined.

	SPRINGBOARD		1 M	ETER		3 METER				
	SPRINGBOARD	STR	PIKE	TUCK	FREE	STR	PIKE	TUCK	FREE	
Forward Group		A	В	С	D	Α	В	С	D	
101	Forward Dive	1.4	1.3	1.2	-	1.6	1.5	1.4	-	
102	Forward Somersault	1.6	1.5	1.4	-	1.7	1.6	1.5	-	
103	Forward 1½ Somersaults	2.0	1.7	1.6	-	1.9	1.6	1.5	-	
104	Forward 2 Somersaults	2.6	2.3	2.2	-	2.4	2.1	2.0	-	
105	Forward 2½ Somersaults	-	2.6	2.4	-	2.8	2.4	2.2	-	
106	Forward 3 Somersaults	-	3.2	2.9	-	-	2.8	2.5	-	
107	Forward 3½ Somersaults	-	3.3	3.0	-	-	3.1	2.8	-	
108	Forward 4 Somersaults	-	-	4.0	-	-	3.8	3.4	-	
109	Forward 4½ Somersaults	-	-	4.3	-	-	4.2	3.8	-	
112	Forward Flying Somersault	-	1.7	1.6	-	-	1.8	1.7	-	
113	Forward Flying 11/2 Somersaults	-	1.9	1.8	-	-	1.8	1.7	-	
115	Forward Flying 21/2 Somersaults	-	-	-	-	-	2.7	2.5	-	

Table H: Sample scoring sheet from a diving event

Dive Order	Onder & Pos. Ltr. Level		Position	D.D.				Judg	jes' Awa	ards				Judges	Cumulative	
O do			(S,P,T,F)		1	2	3	4	5	6	7	8	9	Total	Total Award	
1	103B		Forward 1.5 Somersault	Р	1.6	6.5	7.0	6.5	7.0	6.5	0.0	0.0	0.0	0.0	20.0	32.00 32.00
2	105C		Forward 2.5 Somersault	Т	2.2	6.0	5.5	6.5	6.5	6.5	0.0	0.0	0.0	0.0	19.0	41.80 73.80
3	201B		Back Dive	Р	1.8											

Table I: Text descriptions for individual steps in MTL-AQA [31]

Subaction	Description
	In this position, athletes have the freedom to perform
E	any combination of dives from various categories
Free	without any restrictions or limitations
	In this position, athletes bring their knees to their
m 1	chest and hold onto their shins while maintaining a
Tuck	compact shape throughout their dive
	In this position, athletes maintain a straight body
Pike	with their legs extended and their toes pointed out
r ike	while bending at the waist to bring their hands to-
	ward their toes
	In this position, athletes start by standing on their
Armstand	hands on the edge of the diving board and perform
Allistand	their dive while maintaining this handstand position
	In this rotation type, athletes perform a forward-
Inwards	facing takeoff and rotate inward toward the diving
Inwards	board as they execute their dive
	In this rotation type, athletes perform a backward-
Reverse	facing takeoff and rotate backward away from the
Iteverse	diving board as they execute their dive
	In this rotation type, athletes perform a backward-
Backward	facing takeoff and rotate backward toward the diving
Dackward	board as they execute their dive
	In this rotation type, athletes perform a forward-
Forward	facing takeoff and rotate forward away from the
roiward	diving board as they execute their dive
0.5 Somersault	Athletes perform a half rotation in the air during
0.5 Somersaun	their dive
1 Somersault	Athletes perform a full forward or backward rotation
	in the air during their dive
1.5 Somersault	Athletes perform a full rotation and an additional
1.0 Domersaure	half rotation in the air during their dive
2 Somersault	Athletes perform two full forward or backward rota-
2 Domersaure	tions in the air during their dive
	Athletes perform two full rotations and an additional
2.5 Somersault	half rotation in the air during their dive
3 Somersault	Athletes perform three full forward or backward ro-
o comercial	tations in the air during their dive
	Athletes perform three full rotations and an addi-
3.5 Somersault	tional half rotation in the air during their dive
	Athletes perform four full rotations and an additional
4.5 Somersault	half rotation in the air during their dive
0.5 Twist	Athletes perform a half twist in the air during their
	dive
1 Twist	Athletes perform one full twist in the air during their
	dive
1.5 Twist	Athletes perform one and a half twists in the air
	during their dive
2 Twist	Athletes perform two full twists in the air during
2 1 11150	their dive
2.5 Twist	Athletes perform two and a half twists in the air
2.0 1 1120	during their dive
3 Twist	Athletes perform three full twists in the air during
	their dive
3.5 Twist	Athletes perform three and a half twists in the air
2.5 2250	during their dive
Entry	A diving technique involving a entry into the water,
	typically performed at the end of a dive

Table J: Text descriptions for individual steps in FineDiving [52]

Subaction	Description
Forward	A diving technique involving a front-facing takeoff and entry
Back	A diving technique involving a back-facing takeoff and entry
Reverse	A diving technique involving a back-facing takeoff and entry while rotating forward
Inward	A diving technique involving a front-facing takeoff and entry while rotating backwards
Arm Forward	A diving technique involving a front-facing takeoff and entry with arms extended and hands meeting above the head
Arm Back	A diving technique involving a back-facing takeoff and entry with arms extended and hands meeting above the head
Arm Reverse	A diving technique involving a back-facing takeoff and entry with arms extended and hands meeting above the head while rotating forward
1 Somersault Pike	A diving technique involving a takeoff and rotating forward to form a pike position with one somersault
1.5 Somersaults Pike	A diving technique involving a takeoff and rotating forward to form a pike position with one and a half somersaults
2 Somersaults Pike	A diving technique involving a takeoff and rotating forward to form a pike position with two somersaults
2.5 Somersaults Pike	A diving technique involving a takeoff and rotating forward to form a pike position with two and a half somersaults
3 Somersaults Pike	A diving technique involving a takeoff and rotating forward to form a pike position with three somersaults
3.5 Somersaults Pike	A diving technique involving a takeoff and rotating forward to form a pike position with three and a half somersaults
4.5 Somersaults Pike	A diving technique involving a takeoff and rotating forward to form a pike position with four and a half somersaults
1.5 Somersaults Tuck	A diving technique involving a takeoff and rotating forward to bend at the waist with one and a half somersaults
2 Somersaults Tuck	A diving technique involving a takeoff and rotating forward to bend at the waist with two somersaults
2.5 Somersaults Tuck	A diving technique involving a takeoff and rotating forward to bend at the waist with two and a half somersaults
3 Somersaults Tuck	A diving technique involving a takeoff and rotating forward to bend at the waist with three somersaults
3.5 Somersaults Tuck	A diving technique involving a takeoff and rotating forward to bend at the waist with three and a half somersaults
4.5 Somersaults Tuck	A diving technique involving a takeoff and rotating forward to bend at the waist with four and a half somersaults
0.5 Twist	A diving technique involving a takeoff and half a twist before entering the water
1 Twist	A diving technique involving a takeoff and one full twist before entering the water
1.5 Twists	A diving technique involving a takeoff and one and a half twists before entering the water
2 Twists	A diving technique involving a takeoff and two full twists before entering the water
2.5 Twists	A diving technique involving a takeoff and two and a half twists before entering the water
3 Twists	A diving technique involving a takeoff with three twists before entering the water
3.5 Twists	A diving technique involving a takeoff with three and a half twists before entering the water
Entry	A diving technique involving a entry into the water, typically performed at the end of a dive
0.5 Somersault Pike	A diving technique involving a take-off with half a somersault in the pike position before entering the water

Table K: Text descriptions for individual steps in JIGSAWS [11]

Phase	Subaction	Description
	G1	Reaching for needle with right hand
	G2	Positioning a needle to adjust its placement in a particular
		location or orientation
	G3	Pushing a needle through tissue which involves applying force
	Go	to the needle in order to penetrate and pass through bodily
		tissue
	G4	Transfer a needle from the left hand to the right hand
		Moving to the center with the needle in grip which involves
Suturing	G5	holding and manipulating the needle to direct it towards the
		central area of a target or site
		To pull a suture with the left hand is to use the left hand
	G6	to apply tension and draw a length of suture thread through
		tissue
	G8	Orienting a needle involves adjusting the position, angle, or
	Go	direction of the needle
	G9	The action of using the right hand to assist in tightening a
		suture which involves using the right hand to apply additional
	C10	tension or pressure to the suture thread
	G10	To loosen additional suture which involves manipulating the
		suture thread in order to reduce the tension or pressure that
		it is exerting on tissue
	G11	Dropping the suture at the end and moving to the end points
	011	which involves releasing the suture thread from one hand and
		repositioning oneself or the needle to prepare for the next
		step in a medical procedure
	G1	Reaching for needle with right hand
	G11	Dropping the suture at the end and moving to the end points
	GII	which involves releasing the suture thread from one hand and
		repositioning oneself or the needle to prepare for the next
		step in a medical procedure
Knot Tying	G12	Reaching for a needle with the left hand involves extending
		the left arm and grasping the needle with the hand
	G10	Making a C-loop around the right hand involves manipulating
	G13	the suture thread in a circular motion to form a loop that
		encircles the fingers or hand of the right hand
		Reaching for a suture with the right hand involves extending
	G14	the right arm and grasping the suture material with the hand
		Pulling a suture with both hands involves using both hands
	G15	to apply tension and draw a length of suture thread through
		tissue
	G1	Reaching for needle with right hand
	G2	Positioning a needle to adjust its placement in a particular
	G2	location or orientation
Needle Passing		Pushing a needle through tissue which involves applying force
	G3	9 9
		to the needle in order to penetrate and pass through bodily
		tissue
	G4	Transfer a needle from the left hand to the right hand which
		involves moving the needle from one hand to the other
	G5	Moving to the center with the needle in grip involves holding
		and manipulating the needle to direct it towards the central
		area of a target or site
	G6	To pull a suture with the left hand is to use the left hand
		to apply tension and draw a length of suture thread through
		tissue
	G8	Orienting a needle involves adjusting the position, angle, or
		direction of the needle
	G11	Dropping the suture at the end and moving to the end points
	GII	which involves releasing the suture thread from one hand and
		repositioning oneself or the needle to prepare for the next
		step in a medical procedure
	l	x x

Table L: Text descriptions for individual steps in Cataract [38]

Description
A sharp blade is used to create a precise cut through the
cornea, which provides intraocular access for instruments. The
paracentesis is followed by a clear cornea incision which is
less than 3 mm wide and is large enough to insert the phaco
handpiece
Viscous agent is injected to widen the anterior chamber and to
protect the corneal endothelium and the intraocular structures.
This is repeated before Phase 8 but is indistinguishable
The anterior capsule of the lens is opened. The surgeon begins
with a central radial cut. At the end of the cut, a tear is built
and allows the anterior capsule to fold over itself. This tear
is grasped and a flap is carried around in a circular way
The surgeon injects electrolyte solution and epinephrin under
the rhexis to separate the peripheral cortex of the lens from
the capsule. This facilitates the rotation of the nucleus and
hydrates the peripheral cortex
With ultrasound power, the phaco tip emulsifies the anterior
central cortex. A deep central linear groove through the nu-
cleus is made and the lens is cracked into two parts. The lens
is rotated and chopped into pieces, which can be emulsified.
During this procedure, it is essential to keep the posterior
capsule intact
Remaining parts of the cortex are extracted
The posterior capsule is polished in order to avoid opacifica-
tion of the capsule
The folded artificial lens is inserted. The lens is slowly unfold-
ing and is pushed into the capsular bag
Viscous elastic agent is removed from anterior chamber and
capsule bag
The corneal incision is hydrated with electrolyte solution
and antibiotics are injected. This induces temporary stromal
swelling and closure of incision. Only if it leaks, a suture is
required

References

- 1. Alemi, A.A., Fischer, I., Dillon, J.V., Murphy, K.: Deep variational information bottleneck. In: International Conference on Learning Representations (2016)
- Bai, Y., Zhou, D., Zhang, S., Wang, J., Ding, E., Guan, Y., Long, Y., Wang, J.: Action quality assessment with temporal parsing transformer. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IV. pp. 422–438. Springer (2022)
- Battaglia, P.W., Hamrick, J.B., Bapst, V., Sanchez-Gonzalez, A., Zambaldi, V., Malinowski, M., Tacchetti, A., Raposo, D., Santoro, A., Faulkner, R., et al.: Relational inductive biases, deep learning, and graph networks. arXiv preprint arXiv:1806.01261 (2018)
- 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)
- 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. Human factors 55(2), 298–308 (2013)
- 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)
- Chung, H.W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, E., Wang, X., Dehghani, M., Brahma, S., Webson, A., Gu, S.S., Dai, Z., Suzgun, M., Chen, X., Chowdhery, A., Narang, S., Mishra, G., Yu, A., Zhao, V., Huang, Y., Dai, A., Yu, H., Petrov, S., Chi, E.H., Dean, J., Devlin, J., Roberts, A., Zhou, D., Le, Q.V., Wei, J.: Scaling instruction-finetuned language models (2022)
- Chung, H.W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., Webson, A., Gu, S.S., Dai, Z., Suzgun, M., Chen, X., Chowdhery, A., Castro-Ros, A., Pellat, M., Robinson, K., Valter, D., Narang, S., Mishra, G., Yu, A., Zhao, V., Huang, Y., Dai, A., Yu, H., Petrov, S., Chi, E.H., Dean, J., Devlin, J., Roberts, A., Zhou, D., Le, Q.V., Wei, J.: Scaling instruction-finetuned language models. Journal of Machine Learning Research 25(70), 1–53 (2024), http://jmlr.org/papers/v25/23-0870.html
- 9. Duvenaud, D.K., Maclaurin, D., Iparraguirre, J., Bombarell, R., Hirzel, T., Aspuru-Guzik, A., Adams, R.P.: Convolutional networks on graphs for learning molecular fingerprints. Advances in neural information processing systems 28 (2015)
- 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
- Gao, Y., Vedula, S.S., Reiley, C.E., Ahmidi, N., Varadarajan, B., Lin, H.C., Tao, L., Zappella, L., Béjar, B., Yuh, D.D., 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)
- 12. Gordon, A.S.: Automated video assessment of human performance. In: Proceedings of AI-ED. vol. 2 (1995)
- 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)

- Ilharco, G., Wortsman, M., Wightman, R., Gordon, C., Carlini, N., Taori, R., Dave,
 A., Shankar, V., Namkoong, H., Miller, J., et al.: Openclip. Zenodo 4, 5 (2021)
- 15. International Swimming Federation (FINA): Fina high diving officials manual. https://resources.fina.org/fina/document/2021/02/03/916b4d2d-1ac9-4128-9a42-27abe131b77b/2019-10-14_fina_high_diving_officials_manual.pdf (2019), accessed on March 12, 2024
- Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., Suleyman, M., Zisserman, A.: The kinetics human action video dataset (2017)
- 17. Kendall, M.G.: A new measure of rank correlation. Biometrika 30(1/2), 81–93 (1938)
- 18. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. In: International Conference on Learning Representations (2014)
- 19. 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
- 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)
- 21. Likert, R.: A technique for the measurement of attitudes. Archives of psychology (1932)
- Liu, D., Li, Q., Jiang, T., Wang, Y., Miao, R., Shan, F., Li, Z.: Towards unified surgical skill assessment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9522–9531 (2021)
- 23. Liu, D., Li, Q., Jiang, T., Wang, Y., Miao, R., Shan, F., Li, Z.: Towards unified surgical skill assessment (2021)
- Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019 (2019)
- 25. Martin, J., Regehr, G., Reznick, R., Macrae, H., Murnaghan, J., Hutchison, C., Brown, M.: Objective structured assessment of technical skill (osats) for surgical residents. British journal of surgery 84(2), 273–278 (1997)
- 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)
- 27. 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)
- 28. 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
- 29. OpenAI: GPT-4 technical report (2023)
- 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)
- 31. 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)

- 32. 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)
- Pirsiavash, H., Vondrick, C., Torralba, A.: Assessing the quality of actions. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI 13. pp. 556–571. Springer (2014)
- 34. Prassas, S., Kwon, Y.H., Sands, W.A.: Biomechanical research in artistic gymnastics: a review. Sports Biomechanics 5(2), 261–291 (2006)
- 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. Biomedical Signal Processing and Control 72, 103323 (2022)
- 36. Santoro, A., Raposo, D., Barrett, D.G., Malinowski, M., Pascanu, R., Battaglia, P., Lillicrap, T.: A simple neural network module for relational reasoning. Advances in neural information processing systems **30** (2017)
- 37. Scarselli, F., Gori, M., Tsoi, A.C., Hagenbuchner, M., Monfardini, G.: The graph neural network model. IEEE transactions on neural networks **20**(1), 61–80 (2008)
- 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)
- Shi, Y., Jain, A.K.: Probabilistic face embeddings. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6902–6911 (2019)
- Sohn, K., Lee, H., Yan, X.: Learning structured output representation using deep conditional generative models. Advances in neural information processing systems 28 (2015)
- 41. Sun, J.J., Zhao, J., Chen, L.C., Schroff, F., Adam, H., Liu, T.: View-invariant probabilistic embedding for human pose. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16. pp. 53–70. Springer (2020)
- 42. Tang, Y., Ni, Z., Zhou, J., Zhang, D., Lu, J., Wu, Y., Zhou, J.: 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)
- 43. TISHBY, N.: The information bottleneck method. In: Proc. of the 37th Allerton Conference on Communication and Computation, 1999 (1999)
- 44. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
- 45. Vilnis, L., McCallum, A.: Word representations via gaussian embedding. In: International Conference on Learning Representations (2015)
- Wang, L., Yang, N., Huang, X., Jiao, B., Yang, L., Jiang, D., Majumder, R., Wei,
 F.: Text embeddings by weakly-supervised contrastive pre-training (2022)
- 47. Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., Van Gool, L.: Temporal segment networks: Towards good practices for deep action recognition. In: European conference on computer vision. pp. 20–36. Springer (2016)
- 48. 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)
- 49. Waters, T.R., Putz-Anderson, V., Garg, A.: Applications manual for the revised niosh lifting equation (1994)

- 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)
- 51. Xu, C., Fu, Y., Zhang, B., Chen, Z., Jiang, Y.G., Xue, X.: Learning to score figure skating sport videos. IEEE transactions on circuits and systems for video technology **30**(12), 4578–4590 (2019)
- 52. 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)
- 53. 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
- 54. XXL, F.T.: (Nov 14th), available at: https://huggingface.co/google/flan-t5-xxl (Nov. 2023)
- 55. 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)
- 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 (oct 2021)
- 57. Zhang, B., Chen, J., Xu, Y., Zhang, H., Yang, X., Geng, X.: Auto-encoding score distribution regression for action quality assessment. Neural Computing and Applications pp. 1–14 (2023)
- 58. Zhang, J., Bargal, S.A., Lin, Z., Brandt, J., Shen, X., Sclaroff, S.: Top-Down Neural Attention by Excitation Backprop. International Journal of Computer Vision 126(10), 1084–1102 (Oct 2018)
- Zhou, C., Huang, Y.: Uncertainty-driven action quality assessment. arXiv preprint arXiv:2207.14513 (2022)
- 60. Zhou, K., Ma, Y., Shum, H.P.H., Liang, X.: Hierarchical graph convolutional networks for action quality assessment. IEEE Transactions on Circuits and Systems for Video Technology pp. 1–1 (2023)
- 61. 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)