# Vision-Language Action Knowledge Learning for Semantic-Aware Action Quality Assessment

Huangbiao Xu[1,2], Xiao Ke[1,2,✉], Yuezhou Li[1,2], Rui Xu[1,2], Huanqi Wu[1,2], Xiaofeng Lin[1,2], and Wenzhong Guo[1,2]

[1] College of Computer and Data Science, Fuzhou University, Fuzhou, 350108, China
[2] Engineering Research Center of Big Data Intelligence, Ministry of Education, China
{kex,guowenzhong}@fzu.edu.cn,{huangbiaoxu.chn,liyuezhou.cm}@gmail.com

**Abstract.** Action quality assessment (AQA) is a challenging vision task that requires discerning and quantifying subtle differences in actions from the same class. While recent research has made strides in creating fine-grained annotations for more precise analysis, existing methods primarily focus on coarse action segmentation, leading to limited identification of discriminative action frames. To address this issue, we propose a Vision-Language Action Knowledge Learning approach for action quality assessment, along with a multi-grained alignment framework to understand different levels of action knowledge. In our framework, prior knowledge, such as specialized terminology, is embedded into video-level, stage-level, and frame-level representations via CLIP. We further propose a new semantic-aware collaborative attention module to prevent confusing interactions and preserve textual knowledge in cross-modal and cross-semantic spaces. Specifically, we leverage the powerful cross-modal knowledge of CLIP to embed textual semantics into image features, which then guide action spatial-temporal representations. Our approach can be plug-and-played with existing AQA methods, frame-wise annotations or not. Extensive experiments and ablation studies show that our approach achieves state-of-the-art on four public short and long-term AQA benchmarks: FineDiving, MTL-AQA, JIGSAWS, and Fis-V.

**Keywords:** Action quality assessment · Vision-language pre-training · Semantic-aware learning

## 1 Introduction

Action quality assessment (AQA) is an emerging video analysis technique that aims to quantitatively assess how well actions are performed from the same class. AQA has gained growing attention in the computer vision community due to its rich real-world applications, such as sports video analysis [4, 12, 33, 35, 36, 39, 46, 47, 49], healthcare [11, 31, 48, 54, 55], artistic performances, and others [10, 34]. However, since the action differences between the same class are always subtle, it is difficult to capture and understand the subtle differences and corresponding large score variations. Therefore, AQA is regarded as a challenging task.

✉ Corresponding Author.

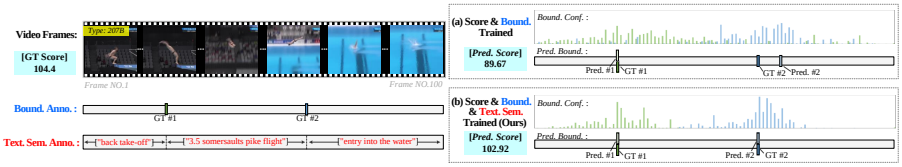



**Fig. 1:** (a) Existing methods solely employ precious frame-wise annotations for stage boundary prediction, often struggling to effectively discern action semantics, relying on rare distinctive frames for boundary selection. (b) Our multi-grained vision-language alignment framework introduces cheaper textual semantics of professional action knowledge to facilitate the understanding of action meanings, bringing satisfactory results.

Existing AQA methods primarily model the mapping of relationships between deep features and scores. However, most of them [21,39,41,47,52] directly analyze coarse-grained video-level and clip-level representations, which is insufficient for AQA that needs to discern subtle action differences. Recently, some works [15,24,46,49] have constructed fine-grained datasets that introduce multi-stage and even frame-wise semantic information for AQA. These are valuable gifts for more accurate, robust and interpretable AQA research. Yet, these fine-grained annotations are still used in a rough manner that only assist in the segmentation of stage actions. This manner may only learn to recognize differences between sub-stage actions to determine the boundaries of segmentation, without further understanding of the action semantics of different stages. As shown in Fig. 1(a), this may not identify the exact stage boundary. Furthermore, fine-grained datasets are extremely expensive to acquire. Particularly for the AQA task that explores inter-class differences in an occupational field, the fine-grained annotation process requires personnel with prior knowledge of the field (e.g., the FineDiving [46] dataset is constructed by six professionals in about 120 hours). Therefore, we raise a pressing question: *Is there available prior knowledge with low annotation overhead, to help understand fine-grained semantic information?*

While actively seeking answers, we realize an indispensable element: Language! Looking to AQA, which assesses the actions of one professional field. Obviously, a key point of AQA is to understand the specific semantics of the field. Expert judges are capable of providing convincing scores due to their abundant prior knowledge in the field, such as the action's meaning and challenges. Similar to the crucial role of acquiring prior knowledge for human judges, language also holds equal significance for AQA in comprehending the action semantics and knowledge. In practice, professional domains often employ specialized language to convey domain-specific knowledge. For instance, diving events are typically divided into "take-off", "flight", and "entry". Furthermore, terms like "forward", "back", and "reverse" can be used to further express the meanings of the athlete's "take-off" actions. These languages are highly valuable for AQA. To incorporate language knowledge, we naturally turn to the prevalent Vision-Language Pre-training (VLP) [16,37], which has achieved great success. VLP has acquired a vast amount of cross-modal knowledge and enriched latent semantic information

from millions of image-text pairs, demonstrating strong generalization and transfer capabilities. Hence, VLP like CLIP [37], which can bridge language semantics with visual features across modalities, is highly suitable for AQA.

To fill the research gap of learning action priors using language in AQA, we first propose a novel multi-grained vision-language alignment framework (MVLA). The key idea lies on leveraging textual features to guide visual features at various levels, thus mining potential commonalities in the feature space. As shown in Fig. 1(b), introducing our textual semantics allows for better discrimination of action variations, and identifying accurate stage boundaries. Specifically, we retain the Classification-Based Pre-training (CBP) [5, 40] model commonly used by AQA methods to extract discriminative action information. Then, we employ the VLP model to extract rich textual semantics and foreground information [18]. These textual semantics are encoded from both stage-level and video-level action texts that we constructed based on specialized terminology. Textual semantics are used to guide the understanding of action semantics at these levels: global video type, sub-stage type, and frame-wise type.

We further propose a semantic-aware collaborative attention (referred to as SCA) to bridge the gap in semantic space between CBP and VLP, ensuring that textual semantics rightly guide action comprehension. Specifically, in the VLP model, we first perform cross-modal alignment between its textual features (serving as "key" and "value") and visual features (serving as "query"), embedding textual semantics into visual action information. Subsequently, we utilize these aligned features enriched with action knowledge (serving as "key" and "value") to guide CBP's visual features (serving as "query"). By doing so, it prevents the confusing cross-modal, cross-semantic space interaction between CBP and textual features. It is worth mentioning that SCA can flexibly incorporate various levels of textual semantics. We only need to construct text annotations for representative fine-grained action types based on professional knowledge, which is much cheaper than frame-wise annotations. This allows our approach to plug-and-play the textual semantics into existing methods. On four public AQA benchmarks: FineDiving [46] (with frame-wise annotations), and MTL-AQA [33], JIGSAWS [11], and Fis-V [45] (without frame-wise annotations), our approach significantly improves the performance of existing methods for action segmentation and assessment, reaching new state-of-the-art results.

The main contributions of this work are summarized as:

- We propose a novel multi-grained vision-language alignment framework for action quality assessment, aiming to leverage textual semantics to extract specialized knowledge for facilitating action understanding.
- We build cheaper multi-grain text annotations and design a flexible semantic-aware collaborative attention. This allows our approach to be plug-and-play with existing methods, frame-wise annotations or not.
- Extensive experimental results and ablation studies reveal the significance of learning language knowledge and the state-of-the-art performance of our semantic-aware approach in both short and long-term AQA.

## 2 Related Work

### 2.1 Action Quality Assessment

**Frame-Wise Unannotated Methods.** Early on, Gordon [12] pioneered the idea of using machines to automatically assess action quality, and Pirsiavash *et al*. [36] first formulated the AQA task and explored several viable approaches. Subsequently, due to its wide applicability, AQA has attracted extensive research in the community. Many works [29, 32, 33, 35, 39, 41, 44, 45, 52] formulated AQA as a regression task, relying on human pose and visual features to predict the quality scores of a single video by fitting expert judges' score labels. Some methods [4, 9, 10, 48] formulated AQA as a pair-wise ranking task, comparing the goodness of videos. Recently, a novel perspective [14, 19, 21, 47] emerged, redefining AQA as a contrastive task by learning to discern subtle action differences through contrasting input and exemplar videos. For example, Yu *et al*. [47] proposed a contrastive regression framework to predict relative scores between two videos. The framework introduces additional reference action information and reduces the range and difficulty of score regression. This idea has shown significant performance improvements, which has spurred researchers to contemplate the crucial role of mining fine-grained semantic information for AQA. Thus, some works [2, 7] had begun to explore finer-grained representations instead of video-level or clip-level representations.
**Frame-Wise Annotated Methods.** Furthermore, some works [15, 24, 46, 49] constructed valuable fine-grained datasets for AQA, aiming to introduce multi-stage and even frame-wise semantic information. **Note that** prior work TSA [46] has demonstrated that action segmentation helps AQA. However, TSA relies on precious frame-wise annotations and only employs real stage boundaries to guide action segmentation, which may not effectively capture the action semantics of sub-stages. In this work, we construct lower-cost textual annotations and explore methods that fully exploit textual semantics, which significantly improves action segmentation and facilitates more accurate quality assessment.

### 2.2 Vision-Language Pre-training

Vision-language pre-training using large-scale image-text pairs has achieved great success, *e.g.*, ALIGN [16] and CLIP [37]. These models learn rich cross-modal knowledge and latent semantic information of open-vocabulary scenes from millions of image-text pairs, demonstrating strong generalization and transfer capabilities. Although VLP can be effectively transferred to downstream applications such as object detection [3, 13, 53], image segmentation [6, 22, 26], and few-shot and zero-shot recognition [1, 50, 51]. Adapting VLP models to the video domain is still challenging due to the lack of temporal prior in image-level pre-training. Therefore, some recent works [17, 18, 23, 27, 28, 30, 38, 42] began to explore the extension of CLIP into video applications such as action recognition and temporal localization. For example, [17, 23, 28, 30] equipped the video domain with

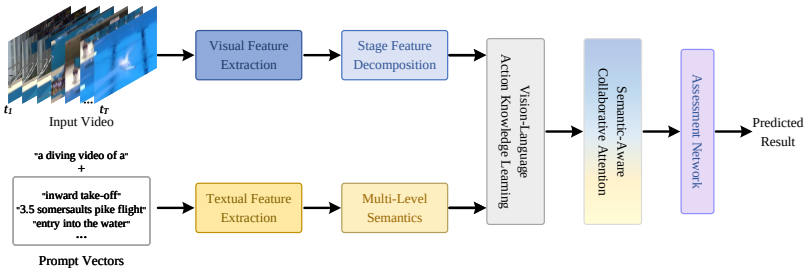



**Fig. 2:** The pipeline of our multi-grained vision-language alignment framework.

additional transformers or attention mechanisms to encode temporal information. [27, 42] incorporated specifically designed prompt learning to efficiently retrieve and detect video information. Rasheed *et al.* [38] explored to only simply fine-tune CLIP to fit the video domain. While Ju *et al.* [18] first explored the application of CLIP to weakly-supervised temporal action localization for downstream video tasks. In this work, we are the first to explore how to leverage VLP to achieve more accurate, reliable, and interpretable action quality assessment.

## 3 Approach

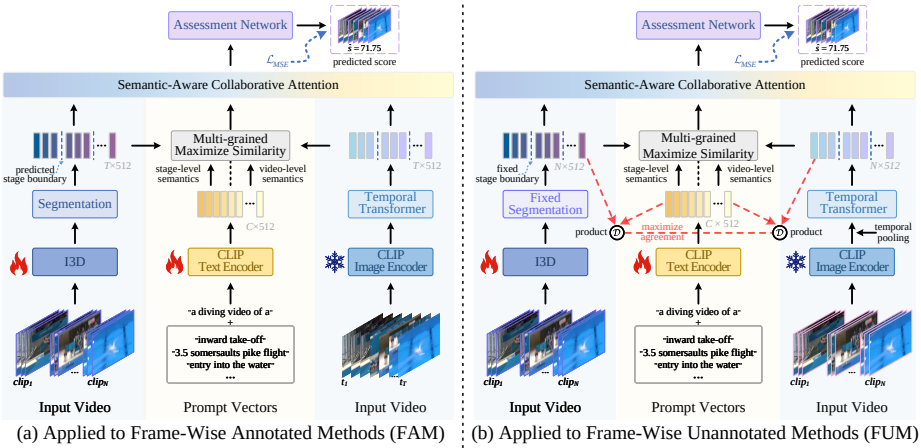
The simplified pipeline of our approach is illustrated in Fig. 2. In this section, we introduce our approach in detail. The primary concept is to incorporate representative professional knowledge into AQA using textual semantics, enhancing the comprehension of action meanings for accurate assessments.

### 3.1 Overview

The inputs of our framework are an action video and the knowledge texts involved in the current action domain. For an input video $V_T$ with $T$ frames, we first perform visual feature extraction using the CLIP image encoder $E_{\text{vis}}$ to extract the frame-level visual features $F_{\text{vis}} \in \mathbb{R}^{T\times D}$, where $D$ refers to feature dimension. We then divide the $T$ frames into $N$ clips ($V_N$) with 16 frames and extract the spatial-temporal representations $F_{\text{i3d}} \in \mathbb{R}^{N\times D}$ using I3D [5] as the backbone $E_{\text{i3d}}$. Meanwhile, for the textual feature extraction, we construct multi-level textual annotations representing action knowledge based on professional terminology and use the CLIP text encoder $E_{\text{text}}$ to extract the multi-level textual semantics $F_{\text{text}} \in \mathbb{R}^{C\times D}$, where $C$ means the number of action types. Formally, with the introduction of the action type name text $C_{\text{name}}$:

$$F_{\text{i3d}} = E_{\text{i3d}}\left(V_T\right), F_{\text{vis}} = E_{\text{vis}}\left(V_N\right), F_{\text{text}} = E_{\text{text}}\left(C_{\text{name}}\right). \tag{1}$$

To obtain fine-grained features, we perform stage feature decomposition to segment $F_{\text{i3d}}$ and $F_{\text{vis}}$ into sub-action stages. Depending on the characteristics of datasets (frame-wise annotations or not), we perform different segmentation




(a) Applied to Frame-Wise Annotated Methods (FAM) (b) Applied to Frame-Wise Unannotated Methods (FUM)

**Fig. 3:** The architecture of our multi-grained vision-language alignment framework. Given an input video, visual features are extracted using the I3D backbone and CLIP image encoder. We design two schemes: frame-wise annotated (a) and frame-wise unannotated (b), based on the availability of frame-wise annotations. These schemes extract multi-grained visual features. Next, we construct multi-level textual annotations based on professional terminology, maximizing similarity with visual features. The semantic-aware collaborative attention mechanism integrates textual semantics into visual representations. Finally, these representations are sent to an assessment network to predict scores. Our approach can be plug-and-played with existing methods to achieve more accurate and interpretive assessments.

operations, namely, frame-level segmentation and clip-level segmentation. With this operation, we obtain rich multi-grain visual features such as frame-level, stage-level and video-level features. These visual features are then aligned with the corresponding multi-level textual semantics for vision-language action knowledge learning. Afterwards, we propose a semantic-aware collaborative attention $\mathcal{S}$ to embed the textual semantics into the visual feature space, which allows to obtain discriminative features with action knowledge. What action the athlete performs is precisely the key factor considered in the assessment. Finally, an assessment network $\mathcal{R}$ is used to regress the predicted score $\hat{s}$. Formally,

$$\hat{s} = \mathcal{R}\left(\mathcal{S}\left(\left(F_{\text{i3d}}, F_{\text{vis}}, F_{\text{text}}\right)|\Phi\right)|\Theta\right), \tag{2}$$

where $\Phi$ and $\Theta$ are the learnable parameters of $\mathcal{S}$ and $\mathcal{R}$. Below we describe the detailed structure of our approach.

### 3.2 Visual Feature Extraction

As discussed in Sec. 2.1, we categorize existing methods into two groups based on whether frame-wise annotations are used: frame-wise annotated methods (FAM) and frame-wise unannotated methods (FUM). To ensure the broad applicability of our approach to existing methods, we accordingly develop two sets of

implementation strategies. The specific details are illustrated in Fig. 3. Since our framework uses both CBP and VLP branches, this may cause significant computational overhead. To address this issue, we freeze the largest CLIP image encoder. In addition, the image-text pre-training lacks the temporal prior, which is important for video understanding. Therefore, we use simple temporal transformer layers $\mathcal{T}$ to construct action temporal relationships. The frozen encoder preserves the action prior knowledge and rich foreground information [18] from pre-training. When combined with straightforward temporal modeling, it can adapt to the current motion context, providing abundant action semantics while saving memory footprint. Acquiring visual features $F_{\text{vis}}$ can be rewritten as:

$$F_{\text{vis}} = \mathcal{T}\left(E_{\text{vis}}\left(V_N\right)|\Psi\right), \tag{3}$$

where $\Psi$ is the learnable parameter of $\mathcal{T}$.

As shown in Fig. 3(a), for FAM, we follow previous work [46] by utilizing a segmentation module consisting of several convolutional blocks to convert the features $F_{\text{i3d}}$ from clip-level to frame-level $F_{\text{i3d}}^{fra}$. $F_{\text{i3d}}^{fra}$ is further converted by an MLP block into a probability distribution $\mathrm{P} = \{\hat{p}_t\}_{t=1}^{T} \in \mathbb{R}^{T\times L}$, where $L$ denotes the number of stage boundaries, and $\hat{p}_t$ is the confidence of whether the $t$-th frame is one of the $L$ stage boundaries. We utilize the frames with the highest confidence as stage boundaries and divide $F_{\text{i3d}}^{fra}$ into $L+1$ stage features, i.e., $F_{\text{i3d}}^{sta}$. Based on the ground-truth $\{p_t\}_{t=1}^{T}$ obtained from the frame-wise annotations, we predict the stage boundaries by minimizing $\mathcal{L}_{\text{BCE}}$, which can be expressed as:

$$\mathcal{L}_{\text{BCE}} = -\sum\nolimits_t \left[p_t \log \hat{p}_t + (1-p_t)\log\left(1-\hat{p}_t\right)\right]. \tag{4}$$

Additionally, we obtain video-level features $F_{\text{i3d}}^{vid}$ from $F_{\text{i3d}}$ using average pooling. Similarly, we perform the above operation for $F_{\text{vis}}$ of the VLP branch. Thus, we obtain visual information at three granularities, namely $F_x^{fra}$, $F_x^{sta}$, and $F_x^{vid}$, where $x \in \{\text{i3d}, \text{vis}\}$.

For FUM, without frame-level labels as support, effective implementation of action stage segmentation is not feasible. We therefore adopt a different setting, as shown in Fig. 3(b). Specifically, we input video clips composed of continuous frames into the VLP branch, similar to the input of the CBP branch, and employ temporal pooling on the frame-level features to obtain clip-level features. Then, we manually set fixed stage boundaries specific to the sports scene to partition the features from clip-level to stage-level. Thus, we similarly obtain multi-grained visual representations, namely $F_x^{cli}$, $F_x^{sta}$, and $F_x^{vid}$, where $x \in \{\text{i3d}, \text{vis}\}$.

### 3.3 Vision-Language Action Knowledge Learning

**Multi-Level Textual Annotation.** To incorporate action knowledge, we construct representative textual annotations based on professional terminology. Specifically, the types of actions performed by athletes in sports competitions are represented using professional terms. For example, in diving events, a complete action instance is denoted by a number such as “207B” or “5152B”. These dive

numbers are known to the judges in advance, and they specify which sub-action types will constitute the athlete's performance. For instance, for an action type "407B", the athlete will sequentially perform sub-action types "inward take-off", "3.5 somersaults pike flight", and "entry into the water". We directly utilize these action terms for sub-action types to construct stage-level textual labels. These sub-action texts can further be combined into text descriptions of complete instances to obtain video-level textual labels, such as "inward take-off, 3.5 somersaults tuck flight, and entry into the water". This allows us to obtain cost-effective prior knowledge incorporating professional semantics at a significantly lower cost than frame-wise annotations. Moreover, this cheap annotation approach can be easily applied to existing datasets such as FineDiving [46] and MTL-AQA [33], as well as to other action assessment scenes [11,45]. Please refer to the Appendix for the complete multi-level textual annotation.

**Multi-Grained Maximize Similarity.** To encode text into the semantic space in a way that aligns with sports scenes, we construct a set of manually crafted prompt templates like '`a diving video of a [category]`'. We encode these prompt vectors using the CLIP text encoder to obtain stage-level text embeddings $F_{\text{text}}^{sta}$ and video-level text embeddings $F_{\text{text}}^{vid}$. Then, the cosine similarity $sim\left(\cdot\right)$ between visual embeddings at different levels and their corresponding text embeddings is computed, and multi-grained maximized through cross-entropy (CE) loss with a temperature parameter $\tau$. This aims to understand the meanings of actions under different temporal structures in a holistic manner, mining subtle intra-class differences. For one grain of visual embedding $f_{\text{vis}}$ and text embedding $f_{\text{text}}$, the learning process can be represented as:

$$\mathcal{L}_{\text{CE}}\left(f_{\text{vis}}, f_{\text{text}}\right) = -\sum_{i} \log \frac{\exp\left(sim\left(f_{\text{vis}}^{i}, f_{\text{text}}^{i}\right)/\tau\right)}{\sum_{j} \exp\left(sim\left(f_{\text{vis}}^{i}, f_{\text{text}}^{j}\right)/\tau\right)}. \tag{5}$$

For FAM, our multi-grained maximize similarity includes stage-level and video-level visual embeddings with their corresponding text embeddings. In addition, we also minimize the distance between stage-level text embeddings and each frame-level visual embedding in that stage. Different from the two-grain loss of [8], we explore novel interactions between three-grain semantics for challenging AQA. The objective function is:

$$\begin{aligned} \mathcal{J} = & \mathcal{L}_{\text{CE}}\left(F_{x}^{fra}, F_{\text{text}}^{sta}\right) \\ & + \mathcal{L}_{\text{CE}}\left(F_{x}^{sta}, F_{\text{text}}^{sta}\right) + \mathcal{L}_{\text{CE}}\left(F_{x}^{vid}, F_{\text{text}}^{vid}\right). \end{aligned} \tag{6}$$

For FUM, since there is no frame-wise annotations, multi-grained maximize similarity only performed on stage-level and video-level embeddings. We further perform semantic alignment of the clip-level representations of the two branches to compensate for the lack of frame-level semantics. Specifically, we compute the similarity curves between the clip-level embeddings and the stage-level textual semantics corresponding to that sample, minimizing the Kullback-Leibler ($KL$)

divergence between the curves of the two branches. The objective function is:

$$\begin{aligned}\mathcal{J} = KL\left(sim\left(F_{\text{i3d}}^{cli}, F_{\text{text}}^{sta}\right) || sim\left(F_{\text{vis}}^{cli}, F_{\text{text}}^{sta}\right)\right) \\ + \mathcal{L}_{\text{CE}}\left(F_{x}^{sta}, F_{\text{text}}^{sta}\right) + \mathcal{L}_{\text{CE}}\left(F_{x}^{vid}, F_{\text{text}}^{vid}\right).\end{aligned} \tag{7}$$

### 3.4 Semantic-Aware Collaborative Attention

Although VLP brings rich action knowledge, there is still a challenge to effectively incorporate it into the CBP branch with different modality and semantic space. To tackle this issue, we design a semantic-aware collaborative attention (SCA), which leverages the powerful sequence modeling capabilities of transformer to bridge the gap between CBP and VLP. For the VLP branch, visual embeddings $F_{\text{vis}}$ serve as "query", and text embeddings $F_{\text{text}}$ serve as "key" and "value". SCA first leverages the strong cross-modal prior knowledge from CLIP to integrate textual semantics into the visual semantic space, generating new spatial-temporal representations $\mathcal{X}$ enriched with action knowledge. Then, SCA employs $\mathcal{X}$ as an intermediary to embed action knowledge into the visual features $F_{\text{i3d}}$ of the CBP branch, with $\mathcal{X}$ serving as the "key" and "value", and $F_{\text{i3d}}$ as the "query". The module prevents the confusing cross-modal, cross-semantic space interaction between CBP and textual semantics. Formally,

$$\begin{aligned}\mathcal{X}' = Softmax\left(F_{\text{vis}} w_q (F_{\text{text}} w_k)^T / \sqrt{d}\right) F_{\text{text}} w_v + F_{\text{vis}}, \\ \mathcal{X} = MLP\left(\mathcal{X}'\right) + \mathcal{X}', \\ \mathcal{Z}' = Softmax\left(F_{\text{i3d}} W_q (\mathcal{X} W_k)^T / \sqrt{d'}\right) \mathcal{X} W_v + F_{\text{i3d}}, \\ \mathcal{Z} = MLP\left(\mathcal{Z}'\right) + \mathcal{Z}',\end{aligned} \tag{8}$$

where $w_q$, $w_k$, $w_v$, $W_q$, $W_k$, and $W_v$ are the learnable weights, $\sqrt{d}$ and $\sqrt{d'}$ are the normalization factors. The MLP module consists of two fully connected layers with a GELU non-linearity. SAC is used for stage-level and video-level semantic collaboration. Finally, an assessment network $\mathcal{R}$ regresses $\mathcal{Z}$ to obtain the predicted score $\hat{s}$ of the input video. $\mathcal{R}$ is optimized by minimizing the mean squared error between $\hat{s}$ and the ground-truth score label $s$, as follows:

$$\mathcal{L}_{\text{MSE}} = \|\hat{s} - s\|^2. \tag{9}$$

## 4 Experiments

### 4.1 Datasets and Experiment Settings

**Datasets.** We perform experiments on four public AQA benchmarks: FineDiving [46] (with frame-wise annotations), and MTL-AQA [33], JIGSAWS [11], and Fis-V [45] (without frame-wise annotations). We adhere to the criteria established by the datasets and previous studies. See more details in the Appendix.

**Evaluation Metrics.** To compare prior works [2, 46, 47, 52], we adopt two metrics to evaluate the AQA performance of our method, the Spearman's rank correlation ($\rho$) and relative L2-distance (R-$\ell_2$). In addition, we evaluate the mean square error (MSE) in Fis-V, following [47]. $\rho$ is used to measure the rank correlation between the predicted series $\hat{q}$ and the ground-truth series $q$, which is defined as follows:

$$\rho = \frac{\sum_i (q_i - \bar{q})\left(\hat{q}_i - \bar{\hat{q}}\right)}{\sqrt{\sum_i (q_i - \bar{q})^2 \sum_i \left(\hat{q}_i - \bar{\hat{q}}\right)^2}}. \tag{10}$$

R-$\ell_2$ is used to measure the numerical difference between the predicted scores $\hat{s}$ and the ground-truth scores $s$, which is defined as:

$$\text{R-}\ell_2 = \frac{1}{N}\sum_n^N \left(\frac{|s_n - \hat{s}_n|}{s_{\max} - s_{\min}}\right)^2. \tag{11}$$

Following prior work [46], we adopt the average Intersection over Union (AIoU) metric to evaluate the action segmentation performance of our method. The metric converts the predicted boundaries into a set of 1D bounding boxes $\hat{\mathcal{B}}$ and computes the Intersection over Union with the ground-truth bounding boxes $\mathcal{B}$, i.e., IoU $= \left|\hat{\mathcal{B}} \cap \mathcal{B}\right| / \left|\hat{\mathcal{B}} \cup \mathcal{B}\right|$. For a threshold $d$, AIoU@$d$ is defined as:

$$\text{AIoU@}d = \frac{1}{N}\sum_n^N \mathcal{I}\left(\text{IoU}_n \geq d\right), \tag{12}$$

where $\mathcal{I}(\cdot)$ is an indicator that yields 1 when $\text{IoU}_n \geq d$, and 0 otherwise.

**Implementation Details.** For fair comparisons, we adopt the I3D backbone pre-trained on the Kinetics [5] and use VST [25] on Fis-V only. The CLIP image encoder and text encoder are both adopted from ViT-B/16. The Adam optimizer is used with a learning rate 1e-4 for visual feature extraction, and 1e-3 for SCA and assessment network. We set the weight decay to 0. For the data preprocessing of datasets, we follow the original settings of the existing methods when our method is plugged into them. In most cases, for FineDiving, we sample 96 frames and divide them into 9 clips of 16 frames each. While for MTL-AQA, videos are sampled as 103 frames and divided into 10 overlapping 16-frame clips. For JIGSAWS, 160 frames are uniformly sampled into 10 non-overlapping 16-frame clips. And for Fis-V, each non-overlapping clip contains 32 consecutive frames. The dimension $D$ of visual features is 512. When the feature dimensions of the existing methods do not match the CLIP features, we only use a simple linear layer to project the features, which can achieve effective results.

### 4.2 Comparison to State-of-the-art

To ensure fair comparisons, we implement all methods and integrate our plug-in approach in the same experimental environment (Two GeForce RTX 3090 GPUs with Pytorch 1.12.0). We first validate our method on two large-scale diving

**Table 1:** Comparisons of performance with existing AQA methods on FineDiving. (w/o DN) indicates random selection, while (w/ DN) indicates using the dive numbers to select exemplars. / denotes without action segmentation. Best results are in bold.

| Method | w/o DN | | | | w/ DN | | | |
|---|---|---|---|---|---|---|---|---|
| | Segmentation | | Assessment | | Segmentation | | Assessment | |
| | AIoU@0.5↑ | AIoU@0.75↑ | $\rho$↑ | R-$\ell_2$ (×100)↓ | AIoU@0.5↑ | AIoU@0.75↑ | $\rho$↑ | R-$\ell_2$ (×100)↓ |
| USDL [39] | / | / | 0.8302 | 0.5927 | / | / | 0.8913 | 0.3822 |
| MUSDL [39] | / | / | 0.8427 | 0.5733 | / | / | 0.8978 | 0.3704 |
| CoRe [47] | / | / | 0.8631 | 0.5565 | / | / | 0.9061 | 0.3615 |
| TSA [46] | 80.71 | 30.17 | 0.8925 | 0.4782 | 82.51 | 34.31 | 0.9203 | 0.3420 |
| **TSA-MVLA (Ours)** | **95.46** | **58.34** | **0.9089** | **0.4242** | **98.66** | **67.16** | **0.9419** | **0.2840** |

**Table 2:** Comparisons of performance with existing AQA methods on MTL-AQA. (w/o DD) indicates random selection, while (w/ DD) indicates using the degree of difficulty to select exemplars. The bold indicates the better, and red indicates the best.

| Method | w/o DD | | w/ DD | |
|---|---|---|---|---|
| | $\rho$↑ | R-$\ell_2$ (×100)↓ | $\rho$↑ | R-$\ell_2$ (×100)↓ |
| TSA-Net [41] (ACM MM'21) | 0.9393 | – | – | – |
| PCLN [21] (ECCV'22) | 0.9230 | – | – | – |
| USDL [39] (CVPR'20) | 0.8861 | 0.774 | 0.9225 | 0.424 |
| **USDL-MVLA (Ours)** | **0.9142** | **0.545** | **0.9267** | **0.401** |
| MUSDL [39] (CVPR'20) | 0.9031 | 0.516 | 0.9244 | 0.440 |
| **MUSDL-MVLA (Ours)** | **0.9233** | **0.449** | **0.9362** | **0.356** |
| CoRe [47] (ICCV'21) | 0.9347 | 0.373 | 0.9519 | 0.266 |
| **CoRe-MVLA (Ours)** | **0.9410** | **0.316** | **0.9577** | **0.243** |
| TSA [46] (CVPR'22) | 0.9266 | 0.543 | 0.9470 | 0.307 |
| **TSA-MVLA (Ours)** | **0.9387** | **0.344** | **0.9615** | **0.257** |
| TPT [2] (ECCV'22) | 0.9431 | 0.348 | 0.9576 | 0.265 |
| **TPT-MVLA (Ours)** | **0.9529** | **0.281** | **0.9655** | **0.224** |
| HGCN [52] (TCSVT'23) | 0.9317 | 0.379 | 0.9555 | 0.260 |
| **HGCN-MVLA (Ours)** | **0.9375** | **0.343** | **0.9597** | **0.229** |

datasets: FineDiving (with frame-wise annotations) and MTL-AQA (without frame-wise annotations). Then, we validate the feasibility of our method without the aid of frame-wise annotations and sub-action terminology on the surgical dataset JIGSAWS and long-term figure skating dataset Fis-V. Specifically, in the FUM scheme, we use action names with quality descriptions as semantic knowledge at different grades to replace sub-action terminologies, *e.g.*, "a video of a poor/fair/good/excellent knot tying". Here, the VLKL operation remains consistent, while in the SCA module, visual features are coherently fused with the highest-similarity textual semantics.

**Results on FineDiving dataset.** As shown in Tab. 1, we follow previous work [46] to conduct experiments on FineDiving with both settings. Some methods based on contrastive regression require selecting exemplar videos to compare with the query video, 'w/o DN' and 'w/ DN' denote random and use the dive numbers to select exemplars, respectively. We can see that our method achieves a significant improvement over the SOTA method TSA [46]. Specifically, under 'w/ DN', our method achieves 16.15%, 32.85%, 2.16%, and 0.0580 improvements in the AIoU@0.5, AIoU@0.75, Spearman's rank correlation, and R-$\ell_2$ metrics, respectively. Similarly, our method achieves 14.75%, 28.17%, 1.64%, and 0.0540

**Table 3:** Performance comparison on JIGSAWS and Fis-V. The bold indicates the better, and red indicates the best.

**JIGSAWS**

| Method | Spearman Correlation (↑) | | | | R-$\ell_2$ (×100)↓ | | | |
|---|---|---|---|---|---|---|---|---|
| | S | NP | KT | Avg. | S | NP | KT | Avg. |
| I3D+MLP [47] | 0.61 | 0.68 | 0.66 | 0.65 | 4.795 | 11.225 | 6.120 | 7.373 |
| USDL [39] | 0.63 | 0.69 | 0.68 | 0.67 | 8.827 | 13.232 | 6.865 | 9.641 |
| +**MVLA (Ours)** | **0.65** | **0.73** | **0.77** | **0.72** | **6.806** | **9.081** | **5.806** | **7.231** |
| MUSDL [39] | 0.71 | 0.75 | 0.73 | 0.73 | 6.595 | 8.603 | 6.169 | 7.122 |
| +**MVLA (Ours)** | **0.74** | **0.78** | **0.76** | **0.76** | **5.249** | **5.304** | **4.991** | **5.181** |
| CoRe [47] | 0.83 | 0.86 | 0.84 | 0.84 | 5.749 | 5.557 | 3.314 | 4.873 |
| +**MVLA (Ours)** | **0.86** | **0.87** | **0.87** | **0.87** | **5.108** | **4.943** | **2.571** | **4.207** |
| TPT [2] | 0.87 | 0.86 | 0.89 | 0.87 | 2.832 | 5.243 | 3.519 | 3.865 |
| +**MVLA (Ours)** | **0.90** | **0.90** | **0.91** | **0.90** | **2.307** | **3.503** | **3.082** | **2.964** |
| HGCN [52] | 0.87 | 0.87 | 0.86 | 0.87 | 4.685 | 4.887 | 3.804 | 4.459 |
| +**MVLA (Ours)** | **0.90** | **0.91** | **0.88** | **0.90** | **3.817** | **3.265** | **3.332** | **3.471** |

**Fis-V**

| Method | Sp. Corr. (↑) | | | MSE (↓) | | |
|---|---|---|---|---|---|---|
| | TES | PCS | Avg. | TES | PCS | Avg. |
| M-BERT (Late) [20] | 0.530 | 0.720 | 0.634 | 27.73 | 12.38 | 20.06 |
| MS-LSTM [45] | 0.651 | 0.786 | 0.725 | 19.86 | 8.75 | 14.31 |
| +**MVLA (Ours)** | **0.693** | **0.802** | **0.753** | **19.02** | **8.02** | **13.52** |
| GDLT [44] | 0.658 | 0.838 | 0.762 | 21.14 | **8.51** | 14.88 |
| +**MVLA (Ours)** | **0.681** | **0.849** | **0.779** | **19.16** | 8.54 | **13.85** |
| CoRe [47] | 0.667 | 0.813 | 0.749 | 23.23 | 9.38 | 16.31 |
| +**MVLA (Ours)** | **0.686** | **0.833** | **0.770** | **20.58** | **8.58** | **14.58** |
| TPT [2] | 0.588 | 0.762 | 0.685 | 26.71 | 13.05 | 19.88 |
| +**MVLA (Ours)** | **0.643** | **0.781** | **0.719** | **25.20** | **11.14** | **18.17** |
| MLP-Mixer [43] | 0.674 | 0.823 | 0.758 | 19.96 | 8.06 | 14.01 |
| +**MVLA (Ours)** | **0.705** | **0.842** | **0.783** | **18.91** | **7.20** | **13.06** |

improvements on these metrics without using dive numbers labels. These overwhelming improvements, especially the significant effect on action segmentation, demonstrate that our semantic-aware approach greatly facilitates the model's understanding of action meanings and discernment of action differences.

**Results on MTL-AQA dataset.** In Tab. 2, our MVLA is integrated into six existing methods (USDL, MUSDL, CoRe, TSA, TPT, and HGCN) encompassing direct regression and contrastive regression frameworks. 'DD' indicates the difficulty degree labels. Our method improves 2.81%, 2.02%, 0.63%, 1.21%, 0.98%, and 0.58% in Sp. Corr. compared to these methods without using DD labels. Meanwhile, we achieve 0.229, 0.067, 0.057, 0.198, 0.067, and 0.036 improvements on R-$\ell_2$. Under 'w/ DD', our approach remains effective, with an average improvement of 0.81% and 0.042 on the two metrics. Note that TSA's performance is suboptimal due to the absence of frame-wise annotations for action segmentation on MTL-AQA. However, our method significantly improves performance by introducing textual semantics enriched with action knowledge. This indicates the feasibility of incorporating textual annotations, which are much more cost-effective compared to frame-wise annotations.

**Results on JIGSAWS dataset.** Following previous works [39, 47], we perform four-fold cross-validation on JIGSAWS. We plug our approach into five state-of-the-art methods. The experimental results are shown in Tab. 3. MVLA significantly improves the performance of existing methods on two metrics. In particular, MVLA respectively improves the Avg. Corr. and Avg. R-$\ell_2$ by 3.40% and 1.381, showing the strong generalization of our method to other actions.

**Results on Fis-V dataset.** We finally validate the performance of our method in long-term action assessment on the figure skating dataset Fis-V, as shown in Tab. 3. It can be seen that our method improves the performance of existing methods in various classes and metrics by a large margin, achieving new state-of-the-art Avg. Corr. (0.783) and Avg. MSE (13.06). This indicates that learning fine-grained semantics is equally important for long-term action understanding. The extensive experimental results from Tab. 1 to Tab. 3 demonstrate the robust

**Table 4:** Different granularities on maximizing similarity. ✓ means using this granularity. $\mathcal{G}_{fra}$, $\mathcal{G}_{std}$, and $\mathcal{G}_{vid}$ are frame-level, stage-level, and video-level maximize similarity.

| $\mathcal{G}_{fra}$ | $\mathcal{G}_{std}$ | $\mathcal{G}_{vid}$ | AIoU@↑ | | $\rho$↑ | R-$\ell_2$↓ (×100) |
|---|---|---|---|---|---|---|
| | | | 0.5 | 0.75 | | |
| × | × | × | 90.79 | 42.99 | 0.9340 | 0.3178 |
| ✓ | × | × | 98.00 | 66.22 | 0.9377 | 0.3222 |
| × | ✓ | × | 95.19 | 58.48 | 0.9330 | 0.3109 |
| × | × | ✓ | 93.72 | 51.54 | 0.9243 | 0.3804 |
| ✓ | ✓ | × | 96.66 | 62.48 | 0.9383 | 0.2913 |
| ✓ | × | ✓ | 95.33 | 55.94 | 0.9332 | 0.3145 |
| × | ✓ | ✓ | 97.60 | 62.08 | 0.9381 | 0.2982 |
| ✓ | ✓ | ✓ | **98.66** | **67.16** | **0.9419** | **0.2840** |

**Table 5:** Different components and backbones on FineDiving. ✓ means using the ground-truth stage boundaries. † can be regarded as an oracle for methods.

| Method (w/ DN) | CBP | VLP | AIoU@↑ | | $\rho$↑ | R-$\ell_2$↓ (×100) |
|---|---|---|---|---|---|---|
| | | | 0.5 | 0.75 | | |
| Baseline (TSA) | I3D | - | 82.51 | 34.31 | 0.9203 | 0.3420 |
| | - | CLIP-Image | 87.35 | 45.53 | 0.9259 | 0.3266 |
| | I3D | CLIP-Image | 88.22 | 47.06 | 0.9266 | 0.3248 |
| +VLKL | I3D | CLIP | 92.52 | 50.60 | 0.9295 | 0.3196 |
| +SCA | I3D | CLIP | 90.79 | 42.99 | 0.9340 | 0.3178 |
| MVLA (Ours) | I3D | CLIP | 98.66 | 67.16 | 0.9419 | 0.2840 |
| + Quality Text | I3D | CLIP | 98.80 | 67.88 | 0.9438 | 0.2816 |
| Freeze CLIP-Text | I3D | CLIP | 97.75 | 64.83 | 0.9390 | 0.2978 |
| MVLA (Ours) | I3D | ViFi-CLIP | 98.93 | 68.19 | 0.9443 | 0.2819 |
| MVLA (Ours) | VST | CLIP | **99.01** | **69.24** | **0.9457** | **0.2804** |
| TSA† | I3D | - | ✓ | ✓ | 0.9310 | 0.3260 |
| MVLA† (Ours) | I3D | CLIP | ✓ | ✓ | 0.9470 | 0.2726 |

effectiveness of our method, which can be flexibly plugged into existing methods for both short and long-term AQA, frame-wise annotations or not.

### 4.3 Ablation Study

Our MVLA comprises vision-language action knowledge learning (VLKL) and semantic-aware collaborative attention (SCA). All the ablation studies are performed on FineDiving under 'w/ DN' setting. We use TSA [46] as the Baseline.

**Effects of granularity to maximize similarity.** To enhance comprehensive action semantic understanding, we use multi-grained maximize similarity in our proposed VLKL, which aligns visual embeddings and textual semantics at frame-level ($\mathcal{G}_{fra}$), stage-level ($\mathcal{G}_{std}$), and video-level ($\mathcal{G}_{vid}$). The results are reported in Tab. 4. It can be observed that the different granularities almost all bring a positive impact on the performance, the best results are obtained with using all three granularities. It is worth mentioning that AQA performance is reduced when only using the video-level maximize similarity. This may be because our video-level textual annotations are relatively long, and there is a lot of shared textual semantics. For example, the video-level text between action types "403B" and "405B" differs by only a numerical distinction in tens of characters, such as a single-digit difference between "1.5 somersaults" and "2.5 somersaults". However, our VLKL guides the alignment of visual semantics with textual semantics from coarse- to fine-grained, addressing this issue to a certain extent.

**Effects of our proposed components.** As shown in Tab. 5, when using only VLKL, the model focuses more on understanding semantics to segment actions, significantly improving AIoU@{0.5, 0.75} from {88.22, 47.06} to {92.52, 50.60}. While when using only SCA, the model integrates professional semantics into visual features for more reliable quality assessment, improving by 1.37% and 0.0242 on Sp. Corr. and R-$\ell_2$. Combining the two components achieves the best result, demonstrating the effectiveness of our components. Moreover, we follow [46] to explore the impact of using ground-truth stage boundaries as an oracle for our

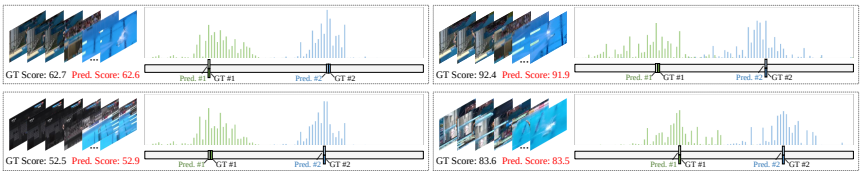



**Fig. 4:** Qualitative results of our method in several video instances. The left half shows the input videos, the ground-truth and predicted scores. The right half shows the frame-level action boundary probabilities, the ground-truth and our predicted boundary.

approach. MVLA achieves better performance than TSA† without using ground-truth labels, and further improves performance with labels, demonstrating that our method can understand action knowledge more efficiently. Adding quality descriptions to multi-level texts only slightly improves performance, possibly due to the fact that the highest-similarity semantics are not necessarily correct.

**Effects of backbones.** The performance increases only slightly after introducing the visual features of CLIP-Image, while the performance improves significantly from 0.9266/0.3248 to 0.9419/0.2840 using our MVLA, as shown in Tab. 5. This shows that learning action semantics is important for AQA. When freezing CLIP-Text, the overhead can be further saved, with only slight performance decreased. Now, with VLP backbone fully frozen, the additional cost is 11.25M and 2.96 GFLOPs. Furthermore, performance can be further enhanced with superior backbones, *e.g.*, VST [25] and ViFi-CLIP [38].

### 4.4 Qualitative Results

To intuitively demonstrate the performance of our approach, we visualize the action segmentation and quality assessment results for several video samples in Fig. 4. These results are obtained on FineDiving with 'w/ DN' setting. Our method predicts accurate stage boundaries and quality scores, demonstrating that the proposed semantic-aware approach can effectively discriminate action differences and understand the relationship between actions and scores.

## 5 Conclusion

This work proposes a novel multi-grained vision-language alignment framework for action quality assessment. To introduce prior knowledge, we construct multi-level textual annotations enriched with action knowledge, like specialized terminology, and pull in the distance between visual representations and corresponding textual semantics at multiple granularities. To bridge the modal and semantic spatial differences between CBP and VLP branches, we propose a new semantic-aware collaborative attention. We use the VLP with cross-modal prior knowledge as a bridge to encode textual semantics into visual representations. Our methods can be flexibly plugged into existing methods and achieve state-of-the-art results in both short and long-term AQA, frame-wise annotations or not.

## Acknowledgments

This work was partially supported by the National Natural Science Foundation of China (No. 61972097, U21A20472), the National Key Research and Development Plan of China (No. 2021YFB3600503), the Natural Science Foundation of Fujian Province (No. 2021J01612, 2020J01494), and the Major Science and Technology Project of Fujian Province (No. 2021HZ022007).

## References

1. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. In: NeurIPS. pp. 23716–23736 (2022)
2. Bai, Y., Zhou, D., Zhang, S., Wang, J., Ding, E., Guan, Y., Long, Y., Wang, J.: Action quality assessment with temporal parsing transformer. In: ECCV. pp. 422–438 (2022)
3. Bangalath, H., Maaz, M., Khattak, M.U., Khan, S.H., Shahbaz Khan, F.: Bridging the gap between object and image-level representations for open-vocabulary detection. In: NeurIPS. pp. 33781–33794 (2022)
4. Bertasius, G., Soo Park, H., Yu, S.X., Shi, J.: Am i a baller? basketball performance assessment from first-person videos. In: ICCV. pp. 2177–2185 (2017)
5. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: CVPR. pp. 6299–6308 (2017)
6. Ding, J., Xue, N., Xia, G.S., Dai, D.: Decoupling zero-shot semantic segmentation. In: CVPR. pp. 11583–11592 (2022)
7. Dong, L.J., Zhang, H.B., Shi, Q., Lei, Q., Du, J.X., Gao, S.: Learning and fusing multiple hidden substages for action quality assessment. KBS **229**, 107388 (2021)
8. Dong, S., Hu, H., Lian, D., Luo, W., Qian, Y., Gao, S.: Weakly supervised video representation learning with unaligned text for sequential videos. In: CVPR. pp. 2437–2447 (2023)
9. Doughty, H., Damen, D., Mayol-Cuevas, W.: Who's better? who's best? pairwise deep ranking for skill determination. In: CVPR. pp. 6057–6066 (2018)
10. Doughty, H., Mayol-Cuevas, W., Damen, D.: The pros and cons: Rank-aware temporal attention for skill determination in long videos. In: CVPR. pp. 7862–7871 (2019)
11. 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: MICCAIW. vol. 3 (2014)
12. Gordon, A.S.: Automated video assessment of human performance. In: AI-ED. vol. 2 (1995)
13. Gu, X., Lin, T.Y., Kuo, W., Cui, Y.: Open-vocabulary object detection via vision and language knowledge distillation. In: ICLR (2021)
14. Jain, H., Harit, G., Sharma, A.: Action quality assessment using siamese network-based deep metric learning. IEEE TCSVT **31**(6), 2260–2273 (2020)
15. Ji, Y., Ye, L., Huang, H., Mao, L., Zhou, Y., Gao, L.: Localization-assisted uncertainty score disentanglement network for action quality assessment. In: ACM MM. pp. 8590–8597 (2023)

16. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: ICML. pp. 4904–4916 (2021)
17. Ju, C., Han, T., Zheng, K., Zhang, Y., Xie, W.: Prompting visual-language models for efficient video understanding. In: ECCV. pp. 105–124 (2022)
18. Ju, C., Zheng, K., Liu, J., Zhao, P., Zhang, Y., Chang, J., Tian, Q., Wang, Y.: Distilling vision-language pre-training to collaborate with weakly-supervised temporal action localization. In: CVPR. pp. 14751–14762 (2023)
19. Ke, X., Xu, H., Lin, X., Guo, W.: Two-path target-aware contrastive regression for action quality assessment. Inf. Sci. **664**, 120347 (2024)
20. Lee, S., Yu, Y., Kim, G., Breuel, T., Kautz, J., Song, Y.: Parameter efficient multimodal transformers for video representation learning. In: ICLR (2020)
21. Li, M., Zhang, H.B., Lei, Q., Fan, Z., Liu, J., Du, J.X.: Pairwise contrastive learning network for action quality assessment. In: ECCV. pp. 457–473 (2022)
22. Liang, F., Wu, B., Dai, X., Li, K., Zhao, Y., Zhang, H., Zhang, P., Vajda, P., Marculescu, D.: Open-vocabulary semantic segmentation with mask-adapted clip. In: CVPR. pp. 7061–7070 (2023)
23. Lin, Z., Geng, S., Zhang, R., Gao, P., de Melo, G., Wang, X., Dai, J., Qiao, Y., Li, H.: Frozen clip models are efficient video learners. In: ECCV. pp. 388–404 (2022)
24. Liu, Y., Cheng, X., Ikenaga, T.: A figure skating jumping dataset for replay-guided action quality assessment. In: ACM MM. pp. 2437–2445 (2023)
25. Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., Hu, H.: Video swin transformer. In: CVPR. pp. 3202–3211 (2022)
26. Luo, H., Bao, J., Wu, Y., He, X., Li, T.: Segclip: Patch aggregation with learnable centers for open-vocabulary semantic segmentation. In: ICML. pp. 23033–23044 (2023)
27. Nag, S., Zhu, X., Song, Y.Z., Xiang, T.: Zero-shot temporal action detection via vision-language prompting. In: ECCV. pp. 681–697 (2022)
28. Ni, B., Peng, H., Chen, M., Zhang, S., Meng, G., Fu, J., Xiang, S., Ling, H.: Expanding language-image pretrained models for general video recognition. In: ECCV. pp. 1–18 (2022)
29. Pan, J.H., Gao, J., Zheng, W.S.: Action assessment by joint relation graphs. In: ICCV. pp. 6331–6340 (2019)
30. Pan, J., Lin, Z., Zhu, X., Shao, J., Li, H.: St-adapter: Parameter-efficient image-to-video transfer learning. In: NeurIPS. pp. 26462–26477 (2022)
31. Pandey, P., AP, P., Kohli, M., Pritchard, J.: Guided weak supervision for action recognition with scarce data to assess skills of children with autism. In: AAAI. pp. 463–470 (2020)
32. Parmar, P., Morris, B.: Action quality assessment across multiple actions. In: WACV. pp. 1468–1476 (2019)
33. Parmar, P., Morris, B.T.: What and how well you performed? a multitask learning approach to action quality assessment. In: CVPR. pp. 304–313 (2019)
34. Parmar, P., Reddy, J., Morris, B.: Piano skills assessment. In: MMSP. pp. 1–5 (2021)
35. Parmar, P., Tran Morris, B.: Learning to score olympic events. In: CVPRW. pp. 20–28 (2017)
36. Pirsiavash, H., Vondrick, C., Torralba, A.: Assessing the quality of actions. In: ECCV. pp. 556–571 (2014)
37. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML. pp. 8748–8763. PMLR (2021)

38. Rasheed, H., Khattak, M.U., Maaz, M., Khan, S., Khan, F.S.: Fine-tuned clip models are efficient video learners. In: CVPR. pp. 6545–6554 (2023)
39. Tang, Y., Ni, Z., Zhou, J., Zhang, D., Lu, J., Wu, Y., Zhou, J.: Uncertainty-aware score distribution learning for action quality assessment. In: CVPR. pp. 9839–9848 (2020)
40. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: ICCV. pp. 4489–4497 (2015)
41. Wang, S., Yang, D., Zhai, P., Chen, C., Zhang, L.: Tsa-net: Tube self-attention network for action quality assessment. In: ACM MM. pp. 4902–4910 (2021)
42. Wu, W., Sun, Z., Ouyang, W.: Revisiting classifier: Transferring vision-language models for video recognition. In: AAAI. vol. 37, pp. 2847–2855 (2023)
43. Xia, J., Zhuge, M., Geng, T., Fan, S., Wei, Y., He, Z., Zheng, F.: Skating-mixer: Long-term sport audio-visual modeling with mlps. In: AAAI. vol. 37, pp. 2901–2909 (2023)
44. Xu, A., Zeng, L.A., Zheng, W.S.: Likert scoring with grade decoupling for long-term action assessment. In: CVPR. pp. 3232–3241 (2022)
45. Xu, C., Fu, Y., Zhang, B., Chen, Z., Jiang, Y.G., Xue, X.: Learning to score figure skating sport videos. IEEE TCSVT **30**(12), 4578–4590 (2019)
46. Xu, J., Rao, Y., Yu, X., Chen, G., Zhou, J., Lu, J.: Finediving: A fine-grained dataset for procedure-aware action quality assessment. In: CVPR. pp. 2949–2958 (2022)
47. Yu, X., Rao, Y., Zhao, W., Lu, J., Zhou, J.: Group-aware contrastive regression for action quality assessment. In: ICCV. pp. 7919–7928 (2021)
48. Zhang, Q., Li, B.: Relative hidden markov models for video-based evaluation of motion skills in surgical training. IEEE TPAMI **37**(6), 1206–1218 (2014)
49. Zhang, S., Dai, W., Wang, S., Shen, X., Lu, J., Zhou, J., Tang, Y.: Logo: A long-form video dataset for group action quality assessment. In: CVPR. pp. 2405–2414 (2023)
50. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Conditional prompt learning for vision-language models. In: CVPR. pp. 16816–16825 (2022)
51. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. IJCV **130**(9), 2337–2348 (2022)
52. Zhou, K., Ma, Y., Shum, H.P., Liang, X.: Hierarchical graph convolutional networks for action quality assessment. IEEE TCSVT (2023)
53. Zhou, X., Girdhar, R., Joulin, A., Krähenbühl, P., Misra, I.: Detecting twenty-thousand classes using image-level supervision. In: ECCV. pp. 350–368 (2022)
54. Zia, A., Sharma, Y., Bettadapura, V., Sarin, E.L., Clements, M.A., Essa, I.: Automated assessment of surgical skills using frequency analysis. In: MICCAI. pp. 430–438 (2015)
55. Zia, A., Sharma, Y., Bettadapura, V., Sarin, E.L., Essa, I.: Video and accelerometer-based motion analysis for automated surgical skills assessment. IJCARS **13**, 443–455 (2018)