MoAKE: Toward Unified All-in-One Action Quality Assessment via Mixture of Action Knowledge Experts

Huangbiao Xu[1,2], Huanqi Wu[1,2], Xiao Ke[1,2,✉], Jiaxin Cai[1], Junyi Wu[1,2], and Jinglin Xu[3]

[1] Fujian Provincial Key Laboratory of Networking Computing and Intelligent Information Processing, College of Computer and Data Science, Fuzhou University, Fuzhou 350108, China
[2] Engineering Research Center of Big Data Intelligence, Ministry of Education, Fuzhou 350108, China
[3] School of Intelligence Science and Technology, University of Science and Technology Beijing, Beijing 100083, China
kex@fzu.edu.cn, {huangbiaoxu.chn, wuhuanqi135, xujinglinlove}@gmail.com, jiaxincai528@163.com, junyi.wu-1@outlook.com

Abstract. Action Quality Assessment (AQA) aims to objectively evaluate performance quality from action videos. Most existing methods follow a "one-by-one" paradigm, training a separate model for each action type. This setting limits real-world deployment, as it requires prior action-type knowledge to select the corresponding model and suffers from poor generalization across diverse actions. To address these limitations, we study the challenging task of **all-in-one AQA**, which aims to assess heterogeneous actions within a single unified model. We propose a novel **Mixture of Action Knowledge Experts (MoAKE)** framework, designed to mitigate negative knowledge transfer caused by large semantic discrepancies among actions. MoAKE learns complementary experts that capture diverse action patterns within a shared semantic space and dynamically aggregates their knowledge to adapt the assessment to the input action. Each expert is tailored with segment-aware prototypes to handle varying temporal lengths, together with an **Adaptive Intra- and Inter-Segment Relationship Modeling (AIISRM)** module to model multi-granularity temporal dynamics. Furthermore, we establish **comprehensive benchmarks** for **all-in-one** as well as **zero/few-shot AQA**. Extensive experiments on three long-term datasets demonstrate that MoAKE significantly outperforms existing methods in the all-in-one setting, while also achieving consistent generalization on three short-term datasets under zero/few-shot evaluation. Code is available at https://github.com/XuHuangbiao/MoAKE.

Keywords: Action quality assessment · Unified all-in-one model · Mixture of experts · Zero/Few-shot learning · Video understanding

✉ Corresponding Author.

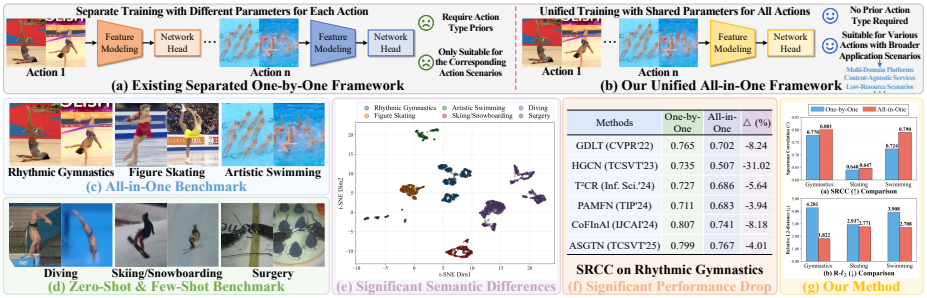

Fig. 1: Illustration of paradigm comparisons, benchmarks, and challenges. (a) Existing separated one-by-one framework trains specific models for each action type. (b) Our unified all-in-one framework assesses heterogeneous actions using a single model. (c) Our all-in-one AQA benchmark. (d) Our zero/few-shot AQA benchmark. (e) Significant semantic differences across actions pose a major challenge. (f) Naively training all-in-one models leads to negative transfer. (g) Our method enables positive transfer.

1 Introduction

Action Quality Assessment (AQA) has made significant strides in recent years across various domains such as sports [47–49,61], medical rehabilitation [3,22,57], and skill determination [42,50]. The primary goal of AQA is to learn a mapping from motion semantics to expert-provided scores. To tackle this challenging problem, recent research often attempts to model fine-grained semantics [47–49, 60] or incorporate additional semantic cues from other modalities [37, 40, 41, 43, 53]. However, these advances are mostly developed under per-action settings.

Despite their success, the prevailing approaches remain strictly confined to a "one-by-one" paradigm (Fig. 1 (a)), where separate models are trained and deployed for each specific action type. While effective in isolated benchmarks, this fragmentation creates major barriers to real-world scalability across various applications. *For example*: *first*, it relies on reliable metadata (*i.e.*, known action types) to select the correct model, complicating deployment in **open-domain platforms** [10] where diverse user-uploaded content lacks reliable labels; *second*, it hinders **content-agnostic services** [14] that must process heterogeneous or even mixed actions without manual intervention; and *third*, it suffers in **low-resource scenarios** [26], where training and maintaining high-capacity models for new or rare actions is data- and compute-intensive. Motivated by these practical needs, and inspired by prior AQA studies [27, 43, 47] that suggest the value of cross-action transfer, we study a more generalizable paradigm: **All-in-One AQA** (Fig. 1 (b)), where a single unified model assesses heterogeneous actions.

However, the path to an all-in-one AQA model is fraught with difficulties. The core challenge stems from the substantial semantic discrepancies across various actions, as illustrated in Fig. 1 (e). Directly applying existing methods to all-in-one training often leads to "**negative transfer**," where knowledge from disparate actions interferes with each other [12]. This results in significant per-

formance degradation (an average SRCC drop of 10.17%), as shown in Fig. 1 (f). This decline suggests that existing models lack mechanisms to adaptively identify relevant assessment patterns for each input action and fail to leverage the rich, albeit varied, information from multiple domains. However, these diverse action cues can be complementary when properly utilized. As shown in Fig. 1 (g), our all-in-one approach effectively mixes multi-action knowledge and demonstrates positive transfer (details in Sec. 4.1). *The key insight, therefore, is to transform this challenge into an opportunity by designing an all-in-one model that can extract and synthesize complementary knowledge from various actions.*

To this end, we propose the **M**ixture **o**f **A**ction **K**nowledge **E**xperts (**MoAKE**), a novel framework designed to mitigate negative transfer and enable adaptive assessment. Instead of forcing a single, monolithic representation, MoAKE learns a set of discriminative action knowledge experts, each specializing in capturing complementary action patterns and mapping them to a shared semantic space. Within each expert, fixed-length, segment-aware prototypes are utilized to learn and aggregate segment semantics, adapting to action scenarios of varying lengths. This is complemented by a novel **A**daptive **I**ntra- and **I**nter-**S**egment **R**elationship **M**odeling (AIISRM) module. AIISRM empowers the experts to capture multi-granularity relationships and adaptively model segment-level patterns, transforming common knowledge into action-specific characteristics. By dynamically mixing the insights from different experts, MoAKE tailors its evaluation to the specific action instance, effectively mitigating knowledge interference and enabling robust score predictions.

To demonstrate the effectiveness and generalization capabilities of our approach, we establish new comprehensive benchmarks, including both all-in-one and zero/few-shot AQA tasks. As depicted in Fig. 1 (c)-(d), we train a unified model on three mainstream long-term action types (rhythmic gymnastics [52], figure skating [39], and artistic swimming [56]) and evaluate its zero/few-shot generalization capabilities on three short-term actions (diving [28], snowboarding/skiing [27], and surgery [9]). Extensive experiments and ablation studies reveal the importance of leveraging the complementarity of different action knowledge for all-in-one AQA, demonstrating that our method outperforms state-of-the-art methods. This work, to our knowledge, pioneers a new direction by enabling the assessment of multiple action types within an all-in-one framework.

Our main contributions are summarized as follows:

- We propose a novel Mixture of Action Knowledge Experts framework for all-in-one AQA, which dynamically fuses knowledge from multiple experts to bridge the semantic gap between diverse action types, adapting its assessment to the input action.
- We introduce action knowledge experts that model representative action patterns within a shared semantic space, and an Adaptive Intra- and Inter-Segment Relationship Modeling module to dynamically capture crucial segment-aware contextual correlations.
- We establish a novel comprehensive benchmark for all-in-one and zero/few-shot AQA, validating the state-of-the-art performance of our method on

six mainstream long-term and short-term action datasets in both unified assessment and cross-action generalization.

2 Related Work

2.1 Action Quality Assessment

Action Quality Assessment (AQA) aims to learn a complex mapping from motion semantics to expert evaluation. To address this, recent research often follows two directions: fine-grained semantic modeling, which enhances discriminative power by focusing on finer sub-actions [2,46,48,49], hierarchical structures [15,60], uncerainty modeling [32], or discrete quality grades [23,38,59,61]; and multi-modal learning, which leverages auxiliary information from audio [37], flow [53, 58], text [25, 41, 43], or pose [3]. As multimodal learning [4, 5, 8, 11, 16, 34, 36, 55] has become increasingly mature, some studies [44, 45] have begun to explore incomplete multimodal AQA that better reflects real-world scenarios. Specifically, MCMoE [45] pioneers the adoption of a mixture-of-experts (MoE) [20,21] approach to investigate stable AQA models when modalities are missing. Meanwhile, LIMSSR [44] further explores the challenging scenario of training without access to complete modal data.

However, these methods adhere to a "one-by-one" paradigm, training specialized models for individual action types. This limits real-world applicability, as it requires prior knowledge of the action type and struggles to generalize across significant semantic gaps. Consequently, existing methods are difficult to deploy in scenarios with unknown action types, such as open-domain platforms, content-agnostic services, and low-resource environments. *In contrast, our MoAKE is designed for challenging all-in-one AQA, dynamically mixing knowledge from multiple specialized experts to adaptively assess diverse actions within a unified model.*

2.2 All-in-One Models

The "all-in-one" model, which handles multiple tasks or domains with a single set of weights, has gained widespread attention for its efficiency and generality in fields like image restoration [1, 13, 17, 29], motion understanding [18, 62], and vision-language learning [35, 54]. However, this paradigm presents unique challenges for AQA. Unlike other tasks, AQA requires sensitivity to subtle execution details that significantly impact scores. This difficulty is amplified when a single model must contend with the vast semantic differences across action types. Thus, existing all-in-one methods are ill-suited for AQA because they are not designed to navigate the nuanced quality assessment space. *To our knowledge, our MoAKE represents an early attempt to systematically explore all-in-one and zero/few-shot AQA, leveraging a mixture of discriminative experts to adapt to different assessment patterns and effectively utilize cross-domain knowledge.*

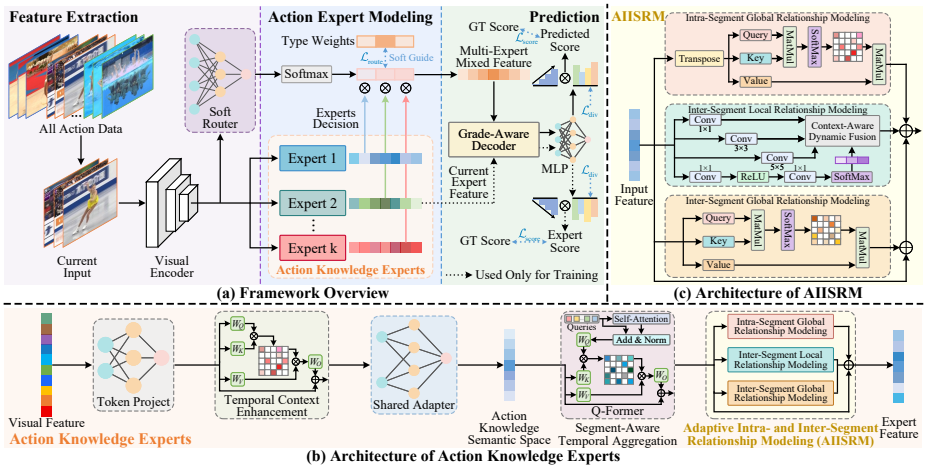

Fig. 2: Overview of our MoAKE framework. (a) MoAKE processes video features via multiple action knowledge experts, dynamically fusing their outputs for score prediction. (b) Each expert maps features to a shared space, aggregates semantics using segment-aware prototypes, and refines them via our AIISRM. (c) AIISRM adaptively models intra- and inter-segment relationships to capture multi-granularity context.

3 Method

In this section, we detail our Mixture of Action Knowledge Experts (MoAKE) framework, designed for the challenging all-in-one AQA, as shown in Fig. 2.

3.1 Overview

The primary challenge in all-in-one AQA is assessing diverse actions with significant semantic differences using a single model. Naively training on diverse actions often causes negative transfer, where conflicting action patterns hinder performance. To address this, our MoAKE framework (Fig. 2 (a)) leverages a Mixture-of-Experts (MoE) architecture. *Unlike traditional MoE designs used primarily for parameter scaling via token-level sparse routing [7,30,31], our MoAKE is specifically tailored for AQA.* First, it operates at the macro action level with a lightweight design to maintain practical computational costs for real-world deployment (as evidenced in Tab. 5). Second, instead of hard routing that strictly isolates knowledge, we employ a soft routing mechanism. This partitions the complex problem space into manageable sub-problems while adaptively blending complementary knowledge, effectively transforming the semantic gap into positive transfer.

Our MoAKE is trained on a consolidated dataset comprising K different action types. For a given input video, we extract a sequence of visual features $\mathbf{X} \in \mathbb{R}^{T \times D}$ using a pre-trained visual encoder, where T is the number of snippets and D is the feature dimension. Note that T varies across different action types due

to their temporal characteristics. These features are then processed in parallel by K distinct action knowledge experts $\{E_k\}_{k=1}^{K}$, each designed to capture unique discriminative semantic patterns.

Concurrently, a lightweight soft router, $R(\cdot)$, dynamically computes a set of weights based on the features $\mathbf{X}$, determining the contribution of each expert. Finally, action-adaptive multi-expert mixed feature $\mathbf{F}_{\text{mix}}$ is obtained by a weighted sum of the individual expert outputs $\mathbf{F}_k = E_k(\mathbf{X})$:

$$\{w_k\}_{k=1}^{K} = R(\mathbf{X}), \quad \mathbf{F}_{\text{mix}} = \sum_{k=1}^{K} w_k \mathbf{F}_k, \tag{1}$$

where w_k is the expert weights produced by the soft router, satisfying $\sum_k w_k = 1$. This mechanism allows the model to adaptively emphasize the most relevant expert knowledge for the specific action being evaluated.

For score prediction, we adopt a grade-aware decoder inspired by recent grade-based regressors [23,38,41,59,61]. The decoder, $De(\cdot)$, uses a set of learnable grade queries to transform the mixed feature $\mathbf{F}_{\text{mix}}$ into M grade-aware representations $\{\boldsymbol{r}_m\}_{m=1}^{M}$, which is then mapped to a final score $\hat{y}$.

$$\{\boldsymbol{r}_m\}_{m=1}^{M} = De(\mathbf{F}_{\text{mix}}), \quad \hat{y} = \text{MLP}(\{\boldsymbol{r}_m\}_{m=1}^{M}). \tag{2}$$

To ensure that each expert learns discriminative knowledge, we encourage each expert to learn score predictions for corresponding actions during training. The entire framework is optimized end-to-end with a composite loss function:

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{score}} + \lambda_2 \mathcal{L}_{\text{route}} + \lambda_3 \mathcal{L}_{\text{div}}, \tag{3}$$

where $\mathcal{L}_{\text{score}}$ is the primary score regression loss, $\mathcal{L}_{\text{route}}$ is a soft routing loss to guide the router using ground-truth action category labels, and $\mathcal{L}_{\text{div}}$ is a diversity loss to ensure different grade patterns focus on distinct action quality. The hyperparameters λ_1, λ_2 and λ_3 balance these objectives.

3.2 Action Knowledge Experts

The core of our MoAKE lies in the action knowledge experts. Crucially, unlike standard MoE frameworks where experts are typically simple feed-forward networks [7,30,31], our experts are custom-designed to handle the unique temporal demands of AQA. As illustrated in Fig. 2 (b), each expert E_k is responsible for learning a distinct subset of action patterns. A key challenge is to create a unified representation space that can handle the diverse temporal lengths $(T_1, T_2, \ldots, T_K)$ and semantic differences of actions. To this end, each expert employs a tailored two-stage process: (1) adaptive temporal aggregation using segment-aware prototypes to normalize variable-length inputs, and (2) multi-granularity relationship modeling via our specifically proposed AIISRM module.

Initially, the input features $\mathbf{X} \in \mathbb{R}^{T \times D}$ are projected into a lower-dimensional space $\mathbb{R}^{T \times d}$ and enhanced with a self-attention layer [33] to capture initial temporal context. To bridge the gap between different action types, the features

are then mapped into a shared semantic space (denoted as $\mathbf{X}' \in \mathbb{R}^{T\times d}$) using a shared adapter, which consists of a two-layer multilayer perceptron (MLP) whose weights are shared across all experts. This encourages the experts to learn a common language for action quality.

Segment-Aware Temporal Aggregation. To handle variable-length action videos, we introduce a segment-aware temporal aggregation (SATA) mechanism employing N learnable, segment-aware prototypes $\mathbf{P}_k \in \mathbb{R}^{N\times d}$ specific to each expert E_k. Unlike shared ones, each expert utilizes specialized prototypes to capture action-specific temporal patterns and semantic characteristics. This design allows each expert to develop domain-specific knowledge for different action types while processing variable-length inputs. These prototypes $\mathbf{P}_k$ act as learnable queries to distill salient information from $\mathbf{X}'$ into fixed-length representations. Inspired by Q-Former [19], this process uses cross-attention where the "queries" are $\mathcal{Q} = \mathbf{W}_{\mathcal{Q}}^{\text{seg}}\mathbf{P}_k$ and the "keys/values" are $\mathcal{K} = \mathbf{W}_{\mathcal{K}}^{\text{seg}}\mathbf{X}', \mathcal{V} = \mathbf{W}_{\mathcal{V}}^{\text{seg}}\mathbf{X}'$:

$$\mathbf{F}_{\text{seg}}^{k} = \text{Softmax}\left(\mathcal{Q}(\mathcal{K})^{\text{T}}/\sqrt{d}\right)\mathcal{V} \in \mathbb{R}^{N\times d}, \tag{4}$$

where $\sqrt{d}$ is a normalization factor, and $\mathbf{W}_{\mathcal{Q}}^{\text{seg}}$, $\mathbf{W}_{\mathcal{K}}^{\text{seg}}$, $\mathbf{W}_{\mathcal{V}}^{\text{seg}}$ are learnable weights. $\mathbf{F}_{\text{seg}}^{k}$ is a fixed-length sequence of N tokens summarizing the entire action regardless of its original duration T. This standardized representation makes subsequent modules robust to temporal variations.

Expert-Specific Feature Refinement. The above temporal aggregation produces discriminative fixed-length representations $\mathbf{F}_{\text{seg}}^{k}$ that capture action-specific patterns. However, effective AQA requires understanding complex temporal relationships at multiple granularities - both within individual action segments (intra-segment) and across different segments (inter-segment). These relationships are crucial for assessing short-term execution details and long-term action coherence [59, 60]. To address this need, we introduce the Adaptive Intra- and Inter-Segment Relationship Modeling (AIISRM) module for each expert E_k.

AIISRM processes the $\mathbf{F}_{\text{seg}}^{k}$ to capture multi-granularity contextual correlations while incorporating residual connections to preserve discriminative characteristics captured by the segment-aware prototypes and stabilize training. The final expert output $\mathbf{F}_k$ is computed as:

$$\mathbf{F}_k = \text{AIISRM}_k(\mathbf{F}_{\text{seg}}^{k}). \tag{5}$$

3.3 Adaptive Segment Relationship Modeling

As introduced, the AIISRM module refines the fixed-length segment-aware features $\mathbf{F}_{\text{seg}}^{k}$ by capturing complex temporal dependencies from multiple perspectives. To achieve this, AIISRM, as depicted in Fig. 2 (c), innovatively processes its input through three parallel branches, each targeting a distinct relationship. The effectiveness of this multi-branch design is validated in our ablation studies.

Intra-Segment Global Relationship Modeling. To capture the temporal evolution patterns from a feature-centric perspective, this branch models global correlations along the feature dimension. We first transpose the input features to

$(\mathbf{F}^k_{\text{seg}})^{\text{T}} \in \mathbb{R}^{d\times N}$ and then apply a standard self-attention mechanism. This allows the model to learn how the values of individual semantic attributes (e.g., velocity, posture cues) evolve and interrelate across the entire sequence of segments. By focusing on these intra-segment temporal dynamics, this branch captures short-term execution details and a complementary view of long-term action patterns:

$$\mathbf{F}^k_{\text{intra}} = \text{Softmax}\left(\frac{\mathbf{W}^{\text{intra}}_{\mathcal{Q}}(\mathbf{F}^k_{\text{seg}})^{\text{T}}\mathbf{W}^{\text{intra}}_{\mathcal{K}}\mathbf{F}^k_{\text{seg}}}{\sqrt{d}}\right)\mathbf{W}^{\text{intra}}_{\mathcal{V}}(\mathbf{F}^k_{\text{seg}})^{\text{T}}, \tag{6}$$

where $\mathbf{W}^{\text{intra}}_{\mathcal{Q}}$, $\mathbf{W}^{\text{intra}}_{\mathcal{K}}$, and $\mathbf{W}^{\text{intra}}_{\mathcal{V}}$ are learnable matrices.

Inter-Segment Local Relationship Modeling. Assessed actions in AQA often consist of multiple sub-actions, making the local temporal correlations between them complex. To address this, this branch employs convolutions with varying kernel sizes to dynamically model the inter-segment local relationships. Specifically, we use three parallel 1D convolutions with kernel sizes of 1, 3, and 5 to extract multi-scale local features. A lightweight attention mechanism, implemented with a two-layer convolutional block, generates context-aware weights to dynamically fuse these features. This allows the model to adaptively focus on the most relevant local patterns for the assessed actions. Formally,

$$\mathbf{F}^k_{\text{local},i} = \text{Conv1D}_i(\mathbf{F}^k_{\text{seg}}), i \in \{1,3,5\}, \tag{7}$$

$$\boldsymbol{w}_i = \text{Softmax}\left(\text{ConvBlock}(\mathbf{F}^k_{\text{seg}})\right), \tag{8}$$

$$\mathbf{F}^k_{\text{local}} = \sum_{i\in\{1,3,5\}} \boldsymbol{w}_i \cdot \mathbf{F}^k_{\text{local},i}, \tag{9}$$

where Conv1D_i denotes a 1D convolution with kernel size i, $\boldsymbol{w}_i$ are the dynamically generated attention weights.

Inter-Segment Global Relationship Modeling. To model long-range dependencies, this branch applies self-attention to the segment-aware features $\mathbf{F}^k_{\text{seg}}$. This relates distant segments, which is vital for assessing properties like overall rhythm, consistency, and coherence throughout the performance:

$$\mathbf{F}^k_{\text{inter}} = \text{Softmax}\left(\frac{\mathbf{W}^{\text{inter}}_{\mathcal{Q}}\mathbf{F}^k_{\text{seg}}\mathbf{W}^{\text{inter}}_{\mathcal{K}}(\mathbf{F}^k_{\text{seg}})^{\text{T}}}{\sqrt{d}}\right)\mathbf{W}^{\text{inter}}_{\mathcal{V}}\mathbf{F}^k_{\text{seg}}, \tag{10}$$

where $\mathbf{W}^{\text{inter}}_{\mathcal{Q}}$, $\mathbf{W}^{\text{inter}}_{\mathcal{K}}$, and $\mathbf{W}^{\text{inter}}_{\mathcal{V}}$ are learnable matrices. The outputs from these three branches are aggregated through element-wise addition and then combined with the initial segment-aware feature $\mathbf{F}^k_{\text{seg}}$ via a residual connection:

$$\mathbf{F}_k = \mathbf{F}^k_{\text{seg}} + \mathbf{F}^k_{\text{intra}} + \mathbf{F}^k_{\text{local}} + \mathbf{F}^k_{\text{inter}}. \tag{11}$$

This multi-granularity modeling enables experts to better understand temporal action structures, capturing fine-grained details and long-range dependencies.

3.4 Score Prediction and Optimization

The final mixed feature $\mathbf{F}_{\text{mix}}$, which adaptively integrates knowledge from all experts, is fed into a grade-aware decoder for score prediction. Following recent

works [23,41], this decoder $De(\cdot)$, implemented as a Transformer decoder, uses M learnable queries to generate M distinct grade-aware representations $\{\boldsymbol{r}_m\}_{m=1}^{M}$. These are processed by an MLP to predict a distribution over M predefined grade levels. The final score $\hat{y}$ is then computed as the expectation over these grades, where the grade values are fixed as $G_m = \frac{m-1}{M-1}$. To encourage expert specialization, we also introduce auxiliary predictions during training: for input action type k, the corresponding expert feature $\mathbf{F}_k$ is used to predict an auxiliary score $\hat{y}_k$, incentivizing each expert to learn discriminative patterns.

Our MoAKE framework is optimized end-to-end with a composite loss function, as in Eq. (3). Specifically, $\mathcal{L}_{\text{score}}$ is a Mean Square Error (MSE) loss that supervises both the final score $\hat{y}$ and auxiliary score $\hat{y}_k$. $\mathcal{L}_{\text{route}}$ is a soft cross-entropy loss that guides the router. Instead of using hard one-hot supervision which forces strict expert isolation, we intentionally employ label smoothing (ϵ) to provide soft supervision during training. **Crucially, during inference, the router dynamically computes weights based solely on visual features without requiring any action labels.** This encourages the router to discover shared, complementary patterns across different action types, preventing rigid boundaries and fostering positive knowledge transfer. $\mathcal{L}_{\text{div}}$ is a triplet loss on the grade-aware representations $\{\boldsymbol{r}_m\}_{m=1}^{M}$ to ensure they capture distinct quality levels. The loss components can be rewritten as:

$$\mathcal{L}_{\text{score}} = \text{MSE}(\hat{y}, y) + \text{MSE}(\hat{y}_k, y), \tag{12}$$

$$\mathcal{L}_{\text{route}} = -\sum_{k=1}^{K} \left(c_k(1-\epsilon) + \frac{\epsilon}{K} \right) \log(w_k), \tag{13}$$

$$\mathcal{L}_{\text{div}} = \sum_{m} \left[\max(\cos(\boldsymbol{r}_m, \boldsymbol{r}_j)) - \min(\cos(\boldsymbol{r}_m, \boldsymbol{r}_j)) + \alpha \right]_+, \tag{14}$$

where y is the ground-truth score, c_k is a one-hot vector indicating the ground-truth action type, ϵ is a label smoothing hyperparameter, $j \neq m$, $\cos(\cdot, \cdot)$ is cosine similarity, $[\cdot]_+$ means $\max(0, \cdot)$, and α is a margin.

4 Experiments

Benchmarks and Metrics. To rigorously evaluate all-in-one AQA, we establish benchmarks designed to facilitate future research in this direction. These benchmarks are built upon six popular AQA datasets, covering a wide range of action types, semantic complexities, and temporal lengths.

- **All-in-One AQA**: We first adapt three long-term action datasets: Rhythmic Gymnastics (RG) [52], Figure Skating (Fis-V) [39], and Artistic Swimming (LOGO) [56]. A single model is trained on the combined training sets of these three datasets and evaluated on each test set separately. That is, the testing follows prior works [41, 59].
- **Zero/Few-Shot AQA**: To evaluate generalization to unseen action types, we use three short-term datasets: Diving (MTL-AQA) [28], Skiing, etc. (AQA-7) [27], and Surgery (JIGSAWS) [9]. For **Zero-shot**, the all-in-one model is

directly evaluated on target datasets without any parameter updates. For **Few-shot**, following standard adaptation protocols [1], we perform full-network few-shot fine-tuning using a minimal subset of data. We use 5% of the training data for MTL-AQA/AQA-7, and 10% (two samples per class) for the small-scale JIGSAWS dataset. While tuning only the final prototypes might seem intuitive, full fine-tuning is adopted to ensure the shared semantic space properly adapts to the entirely new domain distributions.

Following standard AQA evaluation protocols [15, 51, 61], we report Spearman's Rank Correlation (SRCC, ρ) and Relative L2-distance (R-ℓ_2). SRCC measures the monotonic relationship between predicted and ground-truth scores, while R-ℓ_2 quantifies the numerical prediction error. Higher SRCC and lower R-ℓ_2 indicate better performance.

Implementation Details. All experiments were conducted on an RTX 3090 GPU with PyTorch 2.4.1. Following prior works [59,60], we use pre-trained VST [24] and I3D [6] features for the three long-term and three short-term datasets. For RG/Fis-V/LOGO, we extract 68/124/48 snippets of 32 frames. For the short-term datasets, we extract 10 snippets of 16 frames. Feature dimensions D and d are 1024 and 256. We set the number of experts $K = 3$, grade $M = 4$, and segment-aware prototypes $N = 64$. The model is optimized using Adam and a batch size of 32 for 380 epochs. The learning rate is initialized to 3.4e-4 and decayed to 3.4e-6 using a cosine annealing schedule. We apply a dropout rate of 0.3 and a weight decay of 0.0001. The loss weights $\lambda_1, \lambda_2, \lambda_3$ are set to 6, 1, and 1, respectively. The label smoothing parameter ϵ is 0.3. To avoid overfitting, we adopt a data augmentation strategy where, with 0.6 probability, we replace 10% of snippets with those from other two types, using 90% of the original score as a pseudo-label. More details are in the Appendix.

4.1 Comparison with State-of-the-Art Methods

We compare MoAKE with several state-of-the-art (SOTA) AQA works, including visual-only GDLT [38], HGCN [60], T^{2}CR [15], PAMFN [53], CoFInAl [59], and ASGTN [23], and language-guided QGVL [43] and MLAVL [41]. Additionally, to evaluate an intuitive all-in-one AQA approach, we construct a strong **Conditional Baseline** (Cond-Base) featuring a shared backbone with an action classifier and per-action regression heads. Meanwhile, we also compare our model with MCMoE [45], an AQA model that recently adopted a mixture-of-experts (MoE) architecture. Since this model constructs experts for different modalities, and obtaining consistent modal information across different datasets is challenging, we evaluate MCMoE exclusively in visual scenarios and use visual features at different scales to simulate the multi-source inputs it requires. We evaluate the average performance across four action types on RG, the 'Total Element Score (TES)' closely related to visual performance on Fis-V, and the total score on LOGO. For fair comparisons, we re-implement these methods using their official codes under two settings: (1) **One-by-One**, where a separate model is trained for each dataset, representing the standard paradigm; (2)

Table 1: Comparisons on three long-term benchmarks. † indicates using textual semantics to guide visual modeling, while other methods utilize only the visual modality. The **bold**/underline is the best/second-best results. "Average" shows the mean performance across the three actions. One-by-one results are reported from the original papers and prior works [41, 61], with missing metrics reproduced by us.

Type	Methods	Rhythmic Gymnastics (RG)		Figure Skating (Fis-V)		Artistic Swimming (LOGO)		**Average**	
		SRCC ↑	R-ℓ_2 ↓	SRCC ↑	R-ℓ_2 ↓	SRCC ↑	R-ℓ_2 ↓	SRCC ↑	R-ℓ_2 ↓
One by One	GDLT [38] (CVPR'22)	0.765	2.401	0.685	3.717	0.647	4.148	0.703	3.422
	HGCN [60] (TCSVT'23)	0.735	4.341	0.246	12.628	0.671	6.564	0.583	7.844
	T²CR [15] (Inf. Sci.'24)	0.727	4.054	0.652	5.113	0.681	5.973	0.688	5.047
	PAMFN [53] (TIP'24)	0.711	4.561	0.665	7.235	0.659	13.979	0.679	8.592
	CoFInAl [59] (IJCAI'24)	**0.807**	4.028	**0.716**	2.875	0.698	4.019	0.744	3.641
	ASGTN [23] (TCSVT'25)	0.799	3.027	0.703	3.039	0.704	5.695	0.739	3.920
AinO	**MoAKE** (This Paper)	0.801	**1.822**	0.647	**2.771**	**0.790**	**2.708**	**0.753**	**2.434**
All in One	GDLT [38] (CVPR'22)	0.702	2.716	0.377	4.681	0.447	5.499	0.525	4.299
	HGCN [60] (TCSVT'23)	0.507	4.621	0.441	2.935	0.358	6.181	0.437	4.579
	T²CR [15] (Inf. Sci.'24)	0.686	3.164	0.594	3.569	0.704	20.828	0.664	9.187
	PAMFN [53] (TIP'24)	0.683	3.399	0.502	2.849	0.690	13.082	0.632	6.443
	CoFInAl [59] (IJCAI'24)	0.741	2.834	0.496	3.032	0.603	6.580	0.624	4.149
	ASGTN [23] (TCSVT'25)	0.767	3.558	0.602	3.790	0.545	4.809	0.649	4.052
	MCMoE [45] (AAAI'26)	0.771	3.426	0.625	3.608	0.763	4.816	0.726	3.950
	Cond-Base (Baseline)	0.748	2.891	0.621	3.900	0.769	4.924	0.718	3.905
	QGVL† [43] (TMM'25)	0.736	2.240	0.592	4.377	0.746	3.393	0.697	3.337
	MLAVL† [41] (CVPR'25)	0.755	2.655	0.632	2.919	0.765	5.043	0.722	3.539
	MoAKE (This Paper)	**0.801**	**1.822**	**0.647**	**2.771**	**0.790**	**2.708**	**0.753**	**2.434**

All-in-One, where a single model is trained on the combined datasets. Notably, as HGCN and ASGTN construct graph nodes based on video length, we incorporate the same segment-aware temporal aggregation to produce fixed-length features, enabling all-in-one training.

Results on All-in-One. As shown in Tab. 1, MoAKE significantly outperforms all methods in the all-in-one setting. Notably, it surpasses the strong Conditional Baseline (0.753 vs. 0.726 in Avg. SRCC), proving that simply branching at the regression head is insufficient to overcome cross-action semantic gaps; adaptive feature-level routing is essential. Furthermore, MoAKE exceeds SOTA visual-only (MCMoE) and multi-modal (MLAVL) methods by large margins, demonstrating its superior capacity for diverse actions. Remarkably, our all-in-one MoAKE even outperforms SOTA models trained in the standard one-by-one paradigm, with an average SRCC improvement of 1.2% and a remarkable 28.9% reduction in R-ℓ_2. The substantial gain in R-ℓ_2 highlights that its ability to generate highly accurate scores, not just correct rankings. It is worth noting that both QGVL and MLAVL introduce additional textual semantics to assist visual modeling, which is orthogonal to the main issue of negative transfer across scenes addressed in our work. Consequently, we expect that our method can be further improved by incorporating action knowledge texts as in [41].

An interesting observation is that naively applying existing methods to all-in-one training often results in "negative transfer," with performance dropping significantly compared to their one-by-one counterparts (e.g., an average SRCC drop of 10.17% for baselines on RG). However, in some cases, all-in-one training can bring benefits, such as PAMFN on the LOGO and HGCN on the Fis-V. This suggests that complementary knowledge exists across different action domains, which can be beneficial, especially for datasets with limited samples. This insight

Table 2: Comparisons of generalization performance on three short-term benchmarks. The **bold** and underline indicate the best and second-best results. † indicates using text semantics.

Type	Methods	Diving		Skiing, etc.		Surgery	
		SRCC↑	R-ℓ_2↓	SRCC↑	R-ℓ_2↓	SRCC↑	R-ℓ_2↓
Zero Shot	GDLT [38]	0.088	4.564	-0.024	23.615	0.133	15.917
	HGCN [60]	0.093	9.285	0.093	32.383	-0.032	30.229
	T^{2}CR [15]	0.125	6.351	-0.008	19.809	-0.017	15.249
	PAMFN [53]	0.090	6.768	0.013	28.522	0.001	24.232
	CoFInAl [59]	0.103	5.421	0.023	28.460	0.079	31.255
	ASGTN [23]	0.198	8.793	0.143	37.896	-0.055	28.787
	MCMoE [45]	0.201	6.174	0.136	30.479	0.101	17.987
	Cond-Base	0.207	5.437	0.138	27.538	0.094	19.467
	QGVL† [43]	0.205	4.798	0.133	22.307	0.114	15.387
	MLAVL† [41]	0.211	5.017	0.148	18.249	0.122	16.053
	MoAKE (Ours)	**0.319**	**3.928**	**0.210**	**12.640**	**0.219**	**12.367**
Few Shot	GDLT [38]	0.605	2.319	0.463	10.426	0.602	12.568
	HGCN [60]	0.387	2.894	0.383	12.913	0.498	15.097
	T^{2}CR [15]	0.513	3.277	0.417	13.618	0.543	16.675
	PAMFN [53]	0.517	2.913	0.405	12.579	0.581	15.703
	CoFInAl [59]	0.500	3.232	0.457	13.441	0.595	16.153
	ASGTN [23]	0.577	2.199	0.443	10.094	0.576	12.082
	MCMoE [45]	0.598	2.311	0.452	10.732	0.603	12.447
	Cond-Base	0.602	2.507	0.455	11.791	0.608	13.472
	QGVL† [43]	0.589	2.117	0.469	10.132	0.611	14.578
	MLAVL† [41]	0.618	2.244	0.458	11.247	0.620	13.047
	MoAKE (Ours)	**0.664**	**2.058**	**0.513**	**9.138**	**0.651**	**11.168**

Table 3: Transfer gains (Δ (%)) of our MoAKE from one-by-one (ObyO) to all-in-one (AinO). Metrics: SRCC↑ / R-ℓ_2↓, '+' denotes improvement.

Type	RG	Fis-V	LOGO	**Average**
ObyO	0.778 / 4.281	0.640 / 2.937	0.724 / 3.908	0.719 / 3.709
AinO	0.801 / 1.822	0.647 / 2.771	0.790 / 2.708	0.753 / 2.434
Δ (%)	+**3.0** / +**57.4**	+**1.1** / +**5.7**	+**9.1** / +**30.7**	+**4.7** / +**34.4**

Table 4: Ablations of MoAKE. The metrics are reported on the average of three datasets in the all-in-one setting. The red / blue is performance increase / decrease.

Methods	SRCC ↑	R-ℓ_2 ↓
Simple Baseline (Single Branch)	0.506	5.245
+ SATA (Single Branch)	$0.607^{\uparrow 20\%}$	$4.057^{\downarrow 23\%}$
+ SATA & AIISRM (Single Branch)	$0.655^{\uparrow 8\%}$	$3.439^{\downarrow 15\%}$
+ MoAKE w/ Specific Adapter	$0.742^{\uparrow 13\%}$	$2.628^{\downarrow 24\%}$
+ MoAKE w/ Shared Adapter (Full Model)	$\mathbf{0.753}^{\uparrow 20\%}$	$\mathbf{2.434}^{\downarrow 23\%}$
Full Model w/o AIISRM	$0.720^{\downarrow 4\%}$	$2.905^{\uparrow 19\%}$
Full Model w/o SATA & AIISRM	$0.683^{\downarrow 2\%}$	$3.109^{\uparrow 28\%}$
Full Model w/o Shared Adapter	$0.737^{\downarrow 2\%}$	$2.648^{\uparrow 9\%}$
Full Model w/o Soft Router	$0.529^{\downarrow 30\%}$	$6.081^{\uparrow 150\%}$
Full Model w/o $\mathcal{L}_{\text{route}}$	$0.730^{\downarrow 3\%}$	$2.952^{\uparrow 21\%}$
Full Model w/o $\mathcal{L}_{\text{div}}$	$0.718^{\downarrow 5\%}$	$2.888^{\uparrow 19\%}$
Full Model w/o $\mathcal{L}_{\text{route}}$ & $\mathcal{L}_{\text{div}}$	$0.706^{\downarrow 6\%}$	$3.131^{\uparrow 29\%}$

motivates our work. In contrast, as shown in Tab. 3, our MoAKE consistently shows positive transfer across all datasets, achieving an average improvement of 4.7% in SRCC and 34.4% in R-ℓ_2. This indicates that our expert-based architecture effectively captures and leverages these complementary cues.

Results on Zero/Few-Shot. Tab. 2 presents the generalization performance on unseen short-term action datasets. Zero-shot AQA is extremely challenging as the model must infer quality criteria for new actions without any labeled examples. As expected, most SOTA methods struggle, yielding unsatisfactory results. In contrast, MoAKE demonstrates significantly better generalization, outperforming prior methods by a large margin. This is because MoAKE learns more robust and generalizable representations by modeling complementary patterns from diverse long-term actions. Under the few-shot setting, MoAKE's performance improves dramatically, quickly adapting to the new action types and far exceeding the SOTA baselines. This confirms that our expert-based architecture provides a superior foundational representation for downstream adaptation.

4.2 Quantitative Analysis

Ablation Studies. To validate each component in our MoAKE framework, we conduct ablation studies on the all-in-one benchmark. We start with a **Simple Baseline** model, which consists of token projection, temporal context enhancement layer, adapter and grade-aware regression head. As shown in Tab. 4, we systematically add or remove our key modules: Segment-Aware Temporal Aggregation (SATA), Adaptive Intra- and Inter-Segment Relationship Modeling

Table 5: Comparisons of computational costs under the all-in-one setting.

Methods	GDLT	HGCN	T²CR	PAMFN	CoFInAI	ASGTN	QGVL	MLAVL	Cond-Base	**Ours**
Params (M)	3.20	4.16	0.64	18.06	5.24	6.92	5.58	3.82	3.40	8.01
FLOPs (G)	0.268	0.532	0.534	2.562	0.509	0.870	1.261	0.778	0.460	1.440
Avg. ρ (↑)	0.525	0.437	0.664	0.632	0.624	0.649	0.697	0.722	0.718	**0.753**
Avg. R-ℓ_2 (↓)	4.299	4.579	9.187	6.443	4.149	4.052	3.337	3.539	3.905	**2.434**

(AIISRM), the full Mixture of Action Knowledge Experts (MoAKE) architecture, the soft router, the shared adapter, and the loss components.

The results clearly show that each component contributes positively, with the full MoAKE model achieving the best performance. Notably, upgrading from a single-branch baseline to the MoE architecture ("MoAKE w/ Specific Adapter") yields substantial gains (13% in SRCC and 24% in R-ℓ_2). This confirms that specialized experts effectively capture discriminative patterns, and adaptively fusing their complementary knowledge is crucial for all-in-one AQA. Removing any component leads to a noticeable performance drop, confirming their necessity and effectiveness. Crucially, removing the soft router causes severe degradation (30% in SRCC and 150% in R-ℓ_2). This indicates that naively fusing expert outputs causes feature space confusion due to conflicting action patterns, a core challenge that our adaptive mixing effectively mitigates. Furthermore, even without the routing loss $\mathcal{L}_{\text{route}}$, the model still achieves 0.730 / 2.952 in two metrics. This remains superior to SOTA methods, proving that our framework does not rely on label-conditioned routing, but inherently adapts to diverse action assessment patterns.

Computational Costs. Tab. 5 compares the number of parameters, FLOPs, and average performance under the all-in-one setting. Benefiting from our tailored lightweight MoE design ($K = 3$ experts operating at the macro-action level), MoAKE achieves the best performance while maintaining highly practical computational costs. Compared to heavy multi-modal models (e.g., PAMFN with 18.06M Params/2.562G FLOPs), our MoAKE requires only 8.01M parameters and 1.44G FLOPs. This demonstrates that our framework avoids the massive parameter inflation typical of traditional MoEs in large language models, making it highly efficient and deployable for real-world all-in-one AQA.

4.3 Qualitative Analysis

Semantic Space Visualization. To qualitatively assess how different methods model the semantic space, we use t-SNE to visualize the video-level features learned by our MoAKE and baselines. As shown in Fig. 3, existing methods fail to distinguish the complex semantics across different types. The features from different action types are poorly separated and intermingled, reflecting their inability to overcome negative knowledge transfer. In stark contrast, MoAKE learns a well-structured feature space where videos of the same action type form distinct, compact clusters, while maintaining clear boundaries and appropriate

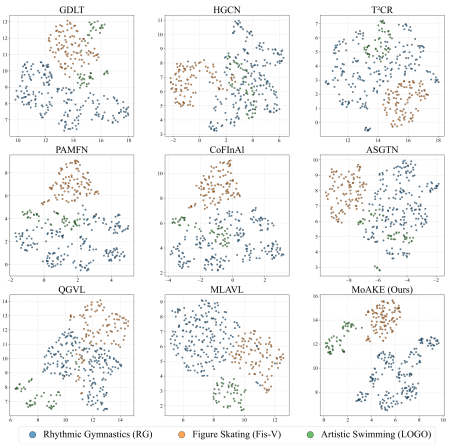

Fig. 3: t-SNE visualization of the features learned by different methods for three actions in the all-in-one setting.

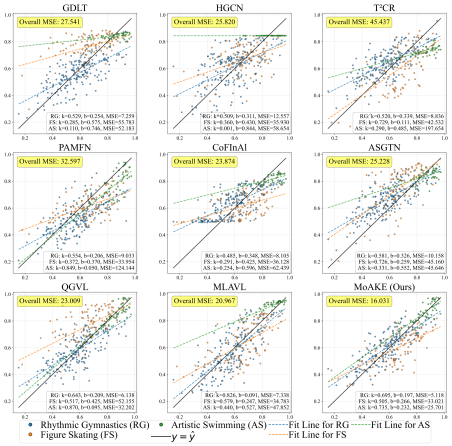

Fig. 4: Scatter plots comparing the prediction results of different methods for various actions under all-in-one AQA.

semantic distances between different types. This intuitively demonstrates that MoAKE effectively learns both shared knowledge and action-specific patterns.

Score Prediction Correlation. To further analyze prediction results, Fig. 4 plots the predicted scores against the ground-truth scores. The scatter plots reveal that MoAKE predicts more accurate scores with tighter correlations to the ground truth compared to other methods, confirming the effectiveness of our unified assessment capability.

Effectiveness of Expert Fusion. To visualize the mechanism of our soft routing, Fig. 5 (a) compares t-SNE distributions of features from individual experts and the final mixed representation. Individual experts specialize in certain action patterns, leading to overlapping clusters. In contrast, the mixed features show clearer inter-class separability. Crucially, as demonstrated by the case studies in Fig. 5 (b), while an action-aligned single expert can outperform others on its specific domain, the dynamically mixed representation consistently yields the most accurate predictions. It is important to note that during inference, our soft router relies solely on visual features without accessing any ground-truth action labels. This confirms that our framework does not simply act as a rigid, label-conditioned switch, but rather adaptively extracts and fuses complementary knowledge across all experts to achieve optimal assessment.

Fine-Grained Grade Pattern Weights. To validate our model's sensitivity to subtle execution details, we visualize the learned grade pattern weights in Fig. 6. The x-axis represents samples sorted by ground-truth scores, and the y-axis denotes progressive quality grades (1-4). A model with strong fine-grained discriminative ability should exhibit a prominent diagonal trend, indicating that higher-scoring samples correctly activate higher-grade patterns. As observed, MoAKE maintains a strict diagonal prominence not only within individual action types

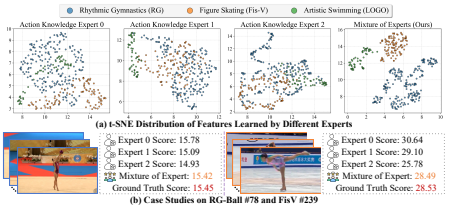

Fig. 5: t-SNE distribution of different experts and case studies.

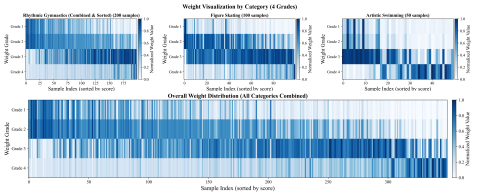

Fig. 6: Heatmap of grade pattern weights for each type (top) and all actions (bottom).

(top) but also across the highly complex all-in-one scenario (bottom). This confirms that despite bridging massive cross-action semantic gaps, our framework effectively disentangles nuanced execution differences, demonstrating exceptional fine-grained modeling capabilities for accurate quality assessment.

5 Conclusion

This paper introduces a novel framework, Mixture of Action Knowledge Experts (MoAKE), to address the challenging task of all-in-one action quality assessment. We identify the critical limitations of the prevailing "one-by-one" paradigm and the "negative transfer" issue that arises when naively training a single model on diverse actions. To mitigate this, MoAKE leverages specialized experts to learn discriminative action patterns within a shared semantic space. Complemented by dynamic routing and adaptive relationship modeling, MoAKE transforms the challenge of action diversity into an opportunity for positive knowledge transfer. To validate our approach, we establish the first comprehensive benchmarks for all-in-one and zero/few-shot AQA. Extensive experiments demonstrate that MoAKE not only sets a new state-of-the-art in all-in-one AQA but also exhibits remarkable generalization capabilities. We believe this work paves the way for developing more generic, adaptive, and practical AQA systems.

Future Work. Although our method shows superior zero-shot results, the current zero-shot scenario remains significantly challenging. Future work could explore ways to further enhance generalization in zero-shot scenarios, such as leveraging self-supervised learning methods or integrating multimodal information to enhance the model's understanding capabilities. Additionally, future work will further expand the all-in-one assessment of more action scenarios.

Acknowledgements

This work was supported by the grants from the National Key Research and Development Plan of China (2021YFB3600503), National Natural Science Foundation of China (61972097, U21A20472, 62522102, 62432001), Funds for the Innovation of Policing Science and Technology, Fujian Province (2025YZ040003, 2024YZ040001), Natural Science Foundation of Fujian Province (2025J01536), and the Beijing Natural Science Foundation (L247006).

References

1. Ai, Y., Huang, H., Zhou, X., Wang, J., He, R.: Multimodal prompt perceiver: Empower adaptiveness generalizability and fidelity for all-in-one image restoration. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 25432–25444 (2024)
2. Bai, Y., Zhou, D., Zhang, S., Wang, J., Ding, E., Guan, Y., Long, Y., Wang, J.: Action quality assessment with temporal parsing transformer. In: Eur. Conf. Comput. Vis. pp. 422–438 (2022)
3. Bruce, X., Liu, Y., Chan, K.C., Chen, C.W.: Egcn++: A new fusion strategy for ensemble learning in skeleton-based rehabilitation exercise assessment. IEEE Trans. Pattern Anal. Mach. Intell. **46**(9), 6471–6485 (2024)
4. Cai, J., Li, Q., Shen, Y., Pan, J., Liu, W.: Efficient semantic segmentation for compressed video. In: IEEE Int. Conf. Robot. Autom. pp. 4266–4272 (2024)
5. Cai, J., Su, J., Li, Q., Yang, W., Wang, S., Zhao, T., He, S., Liu, W.: Keep the balance: A parameter-efficient symmetrical framework for rgb+ x semantic segmentation. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 10587–10598 (2025)
6. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 6299–6308 (2017)
7. Fedus, W., Zoph, B., Shazeer, N.: Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. J. Mach. Learn. Res. **23**(120), 1–39 (2022)
8. Feng, Y., Yan, Z., Jia, Y., Chen, E.Q., Qin, J.: Training-free dense video captioning with large-scale pretrained models. Mach. Intell. Res. pp. 1–16 (2026)
9. Gao, Y., Vedula, S.S., Reiley, C.E., Ahmidi, N., Varadarajan, B., Lin, H.C., Tao, L., Zappella, L., Béjar, B., Yuh, D.D., et al.: Jhu-isi gesture and skill assessment working set (jigsaws): A surgical activity dataset for human motion modeling. In: Med. Image. Comput. Comput. Assist. Interv. Worksh. No. 3 (2014)
10. Hu, Y., Gao, J., Dong, J., Fan, B., Liu, H.: Exploring rich semantics for open-set action recognition. IEEE Trans. Multimedia **26**, 5410–5421 (2023)
11. Huang, C., Su, Y., Xu, H., Ke, X.: Progressive modality-adaptive interactive network for multi-modality image fusion. In: IJCAI. pp. 1161–1169 (2025)
12. Jiang, J., Chen, B., Pan, J., Wang, X., Liu, D., Jiang, J., Long, M.: Forkmerge: Mitigating negative transfer in auxiliary-task learning. In: Adv. Neural Inform. Process. Syst. pp. 30367–30389 (2023)
13. Jiang, J., Zuo, Z., Wu, G., Jiang, K., Liu, X.: A survey on all-in-one image restoration: Taxonomy, evaluation and future trends. IEEE Trans. Pattern Anal. Mach. Intell. **47**(12), 11892–11911 (2025)
14. Kayal, P., Mettes, P., Dehmamy, N., Park, M.: Large language models are natural video popularity predictors. In: Findings of the Association for Computational Linguistics. pp. 11432–11464 (2025)
15. Ke, X., Xu, H., Lin, X., Guo, W.: Two-path target-aware contrastive regression for action quality assessment. Inf. Sci. **664**, 120347 (2024)
16. Lai, X., Ke, X., Xu, H., Wu, S., Guo, W.: Msp: Multimodal self-attention prompt learning. IEEE Trans. Image Process. **34**, 5978–5988 (2025)
17. Li, B., Liu, X., Hu, P., Wu, Z., Lv, J., Peng, X.: All-in-one image restoration for unknown corruption. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 17452–17462 (2022)
18. Li, H., Ye, M., Zhang, M., Du, B.: All in one framework for multimodal re-identification in the wild. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 17459–17469 (2024)

19. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: Int. Conf. Mach. Learn. pp. 19730–19742 (2023)
20. Li, Y., Niu, Y., Xu, H., Da, H., Xu, R., Liu, W.: Ipcmoe: Integrating perceptual cues with mixture-of-experts for joint low-light image enhancement and deblurring. In: ACM Int. Conf. Multimedia. pp. 7644–7652 (2025)
21. Li, Y., Niu, Y., Xu, H., Xu, R., Hu, Y., Da, H., Liu, W., Wei, L.: Integrating perceptual cues with mixture-of-experts for low-light image restoration. Neural Networks p. 108915 (2026)
22. Liu, D., Li, Q., Jiang, T., Wang, Y., Miao, R., Shan, F., Li, Z.: Towards unified surgical skill assessment. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 9522–9531 (2021)
23. Liu, J., Wang, H., Zhou, W., Stawarz, K., Corcoran, P., Chen, Y., Liu, H.: Adaptive spatiotemporal graph transformer network for action quality assessment. IEEE Trans. Circuit Syst. Video Technol. pp. 1–1 (2025)
24. Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., Hu, H.: Video swin transformer. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 3202–3211 (2022)
25. Majeedi, A., Gajjala, V.R., GNVV, S.S.S.N., Li, Y.: Rica^ 2: Rubric-informed, calibrated assessment of actions. In: Eur. Conf. Comput. Vis. pp. 143–161 (2024)
26. NC, G., Ladi, M., Negi, S., Selvaraj, P., Kumar, P., Khapra, M.: Addressing resource scarcity across sign languages with multilingual pretraining and unified-vocabulary datasets. In: Adv. Neural Inform. Process. Syst. pp. 36202–36215 (2022)
27. Parmar, P., Morris, B.: Action quality assessment across multiple actions. In: IEEE Winter Conference on Applications of Computer Vision. pp. 1468–1476 (2019)
28. Parmar, P., Morris, B.T.: What and how well you performed? a multitask learning approach to action quality assessment. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 304–313 (2019)
29. Potlapalli, V., Zamir, S.W., Khan, S.H., Shahbaz Khan, F.: Promptir: Prompting for all-in-one image restoration. Adv. Neural Inform. Process. Syst. **36**, 71275–71293 (2023)
30. Riquelme, C., Puigcerver, J., Mustafa, B., Neumann, M., Jenatton, R., Susano Pinto, A., Keysers, D., Houlsby, N.: Scaling vision with sparse mixture of experts. In: Adv. Neural Inform. Process. Syst. pp. 8583–8595 (2021)
31. Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G., Dean, J.: Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In: Int. Conf. Learn. Represent. (2017)
32. Tang, Y., Ni, Z., Zhou, J., Zhang, D., Lu, J., Wu, Y., Zhou, J.: Uncertainty-aware score distribution learning for action quality assessment. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 9839–9848 (2020)
33. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Adv. Neural Inform. Process. Syst. pp. 5998–6008 (2017)
34. Wang, H., Fei, H., Dai, B., Gui, J.: Multimodal skeleton-based action representation learning via decomposition and composition. Mach. Intell. Res. pp. 1–13 (2026)
35. Wang, J., Ge, Y., Yan, R., Ge, Y., Lin, K.Q., Tsutsui, S., Lin, X., Cai, G., Wu, J., Shan, Y., et al.: All in one: Exploring unified video-language pre-training. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 6598–6608 (2023)
36. Wu, J., Huang, Y., Gao, M., Niu, Y., Chen, Y., Wu, Q.: Enhanced visual-semantic interaction with tailored prompts for pedestrian attribute recognition. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 9570–9579 (2025)

37. Xia, J., Zhuge, M., Geng, T., Fan, S., Wei, Y., He, Z., Zheng, F.: Skating-mixer: Long-term sport audio-visual modeling with mlps. In: AAAI. pp. 2901–2909 (2023)
38. Xu, A., Zeng, L.A., Zheng, W.S.: Likert scoring with grade decoupling for long-term action assessment. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 3232–3241 (2022)
39. Xu, C., Fu, Y., Zhang, B., Chen, Z., Jiang, Y.G., Xue, X.: Learning to score figure skating sport videos. IEEE Trans. Circuit Syst. Video Technol. **30**(12), 4578–4590 (2019)
40. Xu, H., Ke, X., Li, Y., Xu, R., Wu, H., Lin, X., Guo, W.: Vision-language action knowledge learning for semantic-aware action quality assessment. In: Eur. Conf. Comput. Vis. pp. 423–440 (2024)
41. Xu, H., Ke, X., Wu, H., Xu, R., Li, Y., Guo, W.: Language-guided audio-visual learning for long-term sports assessment. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 23967–23977 (June 2025)
42. Xu, H., Ke, X., Wu, H., Xu, R., Li, Y., Xu, P., Guo, W.: Dancefix: An exploration in group dance neatness assessment through fixing abnormal challenges of human pose. In: AAAI. pp. 8869–8877 (2025)
43. Xu, H., Wu, H., Ke, X., Li, Y., Xu, R., Guo, W.: Quality-guided vision-language learning for long-term action quality assessment. IEEE Trans. Multimedia **27**, 7326–7339 (2025)
44. Xu, H., Wu, H., Ke, X., Peng, Y.: Limssr: Llm-driven sequence-to-score reasoning under training-time incomplete multimodal observations. arXiv preprint arXiv:2605.00434 (2026)
45. Xu, H., Wu, H., Ke, X., Wu, J., Xu, R., Xu, J.: Mcmoe: Completing missing modalities with mixture of experts for incomplete multimodal action quality assessment. In: AAAI. pp. 11241–11249 (2026)
46. Xu, J., Rao, Y., Yu, X., Chen, G., Zhou, J., Lu, J.: Finediving: A fine-grained dataset for procedure-aware action quality assessment. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 2949–2958 (2022)
47. Xu, J., Rao, Y., Zhou, J., Lu, J.: Procedure-aware action quality assessment: Datasets and performance evaluation. Int. J. Comput. Vis. **132**(12), 6069–6090 (2024)
48. Xu, J., Yin, S., Peng, Y.: Human-centric fine-grained action quality assessment. IEEE Trans. Pattern Anal. Mach. Intell. **47**(8), 6242–6255 (2025)
49. Xu, J., Yin, S., Zhao, G., Wang, Z., Peng, Y.: Fineparser: A fine-grained spatio-temporal action parser for human-centric action quality assessment. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 14628–14637 (2024)
50. Yu, H., Ke, X., Zeng, Z., Xu, H., Wu, H.: Harp: Hierarchical adaptive ranking with probabilistic modeling for skill determination. In: IEEE Conf. Comput. Vis. Pattern Recog. Findings. pp. 8337–8346 (2026)
51. Yu, X., Rao, Y., Zhao, W., Lu, J., Zhou, J.: Group-aware contrastive regression for action quality assessment. In: Int. Conf. Comput. Vis. pp. 7919–7928 (2021)
52. Zeng, L.A., Hong, F.T., Zheng, W.S., Yu, Q.Z., Zeng, W., Wang, Y.W., Lai, J.H.: Hybrid dynamic-static context-aware attention network for action assessment in long videos. In: ACM Int. Conf. Multimedia. pp. 2526–2534 (2020)
53. Zeng, L., Zheng, W.: Multimodal action quality assessment. IEEE Trans. Image Process. **33**, 1600–1613 (2024)
54. Zeng, Y., Zhang, X., Li, H., Wang, J., Zhang, J., Zhou, W.: $X^2$2-vlm: All-in-one pre-trained model for vision-language tasks. IEEE Trans. Pattern Anal. Mach. Intell. **46**(5), 3156–3168 (2024)

55. Zhang, J., Lin, X., Huang, J., Ye, S., Guo, X., Zhu, D., Hu, R., Guo, D., Liang, Y., Yu, Z., et al.: Multimodal deception detection: A survey. Mach. Intell. Res. **23**(2), 284–307 (2026)
56. Zhang, S., Dai, W., Wang, S., Shen, X., Lu, J., Zhou, J., Tang, Y.: Logo: A long-form video dataset for group action quality assessment. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 2405–2414 (2023)
57. Zhou, K., Cai, R., Ma, Y., Tan, Q., Wang, X., Li, J., Shum, H.P.H., Li, F.W.B., Jin, S., Liang, X.: A video-based augmented reality system for human-in-the-loop muscle strength assessment of juvenile dermatomyositis. IEEE Trans. Vis. Comput. Graph. **29**(5), 2456–2466 (2023)
58. Zhou, K., Li, C., Pan, Q., Wang, L.: Brima: Bridged modality adaptation for multi-modal continual action quality assessment. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 38904–38914 (2026)
59. Zhou, K., Li, J., Cai, R., Wang, L., Zhang, X., Liang, X.: Cofinal: Enhancing action quality assessment with coarse-to-fine instruction alignment. In: IJCAI. pp. 1771–1779 (2024)
60. Zhou, K., Ma, Y., Shum, H.P., Liang, X.: Hierarchical graph convolutional networks for action quality assessment. IEEE Trans. Circuit Syst. Video Technol. **33**(12), 7749–7763 (2023)
61. Zhou, K., Shum, H.P.H., Li, F.W.B., Zhang, X., Liang, X.: Phi: Bridging domain shift in long-term action quality assessment via progressive hierarchical instruction. IEEE Trans. Image Process. **34**, 3718–3732 (2025)
62. Zhou, Z., Wan, Y., Wang, B.: Avatargpt: All-in-one framework for motion understanding planning generation and beyond. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 1357–1366 (2024)