# CoFInAl: Enhancing Action Quality Assessment with Coarse-to-Fine Instruction Alignment

**Kanglei Zhou[1], Junlin Li[2], Ruizhi Cai[1],**
**Liyuan Wang[3], Xingxing Zhang[3] and Xiaohui Liang[1,4]***
[1] State Key Laboratory of Virtual Reality Technology and Systems, Beihang University
[2] China Three Gorges University
[3] Department of Computer Science and Technology, Institute for AI, BNRist Center, Tsinghua-Bosch Joint ML Center, THBI Lab, Tsinghua University
[4] Zhongguancun Laboratory
liang_xiaohui@buaa.edu.cn

## Abstract

Action Quality Assessment (AQA) is pivotal for quantifying actions across domains like sports and medical care. Existing methods often rely on pre-trained backbones from large-scale action recognition datasets to boost performance on smaller AQA datasets. However, this common strategy yields suboptimal results due to the inherent struggle of these backbones to capture the subtle cues essential for AQA. Moreover, fine-tuning on smaller datasets risks overfitting. To address these issues, we propose Coarse-to-Fine Instruction Alignment (CoFInAl). Inspired by recent advances in large language model tuning, CoFInAl aligns AQA with broader pre-trained tasks by reformulating it as a coarse-to-fine classification task. Initially, it learns grade prototypes for coarse assessment and then utilizes fixed sub-grade prototypes for fine-grained assessment. This hierarchical approach mirrors the judging process, enhancing interpretability within the AQA framework. Experimental results on two long-term AQA datasets demonstrate CoFInAl achieves state-of-the-art performance with significant correlation gains of 5.49% and 3.55% on Rhythmic Gymnastics and Fis-V, respectively. Our code is available at https://github.com/ZhouKanglei/CoFInAl_AQA.

## 1 Introduction

Action Quality Assessment (AQA) aims to evaluate the quality of executed actions and is commonly used for analyzing human movements and activities [Wang *et al.*, 2021b]. Its application spans diverse domains, including sports [Liu *et al.*, 2023; Zhou *et al.*, 2023c; Zhou *et al.*, 2024], medical care [Zhou *et al.*, 2023b], surgical training [Ding *et al.*, 2023], *etc.*

The primary challenge in AQA originates in the scarcity of labeled data, resulting in the typically small size of AQA datasets. For instance, the representative MTL-AQA dataset [Parmar and Morris, 2019b] comprises only about 1,000 samples. To improve the performance of AQA, a prevalent strategy [Parmar and Morris, 2019a; Yu *et al.*, 2021; Bai *et al.*, 2022] involves leveraging the backbone pre-trained on large-scale action recognition datasets (*e.g.*, Kinetics 400 [Kay *et al.*, 2017] with over 300,000 samples) to adapt the score regression requirements of small-scale AQA datasets (see Fig. 1(a)). However, this strategy faces two critical issues. First, the **domain shift** between the pre-trained classification task of action recognition and the fine-tuned regression task of AQA renders the pre-trained model suboptimal for AQA. Second, fine-tuning the pre-trained model on small-scale AQA datasets often involves a severe risk of **overfitting**, making it difficult to bridge the domain shift.

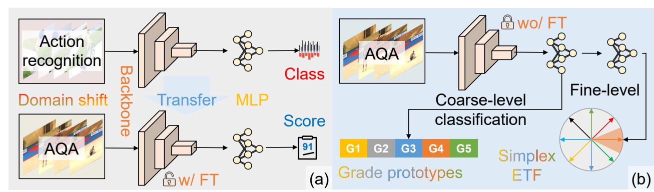


Figure 1: Motivation: (a) Previous methods often fine-tune large-scale pre-trained action recognition backbones, yielding suboptimal performance due to domain shift and overfitting. (b) Our method aligns AQA with broader tasks via coarse-to-fine instruction alignment, employing grade prototype learning and fine-grained sub-grade classification with a simplex Equiangular Tight Frame (ETF).

To overcome the double challenges of domain shift and overfitting that constrain the AQA performance, we propose here an innovative approach named Coarse-to-Fine Instruction Alignment (CoFInAl), which aligns the objectives of pre-training and fine-tuning through characterizing AQA as a coarse-to-fine classification task (see Fig. 1(b)). This coarse-to-fine process mirrors the two-step assessment taken by a judge, initially identifying a coarse grade and subsequently discerning variations within each grade.

At the coarse level, we categorize the performance of actions into different grades, such as excellent, good, or fair, representing varying levels of skill. For instance, in figure skating, a routine may include a flawless triple axel (excel-

*Corresponding author

lent), graceful spins (good), and an awkward landing (fair). To identify the grade of a given sample, we propose the Grade Parsing Module (GPM, Sect. 3.3), which dynamically captures performance criteria for the same action type through a set of grade prototypes. These learned criteria are then utilized to query each action step, resulting in score responses as coarse-grained features. Subsequently, these features are categorized using an MLP classifier to predict the coarse-grained grades. However, relying solely on coarse-level classification is insufficient for effectively distinguishing intra-grade actions, as neural collapse [Papyan *et al.*, 2020] suggests that last-layer features of the same class collapse into their intra-class mean. This motivates us to introduce a fine-grained classification task to discern variations within each grade.

At the fine level, we categorize the performance of actions within the same grade into different sub-grades. The score response associated with action details is first selected from all the score responses by masking redundant information. The selected information, serving as the fine-grained features, is then input into the Fine-Grained Scoring (FGS, see Sect. 3.4) module. Given that the number of sub-grades is typically larger than the number of grades in high-precision assessments, learning an extensive set of sub-grade prototypes might lead to overfitting. To address this concern, the FGS employs a larger set of fixed sub-grade prototypes defined by a simplex Equiangular Tight Frame (ETF) to classify the fine-grained features and predict fine-level sub-grades.

Experimental results demonstrate the significant improvements achieved by CoFInAl compared to state-of-the-art methods with notable gains of 5.49% and 3.55% in correlation on two long-term AQA datasets, Rhythmic Gymnastics and Fis-V, respectively. The effectiveness of individual designs is further validated through an extensive ablation study.

Our contributions can be summarized as:

- We identify the central challenge of current AQA methods as the domain shift and overfitting, and attribute it to the discrepancy between pre-trained action recognition tasks and fine-tuned AQA tasks.
- We propose a coherent coarse-to-fine instruction strategy that aligns AQA tasks with pre-trained models, so as to alleviate the discrepancy.
- Our approach can effectively overcome both domain shift and overfitting, and thus achieves substantial improvements across a variety of AQA benchmarks.

## 2 Related Work

**Action Quality Assessment (AQA)** aims to quantitatively evaluate the performance of executed actions across diverse domains [Zhou *et al.*, 2023c; Zhou *et al.*, 2023b; Ding *et al.*, 2023; Smith *et al.*, 2020; Wang *et al.*, 2020]. Early methods [Pirsiavash *et al.*, 2014] heavily relied on handcrafted features, revealing inherent poor generalization [Wang *et al.*, 2021a]. Recent deep learning-based methods [Parmar and Morris, 2019b; Tang *et al.*, 2020; Yu *et al.*, 2021] have demonstrated improved performance. The primary challenge stems from the relatively small size of existing AQA datasets [Wang *et al.*, 2021b], posing the risk of overfitting. To counter this challenge, pre-trained backbones are commonly employed. Parmar and Morris [2019b] leveraged C3D [Tran *et al.*, 2015] to enhance AQA performance, Pan *et al.* [2019] optimized further with I3D [Carreira and Zisserman, 2017], and Xu *et al.* [2022] explored the integration of VST [Liu *et al.*, 2022] for more powerful feature extraction. However, the challenge of domain shift persists as most AQA methods [Zhou *et al.*, 2023c; Yu *et al.*, 2021] treat the task as a regression problem, with pre-trained backbones invariably trained on action recognition. Dadashzadeh *et al.* [2024] proposed a parameter-efficient adapter to address the domain shift to some extent. However, there is still a performance gap due to potential overfitting during the fine-tuning process. In response to both domain shift and overfitting challenges, we present an innovative solution—**a novel pre-training alignment for AQA by aligning it with the pre-trained task.**

**Neural Collapse** describes an elegant geometric structure within the last-layer features and classifier of a well-trained model [Papyan *et al.*, 2020]. In a simplified model focusing solely on last-layer optimization, it has been demonstrated as the global optimality in the realm of balanced training with both cross-entropy [Graf *et al.*, 2021; Fang *et al.*, 2021; Zhu *et al.*, 2021] and MSE [Zhou *et al.*, 2022; Han *et al.*, 2021; Tirer and Bruna, 2022] loss functions. Recent investigations have extended the understanding of neural collapse to imbalanced training scenarios, either by fixing a classifier [Yang *et al.*, 2022; Zhong *et al.*, 2023] or introducing novel loss functions [Xie *et al.*, 2023]. Notably, Galanti *et al.* [2021] have demonstrated that neural collapse remains valid even when transferring a model to new samples or classes. Rather than the previous work [Yang *et al.*, 2022; Yang *et al.*, 2023], we extend an alignment application of the classifier. To the best of our knowledge, **we are the first to study AQA from the neural collapse perspective, which offers our method sound interpretability.**

## 3 Methodology

In this section, we first introduce the CoFInAl framework, and then elaborate on its core components in detail.

### 3.1 Problem Definition and Framework Overview

**Vanilla AQA.** Given a training dataset $\mathcal{D}_{\text{train}} = \{(\mathbf{x}_i, s_i)\}_{i=1}^{N}$ with $N$ action-score pairs, the goal of AQA is to learn assigning the quality score $s_i \in \mathbb{R}$ to the action $\mathbf{x}_i \in \mathbb{R}^{T\times H\times W\times 3}$ with length $T$, resolution $H \times W$, and 3 channels of video frames. A common practice [Parmar and Morris, 2019a] is to employ a pre-trained backbone as the feature extractor $f(\cdot)$ with a regressor $g(\cdot)$ to enhance the performance. This problem can be formulated as follows:

$$\begin{aligned}\min_{\theta_f,\theta_g}\ \mathcal{L}_{\text{S}} &= \frac{1}{2N}\sum_i (s_i - \hat{s}_i)^2,\\ \text{s.t. } \hat{s}_i &= g(\boldsymbol{h}_i),\ \boldsymbol{h}_i = f(\mathbf{x}_i),\end{aligned} \tag{1}$$

where $\boldsymbol{h}_i \in \mathbb{R}^D$ is the video-level feature, $\hat{s}_i$ is the predicted score, $\mathcal{L}_{\text{S}}$ represents the score regression loss function using MSE, and $\theta_f, \theta_g$ denote the network parameters.

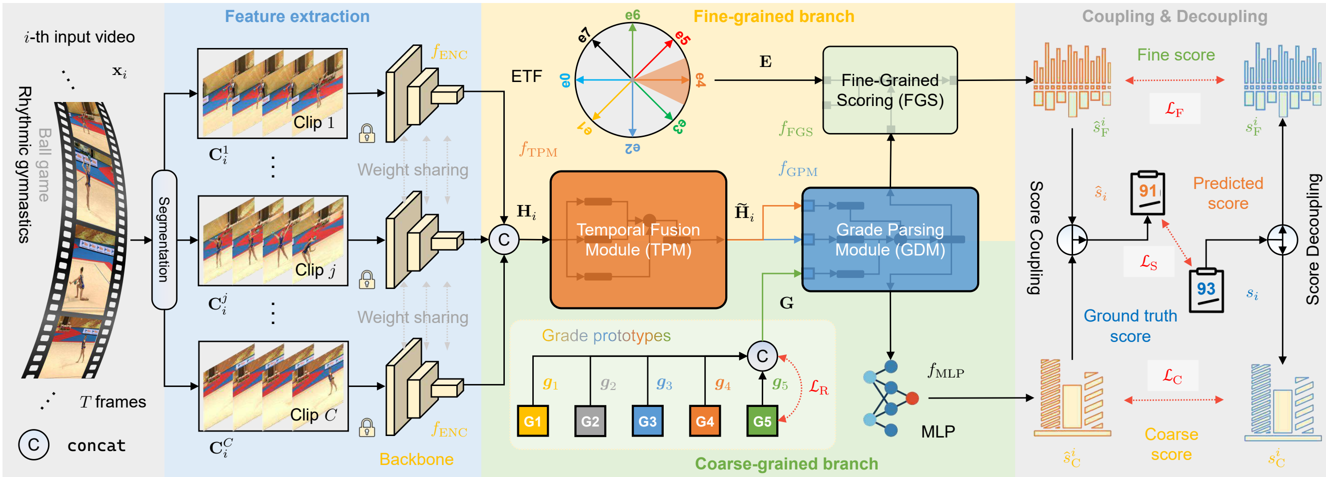


Figure 2: CoFInAl Framework: The input video undergoes segmentation into clips for feature extraction using a shared backbone. The Temporal Fusion Module (TFM, see Sect. 3.2) enhances clip features. The Grade Parsing Module (GPM, see Sect. 3.3) then separates features into coarse-grained and fine-grained components. Predictions for coarse-grained and fine-grained scores are derived from these features through an MLP and the Fine-Grained Scoring (FGS, see Sect. 3.4) module. Finally, the final score is coupled with predicted coarse-grained and fine-grained scores. During training, the ground truth score is decoupled to supervise coarse-to-fine learning.

While using pre-trained models from large-scale action recognition datasets improves small-scale AQA performance, domain shift issues persist. Existing methods, exemplified by Zhou *et al.* [2023c], often fine-tune pre-trained models for AQA adaptation, risking overfitting and suboptimal outcomes. To address this issue, we introduce CoFInAl, which strategically aligns AQA with the pre-trained task. This approach is inspired by recent advances [Zhou *et al.*, 2023a] in large language models that effectively leverage knowledge acquired during pre-training. Specifically, we reformulate AQA as a coarse-to-fine classification problem.

**Formulation of CoFInAl.** Given a type of action, we represent its quality score using $G$ coarse grades to delineate distinct performance levels, with each grade spanning a length of $S_{\mathrm{C}}$. Further, within each grade, we employ $G'$ sub-grades to capture finer variations, and each sub-grade spans a length of $S_{\mathrm{F}}$. Despite the precision error associated with this hierarchical score representation at the fine-grained interval ($S_{\mathrm{F}}$), it allows us to reframe AQA as a classification problem, aligning with the pre-trained task. Compared to directly dividing the score space into sub-grades, our approach significantly reduces space complexity from $\mathcal{O}(G \times G')$ to $\mathcal{O}(G + G')$, where $G = \lceil S/S_{\mathrm{C}} \rceil$ and $G' = \lceil S_{\mathrm{C}}/S_{\mathrm{F}} \rceil$.

Based on such a system, CoFInAl learns a coarse-to-fine instruction by optimizing the following objective:

$$\min_{\Theta} \ \mathcal{L} = \mathcal{L}_{\mathrm{S}} + \lambda_{\mathrm{C}}\mathcal{L}_{\mathrm{C}} + \lambda_{\mathrm{F}}\mathcal{L}_{\mathrm{F}} + \lambda_{\mathrm{R}}\mathcal{L}_{\mathrm{R}}, \tag{2}$$

where $\mathcal{L}_{\mathrm{C}}$ is the coarse-grained loss using the cross-entropy fuction, $\mathcal{L}_{\mathrm{F}}$ is the fine-grained loss (see Eq. (15)), $\mathcal{L}_{\mathrm{R}}$ is the regularization term (see Eq. (8)), $\lambda_{\mathrm{C}}, \lambda_{\mathrm{F}}, \lambda_{\mathrm{R}}$ serve as the loss weights, and $\Theta$ denotes the entire network parameter set.

Next, we provide a brief overview of the entire framework (see Fig. 2) to offer a high-level understanding of CoFInAl.

**Framework Overview.** In line with prior work [Xu *et al.*, 2022], we initially segment the input video $\mathbf{x}_i$ ($i = 1, 2, \cdots, N$) into $C$ clips. To accommodate computational constraints, each clip $\mathbf{c}_i^j$ ($j = 1, 2, \cdots, C$) undergoes independent processing by the backbone to extract the clip feature $\boldsymbol{h}_i^j \in \mathbb{R}^{D_{\mathrm{C}}}$. Subsequently, these clip features are enhanced to yield the enhanced clip feature $\tilde{\mathbf{H}}_i \in \mathbb{R}^{P \times D_{\mathrm{P}}}$, as follows:

$$\tilde{\mathbf{H}}_i = f_{\mathrm{TFM}}(\mathbf{H}_i), \ \mathbf{H}_i = \mathtt{concat}(\boldsymbol{h}_i^1, \boldsymbol{h}_i^2, \cdots, \boldsymbol{h}_i^C), \tag{3}$$

where $\mathtt{concat}(\cdot)$ denotes the concatenation operator, and $f_{\mathrm{TFM}}(\cdot)$ indicates the Temporal Fusion Module (TFM, see Sect. 3.2). Next, $\tilde{\mathbf{H}}_i$ is used for coarse-to-fine instruction.

In the initial coarse step, the incorporation of $G$ learnable grade prototypes $\boldsymbol{g}_1, \boldsymbol{g}_2, \cdots, \boldsymbol{g}_G \in \mathbb{R}^{D_{\mathrm{P}}}$ enables the parsing of enhanced features into two distinct components: coarse-grained and fine-grained features $\mathbf{H}_{\mathrm{C}}^i, \mathbf{H}_{\mathrm{F}}^i \in \mathbb{R}^{G \times D_{\mathrm{S}}}$, *i.e.*,

$$\mathbf{H}_{\mathrm{C}}^i, \mathbf{H}_{\mathrm{F}}^i = f_{\mathrm{GPM}}(\mathbf{G}, \tilde{\mathbf{H}}_i), \tag{4}$$

where $\mathbf{G} = \mathtt{concat}(\boldsymbol{g}_1, \boldsymbol{g}_2, \cdots, \boldsymbol{g}_G)$ is the grade prototype matrix and $f_{\mathrm{GPM}}(\cdot)$ indicates the Grade Parsing Module (GPM, see Sect. 3.3). Then, the coarse-grained feature $\mathbf{H}_{\mathrm{C}}^i$ is directly predicted to a grade through an MLP classifier $f_{\mathrm{MLP}}$, *i.e.*, $\hat{s}_{\mathrm{C}}^i = f_{\mathrm{MLP}}(\mathbf{H}_{\mathrm{C}}^i)$, providing an overarching assessment. In contrast, at the fine step, the utilization of the fine-grained features $\mathbf{H}_{\mathrm{F}}^i$ coupled with a pre-defined ETF matrix facilitates the derivation of the predicted sub-grade $\hat{s}_{\mathrm{F}}^i$, *i.e.*,

$$\hat{s}_{\mathrm{F}}^i = f_{\mathrm{FGS}}(\mathbf{H}_{\mathrm{F}}^i, \mathbf{E}), \tag{5}$$

where $f_{\mathrm{FGS}}(\cdot)$ represents the Fine-Grained Scoring (FGS, see Sect. 3.4) module. This allows for a more detailed analysis of subtle variations within a specific performance level.

Finally, we can obtain the final predicted score $\hat{s}_i$ by coupling the coarse-grained and fine-grained predictions:

$$\hat{s}_i = \hat{s}_{\mathrm{C}}^i \times S_{\mathrm{C}} + \hat{s}_{\mathrm{F}}^i \times S_{\mathrm{F}}. \tag{6}$$

During training, the ground-truth score $s_i$ is decomposed into the grade $s_{\mathrm{C}}^i = \lfloor s_i / S_{\mathrm{C}} \rfloor$ and the sub-grade $s_{\mathrm{F}}^i = \lfloor (s_i - \lfloor s_i / S_{\mathrm{C}} \rfloor) / S_{\mathrm{F}} \rfloor$ for supervision. This decomposition is instrumental in optimizing the network using Eq. (2).

### 3.2 CoFInAl: Temporal Fusion Module (TFM)

Due to the computational intensity of pre-trained backbones [Tran *et al.*, 2015], existing AQA methods [Yu *et al.*, 2021; Zhou *et al.*, 2023c] opt to segment entire video sequences for separate feature extraction. However, this compromise poses challenges for a holistic assessment. Actions typically comprise distinct procedures, each associated with key points contributing to the overall assessment. Thus, the initial segmentation may lead to inaccurate evaluations. To address this, our Temporal Fusion Module (TFM) is designed to strategically fuse temporal cues from individual clips, enabling a more meaningful representation of action procedures.

An inherent challenge in implementing TFM lies in the variability in the number of procedures, which may not always align with the number of clips. To address this, we introduce a procedure-aware self-attention mechanism designed to enhance the representation of individual clips. In the case of an action encompassing $P$ procedures, we initially derive the key, query, and value representations $\mathbf{K}_{\mathrm{T}} \in \mathbb{R}^{C\times D_{\mathrm{P}}}, \mathbf{Q}_{\mathrm{T}} \in \mathbb{R}^{C\times P\times D_{\mathrm{P}}}, \mathbf{V}_{\mathrm{T}} \in \mathbb{R}^{C\times D_{\mathrm{P}}}$ of the initial clip feature $\mathbf{H}_i$ by the linear transformation. Then, we apply average pooling to $\mathbf{Q}_{\mathrm{T}}$ to obtain the new query embedding $\tilde{\mathbf{Q}}_{\mathrm{T}} \in \mathbb{R}^{P\times D_{\mathrm{P}}}$. This enables us to rewrite Eq. (3) as follows:

$$\tilde{\mathbf{H}}_i = \mathtt{softmax}\left(\tilde{\mathbf{Q}}_{\mathrm{T}}\mathbf{K}_{\mathrm{T}}^{\top} \Big/ \sqrt{D_{\mathrm{P}}}\right)\mathbf{V}_{\mathrm{T}}, \tag{7}$$

where $\mathtt{softmax}(\cdot)$ denotes the softmax function.

**Benefits of TFM.** TFM facilitates the creation of $P$ enhanced clip features, with each clip dedicated to representing a meaningful procedure requiring assessment. In contrast to previous approaches such as the GCN-based method used in [Zhou *et al.*, 2023c], TFM adopts a novel Transformer-based encoder for the fusion process. This novel strategy enhances flexibility and expressiveness in fusing temporal information, thereby contributing to improved action assessment.

### 3.3 CoFInAl: Grade Parsing Module (GPM)

In sports assessments, judges meticulously analyze various facets of performance, such as precision, creativity, and overall execution. Similarly, our approach acknowledges the inherent complexity of action assessment and aims to replicate a two-step evaluation process. Analogous to a judge first identifying a coarse grade and then discerning variations within each grade, the Grade Parsing Module (GPM, see Fig. 3) parses the enhanced features $\tilde{\mathbf{H}}_i$ into coarse-grained and fine-grained components $\mathbf{H}_{\mathrm{C}}^i, \mathbf{H}_{\mathrm{F}}^i$, enabling our system to capture both global and detailed performance nuances.

**Graph Regularization.** To ensure that the network learns assessment facets within the same action type, we introduce $G$ grade prototypes. Each grade prototype encapsulates a standardized rule representation corresponding to a specific performance level, acting as learnable parameters for the network. To guide the learning process and align these grade prototypes with the quality space, we introduce a novel quality-aware graph regularization defined as follows:

$$\mathcal{L}_{\mathrm{R}} = \mathtt{KL}(\arccos(\bar{\mathbf{G}}^{\top}\bar{\mathbf{G}}, \mathbf{D})),\ \bar{\mathbf{G}} = \mathbf{G}/\|\mathbf{G}\|_2, \tag{8}$$

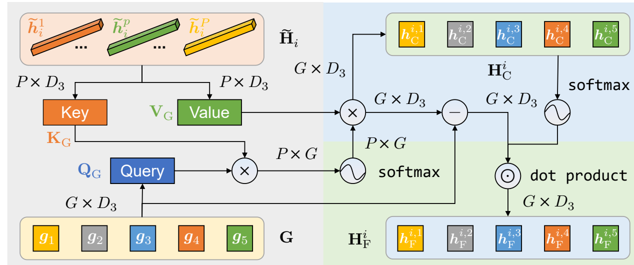


Figure 3: Illustration of Grade Parsing Module (GPM).

where $\mathtt{KL}(\cdot)$ denotes the Kullback–Leibler (KL) divergence, $\|\cdot\|_2$ represents the $\ell_2$ norm, and $\mathbf{D} \in \mathbb{R}^{G\times G}$ is the distance matrix defined on the quality space with $D_{i,j} = |i-j|$. The regularization ensures that the grade prototypes are anchored in a meaningful space, reducing the risk of overfitting and enhancing the model's generalization capabilities.

**Coarse-to-Fine Parsing.** We apply these grade prototypes to query each enhanced clip feature, aggregating the obtained scores to form the coarse-grained feature $\mathbf{H}_{\mathrm{C}}^i$. Different from TFM, this is implemented by a cross-attention mechanism:

$$\mathbf{H}_{\mathrm{C}}^i = \mathtt{softmax}\left(\mathbf{Q}_{\mathrm{G}}\mathbf{K}_{\mathrm{G}}^{\top} \Big/ \sqrt{D_{\mathrm{S}}}\right)\mathbf{V}_{\mathrm{G}}, \tag{9}$$

where $\mathbf{Q}_{\mathrm{G}} \in \mathbb{R}^{G\times D_{\mathrm{S}}}$ is the linear embedding of $\mathbf{G}$, $\mathbf{K}_{\mathrm{G}}, \mathbf{V}_{\mathrm{G}} \in \mathbb{R}^{P\times D_{\mathrm{S}}}$ are the linear embeddings of $\tilde{\mathbf{H}}_i$. Each row in $\mathbf{H}_{\mathrm{C}}^i$ represents the action's responses across different grades. By examining the distribution of these responses, a judge can make a coarse-level assessment of the action. When a finer evaluation is required, the judge can access the responses associated with the specific grade for further analysis. To execute this procedure, we begin by calculating a mask $\mathbf{M} \in \mathbb{R}^{G\times 1}$, which is:

$$\mathbf{M} = \mathtt{softmax}(\mathtt{avgPool}(\mathbf{H}_{\mathrm{C}}^i)), \tag{10}$$

where $\mathtt{avgPool}(\cdot)$ denotes the average pooling function along with the feature dimension axis. To eliminate irrelevant responses, this mask is then applied to the difference between the coarse-grained feature and the grade prototypes $\mathbf{G}$ to obtain the fine-grained feature $\mathbf{H}_{\mathrm{F}}^i$, which is:

$$\mathbf{H}_{\mathrm{F}}^i = \mathbf{M} \odot (\mathbf{H}_{\mathrm{C}}^i - \mathbf{G}), \tag{11}$$

where $\odot$ denotes the element-wise dot product operator.

**Benefits of GPM.** The rationale behind GPM is to replicate the natural two-step evaluation process of judges. This design allows us to reformulate AQA to better align with the pre-trained task. In contrast to the previous work [Xu *et al.*, 2022], the regularization of grade prototypes aligns the quality-aware space, making it more interpretable. Furthermore, compared with existing methods [Yu *et al.*, 2021; Zhou *et al.*, 2023c], the parsing of both coarse and fine features enhances the understanding of global and detailed performance cues, contributing to a holistic action assessment.

### 3.4 CoFInAl: Fine-Grained Scoring (FGS)

For coarse-grained scoring, we employ a basic MLP classifier. However, the last-layer features of the same class are susceptible to collapsing into their class mean, as indicated by neural collapse theory [Papyan *et al.*, 2020]. This collapse tends to eliminate intra-class variations, resulting in inaccurate assessments for AQA. To address this issue, we introduce the Fine-Grained Scoring (FGS) module to maintain such variations within intra-class samples.

Neural collapse theory asserts that the classifier weights on the balanced dataset will converge to a simplex Equiangular Tight Frame (ETF) matrix, defined as:

**Definition 1 (Simplex Equiangular Tight Frame)** *A simplex Equiangular Tight Frame (ETF) refers to a matrix* $\mathbf{E} = [\boldsymbol{e}_1, \boldsymbol{e}_2, \cdots, \boldsymbol{e}_K] \in \mathbb{R}^{d\times K}$*, which satisfies:*

$$\mathbf{E} = \sqrt{\frac{K}{K-1}}\mathbf{U}\left(\mathbf{I}_K - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top\right), \tag{12}$$

*where* $\mathbf{I}_K \in \mathbb{R}^{K\times K}$ *is the identity matrix,* $\mathbf{1}_K \in \mathbb{R}^{K\times 1}$ *is an all-ones vector, and* $\mathbf{U} \in \mathbb{R}^{d\times K}$ *allows a rotation and satisfies* $\mathbf{U}^\top\mathbf{U} = \mathbf{I}_K$*. All column vectors in* $\mathbf{E}$ *have the same* $\ell_2$ *norm and any pair has an inner product of* $-\frac{1}{K-1}$, i.e.,

$$\boldsymbol{e}_i^\top\boldsymbol{e}_j = \frac{K}{K-1}\delta_{ij} - \frac{1}{K-1}, \ \forall i,j \in \{1,2,\cdots,K\}, \tag{13}$$

*where* $\delta_{ij} = 0$ *when* $i \neq j$*, and 1 otherwise.*

In contrast to coarse-grained scoring, our FGS module leverages a non-learnable ETF matrix $\mathbf{E}$ with dimensions $G' \times D_{\mathrm{S}}$. The key advantages of this approach are twofold: (1) The non-learnable nature of ETF mitigates overfitting, especially considering that the number of fine-grained levels typically exceeds that of coarse-grained levels. (2) The predefined matrix embodies the optimality of a classifier. In this way, the model prediction can be simplified to the nearest class centers, *i.e.*, Eq. (5) can be rewritten as follows:

$$\hat{s}_{\mathrm{F}}^i = \arg\max_j \langle \boldsymbol{h}_{\mathrm{F}}^i, \boldsymbol{e}_j\rangle, \ \boldsymbol{h}_{\mathrm{F}}^i = \texttt{avgPool}(\mathbf{H}_{\mathrm{F}}^i), \tag{14}$$

where $\langle\cdot\rangle$ denotes the inner-product operator. Accordingly, the fine-grained loss $\mathcal{L}_{\mathrm{F}}$ in Eq. (2) can be defined as follows:

$$\mathcal{L}_{\mathrm{F}} = \frac{1}{2N}\sum_i \left(\langle \bar{\boldsymbol{h}}_{\mathrm{F}}^i, \hat{\boldsymbol{e}}_i\rangle - 1\right)^2, \tag{15}$$

where $\bar{\boldsymbol{h}}_{\mathrm{F}}^i = \boldsymbol{h}_{\mathrm{F}}^i/\|\boldsymbol{h}_{\mathrm{F}}^i\|_2$ and $\hat{\boldsymbol{e}}_i$ denotes the predicted fixed prototype in $\mathbf{E}$ with respect to the sub-grade $\hat{s}_{\mathrm{F}}^i$.

**Benefits of FGS.** The novel application of the ETF matrix in our FGS module ensures capturing fine-grained details effectively, enhancing the precision of our framework in discerning subtle variations. Importantly, in addressing domain shift, the non-learnable nature of ETF aligns with the pre-trained task, mitigating overfitting issues associated with shifts in data distribution. This strategic alignment contributes to the robustness of our model, setting it apart from existing AQA methods [Yu *et al.*, 2021; Zhou *et al.*, 2023c] and making it well-suited for real-world applications with diverse datasets.

# 4 Experiments

In this section, we first describe the experimental setup, and then present and analyze the experimental results.

## 4.1 Experimental Setups

**Datasets.** In this work, we evaluate all models on two comprehensive long-term AQA datasets. The **Rhythmic Gymnastics (RG)** dataset [Zeng *et al.*, 2020] comprises a total of 1000 videos featuring four distinct rhythmic gymnastics actions performed with various apparatuses, including ball, clubs, hoop, and ribbon. Each video has an approximate duration of 1.6 minutes, and the frame rate is set at 25 frames per second. The dataset is divided into training and evaluation sets, with 200 videos allocated for training and 50 for evaluation in each action category. The **Figure Skating Video (Fis-V)** dataset [Pirsiavash *et al.*, 2014; Parmar and Tran Morris, 2017] consists of 500 videos capturing ladies' singles short programs in figure skating. Each video has a duration of approximately 2.9 minutes, and the frame rate is set at 25 frames per second. Adhering to the official split, the dataset is divided into 400 training videos and 100 testing videos. All videos come with annotations for two scores: Total Element Score (TES) and Total Program Component Score (PCS). Following Xu *et al.* [2019], we develop independent models for different score/action types.

**Metric.** Consistent with previous studies [Yu *et al.*, 2021; Xu *et al.*, 2022], we utilize the Spearman's Rank Correlation Coefficient (SRCC) as the evaluation metric, denoted as $\rho$. The SRCC is defined as the Pearson correlation coefficient between two rank vectors, $\boldsymbol{p}$ and $\boldsymbol{q}$, with respect to predicted and ground-truth scores, which can be formulated as follows:

$$\rho = \frac{\sum_i(p_i-\bar{p})(q_i-\bar{q})}{\sqrt{\sum_i(p_i-\bar{p})^2\sum_i(q_i-\bar{q})^2}}, \tag{16}$$

where $\bar{p}$ and $\bar{q}$ denote the average values of the rank vectors $\boldsymbol{p}$ and $\boldsymbol{q}$, respectively. A higher SRCC indicates a stronger rank correlation between predicted and ground-truth scores. Following Pan *et al.* [2019], we compute the average SRCC across different action types for RG and score types for Fis-V by aggregating individual SRCCs using the Fisher's z-value.

**Implementation Details.** CoFInAl is implemented using PyTorch on a GPU for efficient parallel processing. We employ VST pre-trained on Kinetics 600 [Xu *et al.*, 2022] as the backbone and fixed for AQA. The feature dimensions $D_{\mathrm{C}}, D_{\mathrm{P}}, D_{\mathrm{S}}$ are set to 1024, 512, and 256, respectively. We initially partition the video into non-overlapping 32-frame segments. During training, we randomly determine the start segment, specifically $C = 68$ for RG and $C = 124$ for Fis-V. During testing, all segments are utilized. The number of procedures $P$ is fixed at 5 for all actions. We optimize all models using SGD with a momentum of 0.9. The batch size is 32, and the learning rate starts at 0.01, gradually decreasing to 0.0001 through a cosine annealing strategy. The number of epochs is set to 200. The loss weights $\lambda_{\mathrm{C}}, \lambda_{\mathrm{F}}, \lambda_{\mathrm{R}}$ are set to 1. To further regularize the models, we apply a dropout of 0.3/0.7 for RG/Fis-V, and the weight decay is set to 0.01.

## 4.2 Results and Analysis

We present the primary experiments here. For additional results, please refer to https://arxiv.org/abs/2404.13999.

**Comparisons with the State-of-the-Art.** We conduct a comparison of various state-of-the-art methods, including

| Method | Backbone | RG | | | | | Fis-V | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Ball | Clubs | Hoop | Ribbon | Average | TES | PCS | Average |
| C3D+SVR [Parmar and Tran Morris, 2017] | C3D | 0.357 | 0.551 | 0.495 | 0.516 | 0.483 | 0.400 | 0.590 | 0.501 |
| MS-LSTM [Xu *et al.*, 2019] | C3D | - | - | - | - | - | 0.650 | 0.780 | 0.721 |
| MS-LSTM [Xu *et al.*, 2019] | I3D | 0.515 | 0.621 | 0.540 | 0.522 | 0.551 | - | - | - |
| ACTION-NET [Zeng *et al.*, 2020] | I3D + ResNet | 0.528 | 0.652 | 0.708 | 0.578 | 0.623 | - | - | - |
| GDLT [Xu *et al.*, 2022] | I3D | 0.526 | <u>0.710</u> | <u>0.729</u> | <u>0.704</u> | <u>0.674</u> | 0.260 | 0.395 | 0.329 |
| HGCN* [Zhou *et al.*, 2023c] | I3D | <u>0.534</u> | 0.609 | 0.706 | 0.621 | 0.621 | <u>0.311</u> | <u>0.407</u> | <u>0.360</u> |
| CoFInAl (Ours) | I3D | **0.625** | **0.719** | **0.734** | **0.757** | **0.712** | **0.589** | **0.788** | **0.702** |
| MS-LSTM [Xu *et al.*, 2019] | VST | 0.621 | 0.661 | 0.670 | 0.695 | 0.663 | 0.660 | 0.809 | 0.744 |
| ACTION-NET [Zeng *et al.*, 2020] | VST + ResNet | 0.684 | 0.737 | 0.733 | <u>0.754</u> | 0.728 | <u>0.694</u> | 0.809 | 0.757 |
| GDLT [Xu *et al.*, 2022] | VST | <u>0.746</u> | <u>0.802</u> | <u>0.765</u> | 0.741 | <u>0.765</u> | 0.685 | <u>0.820</u> | <u>0.761</u> |
| HGCN* [Zhou *et al.*, 2023c] | VST | 0.711 | 0.789 | 0.728 | 0.703 | 0.735 | 0.246 | 0.221 | 0.234 |
| CoFInAl (Ours) | VST | **0.809** | **0.806** | **0.804** | **0.810** | **0.807** | **0.716** | **0.843** | **0.788** |

Table 1: Experimental results of SRCC on RG and Fis-V datasets. The best results are presented in **bold**, while the second-best results are <u>underlined</u>. The symbol $\star$ denotes our reimplementation based on the official code. The average SRCC is calculated using the Fisher-z value.

| Method | FLOPs (G) | Parameter (M) | Inference Time (ms) |
|---|---|---|---|
| ACTION-NET [Zeng *et al.*, 2020] | 34.7500 | 28.08 | 305.2474 |
| GDLT [Xu *et al.*, 2022] | 0.1164 | 3.20 | 3.2249 |
| HGCN [Zhou *et al.*, 2023c] | 1.1201 | 0.50 | 6.7830 |
| CoFInAl (Ours) | 0.1178 | 3.70 | 3.8834 |

Table 2: Computational comparison with existing methods.

| Setting | Ball | Clubs | Hoop | Ribbon |
|---|---|---|---|---|
| Ours | 0.809 | 0.806 | 0.804 | 0.810 |
| w/o $\mathcal{L}_{\mathrm{R}}$ | 0.777 $^{\downarrow 4\%}$ | 0.746 $^{\downarrow 7\%}$ | 0.786 $^{\downarrow 4\%}$ | 0.750 $^{\downarrow 7\%}$ |
| w/o TFM | 0.590 $^{\downarrow 27\%}$ | 0.705 $^{\downarrow 13\%}$ | 0.685 $^{\downarrow 15\%}$ | 0.761 $^{\downarrow 6\%}$ |
| w/o GPM | 0.332 $^{\downarrow 59\%}$ | 0.511 $^{\downarrow 36\%}$ | 0.054 $^{\downarrow 93\%}$ | 0.493 $^{\downarrow 39\%}$ |
| w/o FGS | 0.702 $^{\downarrow 13\%}$ | 0.754 $^{\downarrow 6\%}$ | 0.709 $^{\downarrow 12\%}$ | 0.757 $^{\downarrow 7\%}$ |

Table 3: Ablation results on the RG dataset.

C3D+SVR [Parmar and Tran Morris, 2017], MS-LSTM [Xu *et al.*, 2019], ACTION-NET [Zeng *et al.*, 2020], and GDLT [Xu *et al.*, 2022]. The results are reported in Tab. 1.

To demonstrate the effect of different backbones, Tab. 1 compares diverse architectures including C3D [Tran *et al.*, 2015], I3D [Carreira and Zisserman, 2017], ResNet [He *et al.*, 2016], and VST [Liu *et al.*, 2022]. I3D consistently outperforms C3D on RG, showcasing its superior ability to capture temporal dynamics. VST stands out as the most effective backbone, achieving the highest average SRCC on RG and Fis-V. For instance, the results of ACTION-NET with the VST backbone on RG showcase a notable correlation gain of over 16.85% compared to its counterpart with I3D. This underscores that VST outperforms I3D in capturing the subtle temporal dynamics crucial for AQA. Particularly, CoFInAl with I3D and VST achieves excellent results in all categories, while CoFInAl with VST achieves the best results, emphasizing the effectiveness of the proposed alignment framework and the superiority of VST. These findings underscore the significance of advanced backbones in capturing detailed action details for improved AQA performance.

CoFInAl with VST consistently outperforms others across all actions and score types in both datasets, demonstrating its effectiveness in aligning AQA with pre-trained tasks. Notably, it excels in categories like Ball, Hoop, and Ribbon on RG and across TES and PCS on Fis-V by a large margin. For example, CoFInAl achieves a remarkable 8.45% correlation gain in Ball compared to the second-best GDLT. Overall, CoFInAl delivers significant correlation gains of 5.49% and 3.55% on RG and Fis-V, respectively. These results underscore the benefits of CoFInAl in addressing domain shift, making it a promising solution for advancing AQA fields.

We have tested the model efficiency under identical conditions. The results in Tab. 2 highlight CoFInAl's significant computational efficiency compared to ACTION-NET and HGCN. Notably, our method builds upon GDLT, with a mere 0.0014G computation expansion, 0.50M parameters expenditure, and 0.6585ms inference delay. This can be attributed to coarse-to-fine alignment, which enhances performance with minimal computational overhead.

**Ablation Study.** We conduct a comprehensive ablation study to explore the individual contributions of core components within CoFInAl, presenting the results in Tab. 3. Each row in the table corresponds to a specific configuration, and the columns represent different action types on the RG dataset.

The removal of the regularization term ($\mathcal{L}_{\mathrm{R}}$) results in a slight decrease in performance across all actions. This indicates the importance of regularizing the grades to align with the quality-aware space, especially in the context of AQA with limited labeled data. Removing TFM leads to a substantial performance drop, especially in Ball and Hoop, with 27% and 15% correlation drops, respectively. This underscores the importance of capturing temporal dependencies in AQA. The absence of GPM results in severe performance degradation across all actions, with a dramatic 93% correlation drop in Hoop. This emphasizes the necessity of hierarchical processing for effective AQA, aligning with the judging process of providing coarse and fine-grained assessments. Removing FGS leads to a notable decrease in performance, especially in Ball and Hoop, with 13% and 12% correlation drops, respectively. This highlights the significance of discerning variations within grades for enhanced action assessment.

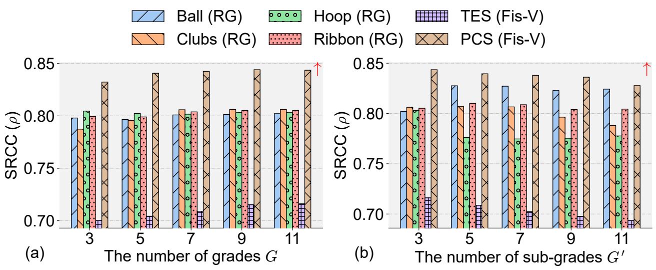


Figure 4: SRCC bars of the number of (a) grades and (b) sub-grades.

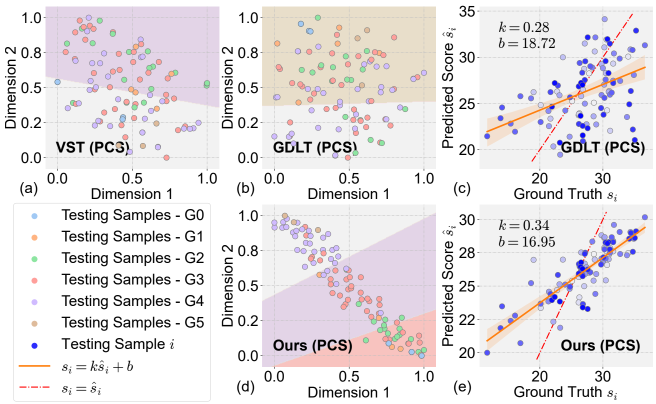


Figure 5: T-SNE feature distribution plots (a, b, d) and correlation comparison plots (c, e) contrasting GDLT with our CoFInAl method.

**Effectiveness of Grades and Sub-Grades.** We systematically varied the number of grades and sub-grades to identify the optimal parameter combination. The results in Fig. 4 indicate that excessively large or small numbers of grades lead to suboptimal outcomes. For instance, when the number of grades is set to 7, the best results are achieved in Clubs (RG). This outcome is attributed to the judicious balance achieved with four grades, ensuring a fine yet comprehensible assessment. Similar observations hold for sub-grades, reinforcing the importance of carefully selecting the number of grades and sub-grades for effective action assessment.

**Effectiveness in Addressing Domain Shifts.** We evaluate the effectiveness of our CoFInAl method in addressing domain shifts by visualizing the feature distribution in the latent space and comparing correlation plots between the predicted and ground truth scores. Leveraging the T-SNE toolbox [Van der Maaten and Hinton, 2008], we present feature and correlation plots in Fig. 5, which are conducted on PCS (Fis-V). For feature distribution plots, an SVC classifier is applied to segment the plane into different areas, enhancing visualization. Samples of the same grade (in the same color shading) occupying the same area indicate better feature distribution for action assessment. Additionally, we include a comparison with GDLT [Xu *et al.*, 2022] in Fig. 5.

Initially, we categorized samples in the entire PCS testing set into six grades, assigning grade labels from 0 to 5. Specifically, Fig. 5(a) illustrates features extracted by the VST backbone, revealing a mixed sample distribution that poses challenges in distinguishment. This highlights the suboptimal suitability of the pre-trained backbone for the AQA task due to domain shifts from broader tasks to AQA. In Fig. 5(b), we observe the feature distribution of GDLT, which appears confused, indicating inappropriate feature learning. In contrast, Fig. 5(d) demonstrates the linear feature distribution of our CoFInAl, facilitating clearer identification of samples in each grade. This underscores the superiority of CoFInAl over existing methods, attributed to the coarse-to-fine instruction aligning AQA with broader pre-trained tasks. Finally, we compare correlation plots between GDLT and CoFInAl. The regressed correlation line ($\hat{s}_i = ks_i + b$) of CoFInAl (see Fig. 5(e)) is closer to the ideal line ($\hat{s}_i = s_i$) than that of GDLT (see Fig. 5(c)), which indicates the higher correlation of CoFInAl and thus further emphasizes its effectiveness.

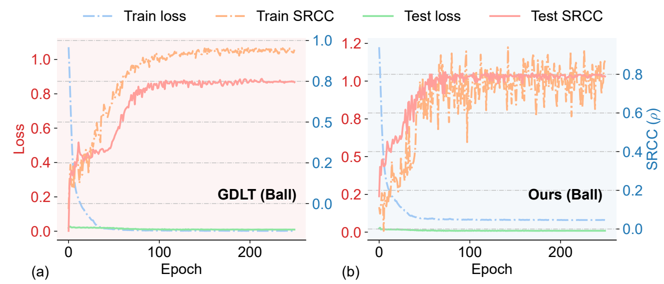


Figure 6: Comparison of loss and SRCC with GDLT on Ball (RG).

**Effectiveness in Mitigating Overfitting.** Fig. 6 provides a performance comparison in terms of loss and SRCC between GDLT [Xu *et al.*, 2022] (see Fig. 6(a)) and our proposed method (see Fig. 6(b)) on the Ball (RG) dataset. In Fig. 6(a), it is observed that during training, the training SRCC surpasses the testing SRCC after approximately 25 epochs, and the gap continues to widen, indicating a severe overfitting issue with GDLT. In contrast, the testing SRCC in Fig. 6(b) remains higher than the training SRCC, suggesting that our method avoids overfitting. These results in Fig. 6 underscore the effectiveness of our approach in mitigating overfitting issues compared to the state-of-the-art GDLT.

## 5 Conclusion

In conclusion, we introduce CoFInAl as a solution to address the key challenges of AQA. Specifically, CoFInAl integrates a two-step learning process: coarse-grained grade prototype learning and fine-grained sub-grade classification using the simplex ETF. By aligning AQA with broader pre-trained tasks, CoFInAl effectively navigates the issues of domain shift and overfitting. Experiments on RG and Fis-V demonstrate the effectiveness of CoFInAl. Its ability to effectively align downstream tasks with pre-training objectives opens avenues for enhanced model generalization and performance in broader machine learning contexts.

**Potential Limitations and Future Work.** The effectiveness of CoFInAl relies on the assumed transferability between pre-trained tasks and AQA. Challenges may arise in specialized scenarios where conflicts between pre-trained features and AQA cues impact performance. Future work should explore advanced adaptation strategies to better capture AQA cues.

## Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Project 62272019 (Liang), and also in part by the International Joint Doctoral Education Fund of Beihang University (Zhou).


## References


[Bai *et al.*, 2022] Yang Bai, Desen Zhou, Songyang Zhang, Jian Wang, Errui Ding, Yu Guan, Yang Long, and Jingdong Wang. Action quality assessment with temporal parsing transformer. In *ECCV*, pages 422–438, 2022.

[Carreira and Zisserman, 2017] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, pages 6299–6308, 2017.

[Dadashzadeh *et al.*, 2024] Amirhossein Dadashzadeh, Shuchao Duan, Alan Whone, and Majid Mirmehdi. Pecop: Parameter efficient continual pretraining for action quality assessment. In *WACV*, pages 42–52, 2024.

[Ding *et al.*, 2023] Xinpeng Ding, Xiaowei Xu, and Xiaomeng Li. Sedskill: Surgical events driven method for skill assessment from thoracoscopic surgical videos. In *MICCAI*, pages 35–45, 2023.

[Fang *et al.*, 2021] Cong Fang, Hangfeng He, Qi Long, and Weijie J Su. Exploring deep neural networks via layer-peeled model: Minority collapse in imbalanced training. *PNAS*, 118(43):e2103091118, 2021.

[Galanti *et al.*, 2021] Tomer Galanti, András György, and Marcus Hutter. On the role of neural collapse in transfer learning. *arXiv preprint arXiv:2112.15121*, 2021.

[Graf *et al.*, 2021] Florian Graf, Christoph Hofer, Marc Niethammer, and Roland Kwitt. Dissecting supervised contrastive learning. In *ICML*, pages 3821–3830, 2021.

[Han *et al.*, 2021] XY Han, Vardan Papyan, and David L Donoho. Neural collapse under mse loss: Proximity to and dynamics on the central path. *arXiv preprint arXiv:2106.02073*, 2021.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[Kay *et al.*, 2017] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[Liu *et al.*, 2022] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *CVPR*, pages 3202–3211, 2022.

[Liu *et al.*, 2023] Yanchao Liu, Xina Cheng, and Takeshi Ikenaga. A figure skating jumping dataset for replay-guided action quality assessment. In *ACM MM*, pages 2437–2445, 2023.

[Pan *et al.*, 2019] Jia-Hui Pan, Jibin Gao, and Wei-Shi Zheng. Action assessment by joint relation graphs. In *ICCV*, pages 6331–6340, 2019.

[Papyan *et al.*, 2020] Vardan Papyan, XY Han, and David L Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *PNAS*, 117(40):24652–24663, 2020.

[Parmar and Morris, 2019a] Paritosh Parmar and Brendan Morris. Action quality assessment across multiple actions. In *WACV*, pages 1468–1476, 2019.

[Parmar and Morris, 2019b] Paritosh Parmar and Brendan Tran Morris. What and how well you performed? a multitask learning approach to action quality assessment. In *CVPR*, pages 304–313, 2019.

[Parmar and Tran Morris, 2017] Paritosh Parmar and Brendan Tran Morris. Learning to score olympic events. In *CVPR workshops*, pages 20–28, 2017.

[Pirsiavash *et al.*, 2014] Hamed Pirsiavash, Carl Vondrick, and Antonio Torralba. Assessing the quality of actions. In *ECCV*, pages 556–571, 2014.

[Smith *et al.*, 2020] Eric Michael Smith, Mary Williamson, Kurt Shuster, Jason Weston, and Y-Lan Boureau. Can you put it all together: Evaluating conversational agents' ability to blend skills. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2021–2030, Online, July 2020. Association for Computational Linguistics.

[Tang *et al.*, 2020] Yansong Tang, Zanlin Ni, Jiahuan Zhou, Danyang Zhang, Jiwen Lu, Ying Wu, and Jie Zhou. Uncertainty-aware score distribution learning for action quality assessment. In *CVPR*, pages 9839–9848, 2020.

[Tirer and Bruna, 2022] Tom Tirer and Joan Bruna. Extended unconstrained features model for exploring deep neural collapse. In *ICML*, pages 21478–21505, 2022.

[Tran *et al.*, 2015] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, pages 4489–4497, 2015.

[Van der Maaten and Hinton, 2008] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[Wang *et al.*, 2020] Tianyu Wang, Yijie Wang, and Mian Li. Towards accurate and interpretable surgical skill assessment: A video-based method incorporating recognized surgical gestures and skill levels. In *MICCAI*, pages 668–678, 2020.

[Wang *et al.*, 2021a] Shunli Wang, Dingkang Yang, Peng Zhai, Chixiao Chen, and Lihua Zhang. Tsa-net: Tube self-attention network for action quality assessment. In *ACM MM*, pages 4902–4910, 2021.

[Wang *et al.*, 2021b] Shunli Wang, Dingkang Yang, Peng Zhai, Qing Yu, Tao Suo, Zhan Sun, Ka Li, and Lihua Zhang. A survey of video-based action quality assessment. In *2021 International Conference on Networking Systems of AI (INSAI)*, pages 1–9, 2021.

[Xie *et al.*, 2023] Liang Xie, Yibo Yang, Deng Cai, and Xiaofei He. Neural collapse inspired attraction–repulsionbalanced loss for imbalanced learning. *Neurocomputing*, 527:60–70, 2023.

[Xu *et al.*, 2019] Chengming Xu, Yanwei Fu, Bing Zhang, Zitian Chen, Yu-Gang Jiang, and Xiangyang Xue. Learning to score figure skating sport videos. *IEEE TCSVT*, 30(12):4578–4590, 2019.

[Xu *et al.*, 2022] Angchi Xu, Ling-An Zeng, and Wei-Shi Zheng. Likert scoring with grade decoupling for long-term action assessment. In *CVPR*, pages 3232–3241, 2022.

[Yang *et al.*, 2022] Yibo Yang, Liang Xie, Shixiang Chen, Xiangtai Li, Zhouchen Lin, and Dacheng Tao. Do we really need a learnable classifier at the end of deep neural network? *arXiv e-prints*, pages arXiv–2203, 2022.

[Yang *et al.*, 2023] Yibo Yang, Haobo Yuan, Xiangtai Li, Zhouchen Lin, Philip Torr, and Dacheng Tao. Neural collapse inspired feature-classifier alignment for few-shot class incremental learning. *arXiv preprint arXiv:2302.03004*, 2023.

[Yu *et al.*, 2021] Xumin Yu, Yongming Rao, Wenliang Zhao, Jiwen Lu, and Jie Zhou. Group-aware contrastive regression for action quality assessment. In *ICCV*, pages 7919–7928, 2021.

[Zeng *et al.*, 2020] Ling-An Zeng, Fa-Ting Hong, Wei-Shi Zheng, Qi-Zhi Yu, Wei Zeng, Yao-Wei Wang, and Jian-Huang Lai. Hybrid dynamic-static context-aware attention network for action assessment in long videos. In *ACM MM*, pages 2526–2534, 2020.

[Zhong *et al.*, 2023] Zhisheng Zhong, Jiequan Cui, Yibo Yang, Xiaoyang Wu, Xiaojuan Qi, Xiangyu Zhang, and Jiaya Jia. Understanding imbalanced semantic segmentation through neural collapse. In *CVPR*, pages 19550–19560, 2023.

[Zhou *et al.*, 2022] Jinxin Zhou, Xiao Li, Tianyu Ding, Chong You, Qing Qu, and Zhihui Zhu. On the optimization landscape of neural collapse under mse loss: Global optimality with unconstrained features. In *ICML*, pages 27179–27202, 2022.

[Zhou *et al.*, 2023a] Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. Lima: Less is more for alignment. *arXiv preprint arXiv:2305.11206*, 2023.

[Zhou *et al.*, 2023b] Kanglei Zhou, Ruizhi Cai, Yue Ma, Qingqing Tan, Xinning Wang, Jianguo Li, Hubert PH Shum, Frederick WB Li, Song Jin, and Xiaohui Liang. A video-based augmented reality system for human-in-the-loop muscle strength assessment of juvenile dermatomyositis. *IEEE TVCG*, 29(5):2456–2466, 2023.

[Zhou *et al.*, 2023c] Kanglei Zhou, Yue Ma, Hubert PH Shum, and Xiaohui Liang. Hierarchical graph convolutional networks for action quality assessment. *IEEE TCSVT*, 2023.

[Zhou *et al.*, 2024] Kanglei Zhou, Liyuan Wang, Xingxing Zhang, Hubert PH Shum, Frederick WB Li, Jianguo Li, and Xiaohui Liang. Magr: Manifold-aligned graph regularization for continual action quality assessment. *arXiv preprint arXiv:2403.04398*, 2024.

[Zhu *et al.*, 2021] Zhihui Zhu, Tianyu Ding, Jinxin Zhou, Xiao Li, Chong You, Jeremias Sulam, and Qing Qu. A geometric analysis of neural collapse with unconstrained features. *NeurIPS*, 34:29820–29834, 2021.