Visual-Semantic Alignment Temporal Parsing for Action Quality Assessment

Kumie Gedamu, Yanli Ji, Yang Yang, *Senior Member, IEEE*, Jie Shao, *Member, IEEE*, and Heng Tao Shen

***Abstract*— Action Quality Assessment (AQA) is a challenging task involving analyzing fine-grained technical subactions, aligning high-level visual-semantic representations, and exploring internal temporal structures that capture the overall meaning of given action sequences. To address these challenges, we propose a Visual-semantic Alignment Temporal Parsing Network (VATP-Net) to understand the high-level visual semantics of subaction sequences and internal temporal structures without explicit supervision for action quality assessment. The proposed approach designs a self-supervised temporal parsing module to generate subaction sequences from the given video by aligning the visual and semantic action features. It captures high-level semantics and the internal temporal dynamics of subaction sequences. Furthermore, a multimodal interaction module is proposed to capture the interaction between different modalities of action features, enabling a comprehensive assessment of fine-grained and scene-invariant action details. The proposed module captures the intricate relationships and encourages interactions between different modalities within an action sequence, enhancing the overall understanding of action assessment. We exhaustively evaluate our proposed approach on the MTL-AQA, Rhythmic Gymnastics (RG), FineFS, and Fis-V datasets. Extensive experimental results demonstrate the effectiveness and feasibility of our proposed approach, which outperforms state-of-the-art methods by a significant margin.**

***Index Terms*— Action quality assessment (AQA), self-supervised learning, temporal parsing, multimodal learning.**

I. Introduction

ACTION Quality Assessment (AQA) plays an important role in video analysis, which is the task of assessing how well an action is performed and quantifying specific action sequences. Recently, AQA has gained significant attention in various real-world applications, including sports activities [1], [2], [3], [4], [5], [6] and medical training skills assessments [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18]. By evaluating the quality of specific professional action sequences, AQA approaches can facilitate informed decision-making in healthcare and athletic training [19]. In medical contexts, AQA allows doctors to monitor patient rehabilitation and adjust treatments based on performance [3], [4], [5], [6]. Similarly, coaches leverage AQA to improve athletes' training efficiency. However, AQA poses unique challenges, particularly when dealing with video features with identical action procedures (e.g. "Forward", "Somersaults", "Twist" and "Entry") which occur in similar backgrounds as aquatic centers [20], [21], [22]. As illustrated in Fig. 1(a), most existing AQA approaches rely on visual input action features with holistic representations [1], [2], [3], but often overlook semantic action features in videos. This leads to limitations in their generalization capability, as they neglect additional semantic action features [20], [21], [22], [23]. Consequently, these approaches may perform sub-optimally in diverse, fine-grained, and scene-invariant action sequences. We propose a multimodal visual-semantic alignment approach to overcome these limitations to understand the high-level representation of subaction sequences and internal temporal structures. Furthermore, the proposed approach captures the interaction between different modalities of action features.

The human ability to recognize actions involves multisensory integration, combining visual processing in the posterior regions with semantic knowledge in the anterior areas. This integration enables quick understanding, vital for accurate action analysis and performance assessment. To achieve a comprehensive understanding of scene-invariant AQA, it's essential to consider both visual and semantic action features. Semantic components represent high-level concepts, while visual elements detail motion dynamics. Using semantic descriptions improves score regression by providing specific contextual features. Furthermore, analyzing the interplay between these modalities offers a deeper insight into the action execution. By merging these aspects, we gain a holistic understanding of the execution of fine-grained and scene-invariant action sequences. Thus, we propose a multimodal interaction module that leverages combined semantic and visual action features to encourage interactions between different modalities and facilitate the refined AQA score prediction. Specifically, we employ a context-aware transformer that uses semantic action features as queries and visual features as

Received 3 August 2024; revised 8 October 2024; accepted 20 October 2024. Date of publication 28 October 2024; date of current version 7 March 2025. This work was supported in part by the Ministry of Science and Technology of China Foreign Young Talents Program under Grant QN2023036003L, in part by Sichuan Science and Technology Program under Grant 2024JDHJ0059, in part by the Natural Science Foundation of China under Grant W2433164, and in part by the Science and Technology Innovation Committee of Shenzhen Municipality Foundation under Grant JCYJ20210324132203007 and Grant JCYJ20230807120000001. This article was recommended by Associate Editor Q. Xu. *(Corresponding author: Yanli Ji.)*

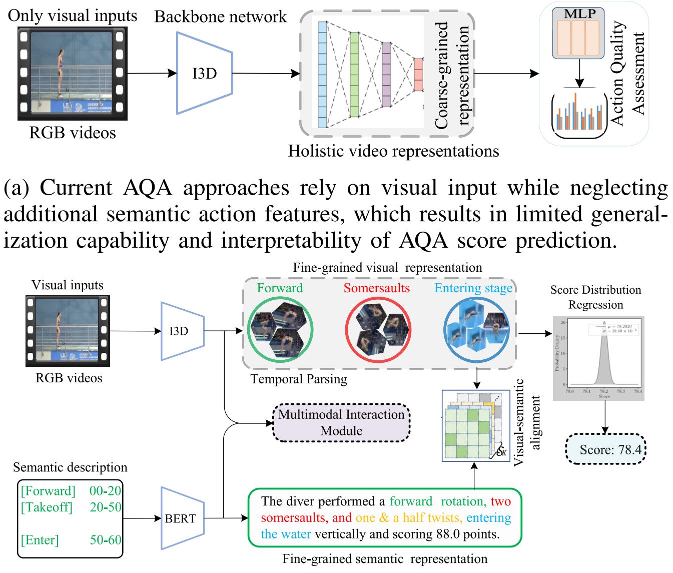
Kumie Gedamu is with Sichuan Artificial Intelligence Research Institute, Yibin 644000, China, and also with the School of Computer Science and Engineering, University of Electronic Science and Technology of China (UESTC), Chengdu 610051, China.

Yanli Ji is with the School of Intelligent Systems Engineering, Sun Yat-sen University, Shenzhen Campus, Shenzhen 518107, China (e-mail: jiyanli82@gmail.com).

Yang Yang and Heng Tao Shen are with the School of Computer Science and Engineering, UESTC, Chengdu 610051, China.

Jie Shao is with the School of Computer Science and Engineering, UESTC, Chengdu 610051, China, and also with Shenzhen Institute for Advanced Study, UESTC, Shenzhen 518028, China.

Digital Object Identifier 10.1109/TCSVT.2024.3487242


Only visual inputs
Backbone network
RGB videos
I3D
Coarse-grained representation
MLP
Action Quality Assessment
Holistic video representations
Fine-grained visual representation
Visual inputs
Forward
Somersaults
Entering stage
Score Distribution Regression
RGB videos
Temporal Parsing
Multimodal Interaction Module
Visual-semantic alignment
Score: 78.4
Semantic description
[Forward] 00-20
[Takeoff] 20-50
[Enter] 50-60
BERT
The diver performed a forward rotation, two somersaults, and one & a half twists, entering the water vertically and scoring 88.0 points.
Fine-grained semantic representation

(a) Current AQA approaches rely on visual input while neglecting additional semantic action features, which results in limited generalization capability and interpretability of AQA score prediction.

(b) Solution in our proposed approach.

Fig. 1. Motivation of our proposed approach. We design a multimodal visual-semantic alignment temporal parsing approach to understand high-level visual-semantic representations and internal structures for accurate AQA score prediction.

memories (keys and values), as shown in Fig. 1(b). Through this structured modeling, the relationships between subactions and semantic features are elucidated, resulting in a better understanding of fine-grained action features.

In long-term action evaluation, it is essential to assess both the overall quality of the response sequence and key fine-grained technical subactions, particularly important in challenging backgrounds like diving [21] and figure skating [24]. As shown in Fig. 1(b), sports activities such as diving involve multiple fine-grained subactions such as "Forward", "Somersaults", "Twist", and "Entry". Understanding these fine-grained technical sub-actions is crucial for enhancing the effectiveness of AQA systems in real-world scenarios. However, it is challenging due to the lack of predefined subaction labels and finely granular subaction sequences with smooth transitions, making boundaries hard to distinguish [22]. These observations motivated us to develop a self-supervised sub-action parsing approach without explicit supervision. Specifically, we design the visual ParseFormer decoder to attend to videos with learnable queries and produce an ordered sequence of technical subactions. We supervised the produced subaction sequences with semantic action features and trained our model using visual-semantic alignment loss with DropDTW [25]. The module evaluates the consistency between an athlete's actions and their semantic context, ensuring accurate analysis of action execution. Moreover, the strong alignment between semantic and visual representations improves our module to capture the nuances of athletes' technical performances, which improves the overall performance of AQA score prediction.

In this paper, we propose a visual-semantic alignment-based temporal parsing network to learn a consistent high-level semantic representation of action features for accurate and interpretable AQA score prediction. As illustrated in Fig. 2, our approach is composed of two main components: self-supervised temporal parsing, and the multimodal interaction module. The former processes video inputs with learnable queries and produces subaction sequences supervised by a visual-semantic alignment loss. The latter evaluates action sequences by capturing visual and semantic action features to encourage interactions between different modalities. In this way, we obtain an accurate representation of action features, which improves the overall performance of AQA. In summary, the major contributions of our proposed approach are:

- We design a visual-semantic alignment Temporal Parsing Network to understand the high-level representation along with internal temporal structures of action features for accurate and interpretability of AQA score prediction.
- We propose a self-supervised visual ParseFormer decoder to generate a sequence of sub-actions supervised by visual-semantic alignment loss that captures fine-grained action features along with their temporal dependencies.
- We propose a multimodal interaction module to capture visual-semantic action features and the interaction between different modalities, enabling a better understanding of scene-invariant action execution.
- We conduct extensive experiments to analyze the effectiveness of our approach over the state-of-the-art methods on the MTL-AQA [1], Rhythmic Gymnastics (RG) [26], FineFS [24] and Fis-V [27] datasets.

The rest of the paper is organized as follows. We review related work in Section II. We then present the proposed approach with its innovative features in Section III. In Section IV, we conduct comprehensive experiments of VATP-Net's performance on benchmark datasets. Finally, Section V gives a conclusion of the manuscript.

II. Related Work

A. Action Quality Assessment

Initially, AQA was approached as a classification task, where actions were categorized into different levels of performance. This was seen in works such as [28] and [29]. Existing approaches in AQA can be categorized into regression-based and pairwise ranking-based approaches.

In the regression formulation, various methods have been proposed to address AQA challenges. Parmar and Tran Morris [1] used spatiotemporal action features and clip-level scoring for estimating AQA scores in sports activities. Pose+DCT [30] employed joint localization and support vector regression for score regression. Jain et al. [31] Proposed reference guided regression, an action scoring system using deep metric Learning for scoring against a reference. Zeng et al. [26] combined static and dynamic action features, considering the contributions of different stages to AQA scores. Zhang et al. [32] introduced a semi-supervised AQA approach using self-supervised learning on unlabeled videos to recover features of masked segments. However, it is limited by contextual biases and holistic representations, restricting its ability to capture subtle variations. Xu et al. [27] introduced a self-attentive multiscale skip convolutional LSTM approach. Tang et al. [33] formulated score regression as a distribution learning problem using KL divergence loss. Zhou et al. [10] introduced a hierarchical GCN for analyzing action procedures and motion units to refine semantic action features, reduce information confusion, and aggregate dependencies. Xu et al. citeFineDiving introduced a procedure-aware representation with a pairwise temporal segmentation attention module. Gedamu et al. [22] presented a spatiotemporal fine-grained representation using a multiscale transformer. Zhou and Huang [34] proposed a distribution auto-encoder to handle aleatoric uncertainty by encoding videos into distributions with the VAE reparameterization trick. Zhang et al. [35] introduced a group-aware attention approach that incorporates contextual information and temporal relations using graph CNNs. Xu et al. [36] introduced FineParser, a spatial-temporal action parser that focuses on target regions to learn human-centric action sequences. The mentioned approaches rely on visual input features, limiting their generalization due to the neglect of semantic action features. In contrast, our approach identifies high-level action features and captures temporal structure by parsing a given action sequence into separate sub-actions without explicit supervision. Furthermore, we explore relationships between visual and semantic action features across different modalities.

On the other hand, in the pairwise ranking formulation, the AQA problem is approached as a pairwise ranking problem when performance scores are unavailable. Doughty et al. [12], [37] employed rank-aware loss functions and discriminative feature learning. Siamese learning [27] focused on overall ranks, while Likert scoring [38] enabled grade quantification. Contrastive Regression (CoRe) [23] emphasized differences between videos for relative score learning. Following this, Fang et al. [39] introduced a parsing transformer to disintegrate the holistic feature into a more fine-grained procedure-wise representation. Li et al. [11] introduced pairwise contrastive learning to learn relative scores between pairs of input videos. Bai et al. [20] presented a temporal parsing transformer that decomposes global features into a fine-grained temporal hierarchical representation. Similarly, An et al. [40] proposed a multi-stage contrastive regression framework for AQA that efficiently extracts spatial-temporal action features. Recently, Zhou et al. [41] proposed coarse-to-fine instruction alignment with broader pre-trained tasks by reformulating it as a coarse-to-fine classification task. In contrast to these approaches, our method exploits solid and consistent representations of action sequences by parsing given action sequences and utilizing multimodal interactions. Furthermore, action parsing and image quality assessment have been studied in related works [8], [42], [43], [44], [45], [46], [47], [48], [49]. Applying these approaches to AQA is challenging due to the need to quantify actions in videos with minor discrepancies in similar backgrounds. Our method, however, identifies sub-action patterns and captures middle-level representations through multimodal interaction without explicit supervision, enabling accurate and interpretable AQA score prediction.

B. Learning From Multimodal Action Features

Integrating visual and semantic action features has demonstrated significant potential in capturing the intricate physical and technical dimensions of challenging human actions [50], [51], [52]. The multimodal approach enhances understanding of fine-grained action features by bridging visual inputs and their contextual meanings. Recent research on visual-semantic action features for AQA and instructional videos [53], [54] highlights their effectiveness in evaluating both the actions performed and their execution quality. The multimodal approaches are now emerging in AQA, leveraging audio as a vital supplementary resource [55]. The inclusion of audio information enriches the contextual backdrop against which actions are assessed. The work of [55] has investigated the incorporation of RGB, optical flow, and audio information, to guide the learning of visual feature representations to assess action quality performance. Furthermore, adaptive knowledge transfer between the semantic and visual action features [54], enhances the model's ability to generalize across different contexts and conditions. The work in [56] learns multimodal representations by modeling audiovisual features, and the proposed MLP-Mixer effectively learns long-term representations through the designed memory recurrent unit. Similarly, [54] utilized semantic attributes for query learning to enhance the assessment of gymnastic routines from video. Furthermore, Zhang et al. [7] proposed an adaptive stage-aware skill transfer framework that transfers assessment skills from source actions to different stages of a target action adaptively. Thus, combining visual and semantic features can better capture the nuanced aspects of complex human actions. However, a key challenge remains in achieving effective visual-semantic alignment. Addressing this misalignment is crucial for fully harnessing the potential of multimodal approaches. By grounding semantic representations in visual action knowledge, our proposed approach aims to enhance the overall understanding of action assessment through the analysis of fine-grained technical subactions and capture the interaction between different modalities.

III. Proposed Approach

In this section, we introduce the multimodal visual-semantic alignment temporal parsing network, which aims to exploit a robust and consistent semantic representation of action features. As illustrated in Fig. 2, the self-supervised temporal parsing module learns fine-grained subaction sequences with their temporal dependencies without explicit supervision. The multimodal interaction module captures high-level visual-semantic relationships between different modalities. This way, we gain a deeper understanding of action execution by integrating visual and semantic action information.

A. Problem Formulation

The AQA is formulated as a regression problem with video input $X_v = \{x_v, y_v\}_{v=1}^{\mathbb{N}}$, where $X_v \in \mathbb{R}^{T \times H \times W \times C}$ (T, H, W, and C represent clip length, height, width, and channels) and the score label y_v to predict the action score $\hat{y}_v$:

$$\hat{y}_v = R_\theta(E_\phi(X_v)) \tag{1}$$

where the regressor network R and the feature learning network E parameterized by θ and ϕ, respectively. The training

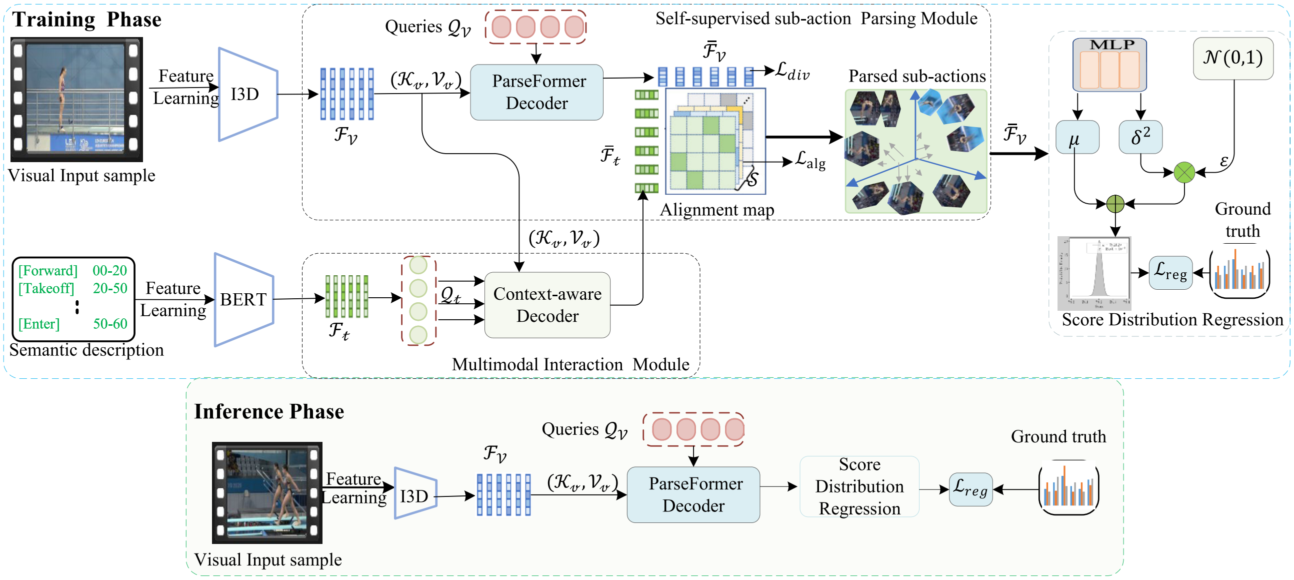


Fig. 2. Overview of our proposed approach. The self-supervised parsing module generates segmented subactions from videos, guided by visual-semantic alignment loss. The multimodal interaction module captures and enhances the alignment of visual-semantic action features.

objective is to minimize the MSE between the predicted and the ground-truth assessment scores. However, most existing approaches struggle with generalizing high-level semantic representations with only visual input.

To overcome these limitations, we propose a multimodal visual-semantic alignment temporal parsing network to understand the high-level representation of subactions along with internal temporal structures, leading to better generalization. Thus, we utilize the I3D backbone [57] and BERT [58] to extract spatio-temporal action features and semantic embeddings of the input features of X_v and X_t, respectively, as shown Eqn. 2.

$$\mathcal{F}_v = E_\phi(X_v), \quad \mathcal{F}_t = E_\phi(X_t) \tag{2}$$

The extracted spatiotemporal action features $\mathcal{F}_v \in \mathbb{R}^{T_v \times D}$ and token embeddings $\mathcal{F}_t \in \mathbb{R}^{T_t \times D}$ are fed into a multimodal visual-semantic alignment network. This network comprehends high-level visual-semantic action sequences and internal temporal structures, enhancing the understanding of fine-grained, and scene-invariant action features.

B. Self-Supervised Temporal Parsing Module

The fine-grained and smooth transitions between subactions in scene-invariant action sequences, such as diving, make it challenging to distinguish their boundaries. Additionally, the lack of predefined subaction labels complicates this further. This motivated us to develop a self-supervised sub-action parsing approach without explicit supervision. Analyzing the internal temporal structures of technical and semantic action sequences improves the overall performance of fine-grained and scene-invariant action execution assessment.

1) Temporal Enhancement Module: The independent extraction of features from each video segment leads to a lack of global context action features [38]. Similarly to previous approaches [38], we adopt a self-attention encoder to aggregate the segment action features. This allows the model to enrich the segment-wise representations with relevant global context action features. The weights for this aggregation are learned based on the correlations between the clip segments. Such context action features are important for long-range video understanding and overall performance improvement.

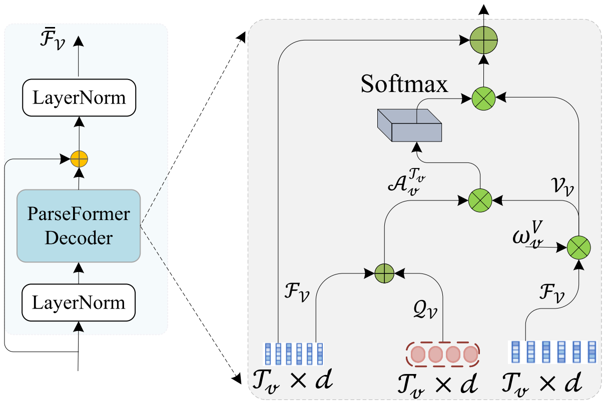


Fig. 3. The detailed structure of visual ParseFormer decoder. The visual feature $\mathcal{F}_v$ serves as memories (keys and values), and the learnable query Q_v serves as queries.

2) Visual ParseFormer Decoder: To achieve our primary objective of identifying concise and representative subaction sequences from a video, we employ a multilayer Transformer decoder inspired by DeiT [20]. The proposed ParseFormer utilizes S learnable queries, denoted as $Q_v \in \mathbb{R}^{T_v \times d}$, alongside video feature $\mathcal{F}_v \in \mathbb{R}^{T_v \times d}$, as illustrated in Fig. 3. We project the input action features $\mathcal{F}_v$ into a matrix V_v using learnable weight parameter ω_v^V. Here, the visual feature $\mathcal{F}_v$ acts as memory (keys and values), while the learnable query Q_v functions as the query. Following the projection, we derive the intermediate tensor V_v through a linear operation, defined as $V_v = \mathcal{F}_v \omega_v^V$. Subsequently, we calculate the attention value for the learnable query Q_v with the temperature τ, which controls the amplification of the inner product across the given clip T_v as follows:

$$A_v^{T_v} = \frac{\exp[(\mathcal{F}_v + Q_v)^\top . V_v/\tau)]}{\sum_{v=1}^{T_v} \exp[(\mathcal{F}_v + Q_v)^\top . V_v/\tau]} \tag{3}$$

where $\tau \in \mathbb{R}$ is used to the learnable temperature to enhance the inner product to make the attentions more discriminative, similar to [20]. The learned attention enhances the action features with the operation, as shown in Eqn. 4.

$$\bar{\mathcal{F}}_v = \sum_{v=1}^{T_v} A_v^{T_v} V_v + \mathcal{F}_v \tag{4}$$

The proposed module consists of multihead cross-attention learning with MLP blocks. In addition, the LayerNorm and residual connections are applied. Thus, the output of ParseFormer is a sequence of contextual vectors, denoted $\bar{\mathcal{F}}_v \in \mathbb{R}^{T_v \times D}$. Each vector in the sequence represents one subaction sequence. In this way, we identify high-level action features and capture fine-grained temporal structures by parsing action sequences into separate sub-actions, which improves the overall performance of AQA score prediction.

C. Multimodal Interaction Action Feature Learning

We investigate how to leverage semantic contexts to refine the visual features extracted by the model. By integrating semantic information, we enhance the model's ability to understand and interpret actions more accurately. Specifically, we propose a Context-aware Decoder (CD) cross-attention mechanism that effectively models the interactions between visual and semantic action features, as illustrated in Fig. 4. The semantic context queries $Q_t \in \mathbb{R}^{T_t \times d}$ are derived from the projection matrix of the semantic action feature $\mathcal{F}_t$. These queries serve as a crucial input to the visual cross-attention mechanism, where they are combined with the visual feature $\mathcal{F}_v \in \mathbb{R}^{T_v \times d}$, as shown in Fig. 4. Thus, the visual feature $\mathcal{F}_v$ serves as memories (keys and values), and the semantic action feature Q_t serves as queries. This integration allows the model to refine the visual feature representation by incorporating relevant semantic contexts. Thus, the proposed context-aware decoder module is formulated with a given semantic action sequence T_t as follows:

$$\mathbf{A}_t^{T_t} = \frac{\exp[(\mathcal{F}_v + Q_t)^\top . V_v/\tau]}{\sum_{v=1}^{T} \exp[(\mathcal{F}_v + Q_t)^\top . V_v/\tau]} \tag{5}$$

where $\mathbf{A}_t^{T_t}$ is the attention matrix that captures the relevance of each visual feature $\mathcal{F}_v$ to the semantic context Q_t. The parameter τ controls the softmax temperature. In this way, the proposed module harnesses the power of interactions between visual and semantic action features to refine visual representation. The refined visual feature effectively integrates relevant semantic contexts through a weighted summation of the visual features, as shown in Eqn. 6.

$$\bar{\mathcal{F}}_t = \sum_{v=1}^{T} \mathbf{A}_t^{T_t} V_v + \mathcal{F}_t \tag{6}$$

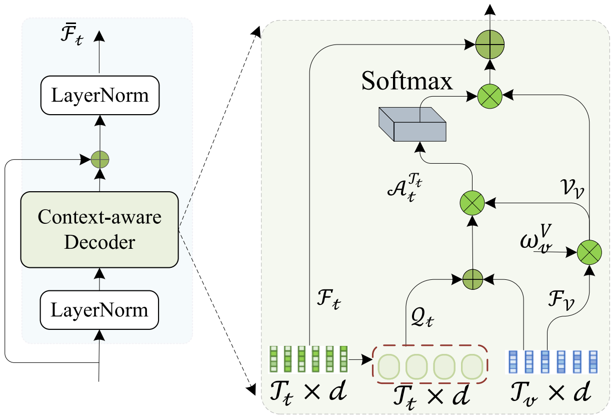


Fig. 4. The detailed procedure of context-aware decoder Transformer. The visual feature $\mathcal{F}_v$ serves as memories (keys and values), and the semantic action feature Q_t serves as queries.

where $\bar{\mathcal{F}}_t$ is the refined visual representation. It's computed by weighting $\mathcal{F}_v$ using $\mathbf{A}_t^{T_t}$ and adding $\mathcal{F}_t$. Integrating visual and semantic information through the multimodal interaction module enhances our model's understanding of action execution. This module refines visual representation by focusing on context-relevant features, forming the basis for the next stage. We then supervise the produced subaction sequences with semantic action features and train our model using visual-semantic alignment loss.

D. Visual-Semantic Action Feature Alignment Learning

Correct alignment is essential for effectively bridging the gap between video content and semantic representations. It enables the model to forge stronger associations between visual subaction features and their corresponding semantic descriptions. These semantic descriptions serve as vital complementary information, significantly enhancing the model's capability to learn more discriminative subaction representation. Since we have no supervised information for aligning visual-semantic action features, we adopt the Drop-DTW [25] solution to assist us in filtering relatively matched visual-semantic representations and use the matching relationship to supervise the visual-semantic alignment learning. The solution automatically detects outliers and removes them from the sequences of two modals [25], [53]. Thus, we design two loss functions to achieve visual-semantic alignment during training, promoting the model to attain accurate alignment. They ensure the learning of discriminative subactions and the smoothing of the temporal continuity of action sequences.

1) Unsupervised Visual-Semantic Alignment: The proposed visual ParseFormer module outputs a subaction sequence that contains visual information. We use Drop-DTW (DD) [25] to identify the correspondences between these visual subaction sequences and the semantic embeddings. These matched visual subactions and semantic embeddings form the positive visual-semantic corresponding pairs, while non-matched pairs are used as negative examples for model training. The operation is illustrated in Eqn. 7.

$$\mathcal{M}_{vt} = DD(\bar{\mathcal{F}}_v, \bar{\mathcal{F}}_t) = \begin{cases} 1, & \Rightarrow \text{used as positive pairs.} \\ 0, & \Rightarrow \text{used as negative pairs.} \end{cases} \tag{7}$$

To simplify the explanation, we use $DD()$ to represent the Drop-DTW (DD) solution, which generates a correspondence matrix, $\mathcal{M}_{vt}$, indicating matched results of visual-semantic feature pairs, $(\bar{\mathcal{F}}_v, \bar{\mathcal{F}}_t)$. The resulting correspondence matrix, $\mathcal{M}_{vt}$, is used to construct positive and negative pairs in a contrastive training setting. Given the correspondence matrix $\mathcal{M}_{vt}$, we perform the supervised subaction parsing by defining loss ℓ_{NCE} [59] to promote the similarity between positive pairs and push apart the non-matching pairs, as defined in Eqn. 8.

$$\ell_{\text{NCE}}(\bar{\mathcal{F}}_v, \bar{\mathcal{F}}_t) = -\log \frac{f(\bar{\mathcal{F}}_v, \bar{\mathcal{F}}_{t^*})}{f(\bar{\mathcal{F}}_v, \bar{\mathcal{F}}_{t^*}) + \sum_{t \neq t^*} f(\bar{\mathcal{F}}_v, \bar{\mathcal{F}}_t)} \tag{8}$$

where $f(\bar{\mathcal{F}}_v, \bar{\mathcal{F}}_t) = \exp(\cos(\bar{\mathcal{F}}_v, \bar{\mathcal{F}}_t)/\gamma)$, γ is a scaling temperature. Here, t^* refers to the index of semantic embeddings which compose positive pairs with subactions $\bar{\mathcal{F}}_v$. Positive pairs are determined by the correspondence matrix, $\mathcal{M}_{vt}$. The complete visual-semantic alignment loss, $\mathcal{L}_{\text{alg}}$, is defined as a combination of two Info-NCE losses, one contrasting the subactions and the other contrasting the semantic embeddings, as illustrated in Eqn. 9.

$$\mathcal{L}_{\text{alg}} = \frac{1}{S} \sum_{v=1}^{S} \ell_{\text{NCE}}(\bar{\mathcal{F}}_v, \bar{\mathcal{F}}_t) + \frac{1}{N} \sum_{t=1}^{L} \ell_{\text{NCE}}(\bar{\mathcal{F}}_t, \bar{\mathcal{F}}_v) \tag{9}$$

where S refers to the number of subaction sequences. The first term of the loss function aims to align the sub-action sequence $\bar{\mathcal{F}}_v$ with the corresponding semantic embeddings $\bar{\mathcal{F}}_t$, while the second term aims to align the semantic embeddings $\bar{\mathcal{F}}_t$ with the sub-actions $\bar{\mathcal{F}}_v$, effectively represents the semantic and visual aspects of the action features.

2) Global Visual-Semantic Contrastive Learning: The visual-semantic alignment loss learns from positive and negative samples from the same videos. However, contrastive learning can greatly benefit from a large and diverse set of negative samples [53], [60]. Thus, we formulate a global contrastive loss that creates negative pairs from the semantics and subactions extracted from different videos. The intuition behind the proposed loss is two-fold: (1). For a given video, the extracted subaction sequences should match the semantic descriptions within that video, regardless of the temporal continuity. (2). The subactions and semantic descriptions extracted from different videos should be different from each other. Specifically, the proposed loss is formulated as shown in Eqn. 10.

$$\mathcal{L}_{\text{gc}} = -\log \frac{1}{M} \sum_{i=1}^{M} \frac{\sum_{t \in \mathcal{P}_v} f(\bar{\mathcal{F}}_v, \bar{\mathcal{F}}_t)}{\sum_{t \in \mathcal{P}_v} f(\bar{\mathcal{F}}_v, \bar{\mathcal{F}}_t) + \sum_{t \in \mathcal{N}_v} f(\bar{\mathcal{F}}_v, \bar{\mathcal{F}}_t)}, \tag{10}$$

where $M = S$, which is the total number of sub-actions across the batch size, $\mathcal{P}_v$ is the set of indexes v that form a positive pair with $\bar{\mathcal{F}}_v$, *i.e.,* the $\mathcal{P}_v$ from the same video, and $\mathcal{N}_v$ indexes the negative pairs, from different videos. By optimizing the proposed contrastive loss and enforcing these criteria, the model learns video representations that reflect each video's structure and the differences in visual and semantic content. This improves nuanced video understanding and overall performance, ensuring that subaction sequences and semantic embeddings from the same video match.

3) Diversity Regularization Loss: To optimize the training process and incorporate prior knowledge into the approach, we added diversity regularization loss. This loss enforces diversity among subactions and improves performance by ignoring duplicated sub-action sequences. The diversity regularization loss ($\mathcal{L}_{\text{div}}$) encourages low cosine similarity among the predicted sub-action sequences [53]. These sequences are extracted from the video $\mathcal{F}_v \in \mathbb{R}^{T_v \times d}$ using the visual ParseFormer decoder. Specifically, the diversity regularization loss is defined as:

$$\mathcal{L}_{\text{div}} = \frac{1}{S(S-1)} \sum_{i=1}^{S} \sum_{j \neq i} \cos(\bar{\mathcal{F}}_v^i, \bar{\mathcal{F}}_v^j), \tag{11}$$

The loss strengthens the generalization ability of our proposed model and makes accurate predictions in complex scenarios. Thus, the approach promotes action sequence diversity and enhances the overall performance of AQA by removing duplicate semantic action features.

E. Score Distribution Regression

Treating the action assessment task as a regression problem ignores the inherent ambiguity in the score labels caused by multiple judges or their subjective appraisals. The uncertainty-aware score distribution learning approaches [33] are typically employed to address this issue, which automatically generates distinguishable variances for different actions. Similar to [10] and [33], we adopt the score distribution regression to predict the score distribution instead of directly regressing the final score. For the video-level feature $\bar{\mathcal{F}}_v$, a probabilistic encoder is first used to encode $\bar{\mathcal{F}}_v$ into a random score variable y_v. The encoded score random variable y_v is subject to the Gaussian distribution, as follows:

$$\mathbb{F} = \text{MLP}(\bar{\mathcal{F}}_v) \tag{12}$$

$$p(y_v, \mathbb{F}) = \frac{1}{\sqrt{2\pi\sigma^2(\mathbb{F})}} \exp\left(-\frac{(y_v - \mu(\mathbb{F})^2}{2\sigma^2(\mathbb{F})}\right) \tag{13}$$

where the mean parameter μ and variance parameter σ^2 concerning the feature representation $\mathbb{F}$ are used to quantify the quality and uncertainty of the action score, respectively. The re-parameterization trick is then applied to sample from the distribution to output the predicted score $\hat{y}_v$:

$$\hat{y}_v = \mu(\mathbb{F}) + \epsilon \cdot \sigma(\mathbb{F}) \tag{14}$$

where $\epsilon \sim \mathcal{N}(0, 1)$ is a standard normal random variable. This differentiable sampling process ensures feasible training.

F. Loss Setting

To supervise the regression of the score distribution, we employ the Mean Squared Error (MSE) to define the regression loss. The loss is defined as follows:

$$\mathcal{L}_{reg} = \frac{1}{\mathbb{N}} \sum_{i=1}^{\mathbb{N}} \left(\hat{y}_v^i - y_v^i \right)^2 \tag{15}$$

Thus, the overall training loss of the proposed approach is given by:

$$\mathcal{L}_{all} = \mathcal{L}_{reg} + \beta \mathcal{L}_{div} + \mathcal{L}_{alg} + \mathcal{L}_{gc} \tag{16}$$

where $\mathcal{L}_{reg}$ is the regression loss, $\mathcal{L}_{div}$ is the diversity regularization loss, $\mathcal{L}_{alg}$ is a self-supervised visual-semantic alignment loss, and $\mathcal{L}_{gc}$ is the global contrastive loss. The β is a hyperparameter for consistency regularization.

IV. Experiment

In this section, we outline the experimental setting and analyze the results. We start by describing the datasets, implementation details, and evaluation metrics. Next, we present an ablation study and compare our results with state-of-the-art methods. We also include qualitative visualizations to highlight our approach's effectiveness and discuss failure cases. Finally, we conclude with future research directions.

A. Datasets and Experiment Settings

1) Datasets: We evaluate our approach on four large-scale public available AQA datasets, such as the MTL-AQA [1], the RG [26], the FineFS [24] and the Fis-V datasets [27].

a) MTL-AQA dataset : The dataset focuses on diving, covering a wide range of actions. There are 1412 samples collected from 16 different world events. Different annotations are available in this dataset such as AQA, action recognition, and comment generation. Additionally, raw score annotations and Degree of Difficulty (DD) are available from multiple judges. Following [2], we divide the dataset into 1059 training samples and 353 test samples.

b) Rhythmic gymnastics dataset (RG) : The dataset contains a total of 1,000 videos with four types of gymnastics: Ball, Clubs, Hoop, and Ribbon. Each routine type consists of 250 videos. We follow the evaluation protocol [26] and partition the dataset into 200 training videos and 50 test videos for each gymnastics routine type.

c) Figure skating video dataset (Fis-V): The Fis-V dataset contains 500 videos of ladies' singles short program figure skating performances, each approximately 2.9 minutes long at 25 frames per second. The dataset is split into 400 training and 100 testing videos, with all videos annotated with Total Element Score (TES) and Total Program Component Score (PCS) according to competition rules. Following [27], we train two independent models to predict these scores.

d) Fine-grained figure skating dataset (FineFS):: The dataset contains 1167 samples where 729 samples are from short program (SP) and 438 samples from free skating (FS). It includes RGB videos, estimated skeleton, fine-grained score labels, and technical subaction categories. Following [24], the dataset is split into training and test sets with a 4:1 ratio, and experiments on SP and FS are performed separately.

2) Implementation: We pretrain the I3D backbone [57] in the Kinetics dataset to extract visual features, with an initial learning rate of 10^{-4} for the MTL-AQA dataset. Subsequently, we extract video features using the Video Swin Transformer (VST) [61] pre-trained on Kinetics 600 for the RG, Fis-V, and FineFS datasets. We utilize the Adam optimizer with the weight decay setting to zero. The initial learning rate of our approach is set to $1e$-3. Following [23], we extract 103 clips in MTL-AQA and multiply predicted scores by the Difficulty Degree (DD). To achieve better convergence, we assign different numbers of epochs for each category Rhythmic Gymnastics dataset (RG), following [26]. We set 300, 400, 500, and 300 epochs for Ball, Clubs, Hoop, and Ribbon, respectively. In addition, the text feature (semantic action features) is extracted by the BERT [58] model in the MTL-AQA dataset. Furthermore, datasets such as RG [26], Fis-V [27], and FineFS [24], which only contain video input without additional text annotations, are fine-tuned with our trained visual ParseFormer decoder module similar to [54], demonstrating the effectiveness of our approach.

3) Evaluation Metrics: To keep alignment with existing approaches, we adopt Spearman's rank correlation (Spr. Corr. ρ, ranges from -1 to 1, the larger value is the better) to measure the difference between predicted and ground-truth scores. This metric can be formulated as:

$$\rho = \frac{\sum_v (y_v - y)(\hat{y}_v - \hat{y})}{\sqrt{\sum_v (y_v - y)^2 \sum_v (\hat{y}_v - \hat{y})^2}} \tag{17}$$

where y_v and $\hat{y}_v$ indicate the rankings of two score sequences, respectively.

B. Ablation Study

In this section, we conduct experiments to understand how well our proposed modules contribute and the impact of each module on the design approach. Specifically, we look at the following different versions of the proposed approach:

- **Baseline (I3D+MLP)**: The baseline I3D+MLP directly feeds the video features into the score distribution module for score prediction. The MSE loss is adopted for optimization.
- Temporal Enhancement only (I3D+TE): This version uses only the Temporal Enhancement, removing self-supervised temporal parsing and multimodal interaction modules to demonstrate its contribution to the I3D backbone, directly feeding the encoder output into the score distribution regression. Furthermore, the Video Swin Transformer with Temporal Enhancement (VST+TE) utilizes the VST as the backbone, focusing solely on the visual encoder and highlighting the contribution of temporal enhancement on top of this vision backbone.
- Visual ParseFormer decoder (I3D+TE+PD): This version includes the visual ParseFormer decoder module with learnable queries but does not use the multimodal interaction branch. The visual ParseFormer output is fed into the score distribution module for score prediction, with the MSE loss adopted for optimization. In addition,

we validate the contribution of the ParseFormer decoder in learning high-level and fine-grained semantic representations on top of the VST backbone (VST+TE+PD)

- **Our approach (VATP-Net)**: This is the complete approach, comparing with other versions, helps us understand the overall effectiveness of our approach and how these modules cooperate within one model.

1) Evaluation on Components of Our Proposed Approach: We explore the effectiveness of the proposed modules in the MTL-AQA and RG datasets through quantitative and qualitative analysis. To demonstrate the effectiveness of each module, we attempt to remove them one by one from our proposed approach and evaluate the performance. For component evaluation, we utilize the I3D [57] and VST [61] backbones, along with self-supervised temporal parsing and multimodal interaction modules, to perform AQA score estimation. The comparison results are shown in Table I and Table II.

a) Effect of temporal enhancement (I3D+TE): This module boosts the performance of the action score prediction model by capturing the internal temporal structure of action sequences. Building on the top of the I3D backbone, adding the temporal enhancement module (I3D+TE) significantly improves the model performance, increasing the Spr. Corr. from 0.9450 to 0.9516 as shown in Table I. Similarly, the addition of temporal enhancements on top of VST significantly improves the model's performance in the RG dataset, as shown in Table II. These results indicate that the proposed module is able to capture important temporal information and use it to enrich the segment-level representation with relevant global context action features.

b) Contribution from visual ParseFormer decoder: We also observe that on top of the I3D backbone, adding our proposed visual ParseFormer decoder (I3D+TE+PD) improves the performance of AQA, which is 0.9531% on Spr. Corr., as shown in Table I. Furthermore, the significance of performance improvement is demonstrated by adding the proposed module on top of VST [61] (VST+TE+PD), as shown in Table II. This is a testament to the effectiveness of the proposed visual ParseFormer module, indicating its potential utility in improving mid-level representations and capturing fine-grained action features to maintain consistent video representation. Moreover, the proposed module is able to understand the high-level semantics and temporal structures of subaction sequences.

c) Visual-semantic alignment action feature learning : Combining visual and semantic action features has a synergistic effect, enhancing fine-grained video representation and boosting AQA performance. To show the contribution of action feature learning from visual-semantic alignment, we experimented and presented the results in Table I and Table II. Specifically, the proposed approach achieved a Spr. Corr. of 0.9588 on the MTL-AQA dataset. Similarly, the proposed approach showed significant contributions by integrating semantic action features for accurate and interpretable score prediction on the RG dataset. These results imply that the semantic action features help the model learn a high-level representation of subaction sequences with internal temporal structure and capture the interactions between different modalities of action features.

TABLE I

EVALUATION ON COMPONENTS OF VATP-NET IN THE MTL-AQA DATASET. THE REP., AND CD REFER TO THE "RE-PARAMETERIZATION TRICK", AND "CONTEXT-AWARE DECODER"

Approaches	Rep.	TE	PD	CD	Spr. Corr. (ρ) ↑
I3D+MLP	-	-	-	-	0.9450
I3D+TE	✓	✓	×	×	0.9516
I3D+PD	✓	×	✓	×	0.9454
I3D+TE+PD	✓	✓	✓	×	0.9531
VATP-Net (Ours)	✓	✓	✓	✓	**0.9588**

TABLE II

EVALUATION ON COMPONENTS OF VATP-NET IN THE RHYTHMIC GYMNASTICS (RG) DATASET, FINE-TUNED WITH OUR TRAINED PARSEFORMER DECODER MODULE

Method	Ball	Clubs	Hoop	Ribbon	Average
VST+TE	0.765	0.770	0.721	0.762	0.754
VST+TE+PD	0.785	0.801	0.778	0.769	0.783
VATP-Net (Ours)	0.800	0.810	0.780	0.769	0.800

TABLE III

EVALUATION ON THE OPTIMAL NUMBER S OF SUB-ACTION SEQUENCES IN THE MTL-AQA DATASET. THE COMPARISON SHOWS THAT THE OPTIMAL VALUE FOR S IS 4

Queries (S)	2	3	4	5	6
Spr. Corr. (ρ) ↑	0.9501	0.9547	**0.9588**	0.9536	0.9494

2) Evaluation on the Optimal Number of Queries: To explore the effect of the number of learnable queries as hyperparameters, we conduct an experiment on the MTL-AQA dataset as shown in Table III. We varied the values of learnable queries and analyzed their impact on the performance of our proposed approach. The experiments revealed the importance of selecting an appropriate value for learnable queries in order to achieve optimal performance in score prediction. A small number of learnable queries may not adequately capture the complexity and nuances of the actions, resulting in suboptimal outcomes. Conversely, an excessive number of learnable queries increases the risk of overfitting, where the model becomes too specialized for the training samples. The ideal learnable queries capture the characteristics of subaction sequences. As shown in Table III, the four sub-action sequences provide the optimal value of Queries (S), enhancing the fine-grained action feature representation and improving the overall performance of AQA score prediction.

3) Evaluation on the Number of Visual ParseFormer Decoder Blocks: To understand the contribution of the ParseFormer decoder block, we conduct experiments and show the results in Table IV. Increasing model complexity without performance gains is not recommended, as it can lead to inefficient and resource-intensive models. The observations in Table IV show that simply increasing the number of transformer blocks does not necessarily improve performance,

TABLE IV

EVALUATION ON THE NUMBER OF VISUAL PARSEFORMER DECODER BLOCKS IN THE MTL-AQA DATASET. THESE RESULTS SHOW THAT SIMPLY INCREASING THE NUMBER OF TRANSFORMER BLOCKS DOES NOT NECESSARILY IMPROVE THE PERFORMANCE

# Blocks	2	4	6	8	10
Spr. Corr. (ρ) ↑	**0.9588**	0.9521	0.953	0.9562	0.9521
Parm. (M)	1.81	2.27	2.74	3.20	3.66

TABLE V

EVALUATION ON LOSS FUNCTIONS IN THE MTL-AQA DATASET. THE RESULTS HIGHLIGHT THE IMPORTANCE OF PROPOSED LOSSES TO LEARN FINE-GRAINED FEATURES WITH TEMPORAL DIVERSITY

Approaches	L_{reg}	l_{div}	L_{gc}	L_{alg}	Spr. Corr. (ρ) ↑
I3D+MLP	✓	×	×	×	0.9450
I3D+TE	✓	×	×	×	0.9516
I3D+TE+PD	✓	×	×	×	0.9531
I3D+TE+PD	✓	✓	×	×	0.9547
VATP-Net	✓	×	×	×	0.9522
VATP-Net	✓	✓	×	×	0.9540
VATP-Net	✓	✓	×	×	0.9554
VATP-Net	✓	✓	✓	×	0.9565
VATP-Net	✓	×	✓	✓	0.9545
VATP-Net (Ours)	✓	✓	✓	✓	**0.9588**

TABLE VI

EVALUATION OF MULTIMODAL INTERACTION MODULE IN MTL-AQA DATASET. THE RESULTS SHOW THAT THE MODULE CAPTURES COMPLEX RELATIONSHIPS AMONG MODALITIES

Method	Modalities		Spr. Corr. (ρ) ↑
	RGB	Semantics	
I3D+MLP	✓	×	0.9450
GDLT [38]	✓	×	0.9531
VATP-Net	✓	×	0.9547
VATP-Net	✓	✓	0.9588

TABLE VII

EVALUATION OF BACKBONE NETWORK FOR ACTION FEATURE IN THE MTL-AQA DATASET. THIS RESULT HIGHLIGHTS THE IMPORTANCE OF BACKBONE IN OPTIMIZING ACTION FEATURE LEARNING ACROSS DATASETS

Method	Backbone Nets			Spr. Corr. (ρ) ↑
	I3D [57]	VST [61]	BERT [58]	
I3D+MLP	✓	×	×	0.9450
GDLT [38]	✓	×	×	0.9531
VATP-Net	✓	×	×	0.9531
VATP-Net	×	✓	×	0.9482
VATP-Net	×	✓	✓	0.9543
VATP-Net	✓	×	✓	0.9588

and can even result in diminishing returns. Using a decoder block with 2 transformer layers appears to be a more efficient and effective approach, compared to increasing the complexity of the model further.

4) Evaluation on Loss Functions: The experimental results shown in Table V demonstrate the contribution of the proposed loss functions. An interesting observation is that the proposed approach performance substantially deteriorates dramatically when excluding either the proposed visual-semantic alignment loss or global contrastive loss.

a) Visual-semantic alignment loss function : First, we integrate visual and semantic action features by fusing the two sets of features, achieving a Spr. Corr. of 0.9522 on the MTL-AQA dataset as shown in Table V. This demonstrates the benefits of combining both types of information for score regression. Secondly, incorporating alignment loss in our VATP-Net model further enhances performance, resulting in a Spr. Corr. of 0.9562 as shown in Table V. Each branch within the VATP-Net module plays a crucial role in the final AQA score prediction. This contributes to a more comprehensive understanding of the underlying dynamics within the videos. Thus, our modules leverage enhanced temporal diversity and maintain consistent video representation by incorporating visual semantic alignment loss for fine-grained AQA score prediction.

b) Global visual-semantic contrastive loss: The proposed loss aims to ensure that extracted subactions match semantic descriptions within a video. This encourages the model to learn to distinguish semantic content across samples. To demonstrate the contribution of the proposed loss, we conduct experiments, and the experimental results are shown in Table V. As indicated in the experimental results, when we add the proposed loss, the performance improves, with Spr. Corr. rising from 0.9565 to 0.9588. These results show that the model can effectively distinguish the semantic content between different video samples, ensuring that the extracted subaction sequences match the semantic descriptions within each video.

c) Evaluation on diversity regularization loss: The main objective of the proposed diversity regularization loss is to enforce diversity among sub-actions and improve performance by ignoring duplicated sub-action sequences. To verify the contribution of the proposed loss function, we conduct experiments and present the results in Table V. The result highlights the significance of diversity learning, achieving a Spr. Corr. of 0.9588. Each proposed loss function is crucial for accurate AQA score prediction, enabling an effective representation of fine-grained and scene-invariant action features.

5) Contribution of multimodal interaction action feature Learning: To demonstrate the contribution of the multimodal interaction action feature learning module, we conducted experiments on the MTL-AQA dataset by removing the semantic branch, thereby utilizing only the RGB modality. The experimental results are presented in Table VI, which compares our approach to various baseline models. As indicated in the table, the removal of the semantic branch significantly degrades the performance of the proposed approach. However, compared to the baseline GDLT [38], our proposed method learns fine-grained action features and achieves a Spr. Corr. of 0.9547. Conversely, by incorporating the semantic action branch, performance improves from 0.9547 to 0.9588 for Spr. Corr. These results illustrate that the proposed module effectively captures the intricate relationships between different modalities within an action sequence, thereby enhancing the overall understanding of action assessment.

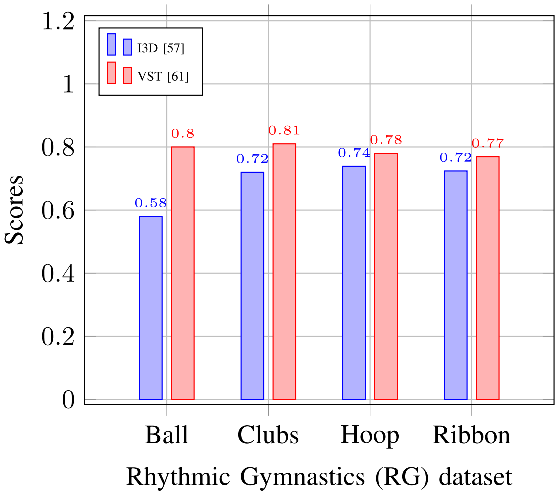


Fig. 5. Evaluation on the backbone network action feature learning in RG dataset.

TABLE VIII

COMPUTATIONAL COMPARISON WITH EXISTING APPROACHES IN MTL-AQA DATASET. THE TOTAL NUMBER OF TRAINABLE PARAMETERS IN OUR PROPOSED APPROACH IS AT A MIDDLE LEVEL

	Approaches	Year	Spr. Corr. (ρ) ↑	Par. (M)
Exemplar Based	CoRe [23]	2021	0.9512	2.05
	TAP [20]	2022	**0.9607**	15.76
	FSPN [22]	2023	0.9601	3.75
	SGN [54]	2023	**0.9607**	1.47
Exemplar Free	USDL [33]	2020	0.9231	0.79
	MUSDL [33]	2020	0.9273	0.79
	GDLT [38]	2022	0.9531	3.20
	H-GCN [10]	2023	0.9563	0.50
	VATP-Net (Ours)	-	**0.9588**	1.81

6) Evaluation on the backbone network action feature learning: We conducted an ablation study to demonstrate the contribution of action feature learning through the backbone networks. The experimental results are presented in Table VII and Fig. 5 for the MTL-AQA and RG datasets, respectively. The results indicate that the I3D backbone [57] achieves superior performance on the MTL-AQA dataset due to its deep network structure, which effectively captures fast and complex action features. In contrast, the VST backbone excels with the RG dataset by effectively capturing long-range action sequences through self-attention mechanism. In addition, we also compare the proposed approach with different backbone feature learning methods on the RG dataset, as shown in Table X. The RG videos are long-range sequences; therefore, the VST [61] is appropriate for feature learning in this context. These results highlight the significance of backbone architecture in optimizing feature learning across different datasets.

C. Model Complexity

Our proposed method requires fewer trainable parameters compared to the existing state-of-the-art approaches. As shown in VIII and Table X, a total number of trainable parameters in our proposed approach is middle-level compared to the existing approaches. It is important to note that the complexity of the backbone architecture is not included when calculating the total number of trainable parameters. By reducing the number of trainable parameters required, our approach is more efficient and easier to deploy, without sacrificing high-performance levels. This makes the proposed approach more practical and scalable for real-world applications where computational resources may be constrained.

TABLE IX

COMPARISON WITH THE STATE-OF-THE-ART APPROACHES IN THE MTL-AQA DATASET. HERE, W/DD AND W/O DD REFER TO WITH AND WITHOUT DEGREE OF DIFFICULTY LABELS, RESPECTIVELY

	Approaches (w/o DD)	Year	Spr. Corr. (ρ) ↑
Exemplar Based	CoRe [23]	2021	0.9341
	TAP [20]	2022	0.9451
	PCLA [11]	2022	0.9230
	SGN [54]	2023	**0.9500**
	FSPN [22]	2023	0.9382
Exemplar Free	C3D-LSTM [62]	2017	0.8489
	MSCADC-MTL [63]	2019	0.8612
	C3D-AVG-MTL [63]	2019	0.9044
	USDL [33]	2020	0.9066
	MUSDL [33]	2020	0.9158
	TSA-Net [13]	2021	0.9422
	H-GCN [10]	2023	0.9390
	VATP-Net (Ours)	-	**0.9452**
	Approaches (w/ DD)	Year	Spr. Corr. (ρ) ↑
Exemplar Based	CoRe [23]	2021	0.9512
	TAP [20]	2022	**0.9607**
	SGN [54]	2023	**0.9607**
	FSPN [22]	2023	0.9601
	FineParser [36]	2024	0.9585
Exemplar Free	USDL [33]	2020	0.9231
	MUSDL [33]	2020	0.9273
	UD-AQA [34]	2022	0.9545
	H-GCN [10]	2023	0.9563
	NAE [64]	2024	0.9430
	VATP-Net (Ours)	-	**0.9588**

D. Comparison With the State-of-the-Art Approaches

The result comparison between our proposed approach and other AQA methods in the MTL-AQA [2], RG [26], FineFS [24] and Fiv-S [27] datasets are shown in Table IX, Table X, Table XII and Table XI, respectively. As shown in the Tables, our approach achieves superior performance compared to state-of-the-art approaches without requiring exemplar videos and fewer trainable parameters. This is due to the fine-grained representation of action features, capturing the interaction between different modalities of action features and accuratelytransferring semantic knowledge, which improves the overall performance of AQA score prediction.

1) Comparison in the MTL-AQA Dataset: The comparison between our proposed approach and other state-of-the-art AQA methods on the MTL-AQA dataset is presented in Table IX. As shown in the table, our approach achieves

TABLE X

COMPARISON WITH THE STATE-OF-THE-ART APPROACHES IN THE RHYTHMIC GYMNASTICS (RG) DATASET, FINE-TUNED WITH OUR TRAINED PARSEFORMER DECODER MODULE. THE AVERAGE SPEARMAN'S RANK CORRELATION IS CALCULATED USING THE FISHER-Z VALUE

Method	Par. (M)	Backbone	Ball	Clubs	Hoop	Ribbon	Avg.
C3D+SVR [2]	-	C3D	0.357	0.551	0.495	0.516	0.483
MS-LSTM [27]	2.66	C3D	-	-	0.650	0.780	0.721
MS-LSTM [27]	2.66	I3D	0.515	0.621	0.540	0.522	0.551
ACTION-NET [26]	28.0	I3D+Resnet	0.528	0.652	0.708	0.578	0.623
GDLT [38]	3.16	I3D	0.526	0.710	0.729	0.704	0.674
HGCN [10]	0.50	I3D	0.534	0.609	0.706	0.621	0.621
CoFInAl [41]	3.70	I3D	0.625	0.719	0.734	0.757	0.712
VATP-Net (ours)	3.70	I3D	0.580	0.720	0.739	0.724	0.709
MS-LSTM [27]	-	VST	0.621	0.661	0.670	0.695	0.663
ACTION-NET [26]	28.0	VST+Resnet	0.684	0.737	0.733	0.754	0.728
GDLT [38]	3.16	VST	0.746	0.802	0.765	0.741	0.765
HGCN [10]	0.50	VST	0.711	0.789	0.728	0.703	0.735
CoFInAl [41]	3.70	VST	**0.809**	0.806	**0.804**	**0.810**	**0.807**
VATP-Net (ours)	1.81	VST	**0.800**	**0.810**	0.780	0.769	**0.800**

superior performance compared with existing methods that only use a single input video. Furthermore, our model achieves comparable performance to recently proposed approaches that require additional exemplar videos [20], [22], [54]. These methods predict score differences by comparing multiple input videos based on exemplars, which limits their effectiveness during inference. In contrast, our approach does not require exemplar videos and relies on fine-grained action feature representation. Our model's visual and semantic encoding enhances its ability to recognize subtle differences, leading to superior AQA performance with fewer trainable parameters.

2) Comparison in the Rhythmic Gymnastics (RG) Dataset: We compare our approach with existing methods using the RG dataset, and the results are presented in Table X. Specifically, our method shows a performance improvement of up to 3% compared to the baseline GDLT [38]. Our proposed approach captures fine-grained action information via visual-semantic alignment and obtains a comprehensive understanding of the temporal structure of action sequences. In comparison to the CoFInAl [41] method, our approach achieves comparable results on the RG dataset with less number of trainable parameters. Thus, our method enhances AQA by incorporating semantic action features, resulting in a more efficient and generalizable representation with a lightweight solution.

3) Comparison in the Figure Skating Video Dataset (Fis-V): We compare the proposed approach with existing methods on the Fis-V dataset [27] for TES and PCS score prediction. The results in Table XI show the proposed method achieves better performance due to fine-grained representation. The knowledge transfer facilitated by multimodal visual-semantic alignment is crucial to the proposed method's improved predictive performance. The model can effectively transfer relevant semantic knowledge to better capture the correspondence relationship between visual and semantic features.

TABLE XI

COMPARISON WITH THE STATE-OF-THE-ART APPROACHES IN THE FIV-S DATASET, FINE-TUNED WITH OUR TRAINED VISUAL PARSEFORMER DECODER MODULE. THE AVERAGE SPEARMAN'S RANK CORRELATION IS CALCULATED USING THE FISHER-Z VALUE

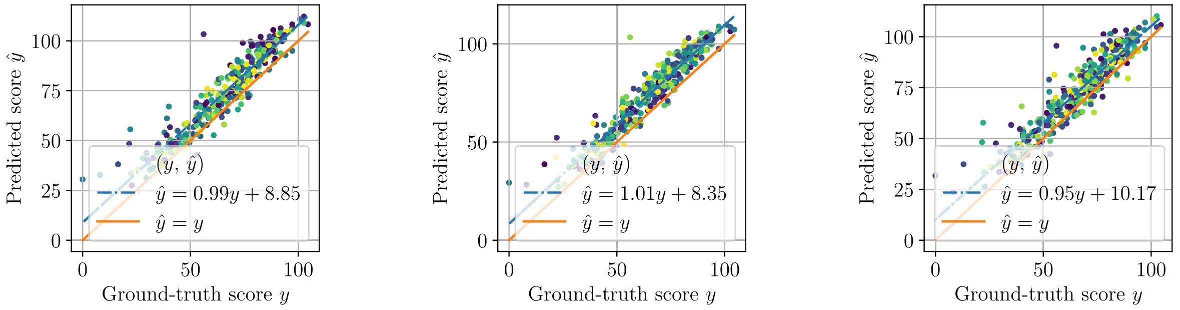
Method	Backbone	Par.(M)	Fis-V		
			TES	PCS	Avg.
C3D+SVR [2]	C3D	-	0.400	0.590	0.501
GDLT [38]	I3D	3.16	0.260	0.395	0.329
HGCN [10]	I3D	0.50	0.311	0.407	0.360
CoFInAl [41]	I3D	3.70	0.589	0.788	0.702
ACTION-NET [26]	VST	28.08	0.694	0.809	0.757
GDLT [38]	VST	3.16	0.685	0.820	0.761
HGCN [10]	VST	0.50	0.246	0.221	0.234
CoRe [23]	VST	2.05	0.660	0.820	0.750
TPT [20]	VST	15.76	0.570	0.760	0.677
SGN [54]	VST	1.47	0.700	0.830	0.772
CoFInAl [41]	VST	3.70	**0.716**	0.843	0.788
VATP-Net (Ours)	VST	1.81	0.702	**0.863**	**0.796**

TABLE XII

COMPARISON WITH THE STATE-OF-THE-ART APPROACHES IN THE FINEFS DATASET, FINE-TUNED WITH OUR TRAINED VISUAL PARSE-FORMERDECODER MODULE

Approach	SP (ρ)			FS (ρ)		
	PCS	TES	Avg.	PCS	TES	Avg.
Ms-LSTM [27]	0.602	0.547	0.575	0.615	0.552	0.584
TSA-Net [13]	0.696	0.657	0.677	0.704	0.663	0.684
TSA [21]	0.776	0.529	0.671	0.779	0.675	0.730
GDLT [38]	0.783	0.644	0.721	0.800	0.633	0.727
LUSD-Net [24]	**0.813**	0.689	**0.758**	0.863	**0.779**	**0.826**
VATP-Net (Ours)	0.809	**0.691**	0.750	**0.868**	0.721	0.795

4) Comparison in the Fine-Grained Figure Skating Dataset (FineFS): We present a comprehensive comparison between our proposed approach and several baseline methods implemented in the FineFS dataset, as shown in Table XII. Our approach achieves impressive average scores of 0.744 and 0.794, outperforming GDLT [38] on the SP and FS subtasks, respectively. These results suggest that our proposed modules have the potential to effectively transfer semantic knowledge to visual features, ultimately leading to improved overall performance in score prediction. Compared to the LUSD-Net method [24], our approach achieves comparable results. Overall, the empirical evaluation presented in Table XII validates the effectiveness of our approach over existing baseline methods for the score prediction of figure skating with less trainable parameters.



(a) I3D+MLP (b) I3D+TE (c) **VATP-Net (Ours)**

Fig. 6. A comparison of predicted score distribution in the MTL-AQA dataset. We show scatter plots of prediction scores, where the orange line illustrates the perfect predictions and the blue line indicates the fitted predictions. As shown, our proposed VATP-Net obtains a more precise predicted score distribution, providing insight into the model's performance.

E. Visualization Results

1) Scatter Plots of Score Prediction Correlation: We visualize the score distributions of the proposed approach in Fig. 6 and compare it with other AQA approaches, namely I3D+MLP and I3D+TE. The scatter plots display the prediction scores, with the orange line representing perfect prediction and the blue line indicating the fitted prediction generated by the model. As shown in Fig. 6, the scatter plot visually demonstrates the effectiveness of our approach. Our proposed approach has a more compact distribution of predicted scores compared to the baseline I3D+MLP and I3D+TE methods. The prediction scores in our approach are much closer to the ground-truth line. This indicates that our approach successfully captures the underlying patterns, leading to a superior understanding of fine-grained action features and an enhancement of the overall performance of AQA score prediction.

2) Cumulative Score Curve: The cumulative score curve at error δ is computed as $\mathrm{CS}(\alpha) = \frac{N_{\delta \leq \alpha}}{N} \times 100\%$, where $N_{\delta \leq \alpha}$ is the number of videos with prediction error δ that do not exceed the threshold α. Fig. 7 shows the curves for our proposed and baseline works. The x-axis measures the absolute difference between prediction and ground truth, and the y-axis shows the proportion of the sample within the current error level. Samples with an absolute difference between prediction and ground truth less than δ are considered positive. A larger curve area indicates higher performance. In any δ, our proposed approach (green) outperforms the baseline approaches and demonstrates the effectiveness of the proposed approach.

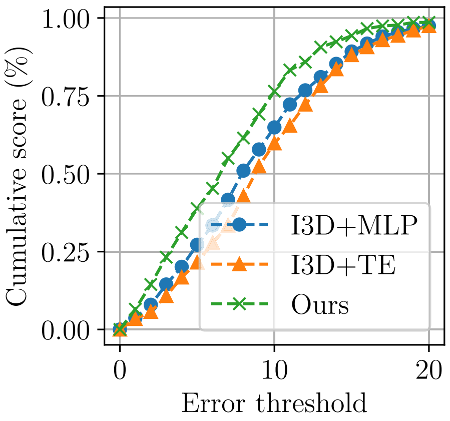


Fig. 7. Cumulative score curve in MTL-AQA dataset. A larger curve area indicates better performance, and our proposed approach outperforms baseline works under any δ value.

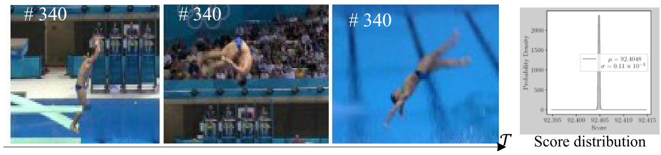


Fig. 8. The failure case study of our approach on the action sequences of diving (#340) in the MTL-AQA dataset. This suggests that future research is needed to improve robustness in handling the AQA for such challenging action scenarios.

F. Failed Case Analysis of Our Proposed Approach

Our proposed approach generally performs well, but it is important to analyze failed cases for future research direction and overall performance improvement. In the MTL-AQA dataset, we found that only one action sample, *i.e.,* #340, had errors greater than 40. Specifically, this outlier error occurred during a diving attempt with a difficulty of 3.1, where the diver's inability to maintain proper rhythm on the springboard and failure to execute necessary movements after takeoff led to the error as shown in Fig. 8. Ultimately, the diver made a large splash when entering the water [10]. As a result of these significant errors in the dive execution, the performance has been judged with a score of 0. Our proposed method, along with other state-of-the-art approaches, such as [10], appears to perform poorly for this particular type of challenging action sequence. This indicates that future research is needed to enhance robustness in detecting fouls and assessing fine-grained action execution in challenging scenarios.

V. Conclusion

This paper proposes a visual-semantic alignment self-supervised temporal parsing network that aims to learn consistent representation and internal temporal structures of action sequences. The proposed approach utilizes a self-supervised temporal parsing module to generate sub-action sequences from a given video, which is supervised by aligning the visual and semantic action features. Moreover, a multimodal interaction module is proposed to capture the interaction between different modalities of action features and encourage interactions between different modalities. Experimental results on four challenging AQA datasets demonstrated the contribution of our proposed approach. The experiments suggest that the proposed approach holds promise in improving AQA performance in practical scenarios with a few trainable parameters.

Future research directions will emphasize few-shot learning to enhance our model's adaptability with limited labeled samples. Additionally, we will integrate more discriminative auxiliary information, such as athlete poses and large language models (LLMs), to effectively model the spatiotemporal dynamics of fine-grained and scene-invariant action features. This integration will address the challenge of accurately detecting fouls in complex scenarios. A key focus will be on improving foul detection in chaotic environments, particularly through the application of few-shot learning with LLMs. This approach will allow our model to achieve better contextual understanding and adapt to new scenarios with limited labeled samples.

References

[1] P. Parmar and B. T. Morris, "What and how well you performed? A multitask learning approach to action quality assessment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 304–313.

[2] P. Parmar and B. T. Morris, "Learning to score Olympic events," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 76–84.

[3] Y. Zhang, W. Xiong, and S. Mi, "Learning time-aware features for action quality assessment," *Pattern Recognit. Lett.*, vol. 158, pp. 104–110, Jun. 2022.

[4] L. Marchesotti, F. Perronnin, D. Larlus, and G. Csurka, "Assessing the aesthetic quality of photographs using generic image descriptors," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 1784–1791.

[5] Y. Tian, H. Zeng, J. Hou, J. Chen, and K.-K. Ma, "Light field image quality assessment via the light field coherence," *IEEE Trans. Image Process.*, vol. 29, pp. 7945–7956, 2020.

[6] Y.-M. Li, L.-A. Zeng, J.-K. Meng, and W.-S. Zheng, "Continual action assessment via task-consistent score-discriminative feature distribution modeling," *IEEE Trans. Circuits Syst. Video Technol.*, early access, May 3, 2024, doi: 10.1109/TCSVT.2024.3396692.

[7] S.-J. Zhang, J.-H. Pan, J. Gao, and W.-S. Zheng, "Adaptive stage-aware assessment skill transfer for skill determination," *IEEE Trans. Multimedia*, vol. 26, pp. 4061–4072, 2024.

[8] W. Zhou, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[9] X. Gao, W. Lu, D. Tao, and X. Li, "Image quality assessment based on multiscale geometric analysis," *IEEE Trans. Image Process.*, vol. 18, no. 7, pp. 1409–1423, Jul. 2009.

[10] K. Zhou, Y. Ma, H. P. H. Shum, and X. Liang, "Hierarchical graph convolutional networks for action quality assessment," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 12, pp. 7749–7763, Dec. 2023.

[11] M. Li, H.-B. Zhang, Q. Lei, Z. Fan, J. Liu, and J.-X. Du, "Pairwise contrastive learning network for action quality assessment," in *Proc. ECCV*, 2022, pp. 457–473.

[12] H. Doughty, D. Damen, and W. Mayol-Cuevas, "Who's better? Who's best? Pairwise deep ranking for skill determination," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6057–6066.

[13] S. Wang, D. Yang, P. Zhai, C. Chen, and L. Zhang, "TSA-Net: Tube self-attention network for action quality assessment," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 4902–4910.

[14] H. Doughty, W. Mayol-Cuevas, and D. Damen, "The pros and cons: Rank-aware temporal attention for skill determination in long videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7854–7863.

[15] A. A. Laghari, R. A. Laghari, and A. Khan, "Quality of experience assessment of online server/cloud gaming," in *Proc. 8th Annu. Int. Conf. Netw. Inf. Syst. Comput. (ICNISC)*, Sep. 2022, pp. 834–837.

[16] A. Laghari, H. He, A. Khan, R. Laghari, S. Yin, and J. Wang, "Crowdsourcing platform for QoE evaluation for cloud multimedia services," *Comput. Sci. Inf. Syst.*, vol. 19, no. 3, pp. 1305–1328, 2022.

[17] A. A. Laghari, H. He, A. Khan, N. Kumar, and R. Kharel, "Quality of experience framework for cloud computing (QoC)," *IEEE Access*, vol. 6, pp. 64876–64890, 2018.

[18] A. A. Laghari, V. V. Estrela, and S. Yin, "How to collect and interpret medical pictures captured in highly challenging environments that range from nanoscale to hyperspectral imaging," *Current Med. Imag.*, vol. 54, no. 36582065, p. 1, 2022.

[19] J.-H. Pan, J. Gao, and W.-S. Zheng, "Adaptive action assessment," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 12, pp. 8779–8795, Dec. 2022.

[20] Y. Bai et al., "Action quality assessment with temporal parsing transformer," in *Proc. ECCV*, 2022, pp. 422–438.

[21] 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 *Proc. CVPR*, 2022, pp. 2949–2958.

[22] K. Gedamu, Y. Ji, Y. Yang, J. Shao, and H. T. Shen, "Fine-grained spatio-temporal parsing network for action quality assessment," *IEEE Trans. Image Process.*, vol. 32, pp. 6386–6400, 2023.

[23] X. Yu, Y. Rao, W. Zhao, J. Lu, and J. Zhou, "Group-aware contrastive regression for action quality assessment," in *Proc. ICCV*, 2021, pp. 7899–7908.

[24] Y. Ji, L. Ye, H. Huang, L. Mao, Y. Zhou, and L. Gao, "Localization-assisted uncertainty score disentanglement network for action quality assessment," in *Proc. 31st ACM Int. Conf. Multimedia*, Oct. 2023, pp. 8590–8597.

[25] N. Dvornik, I. Hadji, K. G. Derpanis, A. Garg, and A. D. Jepson, "Drop-DTW: Aligning common signal between sequences while dropping outliers," in *Proc. NeurIPS*, 2021, pp. 13782–13793.

[26] L.-A. Zeng et al., "Hybrid dynamic-static context-aware attention network for action assessment in long videos," in *Proc. 28th ACM Int. Conf. Multimedia*, 2020, pp. 2526–2534.

[27] C. Xu, Y. Fu, B. Zhang, Z. Chen, Y. Jiang, and X. Xue, "Learning to score figure skating sport videos," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 12, pp. 4578–4590, Dec. 2020.

[28] S. Gattupalli, D. Ebert, M. Papakostas, F. Makedon, and V. Athitsos, "CogniLearn: A deep learning-based interface for cognitive behavior assessment," in *Proc. 22nd Int. Conf. Intell. User Interfaces*, Mar. 2017, pp. 577–587.

[29] A. Zia, Y. Sharma, V. Bettadapura, E. L. Sarin, M. A. Clements, and I. Essa, "Automated assessment of surgical skills using frequency analysis," in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Intervent.*, 2015, pp. 430–438.

[30] H. Pirsiavash, C. Vondrick, and A. Torralba, "Assessing the quality of actions," in *Proc. Eur. Conf. Comput. Vis.*, Cham, Switzerland, 2014, pp. 556–571.

[31] H. Jain, G. Harit, and A. Sharma, "Action quality assessment using Siamese network-based deep metric learning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 6, pp. 2260–2273, Jun. 2021.

[32] S.-J. Zhang, J.-H. Pan, J. Gao, and W.-S. Zheng, "Semi-supervised action quality assessment with self-supervised segment feature recovery," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 9, pp. 6017–6028, Sep. 2022.

[33] Y. Tang et al., "Uncertainty-aware score distribution learning for action quality assessment," in *Proc. CVPR*, 2020, pp. 9839–9848.

[34] C. Zhou, Y. Huang, and H. Ling, "Uncertainty-driven action quality assessment," 2022, *arXiv:2207.14513*.

[35] S. Zhang et al., "Logo: A long-form video dataset for group action quality assessment," in *Proc. CVPR*, 2023, pp. 2405–2414.

[36] 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 *Proc. CVPR*, 2024, pp. 14628–14637.

[37] G. Bertasius, H. S. Park, S. X. Yu, and J. Shi, "Am I a baller? Basketball performance assessment from first-person videos," in *Proc. CVPR*, 2017, pp. 2196–2204.

[38] A. Xu, L.-A. Zeng, and W.-S. Zheng, "Likert scoring with grade decoupling for long-term action assessment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 3222–3231.

[39] H. Fang, W. Zhou, and H. Li, “End-to-end action quality assessment with action parsing transformer,” in *Proc. IEEE Int. Conf. Vis. Commun. Image Process. (VCIP)*, vol. 30, Dec. 2023, pp. 1–5.

[40] Q. An, M. Qi, and H. Ma, “Multi-stage contrastive regression for action quality assessment,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, vol. 33, Apr. 2024, pp. 4110–4114.

[41] K. Zhou, J. Li, R. Cai, L. Wang, X. Zhang, and L. Xiaohui, “CoFInAl: Enhancing action quality assessment with coarse-to-fine instruction alignment,” in *Proc. IJCAI*, 2024, pp. 1771–1779.

[42] D. Shao, Y. Zhao, B. Dai, and D. Lin, “Intra- and inter-action understanding via temporal action parsing,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 727–736.

[43] C. Zhang, A. Gupta, and A. Zisserman, “Temporal query networks for fine-grained video understanding,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 4484–4494.

[44] D. Shao, Y. Zhao, B. Dai, and D. Lin, “FineGym: A hierarchical video dataset for fine-grained action understanding,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2613–2622.

[45] H. Zeng, H. Huang, J. Hou, J. Cao, Y. Wang, and K.-K. Ma, “Screen content video quality assessment model using hybrid spatiotemporal features,” *IEEE Trans. Image Process.*, vol. 31, pp. 6175–6187, 2022.

[46] H. Huang, H. Zeng, J. Hou, J. Chen, J. Zhu, and K.-K. Ma, “A spatial and geometry feature-based quality assessment model for the light field images,” *IEEE Trans. Image Process.*, vol. 31, pp. 3765–3779, 2022.

[47] A. A. Laghari, Y. Sun, M. Alhussein, K. Aurangzeb, M. S. Anwar, and M. Rashid, “Deep residual-dense network based on bidirectional recurrent neural network for atrial fibrillation detection,” *Sci. Rep.*, vol. 13, no. 1, p. 15109, Sep. 2023.

[48] A. K. Jumani et al., “Quality of experience that matters in gaming graphics: How to blend image processing and virtual reality,” *Electronics*, vol. 13, no. 15, p. 2998, Jul. 2024.

[49] M. Javed, Z. Zhang, F. H. Dahri, and A. A. Laghari, “Real-time deepfake video detection using eye movement analysis with a hybrid deep learning approach,” *Electronics*, vol. 13, no. 15, p. 2947, Jul. 2024.

[50] Z.-Y. Dou et al., “An empirical study of training end-to-end vision-and-language transformers,” in *Proc. CVPR*, 2022, pp. 18145–18155.

[51] Z. Fang, J. Wang, X. Hu, L. Wang, Y. Yang, and Z. Liu, “Compressing visual-linguistic model via knowledge distillation,” in *Proc. ICCV*, 2021, pp. 1408–1418.

[52] A. Yang et al., “Vid2Seq: Large-scale pretraining of a visual language model for dense video captioning,” in *Proc. CVPR*, 2023, pp. 10714–10726.

[53] N. Dvornik, I. Hadji, R. Zhang, K. G. Derpanis, R. P. Wildes, and A. D. Jepson, “StepFormer: Self-supervised step discovery and localization in instructional videos,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 18952–18961.

[54] Z. Du, D. He, X. Wang, and Q. Wang, “Learning semantics-guided representations for scoring figure skating,” *IEEE Trans. Multimedia*, vol. 26, pp. 4987–4997, 2024.

[55] L.-A. Zeng and W.-S. Zheng, “Multimodal action quality assessment,” *IEEE Trans. Image Process.*, vol. 33, pp. 1600–1613, 2024.

[56] J. Xia et al., “Skating-Mixer: Long-term sport audio-visual modeling with MLPs,” in *Proc. AAAI Conf. Artif. Intell.*, Jun. 2023, vol. 37, no. 3, pp. 2901–2909.

[57] J. Carreira and A. Zisserman, “Quo vadis, action recognition? A new model and the kinetics dataset,” in *Proc. CVPR*, 2017, pp. 4724–4733.

[58] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in *Proc. NAACL-HLT*, 2019, pp. 4171–4186.

[59] A. van den Oord, Y. Li, and O. Vinyals, “Representation learning with contrastive predictive coding,” 2018, *arXiv:1807.03748*.

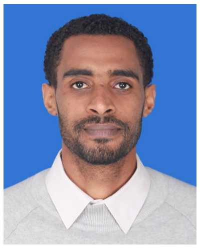
[60] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A simple framework for contrastive learning of visual representations,” in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2020, pp. 1–11.

[61] Z. Liu et al., “Video Swin transformer,” in *Proc. CVPR*, 2022, pp. 3202–3211.

[62] J.-H. Pan, J. Gao, and W.-S. Zheng, “Action assessment by joint relation graphs,” in *Proc. ICCV*, 2019, pp. 6330–6339.

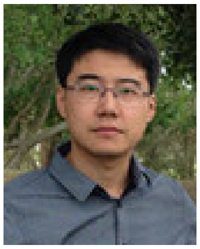
[63] A. Montes, A. Salvador, S. Pascual, and X. Giro-I Nieto, “Temporal activity detection in untrimmed videos with recurrent neural networks,” in *Proc. NIPS*, 2016, pp. 1–5.

[64] S. Zhang et al., “Narrative action evaluation with prompt-guided multimodal interaction,” in *Proc. CVPR*, 2024, pp. 18430–18439.



Kumie Gedamu received the B.Sc. degree in information science from Haramya University, Ethiopia, in 2010, the master’s degree in computer science and system engineering from Andhra University, India, in 2015, and the Ph.D. degree in computer science and technology from the University of Electronic Science and Technology of China, China, in 2022. He is currently a Post-Doctoral Researcher with Sichuan Artificial Intelligence Research Institute, University of Electronic Science and Technology of China. His research interests include computer vision and video quality assessment.

Yanli Ji received the Ph.D. degree from the Department of Advanced Information Technology, Kyushu University, Japan, in 2012. She was a Visiting Researcher with The University of Tokyo from 2018 to 2020. She is currently a Professor with Sun Yat-sen University. Her research interests include human-centric visual understanding, human–robot interaction, and robot intelligence.



Yang Yang (Senior Member, IEEE) received the bachelor’s degree from Jilin University in 2006, the master’s degree from Peking University in 2009, and the Ph.D. degree from The University of Queensland, Australia, in 2012, under the supervision of Prof. Heng Tao Shen and Prof. Xiaofang Zhou. He was a Research Fellow with the National University of Singapore from 2012 to 2014 under the supervision of Prof. Tat Seng Chua. He is currently with the University of Electronic Science and Technology of China. His research interests include computer vision and social media analytics.

Jie Shao (Member, IEEE) received the B.E. degree in computer science from Southeast University, Nanjing, China, in 2004, and the Ph.D. degree in computer science from The University of Queensland, Brisbane, Australia, in 2009. He was a Research Fellow with The University of Melbourne from 2008 to 2011 and the National University of Singapore from 2012 to 2014. He is currently a Professor with the Center for Future Media, School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, China. His research interests include data management and multimedia information retrieval.

Heng Tao Shen received the B.Sc. (Hons.) and Ph.D. degrees from the Department of Computer Science, National University of Singapore, in 2000 and 2004, respectively. Then, he joined The University of Queensland as a Lecturer and a Senior Lecturer, where he became a Professor in 2011. He is currently a Professor of the National “Thousand Talents Plan” and the Dean of the School of Computer Science and Engineering, University of Electronic Science and Technology of China. He is also an Honorary Professor with The University of Queensland. His research interests include multimedia search, computer vision, artificial intelligence, and big data management.