Human-centric Fine-grained Action Quality Assessment

Jinglin Xu, Sibo Yin, and Yuxin Peng

Abstract—Existing action quality assessment (AQA) methods mainly learn deep representations at the video level to score diverse actions. Due to the lack of a fine-grained understanding of actions in videos, they suffer from low credibility and accuracy, thus insufficient for stringent applications, such as competitive sports and sports injury rehabilitation. We argue that a fine-grained understanding of actions requires the model to parse actions in semantics, time, and space, which is the key to the credibility and accuracy of the AQA technique. Based on this insight, we propose a new human-centric fine-grained action quality assessment method named Unified Fine-grained spatial-temporal action Parser, namely **Uni-FineParser**. It learns human-centric foreground action representations by focusing on target action regions within each frame and exploiting their fine-grained alignments in semantics, time, and space, minimizing the impact of invalid backgrounds during the assessment. In addition, we construct human-centric foreground action mask annotations for the FineDiving, AQA-7, and MTL-AQA datasets, respectively called **FineDiving-HM**, **AQA-7-HM**, and **MTL-AQA-HM**. With refined spatio-temporal annotations on diverse target action procedures, Uni-FineParser can provide a potential for human-centric fine-grained action quality assessment with better interpretability. Through extensive experiments, we demonstrate the effectiveness of Uni-FineParser, which outperforms state-of-the-art methods while supporting more tasks of human-centric action understanding.

Index Terms—Action quality assessment, Human-centric foreground action, Spatial-temporal action parser, Fine-grained understanding

1 Introduction

Video understanding is a crucial technique in computer vision that aims to analyze objects, actions, or events in videos automatically. It is essential for many real-world applications, e.g., human-computer interaction [1], [2], [3], [4], medical rehabilitation [5], [6], and sports analysis [7], [8], [9], [10]. An accurate understanding of actions in videos provides critical technique support in action quality assessment (AQA), which considerably impacts sports analysis applications, e.g., helping evaluate athlete performance, designing targeted training programs, and preventing sports injuries.

Unlike general videos, sports videos are sequential processes with explicit procedural knowledge. Athletes have to complete a series of rapid and complex movements. Taking diving as an example, athletes will stretch, curl, and move their limbs and joints to finish different somersaults with three body positions, including straight, pike, and tuck, interspersed with varying twists, such as one twist and 2.5 twists. Then, the referee will assess the scores based on the athletes' take-off, somersault, twists, and entry. To achieve better competitive performance, athletes (1) take off decisively and forcefully at the right angle and with a proper height; (2) perform beautiful body positions, quick somersaults, and twists in the flight; (3) enter the water with a posture perpendicular to the surface, avoiding splashing water around. According to the diving rules, just a few degree differences in the take-off angle/height and the verticality of entry into the water can affect the number of points deducted. The difficulty lies in whether the referees' eyes can accurately discern such subtle differences.

Existing video-based AQA methods [11], [12], [13], [14] lack a fine-grained understanding of actions in videos, which cannot overcome human eye judgment limitations and need more credibility, difficult to apply in rigorous competitive sports. Although recent methods TPT [15] and FSPN [16] parse each action sequence along the temporal axis, they cannot explicitly obtain the spatial action representation due to the lack of spatial annotations. Therefore, it is necessary to have a fine-grained understanding of actions, which requires parsing actions in semantics, time, and space and alignments between their internal structures to develop human-centric foreground action representations with semantic consistency and spatial-temporal correlation.

To this end, we present a new framework for fine-grained action understanding, which learns human-centric foreground action representations by developing a unified fine-grained spatial-temporal action parser named **Uni-FineParser**, as shown in Fig. 1. Uni-FineParser has four core components: spatial action parsing, temporal action parsing, static visual encoding, and fine-grained contrastive regression. Given a query and exemplar videos, Uni-FineParser first models critical action regions in each frame to capture human-centric foreground action representations. The critical action regions in each frame are focused on foreground athletes' bodies, making target action representations of foreground athletes credible and accurate by spatially parsing actions. Uni-FineParser further enhances human-centric foreground action representations by adding static visual features of central adjacent frames of each video snippet into the above target action representations. Furthermore,

- *Jinglin Xu is with the School of Intelligence Science and Technology, University of Science and Technology Beijing, Beijing 100083, China. Email: xujinglinlove@gmail.com.*
- *Sibo Yin and Yuxin Peng are with Wangxuan Institute of Computer Technology, Peking University, Beijing 100871, China. Email: 2401112164@stu.pku.edu.cn; pengyuxin@pku.edu.cn.*
  Yuxin Peng is the corresponding author.

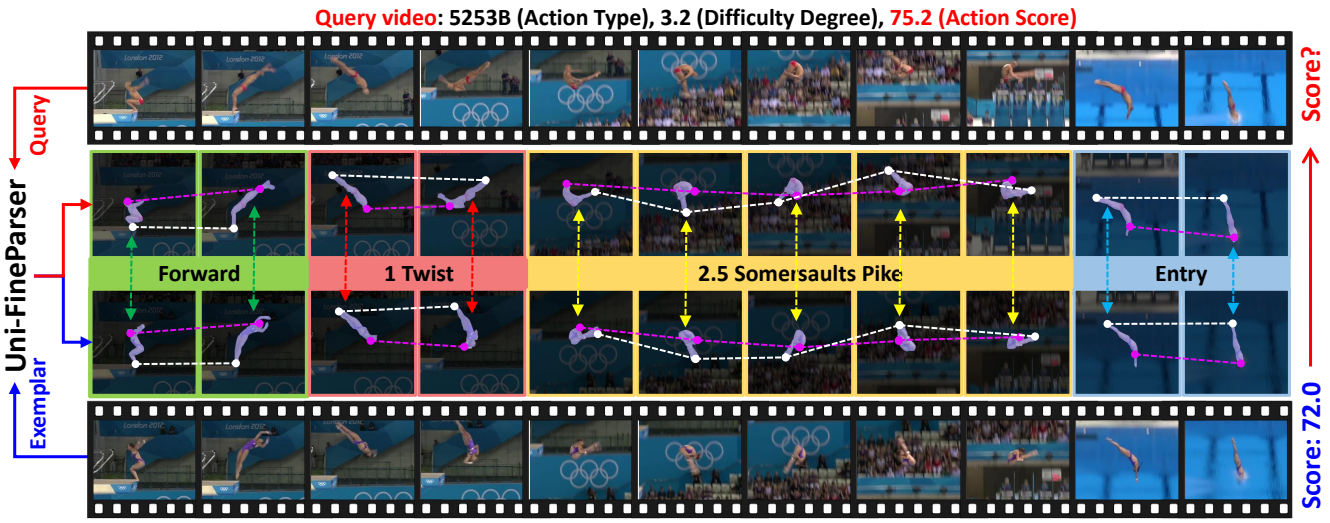

Fig. 1: An overview of unified fine-grained spatial-temporal action parser (**Uni-FineParser**). It enhances human-centric foreground action representations by exploiting fine-grained semantic consistency and spatial-temporal correlation between video frames, improving the AQA performance. Green, red, yellow, and blue dashed lines represent the fine-grained alignment of target actions between query and exemplar videos in semantics, time, and space.

Uni-FineParser continues to parse the action sequences into consecutive steps with explicit semantics to capture semantic and temporal correspondences between videos and obtain spatial-temporal representations of human-centric foreground actions. So far, Uni-FineParser ensures semantic consistency and spatial-temporal correspondence of target action representations across videos. Finally, Uni-FineParser quantifies quality differences in pairwise steps between query and exemplar videos and assesses the action quality at the fine-grained level. To demonstrate the effectiveness of Uni-FineParser extensively, we densely label human-centric foreground action regions of all videos in the FineDiving, AQA-7, and MTL-AQA datasets and construct additional mask annotations, named **FineDiving-HM**, **AQA-7-HM**, and **MTL-AQA-HM**. Experimental results demonstrate that our fine-grained actions understanding framework accurately assesses diverse actions by focusing on critical action regions consistent with human visual understanding.

The contributions of this paper are summarized as follows: (1) We propose a unified fine-grained spatial-temporal action parser, **Uni-FineParser**, improving the AQA performance via fine-grained alignments of human-centric foreground actions in semantics, time, and space. (2) Uni-FineParser models critical action regions in each frame to capture human-centric foreground action representations, minimizing the impact of invalid background on AQA. (3) Uni-FineParser provides human-centric foreground action mask annotations for the FineDiving and AQA-7 datasets, facilitating the evaluation of the credibility and accuracy of the AQA task. (4) Extensive experiments demonstrate that our Uni-FineParser achieves state-of-the-art performance with significant improvements and better interpretability.

The preliminary version of this work appeared in [17], and we improved and expanded it in the following aspects. (1) We improve FineParser (i.e., fine-grained spatial-temporal action parser) to Uni-FineParser (i.e., unified fine-grained spatial-temporal action parser) by redesigning SAP (i.e., spatial action parsing) and TAP (temporal action parsing) of FineParser and embedding them into a unified and compact framework. Besides, we remove the transformer-based interaction between query and examplar features and reduce the complexity of the contrastive score encoder, enabling Uni-FineParser to perform better and more efficiently than FineParser. (2) We present in-depth analyses and detailed ablation studies of Uni-FineParser, demonstrating that the new designs positively affect and promote our Uni-FineParser to become state-of-the-art. (3) We also provide human-centric foreground action mask annotations for the AQA-7 dataset and construct a new benchmark to show that our Uni-FineParser is comparable to the state-of-the-art method in various sports (e.g., gym vault and big skiing).

2 Related Work

Fine-grained Action Understanding. With ongoing advancements in action understanding, analyzing actions in finer granularity has become inevitable. Current endeavors in fine-grained action understanding mainly encompass tasks such as temporal action detection [18], [19], [20], [21], action recognition [22], [23], [24], [25], [26], video question answering [27], [28], [29], [30], and video-text retrieval [31], [32], [33]. Recently, Shao *et al.* [7] constructed FineGym that provides coarse-to-fine annotations temporally and semantically for facilitating action recognition. Chen *et al.* [34] proposed SportsCap that estimates 3D joints and body meshes and predicts action labels. Li *et al.* [8] introduced MultiSports with spatio-temporal annotations of actions from four sports. Zhang *et al.* [28] constructed a temporal query network to answer fine-grained questions about event types and their attributes in untrimmed videos. Li *et al.* [35] presented a hierarchical atomic action network that models actions as combinations of reusable atomic ones to capture the commonality and individuality of actions. Zhang *et al.* [36] introduced a fine-grained video representation learning method to distinguish video processes and capture their temporal dynamics. These methods mainly concentrated on a fine-grained understanding of the temporal dimension. In

contrast, our Uni-FineParser captures human-centric foreground action representations by developing a fine-grained understanding that parses action procedures in both time and space under the same semantics.

Action Quality Assessment. In early pioneering work, Pirsiavash *et al.* [37] formulated the AQA task as a regression problem from action representations to scores, and Parisi *et al.* [38] adopted the correctness of performed action matches to assess action quality. Parmar *et al.* [39] demonstrated the effectiveness of spatio-temporal features for estimating scores in various competitive sports. Recently, Tang *et al.* [13] introduced an uncertainty-aware score distribution learning method to alleviate the ambiguity of judges' scores. Yu *et al.* [14] developed a contrastive regression based on video-level features, enabling the ranking of videos and accurate score prediction. Wang *et al.* [40] introduced TSA-Net to generate action representations using the outputs of the VOT tracker, improving AQA performance. Xu *et al.* [9] contributed to a fine-grained sports video dataset for AQA and proposed a new action procedure-aware method to improve AQA performance. Yang *et al.* [41] developed a temporal parsing transformer to decompose the holistic feature into temporal part-level representations, improving the generalization of capturing fine-grained intra-class variation. Zhang *et al.* [42] proposed a plug-and-play group-aware attention module to enrich clip-wise representations with contextual group information. Gedamu *et al.* [16] propose a fine-grained spatio-temporal parsing network composed of the intra-sequence action parsing module and spatiotemporal multiscale transformer module, which learns fine-grained spatiotemporal sub-action representations. Zhang *et al.* [43] introduced a narrative action evaluation task, which proposes a prompt-guided multimodal interaction multi-task learning framework to generate professional commentary to assess action executions. Xu *et al.* [44] proposed a vision-language action knowledge learning approach for guiding action spatial-temporal representations, plug-and-played with existing AQA methods. Majeedi *et al.* [45] presented a deep probabilistic model that integrates score rubric and accounts for prediction uncertainty for AQA. Due to the lack of human-centric foreground action masks, these methods still have some drawbacks in spatial action parsing and target action representation learning. In contrast, our Uni-FineParser parses actions in space and time under the same semantics to explicitly learn human-centric foreground action representations, improving AQA's reliability and interpretability.

3 Approach

This section presents a unified fine-grained spatial-temporal action parser for human-centric action quality assessment, i.e., **Uni-FineParser**. As illustrated in Fig. 2, Uni-FineParser has four core components: spatial action parsing, temporal action parsing, static visual encoding, and fine-grained contrastive regression, which will be introduced as follows.

3.1 Problem Formulation

Given a pair of query and exemplar videos with the same action type, denoted as $(\mathbf{X}, \mathbf{Z})$, our approach is formulated as a fine-grained understanding framework that predicts the action score of the query video $\mathbf{X}$. Inspired by fine-grained contrastive regression [9], [17], our framework considers fine-grained quality differences between human-centric foreground actions in semantics, time, and space perspectives to model variations in their scores. The core is a new unified fine-grained action parser, Uni-FineParser $\mathcal{F}$, represented as

$$\hat{y}_X = \mathcal{F}(\mathbf{X}, \mathbf{Z}; \mathbf{\Theta}) + y_Z, \tag{1}$$

where $\mathbf{\Theta}$ denotes all learnable parameters of $\mathcal{F}$, and $\hat{y}_X$ denotes the predicted action score of the query video $\mathbf{X}$. $\mathbf{Z}$ is the exemplar video and its ground truth score y_Z.

3.2 Unified Fine-grained Spatio-temporal Action Parser

Uni-FineParser comprises four core components: spatial action parsing, temporal action parsing, static visual encoding, and fine-grained contrastive regression. These components are compactly incorporated into the unified fine-grained spatial-temporal action representation learning framework, which is further achieved for human-centric action quality assessment.

Spatial Action Parsing. Uni-FineParser parses the target action for each input video at the spatial level. In the I3D [46] backbone, we introduce an attention module after the output of the second I3D submodule (i.e., Stage 2) and supervise it by the human-centric foreground action mask to construct a mask-guided action attention module to capture the target action regions and enhance the action representations from the second I3D submodule. Continuing to pass to the third and the fourth I3D submodules (i.e., Stage 3 and Stage 4), we obtain the video features spanning the short-term local features of the first I3D submodule to the short-term global features of the fourth I3D submodule.

Concretely, given the input video $\mathbf{X} = \{\mathbf{X}_i\}_{i=1}^{N}$ composed of N snippets, the first two I3D submodules $\mathcal{B}_1$ and $\mathcal{B}_2$ encode each snippet $\mathbf{X}_i$ to capture short-term local features, denoted as $\mathcal{B}_2(\mathcal{B}_1(\mathbf{X})) = \{\mathcal{B}_2(\mathcal{B}_1(\mathbf{X}_i))\}_{i=1}^{N}$, abbreviated as $\mathcal{B}_{12}(\mathbf{X})$ and $\mathcal{B}_{12}(\mathbf{X}_i)$, where $\mathbf{X}_i$ is with the size of $C \times T \times W \times H$ and $\mathcal{B}_{12}(\mathbf{X}_i)$ is with the size of $C_2 \times T_2 \times W_2 \times H_2$. After that, we present a mask-guided action attention module, denoted as $\mathcal{A}$, to effectively localize the human-centric foreground actions and emphasize discriminative target action regions. In contrast to previous attention mechanisms that did not require supervision, we utilize the human-centric foreground action mask as the ground truth of the attention map to supervise the learning of mask-guided action attention. For each $\mathcal{B}_{12}(\mathbf{X}_i)$, we apply the average and maximum pooling operations to obtain the mean and maximum spatial attention maps, i.e., $\mathbf{A}_{\text{mea}}$ and $\mathbf{A}_{\text{max}}$, calculated by $\mathbf{A}_{\text{mea}}^{t,w,h} = \frac{1}{C_2}\sum_{c=1}^{C_2} \mathcal{B}_{12}(\mathbf{X}_i)_c^{t,w,h}$ and $\mathbf{A}_{\text{max}}^{t,w,h} = \mathcal{B}_{12}(\mathbf{X}_i)_{c^\star}^{t,w,h}$, where $c^\star = \text{argmax}_c(\mathcal{B}_{12}(\mathbf{X}_i)_c^{t,i,j})$ and c denotes the c-th channel. i, j are the spatial coordinates in feature maps, and t is the temporal index of the feature map in each snippet. Then, $\mathbf{A}_{\text{mea}}$ and $\mathbf{A}_{\text{max}}$ are aggregated into the final one-channel attention map $\mathbf{A}$ via the function $f_{\mathcal{A}}$, allowing the network to learn and update the weights between $\mathbf{A}_{\text{mea}}$ and $\mathbf{A}_{\text{max}}$. The activation

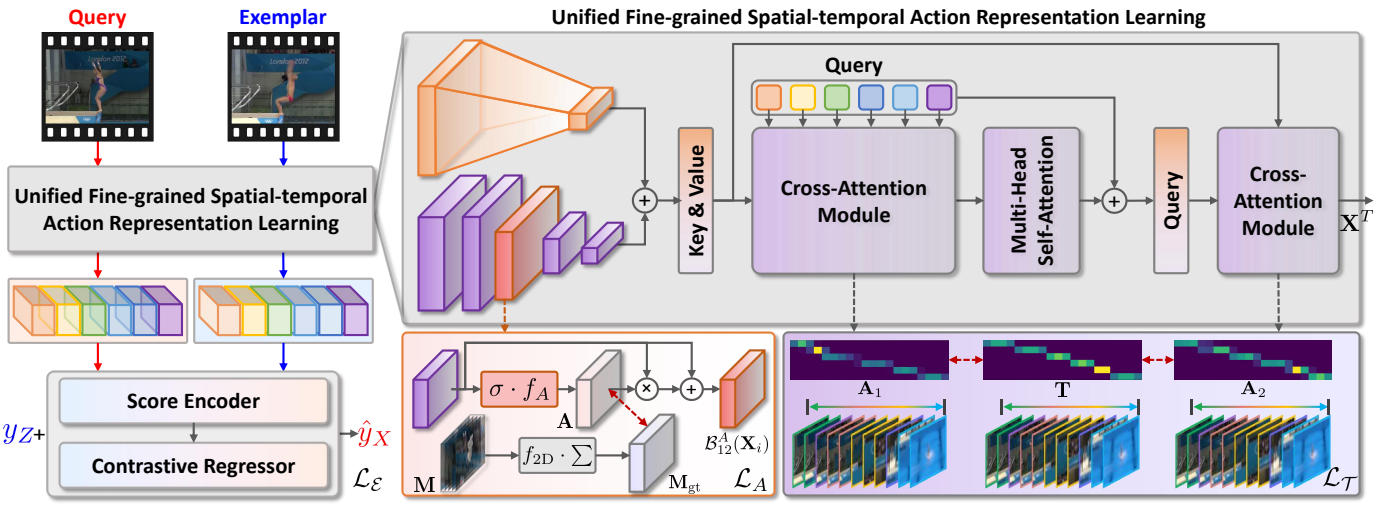

Fig. 2: The architecture of the proposed **Uni-FineParser**. Given a pair of query and exemplar videos, Uni-FineParser compactly incorporates spatial action parsing, temporal action parsing, static visual encoding, and fine-grained contrastive regression into a unified fine-grained spatial-temporal action representation learning framework, achieving human-centric action quality assessment. The red dashed lines represent the loss functions utilized for optimizing the model. The pink and gray dashed lines point to the $\mathcal{A}$ and $\mathcal{T}$ modules, respectively, containing more details of $\mathcal{L}_\mathcal{A}$ and $\mathcal{L}_\mathcal{T}$.

function σ is applied to ensure the values of the attention map fall into [0, 1]. It can be represented as:

$$\mathbf{A} = \sigma(f_\mathcal{A}(\mathbf{A}_{\text{mea}}, \mathbf{A}_{\max})). \tag{2}$$

Consequently, we introduce the human-centric foreground action mask $\mathbf{M}$ to guide learning the attention map $\mathbf{A}$. Concretely, we first sum two adjacent frames and then utilize a 2D adaptive average pooling f_{2D} over the spatial dimension to resize the mask $\mathbf{M}$ to the same size of $\mathbf{A}$, which can be represented as:

$$\mathbf{M}_{\text{gt}}^{w,h} = f_{2D}\left(\sum_{t=1}^{T/2}\left(\mathbf{M}^{2t-1,w,h} + \mathbf{M}^{2t,w,h}\right)\right), \tag{3}$$

where $\mathbf{M}_{\text{gt}}^{w,h}$ and $\mathbf{A}$ have the same size, i.e., $T_2 \times W_2 \times H_2$. Now, a resized ground-truth action mask $\mathbf{M}_{\text{gt}}^{w,h}$ supervises the learning of the attention map $\mathbf{A}$ and the optimization of $\mathbf{A}$ can be formulated by:

$$\mathcal{L}_\mathcal{A} = \frac{1}{T_1 W_1 H_1}\sum_{t=1}^{T_1}\sum_{w=1}^{W_1}\sum_{h=1}^{H_1}\|\mathbf{A}^{t,w,h} - \mathbf{M}_{\text{gt}}^{t,w,h}\|, \tag{4}$$

where w and h are the spatial coordinates in $\mathbf{M}_{\text{gt}}^{w,h}$ and $\mathbf{A}$, and t is the temporal index. The above spatial action parsing process is embedded into the I3D backbone of Uni-FineParser rather than a separate module. After $\mathcal{B}_{12}(\mathbf{X}_i)$ guided by human-centric foreground action mask $\mathbf{M}_{\text{gt}}^{w,h}$, the short-term local features $\mathcal{B}_{12}(\mathbf{X}_i)$ has been enhanced by

$$\mathcal{B}_{12}^A(\mathbf{X}_i) = \mathcal{B}_{12}(\mathbf{X}_i) \odot (\mathbf{A} + \mathbf{1}), \tag{5}$$

where $\odot$ denotes the element-wise multiplication operator and $\mathbf{1}$ indicates the matrix where every entry equals one. $\mathcal{B}_{12}^A(\mathbf{X}_i)$ continues to be fed into the third and the fourth I3D submodules that outputs the short-term global features $\mathcal{B}_4(\mathcal{B}_3(\mathcal{B}_{12}^A(\mathbf{X}_i)))$, denoted as $\mathcal{B}_4^A(\mathbf{X}_i)$.

Temporal Action Parsing. Based on $\mathcal{B}_4^A(\mathbf{X}_i)$, we parse the target action at the temporal level, splitting it into consecutive steps with semantic and temporal correspondences. Inspired by DETR [47] and TPT [15], we construct a DETR-like temporal decoder that introduces learnable queries to model the temporal relationships between consecutive steps of the target action. Our temporal decoder (i.e., $\mathcal{T}$) alternately utilizes the cross-attention and multi-head self-attention layers to capture different temporal attention scores supervised by temporal soft labels created by consecutive step transitions. Finally, our temporal decoder outputs the target action-aware video features with semantic consistency and temporal correspondence.

Concretely, given target action-aware video features $\mathcal{B}_4^A(\mathbf{X}) = \{\mathcal{B}_4^A(\mathbf{X}_i)\}_{i=1}^N$ and static visual features $\mathbf{X}^V = \{\mathbf{X}_i^V\}_{i=1}^N$ (will be depicted in **Static Visual Encoding**), we sum these features as $\mathbf{X}^F = \{\mathcal{B}_4^A(\mathbf{X}_i) + \mathbf{X}_i^V\}_{i=1}^N$ and input them into our temporal decoder $\mathcal{T}$ that consists of two cross-attention layers and a multi-head self-attention layer, alternatively stacking in the manner of *Cross-Attention*$\longrightarrow$*Self-Attention*$\longrightarrow$*Cross-Attention*. First, we set $\mathbf{X}^F \in \mathbb{R}^{N\times C_4}$ as the key and value and take a set of L_q learnable temporal query embeddings $\mathbf{Q} \in \mathbb{R}^{L_q \times C_q}$ as the query for the first cross-attention layer. Each query embedding represents the duration of the target action belonging to a certain step. In the first cross-attention layer, given $\mathbf{Q}_1 = \mathbf{W}_1^Q\mathbf{Q}$, $\mathbf{K}_1 = \mathbf{W}_1^K\mathbf{X}^F$, and $\mathbf{V}_1 = \mathbf{W}_1^V\mathbf{X}^F$, then the first attention scores are $\mathbf{A}_1 = \text{softmax}(\mathbf{Q}_1\mathbf{K}_1^\top/\tau)$ and new generated features are $\mathbf{X}_1^F = \mathbf{A}_1\mathbf{V}_1$. Furthermore, in the multi-head self-attention layer, we set $\mathbf{X}_1^F$ as the query, key, and value, and obtain $\mathbf{X}_2^F = \text{MHSA}(\mathbf{X}_1^F)$. In the second cross-attention layer, we set $\mathbf{X}^F$ as the key and value and $\mathbf{X}_2^{FQ}$ as the query, then the second attention scores are $\mathbf{A}_2 = \text{softmax}(\mathbf{Q}_2\mathbf{K}_2^\top/\tau)$, where $\mathbf{X}_2^{FQ} = \sum(\mathbf{X}_2^F + \mathbf{Q})$, $\mathbf{Q}_2 = \mathbf{W}_2^Q\mathbf{X}_2^{FQ}$, $\mathbf{K}_2 = \mathbf{W}_2^K\mathbf{X}^F$, and $\mathbf{V}_2 = \mathbf{W}_2^V\mathbf{X}^F$. Finally, enhanced target action-aware video features are $\mathbf{X}^T = \mathbf{A}_2\mathbf{V}_2$. The above process can be summarized as $\mathbf{X}^T = \mathcal{T}(\mathbf{X}^F)$.

Consequently, we construct temporal soft labels created by consecutive step transitions to supervise two temporal attention scores $\mathbf{A}_1$ and $\mathbf{A}_2$, with the size of $N \times L_q$, obtained by our temporal decoder $\mathcal{T}$. In particular, given

that the input video $\mathbf{X}$ contains L frames, we construct N snippets via uniform sampling of N frames with every α frame interval, where each sampled frame is regarded as the starting frame, and each snippet consists of β consecutive frames. We convert the ground-truth step transitions in the [1, L] range to corresponding snippets from [1, N]. For instance, if the t_1-th and t_2-th frames are the ground-truth step transitions, then they can be converted to new step transitions $\lfloor t_1//\alpha \rfloor$ and $\lfloor t_2//\alpha \rfloor$ for N snippets. Based on two new transitions, we further divide N snippets into L_q steps, corresponding to the number of learnable temporal query embeddings. We select the central frame of each snippet as the mean and set the standard deviation as 1.3 to model the step transition distribution for the entire video. Therefore, we formulate the temporal soft labels as $\mathbf{T} \in \mathbb{R}^{N \times L_q}$, where the row indicates different snippets and the column indicates which snippets are more focused on in various steps. To ensure each query embedding also focuses on the corresponding step, we introduce the above distribution to supervise two temporal attention scores, $\mathbf{A}_1$ and $\mathbf{A}_2$, and employ the KL loss to optimize two cross-attention layers in our temporal decoder, $\mathcal{T}$, which is

$$\mathcal{L}_{\mathcal{T}} = \mathbf{T}(\log \mathbf{T} - \log \mathbf{A}_1) + \mathbf{T}(\log \mathbf{T} - \log \mathbf{A}_2). \quad (6)$$

Static Visual Encoding. Except for dynamic visual features from the target action, we enhance the target action representation by adding static visual features from each video snippet, especially for high-speed and complex actions in competitive sports.

Concretely, for each video snippet $\mathbf{X}_i$, we select its central adjacent frames and adopt ResNet34 (i.e., $\mathcal{V}$) to extract visual features, where central adjacent frames capture the core visual content described in this snippet. It can be represented as $\mathbf{X}_i^V = \mathrm{Concat}(\mathcal{V}(\mathbf{X}_i[\frac{T}{2} : \frac{T}{2}+2, :]))$, where T is the length of each snippet and Concat denotes the concatenation operation along the channel dimension. Therefore, the static visual features of the input video are denoted as $\mathbf{X}^V = \{\mathbf{X}_i^V\}_{i=1}^N$.

Fine-grained Contrastive Regression. Through the above three components, we obtain enhanced target action-aware video features $\mathbf{X}^T$ for query and exemplar videos. Then, we design a contrastive score encoder $\mathcal{E}_S$ and a contrastive score regressor $\mathcal{E}_R$ to calculate a set of score differences $\{\Delta_m\}_{m=1}^{L_q}$ at the fine-grained level, predicting the action quality score of the query video by referring to that of the exemplar video.

Concretely, we split $\mathbf{X}^T$ into the query video features $\mathbf{X}_X^T$ and the exemplar video features $\mathbf{X}_Z^T$ along the batch dimension and then concatenate them along the channel dimension, denoted as $\mathbf{X}^E$. The contrastive score encoder $\mathcal{E}_S$ contains a two-layer MLP with ReLU non-linearity, and the contrastive score regressor $\mathcal{E}_R$ is a three-layer MLP with ReLU non-linearity. Through $\mathcal{E}_S$, we obtain contrastive score features by encoding video features $\mathbf{X}^E$, capturing differences and similarities between the query and exemplar videos by introducing non-linear transformations through activation functions, denoted as $\mathbf{X}_S^E$. After that, $\mathbf{X}_S^E$ is fed to $\mathcal{E}_R$ that outputs a set of score differences $\mathbf{\Delta} = \{\Delta_m\}_{m=1}^{L_q}$ of L_q steps for the pairwise query and exemplar videos $(\mathbf{X}, \mathbf{Z})$, where $\mathbf{\Delta} = \mathcal{E}_R(\mathcal{E}_S(\mathbf{X}^E))$. $\Delta_m > 0$ indicates the action quality score of the m-step is higher than the exemplar video's, where the score increment is Δ_m. $\Delta_m < 0$ indicates the action quality score of the m-step is lower than the exemplar video's, where the score decrement is Δ_m. Therefore, the final predicted score $\hat{y}_X$ of the query video is calculated by

$$\hat{y}_X = y_Z + \frac{1}{L_q} \sum_{m=1}^{L_q} \Delta_m, \quad (7)$$

where y_Z is the ground truth score of the exemplar video $\mathbf{Z}$.

3.3 Training and Inference

Training. Given a pairwise query and exemplar videos $(\mathbf{X}, \mathbf{Z})$ from the training set, Uni-FineParser is optimized by minimizing the following losses:

$$\mathcal{L} = \mathcal{L}_{\mathcal{A}} + \mathcal{L}_{\mathcal{T}} + \mathcal{L}_{\mathcal{E}}. \quad (8)$$

where $\mathcal{L}_{\mathcal{A}}$ is the attention loss calculated by Eq. (4), which is used to optimize the attention of human-centric foreground action, i.e., attention on the spatial level. $\mathcal{L}_{\mathcal{T}}$ is the Attn Loss calculated by Eq. (6), which is used to optimize the attention of different stages of the target action, i.e., attention on the temporal level. $\mathcal{L}_{\mathcal{E}}$ is the AQA loss, which is used to optimize the predicted score of the query video by minimizing the mean squared error between the ground truth score y_X and the prediction $\hat{y}_X$. $\mathcal{L}_{\mathcal{E}}$ can be formulated as:

$$\mathcal{L}_{\mathcal{E}} = \|\hat{y}_X - y_X\|^2. \quad (9)$$

Inference. For a query video $\mathbf{X}$ from the testing set, the multi-exemplar voting strategy [14] is adopted to select M exemplars $\{\mathbf{Z}_j)\}_{j=1}^M$ from the training set and construct pairwise $\{(\mathbf{X}, \mathbf{Z}_j)\}_{j=1}^M$ with scores $\{y_{Z_j}\}_{j=1}^M$. The inference process can be written as

$$\hat{y}_X = \frac{1}{M} \textstyle\sum_{j=1}^M (\mathcal{F}(\mathbf{X}, \mathbf{Z}_j; \mathbf{\Theta}) + y_{Z_j}). \quad (10)$$

4 Experiments

4.1 Datasets

We evaluated our Uni-FineParser on three widely used AQA datasets: FineDiving, AQA-7, and MTL-AQA. To demonstrate its superiority, we first constructed additional mask annotations for the above three datasets, namely FineDiving-HM, AQA-7-HM, and MTL-AQA-HM.

FineDiving-HM. It provides additional human-centric action mask annotations for the FineDiving dataset [9] and improves it as FineDiving-HM. Like FineDiving, FineDiving-HM contains 3,000 videos across 52 action types, 29 sub-action types, 23 difficulty degree types, fine-grained temporal boundaries, and official action scores. Conversely, FineDiving-HM contains 312,256 action mask annotations, each labeling the target action region to distinguish the human-centric foreground from the background. FineDiving-HM mitigates the problem of requiring frame-level annotations to understand human-centric actions at fine-grained spatial and temporal levels. Fig. 3 shows some examples of human-centric action mask annotations, which precisely focus on foreground target actions. For 312,256 foreground action masks in FineDiving-HM, the number of action masks for individual diving is 248,713, and that for synchronized diving is 63,543. As shown in Fig. 4 (a), the

This article has been accepted for publication in IEEE Transactions on Pattern Analysis and Machine Intelligence. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2025.3556935

Fig. 3: Examples of human-centric action mask annotations for FineDiving. The right column indicates the action type.

largest number of action masks is 35,287, belonging to the action type 107B; the second largest number of action masks is 34,054, belonging to the action type 407C; and the smallest number of action masks is 101, corresponding to the action types 109B, 201A, 201C, and 303C. Coaches and athletes can use the above statistics to develop competition strategies, for instance, to understand what led to the rise of 107B and 407C and how athletes gain a competitive edge.

AQA-7-HM. It provides additional human-centric action mask annotations for the AQA-7 dataset [11] and names it AQA-7-HM. Like AQA-7, AQA-7-HM contains 1,189 videos across seven action types and official action scores. Differently, AQA-7-HM contains 152,711 action mask annotations, each labeling the target action region to distinguish the human-centric foreground from the background. FineDiving-HM mitigates the problem of requiring frame-level annotations to understand human-centric actions at fine-grained spatial and temporal levels. Fig. 5 shows some examples of human-centric action mask annotations, which precisely focus on foreground target actions in various sports scenarios. For 152,711 foreground action masks in AQA-7-HM, the numbers of action masks of seven action types are 37979 (Diving), 14582 (Gym Vault), 16347 (BigSki.), 19704 (BigSnow.), 9244 (Sync.3m), 8996 (Sync.10m), and 45859 (Trampoline), respectively, as shown in Fig. 4 (b).

MTL-AQA-HM. It provides additional human-centric action mask annotations for the MTL-AQA dataset [48] and names it MTL-AQA-HM. MTL-AQA-HM contains 1,369 videos, including 1,059 training videos and 353 testing videos, whose annotations contain the degree of difficulty, scores from each judge (7 judges), action classes, and final scores. MTL-AQA-HM has 142,391 action mask annotations, each labeling the target action region to distinguish the human-centric foreground from the background. MTL-AQA-HM mitigates the problem of requiring frame-level annotations to understand human-centric actions at fine-grained spatial and temporal levels. Fig. 6 shows some examples of human-centric action mask annotations, which precisely focus on foreground target actions. For 142,391 foreground action masks in MTL-AQA-HM, the number of action masks for 3m springboard diving is 29,446, and that for 10m platform diving is 112,945, as shown in Fig. 4 (c).

Parameters	Values	Parameters	Values
C	3	C_2	192
C_4	1024	C_q	256
N	20	M	10
α	5	β	8
$W \times H$	224×224	T	8
$W_2 \times H_2$	56×56	T_2	4

TABLE 1: Parameter configurations of Uni-FineParser.

4.2 Evaluation Metrics

Following previous efforts [9], [13], [14], [48], [49], we utilize Spearman's rank correlation (ρ, the higher, the better) and Relative ℓ_2 distance (R_{ℓ_2}, the lower, the better) for evaluating the AQA task. Concretely, the Spearman's rank correlation ρ is defined as:

$$\rho = \frac{\sum_{i=1}^{N}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2 \sum_{i=1}^{N}(\hat{y}_i - \bar{\hat{y}})^2}}. \tag{11}$$

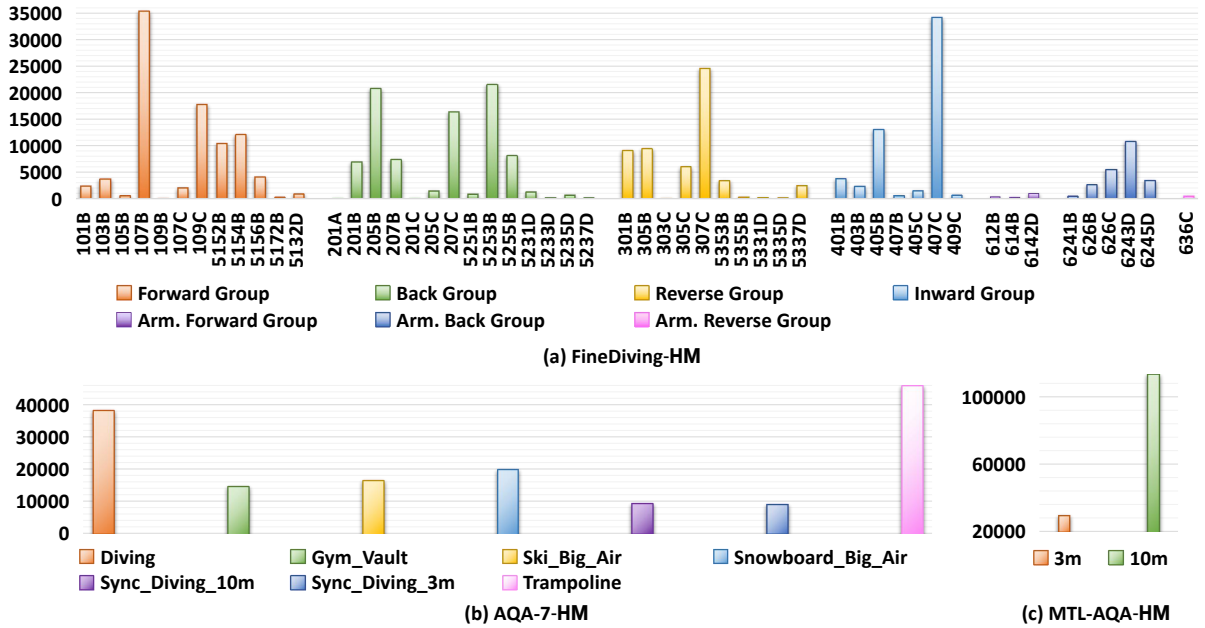

Fig. 4: The distributions of human-centric foreground action masks for the FineDiving, AQA-7, and MTL-AQA datasets.

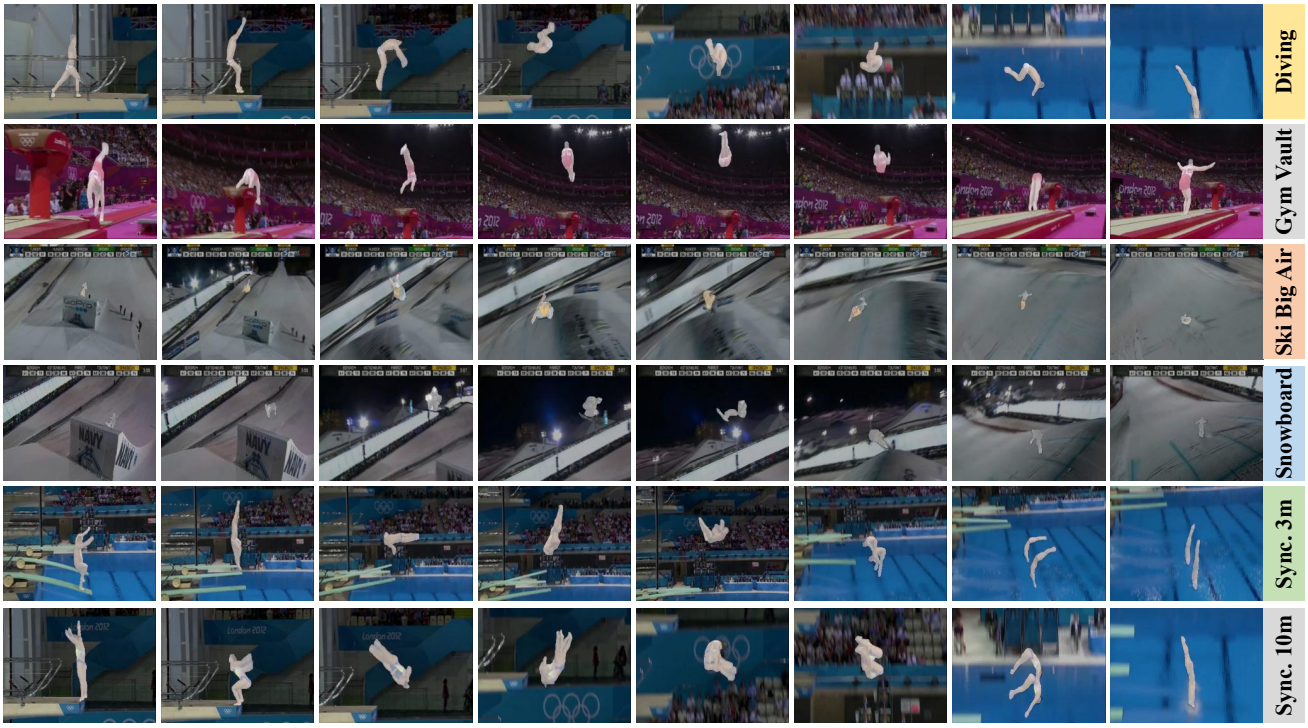

Fig. 5: Examples of human-centric action mask annotations for AQA-7. The right column indicates the action type.

where y_i and $\hat{y}_i$ denote the ranking of two series, respectively. The relative ℓ_2 distance R_{ℓ_2} is defined as:

$$\text{R-}\ell_2 = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{|y_i - \hat{y}_i|}{y_{\max} - y_{\min}}\right) \tag{12}$$

where y_i and $\hat{y}_i$ indicate the ground-truth and predicted scores for the i-th sample, respectively.

4.3 Implementation Details

We adopted the I3D model pre-trained on the Kinetics [46] as the backbone of extracting video features. We set the initial learning rate of the backbone as 10^{-4}. Besides, We adopted ResNet34 as the backbone of encoding static visual features and set the learning rate to 10^{-3}. We extracted 103 frames for each video and split them into 20 snippets, each containing 8 continuous frames with a stride of 5 frames. We followed the multi-exemplar voting strategy used in [9], [14] during inference and the experiment settings of [9], [14] in all the experiments. The function f_A adopted a 1×1 convolution layer. The activation function σ in Eq. (2) used the sigmoid function. In addition, the learnable parameter τ was initially set as 0.07×256 with a learning rate e^{-4}. More parameters are reported in Table 1. We also provided the model parameters, GFLOPs, and convergence rates for Uni-FineParser and FineParser, which were (35.94M, 2444.21

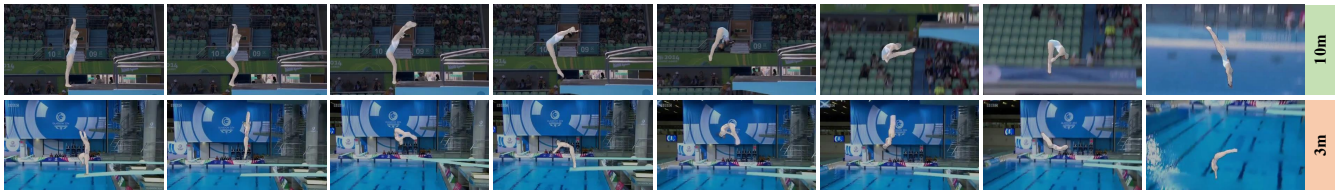

Fig. 6: Examples of human-centric action mask annotations for MTL-AQA. The right column indicates different dives.

Methods	FineDiving		Year
	$\rho\uparrow$	$R\text{-}\ell_2(\times 100)\downarrow$	
C3D-LSTM [39]	0.6969	1.0767	2017
C3D-AVG [48]	0.8371	0.6251	2019
MSCADC [48]	0.7688	0.9327	2019
I3D+MLP [13]	0.8776	0.4967	2020
USDL [13]	0.8830	0.4800	2020
MUSDL [13]	0.9241	0.3474	2020
CoRe [14]	0.9308	0.3148	2021
TSA [9]	0.9324	0.3022	2022
TPT [41]	0.9475	0.2413	2022
STSA [50]	0.9397	0.2707	2024
TSA-MVLA [44]	0.9419	0.2840	2024
RICA^2 [45]	0.9402	0.2838	2024
FineParser [17]	0.9435	0.2602	2024
Uni-FineParser (Ours)	**0.9501**	**0.2294**	2024

TABLE 2: Comparisons of performance with state-of-the-art AQA methods on the FineDiving Dataset. Our result is highlighted in the **bold** format.

GFLOPs, 254 epochs) and (51.76M, 11179.42 GFLOPs, 201 epochs), respectively.

4.4 Results and Analysis

Comparison Results on FineDiving. Table 2 summarizes the experimental results of state-of-the-art AQA methods on the FineDiving dataset. Our Uni-FineParser significantly improves the performance of Spearman's rank correlation and Relative L_2-distance compared to other methods, whose advantages stem from a fine-grained understanding of human-centric foreground actions, i.e., fine-grained parsing actions in time and space. Compared to C3D-LSTM, C3D-AVG, MSCADC, I3D+MLP, USDL, and MUSDL, Uni-FineParser outperforms them significantly and achieved 25.32%, 11.3%, 18.13%, 7.25%, 6.71%, and 2.6% performance improvements in terms of Spearman's rank correlation as well as 0.8473, 0.3957, 0.7033, 0.2673, 0.2506, and 0.118 in Relative ℓ_2-distance. Our Uni-FineParser benefits from the contrastive regression framework to learn the relative scores by making pairwise comparisons, focusing on quantifying the differences between query and exemplar videos for guiding the model to assess the action quality. Compared to CoRe, Uni-FineParser obtains 1.93% and 0.0854 performance improvements on Spearman's rank correlation and Relative ℓ_2-distance since the latter introduces the fine-grained contrastive regression, which finely quantifies the differences between query and exemplar videos by parsing pairwise query and exemplar action instances into consecutive steps with semantic and temporal correspondences for assessment. In addition, Uni-FineParser improves the performance of TSA and its extension STSA on Spearman's rank correlation and Relative ℓ_2-distance. Although STSA tried to introduce spatial motion attention by generating implicit supervision, its performance was still limited due to a lack of annotations of target action masks. Unlike TPT, Uni-FineParser can perform better on Spearman's rank correlation and Relative ℓ_2-distance benefit by explicitly learning human-centric foreground action representations as well as fine-grained alignments in semantics, time, and space. Unlike plug-and-played TSA-MVLA, Uni-FineParser does not need to leverage the language information from additional action knowledge to help learn action spatial-temporal representations. The above action knowledge is less efficient and effective than human-centric action mask annotations. Different from RICA^2, Uni-FineParser considers the prediction uncertainty of step transitions in the TAP module and explicitly introduces human-centric action mask annotations to improve the reliability of AQA. Compared to preliminary FineParser, Uni-FineParser is a unified fine-grained spatial-temporal action parser by redesigning SAP (i.e., spatial action parser) and TAP (temporal action parser) of FineParser and embedding them into a unified and compact learning framework, which performs better and is more efficient than FineParser on Spearman's rank correlation and Relative ℓ_2-distance.

Comparison Results on AQA-7. Table 3 reports the performance of Uni-FineParser compared to other state-of-the-art AQA methods on the AQA-7 dataset. It can be seen that our Uni-FineParser substantially outperforms other methods across different action categories. Our approach can achieve state-of-the-art results compared to other methods for average Spearman's rank correlation, especially in the Diving, Gym Vault, BigSki, and Sync. 10m categories. It is noteworthy that TPT and TSA-Net have slightly superiorities in the BigSnow. and Sync. 3m categories since the foreground target actions' regions in these two sports are difficult to capture better, resulting in performance degradation. In a word, in events that deviate from diving scenarios, such as Gym Vault and BigSki., our Uni-FineParser still performs better, demonstrating the generalization of Uni-FineParser.

Comparison Results on MTL-AQA. Table 4 reports the experimental results of state-of-the-art AQA methods on the MTL-AQA dataset. Our Uni-FineParser outperforms other methods on Spearman's rank correlation and Relative ℓ_2-distance. For instance, our Uni-FineParser achieves better AQA performance than CoRe, RICA^2, and TPT, demonstrating the effectiveness of additional human-centric foreground action masks and the meticulous design of a fine-grained action understanding of Uni-FineParser. Our Uni-FineParser further improves the AQA performance of preliminary FineParser since redesigning SAP and TAP modules and embedding them into a unified and compact learning framework is more beneficial for flexibly capturing target action representations, finely parsing action procedures, and accurately assessing action qualities.

$\rho \uparrow$	Diving	Gym Vault	BigSki.	BigSnow.	Sync. 3m	Sync. 10m	Avg. $\rho \uparrow$	Year
Pose+DCT [37]	0.5300	0.1000	-	-	-	-	-	2014
C3D-LSTM [39]	0.6047	0.5636	0.4593	0.5029	0.7912	0.6927	0.6165	2017
C3D-SVR [39]	0.7902	0.6824	0.5209	0.4006	0.5937	0.9120	0.6937	2017
ST-GCN [51]	0.3286	0.5770	0.1681	0.1234	0.6600	0.6483	0.4433	2018
JRG [49]	0.7630	0.7358	0.6006	0.5405	0.9013	0.9254	0.7849	2019
USDL [13]	0.8099	0.7570	0.6538	0.7109	0.9166	0.8878	0.8102	2020
CoRe [14]	0.8824	0.7746	0.7115	0.6624	0.9442	0.9078	0.8401	2021
TSA-Net [40]	0.8379	0.8004	0.6657	0.6962	0.9493	0.9334	0.8476	2021
TPT [41]	0.8969	0.8043	0.7336	0.6965	0.9456	0.9545	0.8715	2022
FineParser [17]	0.8645	0.7429	0.4615	0.5867	0.9158	0.9112	0.7967	2024
Uni-FineParser (Ours)	**0.9180**	**0.8051**	**0.7614**	**0.6639**	**0.9407**	**0.9605**	**0.8773**	2024
$R\text{-}\ell_2(\times 100) \downarrow$	Diving	Gym Vault	BigSki.	BigSnow.	Sync. 3m	Sync. 10m	Avg. $R\text{-}\ell_2(\times 100) \downarrow$	
C3D-SVR [39]	1.53	3.12	6.79	7.03	17.84	4.83	6.86	2017
USDL [13]	0.79	2.09	4.82	4.94	0.65	2.14	2.57	2020
CoRe [14]	0.64	1.78	3.67	3.87	0.41	2.35	2.12	2021
TPT [41]	0.53	1.69	2.89	3.30	0.33	1.33	1.68	2022
FineParser [17]	0.77	2.08	7.39	4.18	0.74	2.17	2.89	2024
Uni-FineParser (Ours)	**0.43**	**1.86**	**3.28**	**3.37**	**0.54**	**0.81**	**1.72**	2024

TABLE 3: Comparisons of performance with state-of-the-art AQA methods on the AQA-7 dataset. Our result is highlighted in the **bold** format.

Methods	MTL-AQA		Year
	$\rho \uparrow$	$R\text{-}\ell_2(\times 100) \downarrow$	
Pose+DCT [37]	0.2682	/	2014
C3D-SVR [39]	0.7716	/	2017
C3D-LSTM [39]	0.8489	/	2017
C3D-AVG-STL [48]	0.8960	/	2019
C3D-AVG-MTL [48]	0.9044	/	2019
USDL [13]	0.9231	0.4680	2020
MUSDL [13]	0.9273	0.4510	2020
TSA-Net [40]	0.9422	/	2021
CoRe [14]	0.9512	0.2600	2021
TPT [15]	0.9607	0.2378	2022
NAE [43]	0.9430	0.3400	2024
RICA^2 [45]	0.9594	0.2580	2024
FineParser [17]	0.9585	0.2411	2024
Uni-FineParser (Ours)	**0.9622**	**0.2299**	2024

TABLE 4: Comparisons of performance with state-of-the-art AQA methods on the MTL-AQA dataset. Our result is highlighted in the **bold** format.

# of Query Emb. (L_q)	$\rho \uparrow$	$R\text{-}\ell_2(\times 100) \downarrow$
1	0.9447	0.2525
3	0.9446	0.2453
5	0.9473	0.2487
6	**0.9501**	**0.2294**
9	0.9497	0.2375
20	0.9481	0.2390

TABLE 5: Ablation study on different numbers of query embeddings in Uni-FineParser.

4.5 Ablation Studies

We conducted a series of ablation studies on the FineDiving-HM dataset to investigate different configurations and demonstrate the effectiveness of Uni-FineParser.

Effect of Different Numbers of Temporal Query Embeddings. We investigated different numbers of learnable temporal query embeddings in the TAP module on the FineDiving-HM dataset and summarized the experimental results in Table 5. We find that the AQA performance of our Uni-FineParser in the case of $L_q = 6$ is better than in the case of other L_q values (i.e., 1, 3, 5, 9, and 20).

Methods	Modules			$\rho \uparrow$	$R\text{-}\ell_2(\times 100) \downarrow$
	SAP	SVE	TAP		
Base				0.9423	0.2512
Base+SAP	✓			0.9454	0.2442
Base+TAP			✓	0.9473	0.2402
Base+SVE		✓		0.9455	0.2509
w/o SVE	✓		✓	0.9491	0.2421
w/o TAP	✓	✓		0.9433	0.2521
w/o SAP		✓	✓	0.9456	0.2438
Ours	✓	✓	✓	**0.9501**	**0.2294**

TABLE 6: Ablation study on different modules in Uni-FineParser.

Methods	Losses $\mathcal{L}$			$\rho \uparrow$	$R\text{-}\ell_2(\times 100) \downarrow$
	$\mathcal{L}_\mathcal{E}$	$\mathcal{L}_\mathcal{T}$	$\mathcal{L}_\mathcal{A}$		
w/o $\mathcal{L}_\mathcal{T}$+$\mathcal{L}_\mathcal{A}$	✓			0.9444	0.2493
w/o $\mathcal{L}_\mathcal{A}$	✓	✓		0.9472	0.2409
w/o $\mathcal{L}_\mathcal{T}$	✓		✓	0.9487	0.2419
Ours	✓	✓	✓	**0.9501**	**0.2294**

TABLE 7: Ablation study on different loss configurations of Uni-FineParser.

The L_q value determines the granularity of parsing the target action at the temporal level. $L_q = 1$ is an extreme case without action parsing along the temporal axis. In a diving video, the action procedure is usually split into take-off, flight, and entry stages. The model without action parsing cannot correctly understand the internal structure of the action procedure. Therefore, the AQA performance of $L_q = 1$ is worse than that of other L_q values. With L_q increasing, the parsing granularity of the target action is getting finer and finer. $L_q = 6$ can achieve state-of-the-art since it is a proper parsing granularity to split the target action into consecutive steps with semantic and temporal correspondences. However, as L_q continues to increase, the AQA performance degrades. For example, $L_q = 20$ may affect the accuracy of understanding the action procedure

Backbones	$\rho\uparrow$	$R\text{-}\ell_2(\times 100)\downarrow$
ResNet18	0.9475	0.2396
ResNet34	**0.9501**	**0.2294**
ResNet50	0.9448	0.2463
ViT-s/16	0.9458	0.2388

TABLE 8: Ablation study on different backbones of SVE in Uni-FineParser.

Stage	$\rho\uparrow$	$R\text{-}\ell_2(\times 100)\downarrow$
1	0.9468	0.2550
2	**0.9501**	**0.2294**
3	0.9475	0.2380

TABLE 9: Ablation study on embedding target action attention into different stages of the I3D backbone in Uni-FineParser.

Methods	$\rho\uparrow$	$R\text{-}\ell_2(\times 100)\downarrow$
$\mathbf{A}+\mathbf{1}$	**0.9501**	**0.2294**
$\mathbf{A}$	0.9499	0.2329
$\mathbf{A}_1$ (single-head) & $\mathbf{A}_2$ (single-head)	**0.9501**	**0.2294**
$\mathbf{A}_1$ (multi-head) & $\mathbf{A}_2$ (multi-head)	0.9486	0.2352
$\mathbf{A}_1$ (multi-head) & $\mathbf{A}_2$ (single-head)	0.9476	0.2384
$\mathbf{A}_1$ (single-head) & $\mathbf{A}_2$ (multi-head)	0.9498	0.2380

TABLE 10: Ablation study on different designs of $\mathbf{A}$, $\mathbf{A}_1$, and $\mathbf{A}_2$ in Uni-FineParser.

Midpoint	$\rho\uparrow$	$R\text{-}\ell_2(\times 100)\downarrow$
Left	0.9481	0.2406
Middle	**0.9501**	**0.2294**
Right	0.9486	0.2409

TABLE 11: Ablation study on different soft labels in Uni-FineParser.

due to excessive parsing.

Effect of Different Modules of Uni-FineParser. We investigated the effects of different configurations of the model on the AQA performance of Uni-FineParser, summarized in Table 6. *Base* indicates the baseline method that utilizes the pre-trained I3D to extract visual features from the input video, alternately two cross-attention and a multi-head self-attention layers to obtain the target action-aware video features, and equips with a contrastive score encoder and a contrastive score regressor to predict the action score. *Base+SAP* indicates the baseline method that introduces an attention module into I3D to localize the human-centric foreground actions and emphasize discriminative target action regions. *Base+TAP* refers to the baseline method that introduces learnable queries to model the temporal relationships between consecutive steps of the target action. *Base+SVE* denotes the baseline method equipped with a static visual encoder (SVE), where SVE utilizes ResNet34 to extract static visual features to enhance the target action representations. The above three methods improve the AQA performance compared to *Base* by modeling dynamic spatial and temporal information as well as static visual information, respectively. We see that *TAP* has more significant improvement on *Base* than *SAP* and *SVE*. In addition, *w/o SAP* indicates the method that removes the SAP module from our Uni-FineParser, ignoring the construction of a mask-guided action attention module that captures the target action regions and enhances the action representations. *w/o SVE* indicates the method that removes the SVE module from our Uni-FineParser, ignoring the enhancement of the target action representation via encoding static visual features. *w/o TAP* indicates the method that removes the TAP module from our Uni-FineParser, which cannot parse the target action at the temporal level and split it into consecutive steps with semantic and temporal correspondences as well as deeply understand the internal structure of the action procedure. We find that *w/o TAP* leads to more performance degradation than *w/o SAP* and *w/o SVE*, demonstrating the importance and effectiveness of parsing the target action at the temporal level. In contrast, our Uni-FineParser incorporates three modules (i.e., spatial action parsing, temporal action parsing, and static visual encoding) into the unified fine-grained spatial-temporal action representation learning framework to achieve more accurate and interpretable human-centric action quality assessment.

Effect of Different Loss Configurations of Uni-FineParser. We studied the influence of minimizing different losses of Uni-FineParser, as shown in Table 7. w/o $\mathcal{L}_\mathcal{T}$+$\mathcal{L}_\mathcal{A}$ indicates only using the AQA loss $\mathcal{L}_\mathcal{E}$ that is used to optimize a contrastive score encoder $\mathcal{E}_S$ and a contrastive score regressor $\mathcal{E}_R$ by minimizing the mean squared error between the ground truth and the prediction scores, calculated by Eq. (9). w/o $\mathcal{L}_\mathcal{A}$ denotes using the AQA loss $\mathcal{L}_\mathcal{E}$ and the attention loss $\mathcal{L}_\mathcal{T}$, where $\mathcal{L}_\mathcal{T}$ is used to optimize the attention of different stages of the target action, i.e., attention on the temporal level, calculated by Eq. (6). w/o $\mathcal{L}_\mathcal{T}$ denotes using the AQA loss $\mathcal{L}_\mathcal{E}$ and the attention loss $\mathcal{L}_\mathcal{A}$ calculated by Eq. (4), where $\mathcal{L}_\mathcal{A}$ is used to optimize the attention of human-centric foreground action, i.e., attention on the spatial level. Our Uni-FineParser significantly outperforms other loss configurations on FineDiving-HM, demonstrating the effectiveness of Uni-FineParser's loss configuration.

Effect of Different Backbones of Static Visual Encoding. We conducted several experiments on the FineDiving-HM dataset to investigate the effects of different backbones of SVE on the AQA performance. In Table 8, ResNet34 outperforms other ResNet architectures and ViT-s/16. ResNet34 has a deeper network depth than ResNet18, allowing it to capture more global features and high-level semantics, whereas ResNet50 may lead to overfitting on two central adjacent frames. Besides, ViT-s/16 enables the model to capture long-term dependencies among video frames rather than local relationships, which is beneficial to learning target action representations by capturing global features, improving the AQA performance of Uni-FineParser, while slightly inferior to ResNet34 due to only containing 12 layers.

Effect of Embedding Target Action Attention into Different Stages of the I3D backbone. We studied the influence of embedding spatial attention into different stages of I3D on the AQA performance, as shown in Table 9. *Stage* i indicates that we introduce a spatial attention module after the output of the i-th I3D submodule and supervise it with the human-centric foreground action mask to construct a mask-guided action attention module that captures the target action regions and enhances the action representations from the i-th I3D submodule. After that, by passing from the $i+1$ to the last I3D submodules, we can obtain the

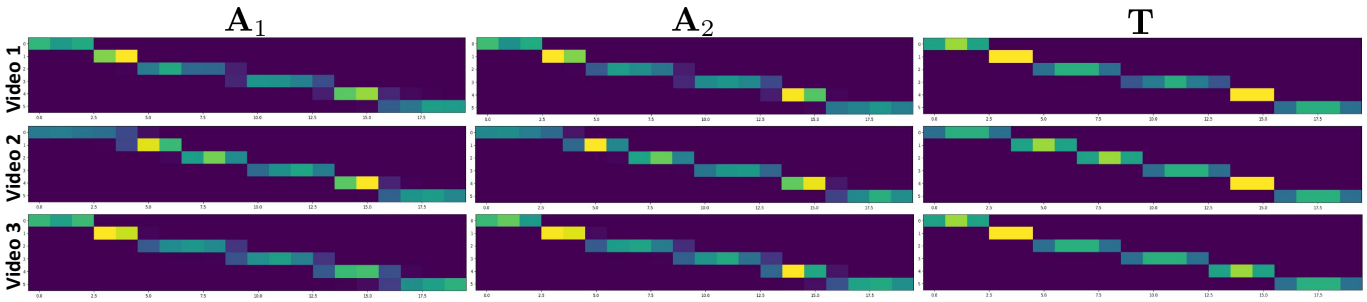

Fig. 7: Visualization of temporal attention scores $\mathbf{A}_1$ and $\mathbf{A}_2$ supervised by temporal soft labels $\mathbf{T}$ in Uni-FineParser.

Exemplar Selection	$\rho \uparrow$	$R\text{-}\ell_2(\times 100) \downarrow$
AN	**0.9501**	**0.2294**
DD	0.9557	0.1849

TABLE 12: Ablation study on different strategies of exemplar choice in Uni-FineParser.

Snippets (N)	FineParser		Uni-FineParser	
	$\rho \uparrow$	$R\text{-}\ell_2(\times 100) \downarrow$	$\rho \uparrow$	$R\text{-}\ell_2(\times 100) \downarrow$
9	0.9435	0.2602	0.9415	0.2569
12	0.9335	0.2916	0.9407	0.2651
15	0.9195	0.3493	0.9427	0.2593
20	0.9222	0.3411	**0.9501**	**0.2294**

TABLE 13: Ablation study on different numbers of snippets in FineParser and Uni-FineParser.

video features spanning the short-term local features of the first I3D submodule to the short-term global features of the last I3D submodule. We see that *Stage 2* can achieve better AQA performance compared to *Stage 1* and *Stage 3*. This is because *Stage 1* focuses on more local features and lacks global features to learn high-level semantics of the target action region. *Stage 2* embeds the spatial attention module that extracts relatively global features and combines them with local features to capture richer information about the target action region. With the deeper stage further, *Stage 3* may miss local features with more details of the target action region, preventing the action representations from being adequately enhanced.

Effect of Different Designs of $\mathbf{A}$, $\mathbf{A}_1$, and $\mathbf{A}_2$. We investigated how different designs of $\mathbf{A}$, $\mathbf{A}_1$, and $\mathbf{A}_2$ influenced AQA performances. For $\mathbf{A}$ in Spatial Action Parsing, we conducted an ablation study to compare the Uni-FineParser performance of using ($\mathbf{A}$+$\mathbf{1}$) versus using $\mathbf{A}$ only, as shown in Table 10, demonstrating the positive impact of the design choice of ($\mathbf{A}$+$\mathbf{1}$). Because adding $\mathbf{1}$ to $\mathbf{A}$ could ensure that even if the attention score $\mathbf{A}$ is very small or zero, the term ($\mathbf{A}$+$\mathbf{1}$) remains non-zero. This prevents division by zero or numerical instability and ensures that the attention mechanism always maintains a baseline attention level. For $\mathbf{A}_1$ and $\mathbf{A}_2$ in Temporal Action Parsing, we conduct experiments comparing single-head attention and multi-head attention to determine if single-head attention restricts the model's ability to capture diverse interactions, as shown in Table 10. We find that single-head attention slightly outperforms multi-head attention in scenarios where efficiency and interpretability are prioritized.

Effect of Different Soft Label Modeling Strategies. In Subsection 3.2, we selected the central frame of each snippet as the mean and set the standard deviation as 1.3 to model the step transition distribution for the entire video. In Table 11, we investigate the effect of different step transition distributions on AQA performance. Specifically, when we select the leftmost frame of each snippet as the mean and keep the same standard deviation to model the step transition distribution, we find that the AQA performance degrades, especially for $R\text{-}\ell_2$. A similar experimental result also appears when selecting the rightmost frame of each snippet as the mean. Therefore, the current soft label modeling strategy used in our Uni-FineParser is a better choice.

Methods	$\rho \uparrow$	$R\text{-}\ell_2(\times 100) \downarrow$
Uni-FineParser (GTM)	**0.9501**	**0.2294**
Uni-FineParser (PM)	0.9495	0.2334
FineParser [17]	0.9435	0.2602

TABLE 14: Ablation study on Uni-FineParser using ground-truth masks and pseudo masks. GTM: ground-truth masks. PM: pseudo masks.

Effect of Different Exemplar Selection Strategies. We investigated the influence of different exemplar selection strategies used in the inference stage described in Subsection 3.3 as shown in Table 12. *AN* indicates utilizing the action number to select exemplar videos during the training and testing stages without difficulty degree annotations. *DD* selects exemplar videos based on difficulty degrees and regresses the raw score. In the *DD* setting, we multiply the raw score by the difficulty degree to obtain the final predicted score. We see that *DD* can bring performance improvements (0.56% on Spearman's rank correlation and 0.0445 on Relative ℓ_2-distance) to Uni-FineParser. Note that the *DD* setting requires additional difficulty degree annotations, while the *AN* setting does not. In this paper, we adopt the *AN* setting in our experiments for fairness to demonstrate the effectiveness of our Uni-FineParser.

Effect of Different Numbers of Snippets. We investigated the influence of different numbers of snippets (N=9, 12, 15, 20), as shown in Table 13. We observe that Uni-FineParser with 20 snippets captures more fine-grained temporal information and achieves a more accurate assessment, which can mainly be attributed to architectural enhancements (e.g., redesigned SAP and TAP, as well as the score encoder) rather than merely increasing the number of snippets. This can be demonstrated by evaluating FineParser and Uni-FineParser under the same experimental setup (N=9, 12, 15, 20). The results of Uni-FineParser with N=20 surpass those of FineParser with N=20.

4.6 Visualization

We showed some visualization results of temporal attention scores obtained by the TAP module in Fig. 7. It can be seen that TAP can accurately learn two temporal attention scores $\mathbf{A}_1$ and $\mathbf{A}_2$ to closely approximate the ground-truth temporal soft labels $\mathbf{T}$. We can see that $\mathbf{A}_2$ from the second cross-attention layer is closer to $\mathbf{T}$ than $\mathbf{A}_1$ from the first cross-attention layer. This helps obtain accurate step transitions in TAP for flexibly and finely parsing the action procedure.

4.7 Discussion

In this subsection, we discuss the meanings of human-centric action masks and provide experimental comparisons to offer insights into our Uni-FineParser. In Table 14, the experimental results reveal two aspects of significant meaning. Firstly, the AQA performance of Uni-FineParser using ground-truth masks (denoted as GTM) outperforms that of Uni-FineParser using pseudo masks (denoted as PM), emphasizing the necessity of manual correction of action masks. PM indicates using the human action detector, comprised of SAM [52] and OSTrack [53], directly detecting foreground target actions in each video frame. GTM denotes manual refinement and correction of human-centric action masks. Secondly, we observe that the performance gap between GTM and PM is not significant, highlighting the actual advancement of Uni-FineParser in a more realistic setting without ideal annotations. These comparisons also open up possibilities for leveraging our Uni-FineParser in various domains, scenarios, and sports with limited annotated data.

5 Conclusion

We presented an end-to-end fine-grained spatial-temporal action parser named Uni-FineParser for the AQA task. It learned fine-grained representations for target actions by integrating spatial action parsing, temporal action parsing, static visual encoding, and fine-grained contrastive regression to achieve state-of-the-art. To understand human-centric actions from a fine-grained spatial level, we also provided human-centric foreground action mask annotations for the FineDiving, AQA-7, and MTL-AQA datasets, named FineDiving-HM, AQA-7-HM, and MTL-AQA-HM, constructing three new benchmarks for fine-grained human-centric action quality assessment. Uni-FineParser could facilitate more tasks that require a fine-grained understanding of sports.

Acknowledgements

This work was supported by grants from the National Natural Science Foundation of China (61925201, 62132001, 62373043, 62432001), the Beijing Natural Science Foundation (L247006, 4252020), and the Young Elite Scientists Sponsorship Program by CAST (2023QNRC001).

References

[1] M. Fieraru, M. Zanfir, E. Oneata, A.-I. Popa, V. Olaru, and C. Sminchisescu, "Three-dimensional reconstruction of human interactions," in *CVPR*, 2020, pp. 7214–7223. 1

[2] E. Ng, D. Xiang, H. Joo, and K. Grauman, "You2me: Inferring body pose in egocentric video via first and second person interactions," in *CVPR*, 2020, pp. 9890–9900. 1

[3] M. Hassan, P. Ghosh, J. Tesch, D. Tzionas, and M. J. Black, "Populating 3d scenes by learning human-scene interaction," in *CVPR*, 2021, pp. 14 708–14 718. 1

[4] J. Wang, H. Xu, J. Xu, S. Liu, and X. Wang, "Synthesizing long-term 3d human motion and interaction in 3d scenes," in *CVPR*, 2021, pp. 9401–9411. 1

[5] C. Wang, J. Kong, and H. Qi, "Areas of research focus and trends in the research on the application of vr in rehabilitation medicine," in *Healthcare*, vol. 11, no. 14, 2023, p. 2056. 1

[6] D. Gupta, K. Attal, and D. Demner-Fushman, "A dataset for medical instructional video classification and question answering," *Scientific Data*, vol. 10, no. 1, p. 158, 2023. 1

[7] D. Shao, Y. Zhao, B. Dai, and D. Lin, "Finegym: A hierarchical video dataset for fine-grained action understanding," in *CVPR*, 2020, pp. 2616–2625. 1, 2

[8] Y. Li, L. Chen, R. He, Z. Wang, G. Wu, and L. Wang, "Multisports: A multi-person video dataset of spatio-temporally localized sports actions," in *ICCV*, 2021, pp. 13 536–13 545. 1, 2

[9] J. Xu, Y. Rao, X. Yu, G. Chen, J. Zhou, and J. Lu, "Finediving: A fine-grained dataset for procedure-aware action quality assessment," in *CVPR*, 2022, pp. 2949–2958. 1, 3, 5, 6, 7, 8

[10] Y. Cui, C. Zeng, X. Zhao, Y. Yang, G. Wu, and L. Wang, "Sportsmot: A large multi-object tracking dataset in multiple sports scenes," in *ICCV*, 2023, pp. 9921–9931. 1

[11] P. Parmar and B. Morris, "Action quality assessment across multiple actions," in *WACV*, 2019, pp. 1468–1476. 1, 6

[12] C. Xu, Y. Fu, B. Zhang, Z. Chen, Y.-G. Jiang, and X. Xue, "Learning to score figure skating sport videos," *TCSVT*, vol. 30, no. 12, pp. 4578–4590, 2019. 1

[13] Y. Tang, Z. Ni, J. Zhou, D. Zhang, J. Lu, Y. Wu, and J. Zhou, "Uncertainty-aware score distribution learning for action quality assessment," in *CVPR*, 2020, pp. 9839–9848. 1, 3, 6, 8, 9

[14] X. Yu, Y. Rao, W. Zhao, J. Lu, and J. Zhou, "Group-aware contrastive regression for action quality assessment," in *ICCV*, 2021, pp. 7919–7928. 1, 3, 5, 6, 7, 8, 9

[15] Y. Bai, D. Zhou, S. Zhang, J. Wang, E. Ding, Y. Guan, Y. Long, and J. Wang, "Action quality assessment with temporal parsing transformer," in *ECCV*, 2022, pp. 422–438. 1, 4, 9

[16] K. Gedamu, Y. Ji, Y. Yang, J. Shao, and H. T. Shen, "Fine-grained spatio-temporal parsing network for action quality assessment," *TIP*, vol. 32, pp. 6386–6400, 2023. 1, 3

[17] J. Xu, S. Yin, G. Zhao, Z. Wang, and Y. Peng, "Fineparser: A fine-grained spatio-temporal action parser for human-centric action quality assessment," in *CVPR*, 2024, pp. 14 628–14 637. 2, 3, 8, 9, 11

[18] A. Piergiovanni and M. S. Ryoo, "Fine-grained activity recognition in baseball videos," in *CVPRW*, 2018, pp. 1740–1748. 2

[19] Y. Liu, L. Wang, X. Ma, Y. Wang, and Y. Qiao, "Fineaction: A fine-grained video dataset for temporal action localization," *arXiv preprint arXiv:2105.11107*, 2021. 2

[20] J. Gao, M. Chen, and C. Xu, "Fine-grained temporal contrastive learning for weakly-supervised temporal action localization," in *CVPR*, 2022, pp. 19 999–20 009. 2

[21] J. Tan, Y. Wang, G. Wu, and L. Wang, "Temporal perceiver: A general architecture for arbitrary boundary detection," *TPAMI*, vol. 45, no. 10, pp. 12 506–12 520, 2023. 2

[22] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, "Temporal segment networks for action recognition in videos," *TPAMI*, vol. 41, no. 11, pp. 2740–2755, 2018. 2

[23] C. Zhu, X. Tan, F. Zhou, X. Liu, K. Yue, E. Ding, and Y. Ma, "Fine-grained video categorization with redundancy reduction attention," in *ECCV*, 2018, pp. 136–152. 2

[24] J. Munro and D. Damen, "Multi-modal domain adaptation for fine-grained action recognition," in *CVPR*, 2020, pp. 122–132. 2

[25] J. Hong, M. Fisher, M. Gharbi, and K. Fatahalian, "Video pose distillation for few-shot, fine-grained sports action recognition," in *ICCV*, 2021, pp. 9254–9263. 2

[26] Y. Gao, J. Lu, S. Li, N. Ma, S. Du, Y. Li, and Q. Dai, "Action recognition and benchmark using event cameras," *TPAMI*, vol. 45, no. 12, pp. 14 081–14 097, 2023. 2

[27] T. Cooray, N.-M. Cheung, and W. Lu, "Attention-based context aware reasoning for situation recognition," in *CVPR*, 2020, pp. 4736–4745. 2

[28] C. Zhang, A. Gupta, and A. Zisserman, "Temporal query networks for fine-grained video understanding," in *CVPR*, 2021, pp. 4486–4496. 2

[29] Z. Yu, L. Zheng, Z. Zhao, F. Wu, J. Fan, K. Ren, and J. Yu, "Anetqa: A large-scale benchmark for fine-grained compositional reasoning over untrimmed videos," in *CVPR*, 2023, pp. 23 191–23 200. 2

[30] J. Xiao, P. Zhou, A. Yao, Y. Li, R. Hong, S. Yan, and T.-S. Chua, "Contrastive video question answering via video graph transformer," *TPAMI*, vol. 45, no. 11, pp. 13 265–13 280, 2023. 2

[31] S. Chen, Y. Zhao, Q. Jin, and Q. Wu, "Fine-grained video-text retrieval with hierarchical graph reasoning," in *CVPR*, 2020, pp. 10 638–10 647. 2

[32] H. Doughty and C. G. Snoek, "How do you do it? fine-grained action understanding with pseudo-adverbs," in *CVPR*, 2022, pp. 13 832–13 842. 2

[33] J. Dong, X. Li, C. Xu, X. Yang, G. Yang, X. Wang, and M. Wang, "Dual encoding for video retrieval by text," *TPAMI*, vol. 44, no. 8, pp. 4065–4080, 2021. 2

[34] X. Chen, A. Pang, W. Yang, Y. Ma, L. Xu, and J. Yu, "Sportscap: Monocular 3d human motion capture and fine-grained understanding in challenging sports videos," *arXiv preprint arXiv:2104.11452*, 2021. 2

[35] Z. Li, L. He, and H. Xu, "Weakly-supervised temporal action detection for fine-grained videos with hierarchical atomic actions," in *ECCV*, 2022, pp. 567–584. 2

[36] H. Zhang, D. Liu, Q. Zheng, and B. Su, "Modeling video as stochastic processes for fine-grained video representation learning," in *CVPR*, 2023, pp. 2225–2234. 2

[37] H. Pirsiavash, C. Vondrick, and A. Torralba, "Assessing the quality of actions," in *ECCV*, 2014, pp. 556–571. 3, 9

[38] G. I. Parisi, S. Magg, and S. Wermter, "Human motion assessment in real time using recurrent self-organization," in *RO-MAN*, 2016, pp. 71–76. 3

[39] P. Parmar and B. Tran Morris, "Learning to score olympic events," in *CVPRW*, 2017, pp. 20–28. 3, 8, 9

[40] S. Wang, D. Yang, P. Zhai, C. Chen, and L. Zhang, "Tsa-net: Tube self-attention network for action quality assessment," in *ACM MM*, 2021, pp. 4902–4910. 3, 9

[41] Y. Bai, D. Zhou, S. Zhang, J. Wang, E. Ding, Y. Guan, Y. Long, and J. Wang, "Action quality assessment with temporal parsing transformer," in *ECCV*, 2022, p. 422–438. 3, 8, 9

[42] S. Zhang, W. Dai, S. Wang, X. Shen, J. Lu, J. Zhou, and Y. Tang, "Logo: A long-form video dataset for group action quality assessment," in *CVPR*, 2023, pp. 2405–2414. 3

[43] S. Zhang, S. Bai, G. Chen, L. Chen, J. Lu, J. Wang, and Y. Tang, "Narrative action evaluation with prompt-guided multimodal interaction," in *CVPR*, 2024, pp. 18 430–18 439. 3, 9

[44] H. Xu, X. Ke, Y. Li, R. Xu, H. Wu, X. Lin, and W. Guo, "Vision-language action knowledge learning for semantic-aware action quality assessment," in *ECCV*, 2024. 3, 8

[45] A. Majeedi, V. R. Gajjala, S. S. S. N. GNVV, and Y. Li, "Rica^ 2: Rubric-informed, calibrated assessment of actions," in *ECCV*, 2024. 3, 8, 9

[46] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *CVPR*, 2017, pp. 6299–6308. 3, 7

[47] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *ECCV*, 2020, pp. 213–229. 4

[48] P. Parmar and B. T. Morris, "What and how well you performed? a ma multitask learningultitask learning approach to action quality assessment," in *CVPR*, 2019, pp. 304–313. 6, 8, 9

[49] J.-H. Pan, J. Gao, and W.-S. Zheng, "Action assessment by joint relation graphs," in *ICCV*, 2019, pp. 6331–6340. 6, 9

[50] J. Xu, Y. Rao, J. Zhou, and J. Lu, "Procedure-aware action quality assessment: Datasets and performance evaluation," *IJCV*, 2024. 8

[51] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *AAAI*, 2018. 9

[52] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," *arXiv preprint arXiv:2304.02643*, 2023. 12

[53] B. Ye, H. Chang, B. Ma, S. Shan, and X. Chen, "Joint feature learning and relation modeling for tracking: A one-stream framework," 2022. 12

Jinglin Xu is now an Associate Professor in the School of Intelligence Science and Technology at the University of Science and Technology Beijing (USTB), a council member, and deputy secretary-general of the Beijing Society of Image and Graphics (BSIG). Before joining USTB, she was a Postdoctoral Fellow in the Department of Automation at Tsinghua University. She received her Ph.D. degree at Northwestern Polytechnical University. Her research interests include computer vision, video understanding, and fine-grained action analysis, where she has authored 20 papers in top-tier journals and conference proceedings.

Sibo Yin is currently pursuing the Ph.D. degree with the Wangxuan Institute of Computer Technology, Peking University, China. He received his B.S. degree in computer science from Peking University, China, in 2024. His research interests lie in fine-grained action quality assessment and image super-resolution.

Yuxin Peng is the Boya Distinguished Professor at Wangxuan Institute of Computer Technology, Peking University. He was a recipient of the National Science Fund for Distinguished Young Scholars of China. He received the Ph.D. degree in computer application technology from Peking University, Beijing, China, in 2003. His research interests mainly include cross-media analysis and reasoning, image and video recognition and understanding, and computer vision. He has authored over 200 papers, including more than 120 papers in top-tier journals and conference proceedings. He has submitted 51 patent applications and has been granted 39 of them. He led his team to win the First Place in the video semantic search evaluation of TRECVID ten times in recent years. He won the First Prize of the Beijing Science and Technology Award in 2016 (ranking first) and the First Prize of the Scientific and Technological Progress Award of the Chinese Institute of Electronics in 2020 (ranking first). He was a recipient of the Best Paper award at MMM 2019 and NCIG 2018, and serves as the associate editor of IEEE TMM, TCSVT, etc.