PHI: Bridging Domain Shift in Long-Term Action Quality Assessment via Progressive Hierarchical Instruction

Kanglei Zhou, Hubert P. H. Shum, *Senior Member, IEEE*, Frederick W. B. Li, Xinxing Zhang, and Xiaohui Liang

***Abstract*—Long-term Action Quality Assessment (AQA) aims to evaluate the quantitative performance of actions in long videos. However, existing methods face challenges due to domain shifts between the pre-trained large-scale action recognition backbones and the specific AQA task, hindering performance. This arises since fine-tuning intensive backbones on small AQA datasets is impractical. We address this by distinguishing domain shifts into task-level, regarding differences in task objectives, and feature-level, regarding differences in important features. For feature-level shifts, which are more detrimental, we propose Progressive Hierarchical Instruction (PHI) with two strategies. First, Gap Minimization Flow (GMF) leverages flow matching to progressively learn a fast flow path that reduces the domain gap between initial and desired features across shallow to deep layers. Additionally, a temporally-enhanced attention module captures long-range dependencies essential for AQA. Second, List-wise Contrastive Regularization (LCR) facilitates coarse-to-fine alignment by comprehensively comparing batch pairs to learn fine-grained cues while mitigating shift. Integrating these, PHI offers an effective solution. Experiments demonstrate that PHI achieves state-of-the-art performance on three representative long-term AQA datasets, proving its superiority in addressing the domain shift issue for long-term AQA.** [1]

***Index Terms*—Action Quality Assessment, Long-Term Action Quality Assessment, Domain Shift, Flow Matching**

I. Introduction

Action Quality Assessment (AQA) [1], [2], [3], [4], [5] aims to evaluate the quantitative performance of actions performed in videos or image sequences. Unlike traditional action recognition, which focuses solely on identifying specific actions, AQA provides a more detailed understanding of how well those actions are executed [6]. This fine-grained evaluation has broad applications across domains such as sports analysis [7], [8], [6], [9], medical rehabilitation [10], [11], and skill assessment [12], [13], [14]. Long-term AQA [15], [16] extends AQA beyond individual snapshots or short clips to encompass extended durations. This broader evaluation is more challenging and practical, as it provides a comprehensive assessment of actions in real-world scenarios.

One of the most significant challenges in long-term AQA is the domain shift issue [17], where the pre-trained backbone is suboptimal for AQA tasks. This challenge arises due to label scarcity and the nature of long video sequences. Firstly, label scarcity contributes to relatively small AQA datasets (e.g., one large-scale long-term AQA dataset for rhythmic gymnastics (balls) with approximately 250 samples). To address this, most AQA methods [18], [8], [19] often leverage backbones pre-trained on large-scale action recognition datasets (e.g.,

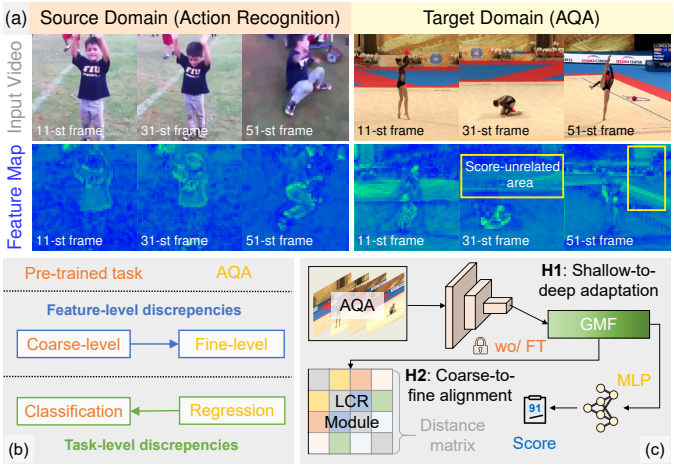


Fig. 1: Illustrations of our main idea: (a) The pre-trained I3D backbone emphasizes coarse features like guardrails (highlighted in yellow boxes), potentially unrelated to scoring for AQA, while it can accurately recognize cartwheeling in the action recognition domain. This discrepancy is primarily due to the pre-trained task's broader focus on coarse-level features, whereas fine-grained features essential for AQA may not be adequately exploited. (b) We identify two distinct domain shift types: task-level discrepancies and feature-level discrepancies. (c) Based on two hypotheses, our approach innovates a shallow-to-deep adaptation using Gap Minimization Flow (GMF), enabling a fast and controllable path to thoroughly minimize the domain gap. Additionally, we introduce a coarse-to-fine alignment mechanism using List-wise Contrastive Regularization (LCR) to enable the model to focus on fine-grained features, essential for AQA, while mitigating domain shift by refining coarse features from the broader pre-trained task.

Manuscript received July 16, 2024; revised March 10, 2025. This work was supported by the National Natural Science Foundation of China under Project 62272019. *(Corresponding author: Xiaohui Liang.)*

Kanglei Zhou is with the State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing 100191, China, and also with the Department of Computer Science, Durham University, DH1 3LE Durham, U.K. (e-mail: zhoukanglei@buaa.edu.cn).

Hubert P. H. Shum and Frederick W. B. Li are with the Department of Computer Science, Durham University, DH1 3LE Durham, U.K. (e-mail: hubert.shum@durham.ac.uk; frederick.li@durham.ac.uk).

Xingxing Zhang is with the Department of Computer Science and Technology, Institute for AI, BNRist Center, Tsinghua-Bosch Joint ML Center, THBI Lab, Tsinghua University. (e-mail: xxzhang1993@gmail.com).

Xiaohui Liang is with the State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing 100191, China, and also with the Zhongguancun Laboratory, Beijing 100190, China (e-mail: liang_xiaohui@buaa.edu.cn).

[1]Our code is included in the supplementary material for examination.

Kinetics 400 [20] with over 300,000 samples). While this strategy enhances performance on small-scale AQA datasets, its performance is restricted by the shift from action recognition to AQA tasks (see Figure 1(a)). Secondly, the computational demands of processing these long sequences within long-term AQA, combined with the complexity of intensive backbones [20], [21], make fine-tuning the backbone impractical with limited computation resources. As a result, existing long-term AQA methods [16], [15] choose to fix the backbone and do not explicitly address the domain shift issue, thereby severely limiting overall performance. Indeed, some methods [16], [6] employing feature aggregation or representation layers, such as Transformers [22], can implicitly mitigate domain shift by leveraging score supervision to encourage the model to capture high-level task-relevant patterns. However, these methods remain ineffective for accurate assessment (refer to results in Figures 7 and 8), as they lack explicit adaptation mechanisms to align domain-specific feature distributions.

Long-term AQA aims to evaluate performance based on important discriminative features, capturing subtle dynamics over extended periods where domain shift significantly hinders AQA performance. As illustrated in Figure 1(b), domain shift arises from both differences in task objectives and variations in data distribution. Accordingly, we categorize domain shift into task-level discrepancies and feature-level discrepancies. Task-level discrepancies arise from transitioning from classification-based pre-training, where class boundaries are discrete, to regression-based scoring for AQA, where scores vary continuously. Feature-level discrepancies occur due to variations in video capture conditions and application domains, leading pre-trained models to focus on coarse, irrelevant patterns rather than the fine-grained scoring cues necessary for precise AQA. The previous work [23] addressed task-level discrepancies by formulating AQA as a coarse-to-fine classification problem (see the alignment arrow "$\leftarrow$" in the bottom of Figure 1(b)), but this resulted in precision loss. Our work instead directly tackles feature-level discrepancies by refining pre-trained coarse features to focus on fine-grained cues (see "$\rightarrow$" in the top of Figure 1(b)). This refines pre-trained coarse features to emphasize the fine-grained cues essential for accurate assessment without sacrificing precision. Experimental results in Tables I to III demonstrate the superior performance of our approach over addressing only task-level discrepancies, showing the increased importance of minimizing feature-level domain shifts for accurate long-term action assessment.

To this end, we introduce the Progressive Hierarchical Instruction (PHI) framework (see Figure 1(c)) to tackle the aforementioned challenges. Built on two major hypotheses regarding shallow-to-deep adaptation and coarse-to-fine alignment, we propose solutions to validate these hypotheses, constituting the core components of PHI. These approaches collectively reduce the domain gap while prioritizing fine-grained features essential for accurate assessment.

Hypothesis 1: A shallow-to-deep adaptation approach through multiple-step control can thoroughly reduce the domain gap between a pre-trained backbone and the AQA task. The rationale behind this lies in the complexity of refining initial features in a single step, whereas the multi-step control facilitates a more progressive and precise gap reduction across shallow to deep layers (see Figure 6).

To validate Hypothesis 1, we introduce Gap Minimization Flow (GMF) to progressively reduce the domain gap, which is motivated by the recent advances in flow models [24], [25]. However, these methods face challenges in constructing training pairs due to the inaccessibility of desired features. To address this, we first integrate temporally-enhanced attention to efficiently estimate desired features, capturing crucial long-range dependencies essential for long-term AQA. This enables GMF to directly and efficiently control the gap reduction, thereby enhancing the model's adaptability.

Hypothesis 2: A coarse-to-fine alignment approach that prioritizes learning representations focusing on fine-grained cues can effectively mitigate domain shift. The rationale behind this is that the pre-trained backbones on large-scale datasets often capture coarse features irrelevant to AQA scoring, whereas AQA depends on fine-grained cues (see Figure 1(a)).

To validate Hypothesis 2, we present List-wise Contrastive Regularization (LCR) to learn the fine-grained cues essential for AQA, which is motivated by the recent advances in contrastive regression [8], [9]. However, these methods are reliant on manual exemplar selection and known exemplar score during inference, restricting the application scope. To address this, LCR comprehensively compares all pairs of batch data, eliminating the need for manual intervention. In addition, this approach ensures a robust evaluation of action quality, especially in scenarios with limited labeled data.

Experimental results demonstrate the significant improvements achieved by PHI compared to state-of-the-art methods that do not specifically address domain shift issues. Notably, PHI achieves gains of 5.88%, 5.65%, and, 24.4% in correlation on three representative long-term AQA datasets, Rhythmic Gymnastics [15], Fis-V [26], and LOGO [3], respectively, compared to shift-unaware methods. Compared to the task-level solution [23], PHI showcases additional correlation gains and significant precision improvements, demonstrating the importance of addressing feature-level discrepancies for accurate long-term AQA. Our main contributions are as follows:

- We define domain shifts from both task and feature levels. Our work achieves additional performance gains in long-term AQA by addressing feature-level discrepancies.
- We propose a novel gap minimization flow module (GMF) to address the domain gap in a shallow-to-deep manner, facilitating efficient gap-reducing control with a fast and straight path.
- We design a novel contrastive regularization module (LCR) to mitigate the domain shift in a coarse-to-fine manner, enabling robust representation learning for performance improvement.

The remainder of this paper is organized as follows: Section II reviews the related work in AQA and flow matching methods; Section III briefly describes the preliminaries in rectified flow; Section IV details the core components of our proposed framework; Section V validates and analyzes the effectiveness of our proposed method; Section VI concludes the whole paper and outlooks the future work.

II. Related Work

In this section, we provide a concise overview of AQA and flow matching, outlining their relevance to our work.

A. Action Quality Assessment (AQA)

AQA focuses on the quantitative performance of performed actions in various application areas such as sports analysis [1], [7], [8], [6], [9], [23], medical rehabilitation [10], [11], and skill assessment [12], [13], [14]. Earlier methods depended heavily on hand-crafted features and heuristics, revealing certain limitations. For example, Pirsiavash et al. [27] leveraged pose features to train a linear SVR model, which was constrained due to the poor pose estimation outcomes [28]. By integrating deep learning, various models have shown improved performance, such as CNNs [10], [6], RNNs [29], [30], and Transformers [19], [23], [31].

Existing AQA datasets [32] are relatively small, risking over-fitting. To mitigate this, pre-trained backbones are commonly employed. Parmar et al. [33] utilized C3D [34] to improve AQA performance. Pan et al. [35] integrated a more robust I3D backbone [20] to further optimize the performance. Xu et al. [16] attempted to incorporate VST [21], aiming to derive more powerful features. However, the computational intensity of such 3D backbones presents a notable issue. As every frame may contain essential AQA cues [26], videos are often segmented into clips for separate processing, which hinders a complete understanding of the action. To this end, Zhou et al. [6] introduced a GCN method to eliminate semantic ambiguities. Features extracted are typically aggregated either by LSTM [29], or more popular average pooling before the score regression. Instead of methods focusing on single-person AQA datasets, Zhang et al. [3] proposed a group-aware diving dataset. Furthermore, by leveraging the unique strengths of different data types, such as video, audio, and skeleton data, multi-modal AQA methods [36], [2], [37], [31] create a more holistic, accurate, and robust assessment system. In this work, we primarily focus on RGB-based single-person AQA.

In the context of long-term AQA, the computational intensity of these backbones, combined with the long sequences, poses significant challenges. Since fine-tuning the backbone to minimize the domain shift between the pre-trained broader task and the AQA task is difficult, existing long-term AQA methods [16], [15] often choose to fix the backbone and overlook the domain shift problem, thereby limiting performance. CoFInAl [23] addressed domain shifts by reformulating AQA into coarse and fine classification tasks, potentially leading to a loss in assessment precision. In contrast, our work identifies and tackles the key challenge of domain shift through the lens of feature-level discrepancies. Our method efficiently resolves this issue without the need for fine-tuning the computational backbone and preserving assessment precision.

B. Flow Matching

Flow Matching (FM) models represent a recent advancement in generative modeling. The term 'flow' refers to a mapping between samples of two distributions, leveraging neural Ordinary Differential Equations (ODEs) to implicitly model the transport plan between simpler and target distributions [38]. These models, such as normalizing flows [39] and continuous normalizing flows [40], have demonstrated impressive capabilities in generating high-quality samples without the need for complex approximate inference techniques [41].

Unlike traditional flow-based models, recent approaches propose training algorithms that require solving the ODE explicitly only during inference, overcoming challenges associated with backpropagating through ODEs during training [38], [25], [39], [24], [42]. FM offers a promising and little-explored avenue for generative modeling, providing a more efficient and effective approach to learning complex distributions. Different from diffusion models [43], [44] that utilize Stochastic Differential Equations (SDEs) and are restricted to Gaussian base distributions [45], FM offers greater flexibility by allowing the choice of base distribution and training with ODEs instead of SDEs, leading to smoother trajectories and improved performance [46].

Our work is inspired by the rectified flow [24] that generates a fast and direct flow path but relies on known target samples, which poses significant challenges in addressing domain shift. To overcome this limitation, we introduce a novel approach that does not depend on any known target distributions. We emphasize that our proposed PHI method is not merely an adjustment to rectified flow but a fundamental extension that introduces a novel hierarchical adaptation framework. By leveraging shallow-to-deep adaptation and coarse-to-fine alignment, PHI achieves self-supervised domain adaptation without explicit target distribution samples, making it highly generalizable to a wide range of real-world applications beyond AQA. Notably, PHI is the first attempt at integrating the concept of flow matching in the realm of AQA.

III. Preliminary: Rectified Flow

Rectified Flow (RF) [24] offers a straightforward solution for finding a transport map between two observed distributions. It involves learning an Ordinary Differential Equation (ODE), also known as the flow model, aiming to traverse straight paths as much as possible. Given empirical observations $\boldsymbol{x}_0 \sim \pi_0, \boldsymbol{x}_1 \sim \pi_1$, RF finds the transport map implicitly by solving the following ODE problem:

$$\mathrm{d}\boldsymbol{z}_t = v(\boldsymbol{z}_t, t)\mathrm{d}t, \quad t \in [0, 1], \tag{1}$$

where $v\colon \mathbb{R}^d \to \mathbb{R}^d$ represents the drift force that converts $\boldsymbol{z}_0$ from distribution π_0 to $\boldsymbol{z}_1$ following distribution π_1.

RF operates by aligning the ODE with the linear interpolation of points from π_0 and π_1. Given $\boldsymbol{x}_0$ and $\boldsymbol{x}_1$, the linear interpolation of $\boldsymbol{x}_0$ and $\boldsymbol{x}_1$ is $\boldsymbol{x}_t = t\boldsymbol{x}_1 + (1-t)\boldsymbol{x}_0$. Observe $\boldsymbol{x}_t$ follows a trivial ODE that already transfers π_0 to π_1, i.e., $\mathrm{d}\boldsymbol{x}_t = (\boldsymbol{x}_1 - \boldsymbol{x}_0)\mathrm{d}t$, where $\boldsymbol{x}_t$ moves following the line direction $(\boldsymbol{x}_1 - \boldsymbol{x}_0)$ with a constant speed. However, this ODE does not solve the problem: it can't be simulated causally, because the update $\boldsymbol{x}_t$ depends on the final state $\boldsymbol{x}_1$, which is not supposed to be known at time $t > 1$.

RF casualizes the interpolation process by projecting it to the space of causally simulatable ODEs in Equation (1). The

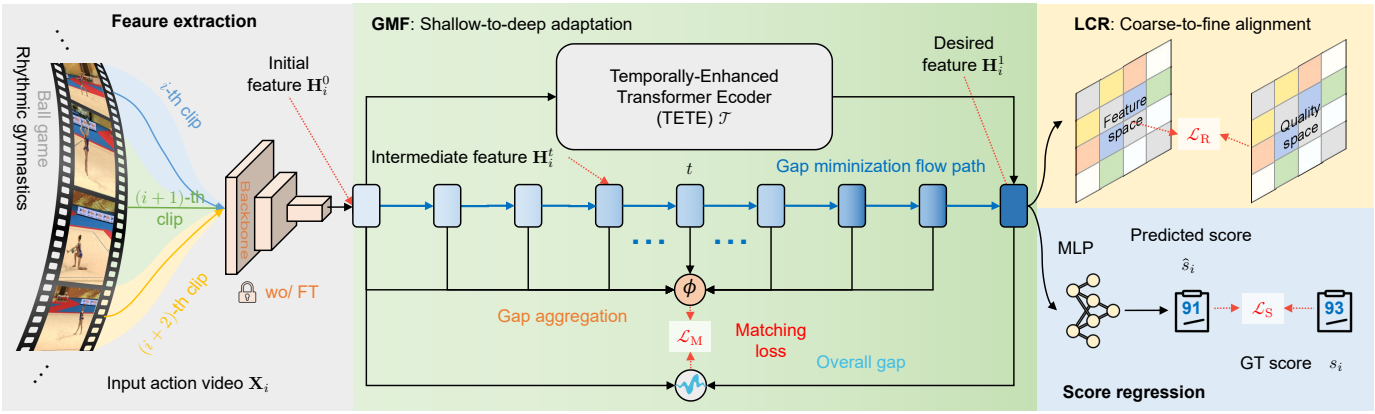


Fig. 2: Framework of PHI: Our PHI framework addresses the domain shift issue through two crucial processes. Firstly, Gap Minimization Flow (GMF) progressively transforms the initial feature into the desired one, minimizing the domain gap. Secondly, List-wise Contrastive Regularization (LCR) guides the model towards subtle variations in actions, facilitating the transition from coarse to fine-grained features crucial for AQA. Finally, the refined feature is used to predict the quality score through an MLP.

drift force v is set to drive the flow to follow the direction $(\boldsymbol{x}_1 - \boldsymbol{x}_0)$ as much as possible. A natural way to the ℓ_2 projection on the velocity field is finding v by solving a simple least squares regression problem:

$$\min_{v} \int_0^1 \mathbb{E}\left[\|(\boldsymbol{x}_1 - \boldsymbol{x}_0) - v(\boldsymbol{x}_t,\ t)\|^2\right] \mathrm{d}t. \tag{2}$$

By fitting v with the direction $(\boldsymbol{x}_1 - \boldsymbol{x}_0)$, RF casualizes the paths of linear interpolation $\boldsymbol{x}_t$, yielding an ODE flow that can be simulated without seeing the future. We can parameterize v with a neural network and solve Equation (2) with any stochastic optimizer.

While RF aims to find a transport map between two observed distributions by projecting the interpolation path onto a space of causally simulatable ODEs, our proposed Gap Minimization Flow module extends this concept to progressively minimize the domain gap between the pre-trained backbone and the target AQA task. Unlike the ODE-based formulation in RF, our approach leverages temporally-enhanced attention to efficiently estimate the desired features, allowing for a more targeted and effective domain gap reduction.

IV. Methodology: PHI

This section first introduces the PHI framework, followed by a detailed explanation of its core components.

a) Problem Definition: Long-term AQA involves assessing the quality or proficiency of actions and activities through extended video recordings, considering temporal factors such as consistency, progression, and the detailed evolution of the full performance over time, rather than just snapshots. This presents significant challenges in addressing the domain shift issue, modeling long-range temporal dependencies, and capturing fine-grained dynamics.

b) Framework Overview: Our PHI method (see Figure 2) addresses the aforementioned challenges through a holistic two-stage methodology involving shallow-to-deep adaptation and coarse-to-fine alignment. Given an action video $\mathbf{X}_i \in \mathbb{R}^{T\times W\times H\times 3}$, representing T frames of resolution $W \times H$ and 3 color channels, the initial feature $\mathbf{H}_i^0 \in \mathbb{R}^{M\times D}$ is extracted using a pre-trained backbone, where M denotes the number of clips. In the shallow-to-deep adaptation stage (see Section IV-A), Gap Minimization Flow (GMF) with temporally-enhanced attention transforms the initial feature $\mathbf{H}_i^0$ into the desired feature representation $\mathbf{H}_i^1 \in \mathbb{R}^{M\times D}$. The loss function $\mathcal{L}_{\mathrm{M}}$ (see Equation (11)) facilitates efficient adjustment of the initial feature to achieve the desired representation that better aligns with AQA. In the coarse-to-fine alignment stage (see Section IV-B), List-wise Contrastive Regularization (LCR) regularizes the feature space. The regularization loss $\mathcal{L}_{\mathrm{R}}$ (see Equation (15)) encourages the adaptation process to prioritize fine-grained features essential for AQA. Ultimately, the desired feature representation $\mathbf{H}_i^1$ is fed into an MLP head network to predict the quality score $\hat{s}_i$, which is supervised by the MSE loss $\mathcal{L}_{\mathrm{S}} = \frac{1}{2}\sum_i (s_i - \hat{s}_i)^2$. Overall, the total loss is:

$$\mathcal{L} = \mathcal{L}_{\mathrm{S}} + \lambda_{\mathrm{M}}\mathcal{L}_{\mathrm{M}} + \lambda_{\mathrm{R}}\mathcal{L}_{\mathrm{R}}, \tag{3}$$

where λ_{M} and λ_{R} are loss weights. During testing, the initial feature only undergoes the flow path to regress the score.

A. GMF: Shallow-to-Deep Adaptation

1) Design Idea:

Hypothesis 1: A shallow-to-deep adaptation approach through multiple-step control can thoroughly reduce the domain gap between a pre-trained backbone and the AQA task.

a) Justification: The challenge of mitigating the domain shift arises from the complexity of transferring complex data between initial and desired features in a single step, often compromising reliability and precision. A multi-step control mechanism involves breaking down the distribution gap into multiple sub-parts and sequentially matching these distributions. This approach enables gradual and accurate feature transformation, facilitating a thorough reduction of the domain gap across shallow to deep layers.

b) Implementation: The core idea of GMF is to utilize the concept of RF [24] to gradually reduce the domain gap, which relies on coupling pairs to train the flow path. However, in our context, we lack the corresponding desired feature for

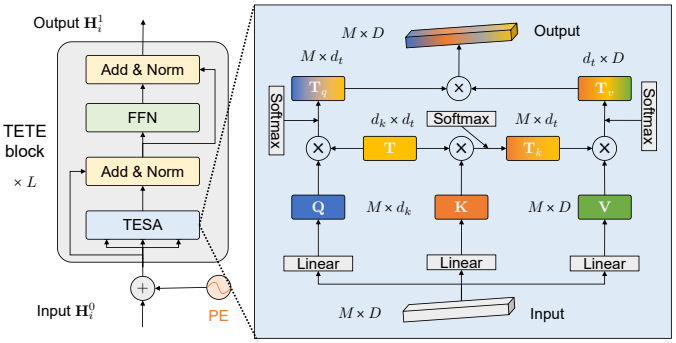


Fig. 3: Illustration of Temporally-Enhanced Transformer Encoder (TETE): Highlighting the attention of Temporally-Enhanced Self-Attention (TESA), we employ the low-rank matrix to reduce the complexity from $\mathcal{O}(M^2D)$ to $\mathcal{O}(Md_tD)$, ensuring efficient modeling of long-term dependencies.

the initial feature, posing a significant challenge in minimizing the domain gap. To address this, we first propose to estimate the desired feature and then progressively minimize the gap. Fortunately, the semantics of the desired feature are known, and represented by the score. Thus, leveraging score loss $\mathcal{L}_{\mathrm{S}}$ allows for the coarse estimation of the desired feature. Our shallow-to-deep adaptation consists of two steps: desired feature estimation and domain gap minimization.

2) Desired Feature Estimation: In long-term AQA, effectively modeling long-range dependencies in sequences is crucial. The previous work [23] employs a simple transformation matrix to aggregate temporal dependencies, which, while effective, cannot capture long-range temporal relationships. While self-attention can model the long-range dependencies, it entails substantial computational overhead. To address this, we propose Temporally-Enhanced Transformer Encoder (TETE, see Figure 3) using low-rank decomposition to optimize computation within attention. TETE can efficiently model long-range dependencies in long-term AQA by reducing computational complexity in attention mechanisms and enabling accurate estimation of desired features.

a) Vanilla Self-Attention: Considering the initial feature $\mathbf{H}_i^0 \in \mathbb{R}^{M\times D}$ of i-th action video, the vanilla self-attention [22] can be represented as:

$$\mathrm{Attention}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \mathrm{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V}, \tag{4}$$

where $\mathbf{Q}\in\mathbb{R}^{M\times d_k}, \mathbf{K}\in\mathbb{R}^{M\times d_k}, \mathbf{V}\in\mathbb{R}^{M\times D}$ are the linear embeddings of the input $\mathbf{H}_i^0$. The computation complexity of Equation (4) is $\mathcal{O}(M^2D)$, where the clip number M is typically large for long-term AQA. Thus, computation increases sharply with longer sequences in real-world scenarios.

b) Temporally-Enhanced Self-Attention (TESA): Our proposed TESA module addresses the computational challenges of long sequences by learning a low-rank matrix $\mathbf{T}\in\mathbb{R}^{d_t\times d_k}$, where d_t is significantly smaller than the feature dimension D, ensuring computational efficiency while preserving temporal dependencies. First, TESA applies $\mathbf{T}$ only to the query $\mathbf{Q}$ and the key $\mathbf{K}$, as they directly determine the attention weights. This enhances the ability of the model to capture long-term dependencies while avoiding redundant transformations on the value $\mathbf{V}$, which does not influence attention weight computation. The transformed queries and keys are computed as:

$$\mathbf{T}_q = \mathrm{Softmax}\left(\mathbf{Q}\mathbf{T}^\top\right)\in\mathbb{R}^{M\times d_t}, \tag{5}$$
$$\mathbf{T}_k = \mathrm{Softmax}(\mathbf{K}\mathbf{T}^\top)\in\mathbb{R}^{M\times d_t}. \tag{6}$$

Next, we use $\mathbf{T}_k$ to query the value $\mathbf{V}$, yielding:

$$\mathbf{T}_v = \mathrm{Softmax}(\mathbf{T}_k\mathbf{V})\in\mathbb{R}^{d_t\times D}. \tag{7}$$

Finally, our attention mechanism can be represented as:

$$\mathrm{Attention}\left(\mathbf{Q},\mathbf{K},\mathbf{V},\mathbf{T}\right) = \mathbf{T}_q\mathbf{T}_v\in\mathbb{R}^{M\times D}. \tag{8}$$

This results in a decent complexity reduction to $\mathcal{O}(Md_tD)$, effectively improving the training efficiency and mitigating overfitting, thus leading to performance improvement (refer to results in Table V).

Finally, the desired feature can be estimated as:

$$\mathbf{H}_i^1 = \mathcal{T}\left(\mathbf{H}_i^0\right), \tag{9}$$

where $\mathcal{T}(\cdot)$ represents the TETE module.

3) Domain Gap Minimization: Inspired by the concept of RF [24], our work aims to learn a fast and controllable path to efficiently mitigate domain shift by progressively reducing the domain gap between initial and desired features.

Suppose the initial feature $\mathbf{H}_i^0$ and the desired feature $\mathbf{H}_i^1$ follow two different empirical distributions, and we sample P intermediate steps. The target representation of the j-th step can be defined as the linear interpolation $\mathbf{H}_i^{j/P} = \frac{j}{P}\mathbf{H}_i^0 + (1-\frac{j}{P})\mathbf{H}_i^1$. During training, our objective is to predict the next step using the previous representation and the step size through a neural network $\phi(\cdot,\cdot)$, indicating the domain gap of the corresponding step. In practice, our experiments demonstrate that a simple MLP is sufficient for this task. The gap at the j-th step can be represented as:

$$\boldsymbol{g}_j = \phi\left(\hat{\mathbf{H}}_i^{(j-1)/P}, \frac{1}{P}\right), \quad j\geq 1, \tag{10}$$

where $\hat{\mathbf{H}}_i^{(j-1)/P}$ denotes the predicted feature of the previous step, calculated as $\boldsymbol{g}_{j-1} + \hat{\mathbf{H}}_i^{(j-2)/P}$ if $j\geq 2$, otherwise $\mathbf{H}_i^0$. The network ϕ generates a driving force that guides the flow in the direction of the overall flow, and we optimize ϕ by:

$$\mathcal{L}_{\mathrm{M}} = \frac{1}{B}\sum_{i=0}^{B-1}\left(\mathcal{L}_{\mathrm{M-global}} + \mathcal{L}_{\mathrm{M-local}}\right), \tag{11}$$

$$\mathcal{L}_{\mathrm{M-global}} = \left\|\left(\mathbf{H}_i^1 - \mathbf{H}_i^0\right) - \sum_{j=1}^{P}\boldsymbol{g}_j\right\|^2, \tag{12}$$

$$\mathcal{L}_{\mathrm{M-local}} = \frac{1}{P}\sum_{j=1}^{P}\left\|\mathbf{H}_i^{j/P} - \hat{\mathbf{H}}_i^{j/P}\right\|^2, \tag{13}$$

where B denotes the batch size, and $\left(\mathbf{H}_i^1 - \mathbf{H}_i^0\right)$ represents the overall flow direction. The first term $\mathcal{L}_{\mathrm{M-global}}$ represents the global flow constraint, while the second term $\mathcal{L}_{\mathrm{M-local}}$ denotes the local flow constraint. These two terms are collectively effective in addressing different aspects of the flow constraints.

4) Benefits of GMF: GMF introduces a fundamentally new perspective on feature transformation in flow-based adaptation, offering several key advantages over existing methods. Unlike traditional flow methods [25], [24], which require explicit access to the target distribution, GMF eliminates this dependency by leveraging an adaptive hierarchical transformation strategy. This enables GMF to handle domain-shifted tasks where the target distribution is either unknown or difficult to obtain, significantly extending the applicability of flow-based learning. Instead of directly aligning source and target distributions, GMF sequentially matches intermediate representations through a multi-stage adaptation process. This hierarchical alignment ensures a smooth and stable transition while preventing abrupt transformations. Unlike stacked or recurrent architectures, GMF dynamically regulates the adaptation trajectory, reducing computational overhead and improving convergence efficiency. Additionally, its implicit knowledge distillation mechanism allows for lightweight inference without compromising performance. These innovations make GMF a scalable and efficient solution for domain-adaptive learning in video-based tasks, surpassing conventional flow-based approaches.

B. LCR: Coarse-to-Fine Alignment

1) Design Idea:

Hypothesis 2: A coarse-to-fine alignment approach that prioritizes learning representations focusing on fine-grained cues can effectively mitigate the domain shift issue in long-term AQA.

a) Justification: As shown in Figure 1(a), the pre-trained broader task often emphasizes coarse-level features that may not directly align with the requirements of AQA. Conversely, AQA tasks rely heavily on fine-grained cues [47], [48] for accurate assessment. Therefore, a method that prioritizes learning these fine-grained representations is likely to be more effective in addressing the domain shift issue.

b) Implementation: Notably, the interpretation of coarse-to-fine alignment differs from CoFInAl [23]. In CoFInAl, coarse-to-fine alignment refers to structuring AQA as a two-stage classification task, which helps address task-level differences but may introduce precision loss. In contrast, PHI adopts a feature refinement approach that progressively enhances pre-trained coarse features to focus on fine-grained scoring cues, ensuring improved AQA performance. In response, our novel LCR module (see Figure 4) employs a coarse-to-fine alignment strategy to further address the domain shift, enhancing the model's ability to capture fine-grained features essential for AQA. By comparing the differences between actions, comparative regression aids in identifying subtle variations or abnormalities that may not be apparent when assessing actions in isolation. Given a batch of data, LCR involves computing a distance matrix where each row encodes the relationships of an action with all other actions in the batch. By aligning the distribution of each row with its ground truth quality score distribution, the model can effectively learn subtle variations.

2) Action Distance Computation: For each action in the batch, we need to calculate its pairwise distance with all other

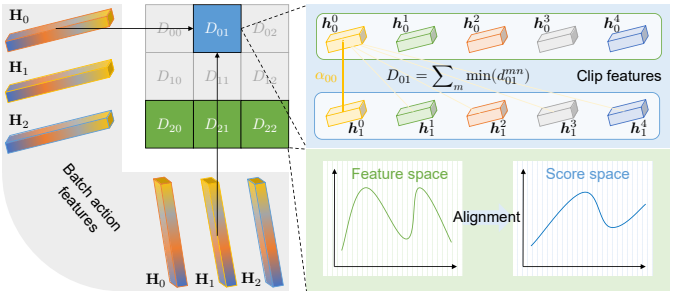


Fig. 4: Illustration of List-wise Contrastive Regularization (LCR): LCR aligns the distance distribution in the feature space with that of the quality score space in a list-wise manner. This ensures comprehensive comparison and alignment across the entire batch of data, leading to robust performance.

actions. However, directly measuring the distance between two actions is inappropriate because their clips may not be temporally or semantically aligned. For instance, clips with similar semantics might appear at different indices in each action sequence. Hence, we need to establish correspondence between clips before computing the action-level distance. (1) To align clips in two actions, For each clip in the first action, we compute its distance to all clips in the second action and select the closest one as its paired clip. Specifically, the pairing is determined based on the minimum ℓ_2 distance. (2) Then, we sum all the clip pair distances to obtain the final distance between the two actions. This process can be represented as:

$$D_{ij} = \sum_{m=0}^{M-1} \left(\min_{n \in [0, M-1]} (d_{ij}^{mn}) \right), \; d_{ij}^{mn} = \|\boldsymbol{h}_i^m - \boldsymbol{h}_j^n\|^2, \quad (14)$$

where d_{ij}^{mn} denotes the distance between the m-th clip of the i-th action video and the n-th clip of the j-th action video using the ℓ_2 distance.

It is interesting to note the implicit temporal order relationship between the paired clip distance calculation. In practice, when considering two clips $\boldsymbol{h}_i^{m_1}$ and $\boldsymbol{h}_i^{m_2}$ $(m_1 \leq m_2)$ from the same video, where m_1 and m_2 represent different time steps, and their paired clips $\boldsymbol{h}_j^{n_1}$ and $\boldsymbol{h}_j^{n_2}$, it is observed that upon convergence, n_1 tends to be less than or equal to n_2. This observation aligns with the inherent temporal ordering within action sequences, where paired clips that are closer in time tend to have smaller indices than those further apart. Therefore, when we added a loss item to constrain $n_1 \leq n_2$, the performance did not change.

3) Distance Matrix Alignment: Aligning the distribution of each row in the distance matrix with the ground truth quality score distribution involves two main steps. (1) We calculate the ground truth quality score distance using $\mathbf{S} = |\boldsymbol{s} - \boldsymbol{s}^\top|$, where $\boldsymbol{s} \in \mathbb{R}^{B \times 1}$ denotes the quality score of the batch data (B is the batch size). (2) Optimizing the alignment process can be challenging, as traditional metrics like MSE may not accurately capture the underlying structure of the data (see results in Table V). To address this issue, we propose incorporating Kullback-Leibler (KL) divergence [49] to enhance the alignment process and better capture the complex relationships

within the data distributions. This process can be optimized by:

$$\mathcal{L}_{\mathrm{R}}=\sum_{i=0}^{B-1}\left(\mathrm{KL}(\boldsymbol{S}_i\|\boldsymbol{D}_i)+\mathrm{KL}(\boldsymbol{D}_i\|\boldsymbol{S}_i)\right). \tag{15}$$

Here, the adoption of a symmetrized and smoothed version of the KL divergence enhances the robustness and stability of the alignment process, leading to accurate assessment. The estimated target features generated by TETE (supervised by $\mathcal{L}_S$) may not fully eliminate the domain shift. Therefore, GMF aligns initial features with the estimated ones and then re-aligns these estimated features with the ideal ones essential for AQA. LCR (with $\mathcal{L}_R$) regularizes this re-alignment, enhancing overall performance.

4) Benefits of LCR: LCR introduces a novel hierarchical alignment strategy that significantly enhances AQA performance by capturing subtle relationships between actions within a batch. Unlike traditional pair-wise contrastive methods [50], [8], [47], LCR employs batch-wise contrastive learning to explicitly consider all pair-wise differences, enabling a more effective capture of fine-grained variations. The loss function $\mathcal{L}_{\mathrm{R}}$ leverages KL divergence for efficient alignment between distributions $\boldsymbol{D}_i$ and $\boldsymbol{S}_i$, ensuring faster convergence and better generalization compared to direct alignment methods like MSE. Additionally, low-rank regularization enforces temporal coherence and robustness, preventing overfitting to spurious correlations. The batch-wise learning and low-rank constraints collectively establish LCR as a theoretically unique and generalizable framework for domain adaptation, applicable not only to AQA but also to tasks like rehabilitation analysis, sports motion scoring, and action recognition. By addressing the limitations of prior methods, LCR represents a significant advancement in feature alignment for temporal and domain-shifted tasks.

V. Experiments

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

A. Experimental Setups

a) Datasets: In our study, we assess all models using three extensive long-term AQA datasets. The first dataset, named the **Rhythmic Gymnastics (RG)** dataset [15], comprises a collection of 1,000 videos showcasing various rhythmic gymnastics actions performed with different 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 split into training and evaluation sets, with 200 videos allocated for training and 50 for evaluation in each action category. The second dataset, referred to as the **Figure Skating Video (Fis-V)** dataset [27], [26], contains 500 videos capturing ladies' singles short programs in figure skating. Each video has a duration of approximately 2.9 minutes, with a frame rate set at 25 frames per second. Following the official split, the dataset is divided into 400 training videos and 100 testing videos. Each video in this dataset comes with annotations for two scores: Total Element Score (TES) and Total Program Component Score (PCS). To align with the previous method [29], we develop separate models for different score/action types. The third dataset, i.e., the **LOng-form GrOup (LOGO)** dataset [3], consists of 150 samples for training and 50 for testing. It captures videos showcasing synchronized swimming group actions, where each video sequence is approximately 3 and a half minutes in length. Currently, LOGO has the longest video duration among all AQA datasets, making it a challenging benchmark for long-term AQA tasks.

b) Evaluation Metrics: We use two evaluation metrics to validate the performance of all the AQA methods.

Consistent with previous long-term AQA methods [16], [15], we utilize Spearman's Rank Correlation Coefficient (SRCC) as the evaluation metric, denoted as ρ. The SRCC is defined as the Pearson correlation coefficient between their ranks, $r(s_i)$ and $r(\hat{s}_i)$, to predicted and ground-truth scores, which can be formulated as follows:

$$\rho=\frac{\sum_{i=1}^{N}\left(r\left(s_i\right)-\bar{r}\right)\left(r\left(\hat{s}_i\right)-\bar{r}\right)}{\sqrt{\sum_{i=1}^{N}\left(r\left(s_i\right)-\bar{r}\right)^2}\sqrt{\sum_{i=1}^{N}\left(r\left(\hat{s}_i\right)-\bar{r}\right)^2}}, \tag{16}$$

where $\bar{r}$ is the average rank. A higher SRCC indicates a stronger rank correlation between predicted and ground-truth scores. Following the previous work [35], we compute the average SRCC across different action types for the RG dataset and score types for the Fis-V dataset by aggregating individual SRCCs using Fisher's z-value.

Compared with the previous work [23], we have added a new stricter metric, the relative ℓ_2 distance (R-ℓ_2) [8], [6]. The purpose of introducing R-ℓ_2 is to measure the relative error of AQA models more precisely without being affected by the score scale. Given the highest and lowest scores for an action $s_{\max}$ and $s_{\min}$, the relative ℓ_2 distance R-ℓ_2 is defined as:

$$\text{R-}\ell_2=\frac{1}{N}\sum_{n}^{N}\left(\frac{|s_n-\hat{s}_n|}{s_{\max}-s_{\min}}\right)^2\times 100, \tag{17}$$

where s_n and $\hat{s}_n$ represent the ground-truth score and prediction for the n-th sample, respectively. Fisher's z-value is used to measure the average performance across actions.

c) Implementation Details: We implemented PHI using PyTorch on an RTX 3090 GPU. We employ VST pre-trained on Kinetics 600 [16] and I3D on Kinetics 400 [20] as the backbone to conduct experiments, respectively. The feature dimensions D, d_k, d_t are set to 1024, 128, and 32, respectively. Following the previous work [16], [3], we initially partition the video into non-overlapping 32-frame segments. During training, we randomly determine the start segment, specifically $M = 68$ for RG, $M = 124$ for Fis-V, and 48 for LOGO, respectively. During testing, all segments are utilized. We optimize all models using SGD with a momentum of 0.9. The batch size B is 32, and the learning rate starts at 0.01, gradually decreasing to 0.0001 through a cosine annealing strategy. For convergence, all the models are trained for 200 epochs. The loss weights $\lambda_{\mathrm{M}}, \lambda_{\mathrm{R}}$ are set to 0.5 and 0.01, respectively. To further optimize the networks, we apply a dropout of 0.3 and a weight decay of 0.01.

TABLE I
RESULTS OF SRCC (↑) AND R-ℓ_2 (↓) ON THE RG DATASET. THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**, WHILE THE SECOND-BEST RESULTS ARE UNDERLINED. THE SYMBOL "⋆" INDICATES OUR REIMPLEMENTATION BASED ON THE OFFICIAL CODE. THE AVERAGE SRCC IS COMPUTED USING THE FISHER-Z VALUE. "–" DENOTES THAT THE METHOD DOES NOT REPORT THIS METRIC. "+" INDICATES THE USE OF ADDITIONAL FEATURES/MODALITIES.

Method	Publisher	Backbone	Shift Aware	Modality	Ball		Clubs		Hoop		Ribbon		Average	
					SRCC	R-ℓ_2	SRCC	R-ℓ_2	SRCC	R-ℓ_2	SRCC	R-ℓ_2	SRCC	R-ℓ_2
C3D+SVR [26]	TPAMI'17	C3D	×	RGB	0.357	–	0.551	–	0.495	–	0.516	–	0.483	–
MS-LSTM [29]	TCSVT'19	I3D	×	RGB	0.515	–	0.621	–	0.540	–	0.522	–	0.551	–
ACTION-NET [15]	ACM MM'20	I3D+	×	RGB	0.528	–	0.652	–	0.708	–	0.578	–	0.623	–
GDLT [16]	CVPR'22	I3D	×	RGB	0.526	2.943	0.710	**2.557**	0.729	8.149	0.704	3.485	0.674	4.284
HGCN⋆ [6]	TCSVT'23	I3D	×	RGB	0.534	6.748	0.609	16.142	0.706	9.270	0.621	9.934	0.621	10.524
VATP-Net [51]	TCSVT'24	I3D	×	RGB+	0.580	–	0.720	–	0.739	–	0.724	–	0.709	–
CoFInAl [23]	IJCAI'24	I3D	✓	RGB	**0.625**	**2.647**	0.719	3.093	**0.734**	**3.892**	**0.757**	**2.607**	**0.712**	**3.060**
PHI (Ours)	–	I3D	✓	RGB	0.598	3.471	**0.732**	3.139	0.731	5.376	0.754	5.674	0.708	4.415
MS-LSTM [29]	TCSVT'19	VST	×	RGB	0.621	–	0.661	–	0.670	–	0.695	–	0.663	–
ACTION-NET [15]	ACM MM'20	VST+	×	RGB	0.684	–	0.737	–	0.733	–	0.754	–	0.728	–
GDLT [16]	CVPR'22	VST	×	RGB	0.746	2.833	0.802	2.179	0.765	**2.012**	0.741	2.579	0.765	2.401
HGCN⋆ [6]	TCSVT'23	VST	×	RGB	0.711	3.030	0.789	3.444	0.728	5.312	0.703	5.576	0.735	4.341
PAMFN [31]	TIP'24	VST	×	RGB	0.636	–	0.720	–	0.769	–	0.708	–	0.711	–
VATP-Net [51]	TCSVT'24	VST	×	RGB+	0.800	–	0.810	–	0.780	–	0.769	–	0.800	–
CoFInAl [23]	IJCAI'24	VST	✓	RGB	0.809	**1.356**	**0.806**	2.453	0.804	9.918	**0.810**	**2.383**	0.807	4.028
PHI (Ours)	–	VST	✓	RGB	**0.818**	2.187	0.803	**2.149**	**0.812**	2.119	0.805	2.744	**0.810**	**2.300**

TABLE II
RESULTS OF SRCC (↑) AND R-ℓ_2 (↓) ON THE FIS-V DATASET. THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**, WHILE THE SECOND-BEST RESULTS ARE UNDERLINED. THE SYMBOL "⋆" INDICATES OUR REIMPLEMENTATION BASED ON THE OFFICIAL CODE. THE AVERAGE SRCC IS COMPUTED USING THE FISHER-Z VALUE. "–" DENOTES THAT THE METHOD DOES NOT REPORT THIS METRIC. "+" INDICATES THE USE OF ADDITIONAL FEATURES/MODALITIES.

Method	Publisher	Backbone	Shift Aware	Modality	TES		PCS		Average	
					SRCC	R-ℓ_2	SRCC	R-ℓ_2	SRCC	R-ℓ_2
C3D+SVR [26]	TPAMI'17	C3D	×	RGB	0.400	–	0.590	–	0.501	–
MS-LSTM [29]	TCSVT'19	C3D	×	RGB	0.650	–	0.780	–	0.721	–
GDLT [16]	CVPR'22	I3D	×	RGB	0.260	5.582	0.395	5.039	0.329	5.311
HGCN⋆ [6]	TCSVT'23	I3D	×	RGB	0.311	4.317	0.407	4.608	0.360	4.463
CoFInAl [23]	IJCAI'24	I3D	✓	RGB	0.589	3.470	0.788	**2.843**	0.702	3.157
PHI (Ours)	–	I3D	✓	RGB	**0.659**	**2.572**	**0.798**	3.073	**0.736**	**2.823**
MS-LSTM [29]	TCSVT'19	VST	×	RGB	0.660	–	0.809	–	0.744	–
ACTION-NET [15]	ACM MM'20	VST+	×	RGB	0.694	–	0.809	–	0.757	–
GDLT [16]	CVPR'22	VST	×	RGB	0.685	3.717	0.820	2.072	0.761	2.895
HGCN⋆ [6]	TCSVT'23	VST	×	RGB	0.246	12.628	0.221	20.531	0.234	16.580
MLP-Mixer [37]	AAAI'23	VST	×	RGB	0.680	–	0.820	–	0.750	–
SGN [36]	TMM'24	VST	×	RGB	0.700	–	0.830	–	0.765	–
PAMFN [31]	TIP'24	VST	×	RGB	0.665	–	0.823	–	0.755	–
VATP-Net [31]	TCSVT'24	VST	×	RGB+	0.702	–	0.863	–	0.796	–
CoFInAl [23]	IJCAI'24	VST	✓	RGB	0.716	2.875	0.843	1.752	0.788	2.314
PHI (Ours)	–	VST	✓	RGB	**0.726**	**2.543**	**0.867**	**1.656**	**0.804**	**2.178**

B. Results and Analysis

We begin with a thorough comparison to state-of-the-art methods, followed by an ablation study that includes parameter sensitivity analysis, and conclude with visualizations.

1) Comparison with the State-of-the-Art: In our comparative analysis, we benchmarked several state-of-the-art methods, consisting of C3D+SVR [26], MS-LSTM [29], ACTION-NET [15], GDLT [16], HGCN [6], and CoFInAl [23]. The comprehensive results are reported in Tables I to III, and the computational performance is listed in Table IV.

a) Comparison with Different Backbones: We employed backbone various architectures, namely C3D [34], I3D [20], ResNet [52], and VST [21], to discern their impact on AQA performance. Among them, I3D consistently outperformed C3D across all categories on the RG dataset, highlighting its proficiency in capturing spatial-temporal dynamics for AQA. Particularly noteworthy was the superior performance of the VST backbone, achieving the highest average SRCC on all the three datasets, as can be seen in Tables I to III. In the RG dataset, ACTION-NET, when coupled with the VST backbone, displayed a remarkable correlation gain of over 16.85% compared to its I3D-based counterpart. This highlights the VST backbone's capability to capture detailed temporal dynamics crucial for AQA. Our proposed method, PHI, leveraging both I3D and VST backbones, showcased outstanding results across all categories, with the VST variant emerging as the top performer. These findings validate the significance of advanced 3D convolutional architectures, such as the I3D and VST backbones, in capturing subtle action details for improved AQA performance, further enhanced by the incorporation of our PHI method.

b) Comparison with Shift-Unaware Methods: Traditional shift-unaware AQA methods [16], [6], [29], [15] employ representation layers that may implicitly mitigate the domain shift issue. However, the persistence of domain shift often leads to suboptimal performance in Tables I to III. PHI with VST consistently outperforms others across all actions

TABLE III
RESULTS OF SRCC (↑) AND R-ℓ_2 (↓) ON THE LOGO DATASET. THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**, WHILE THE SECOND-BEST RESULTS ARE UNDERLINED.

Method	Publisher	Backbone	Shift Aware	SRCC	R-ℓ_2
USDL [53]	CVPR'20	I3D	×	0.426	5.736
CoRe [8]	ICCV'21	I3D	×	0.471	5.402
TSA [48]	CVPR'22	I3D	×	0.452	5.533
CoRe-GOAT [3]	CVPR'23	I3D	×	0.494	5.072
HGCN [6]	TCSVT'23	I3D	×	0.471	4.954
CoFInAl [23]	IJCAI'24	I3D	✓	0.552	4.586
PHI (Ours)	–	I3D	✓	**0.713**	**3.608**
USDL [53]	CVPR'20	VST	×	0.473	5.076
CoRe [8]	ICCV'21	VST	×	0.500	5.960
TSA [48]	CVPR'22	VST	×	0.475	4.778
CoRe-GOAT [3]	CVPR'23	VST	×	0.560	4.763
HGCN [6]	TCSVT'23	VST	×	0.671	6.564
CoFInAl [23]	IJCAI'24	VST	✓	0.698	4.019
PHI (Ours)	–	VST	✓	**0.835**	**2.752**

TABLE IV
COMPUTATIONAL COMPARISON WITH EXISTING OPEN-SOURCE METHODS. "–" DENOTES METHODS WITHOUT OFFLINE MODULES.

Method	Shift Aware	FLOPs (G)	Parameter (M)		Inference Time (ms)
			Online	Offline	
ACTION-NET [15]	×	34.7500	28.08	–	305.2474
GDLT [16]	×	0.1164	3.20	–	3.2249
HGCN [6]	×	1.1201	0.50	–	6.7830
CoFInAl [23]	✓	0.1178	3.70	–	3.8834
PHI-half (Ours)	✓	0.2637	2.80	4.60	5.6870
PHI (Ours)	✓	0.2637	3.00	4.60	5.6870

and score types in all the three datasets, demonstrating its effectiveness in addressing the domain shift issue in long-term AQA. Specifically, it excels in categories like Ball, Hoop, and Ribbon on RG and across TES and PCS on Fis-V by a large margin. For instance, PHI demonstrates notable performance improvements, achieving a remarkable 9.65% and 6.14% correlation gain in the Ball and Hoop categories, respectively, on the RG dataset compared to GDLT [16]. Overall, PHI delivers significant average correlation gains of 5.88%, 5.65%, and 24.44% on RG, Fis-V, and LOGO, respectively. These results validate the benefit of employing the two-stage instruction to address the domain shift issue.

The above statistics focus on unimodal comparisons and exclude multi-modal methods. Below, we compare PHI with recent multi-modal approaches. Notably, PAMFN [31] uses only RGB data, while VATP-Net [51] leverages multi-modal inputs. Despite being a unimodal method, PHI outperforms PAMFN (0.711) by 13.9% and slightly surpasses the multi-modal method [47] on RG, as shown in Table I. On Fis-V, PHI achieves an average SRCC 6.5% higher than PAMFN (0.755) and 1.0% higher than VATP-Net (0.796), as shown in Table II. These results highlight PHI's ability to handle domain shifts without additional modalities, demonstrating that multi-modal methods do not necessarily outperform unimodal methods when domain shifts are effectively addressed. Future work will explore extending PHI to multi-modal settings to further enhance performance and generalization. The comparison with various state-of-the-art methods positions PHI as a promising solution for advancing AQA capabilities.

c) Comparison with Task-Level Discrepancies Solution: This work categorizes domain shifts into two perspectives: task-level discrepancies and feature-level discrepancies (see Figure 1(b)). CoFInAl [23] primarily addressed task-level discrepancies, whereas our work focuses on feature-level discrepancies. Both approaches have demonstrated substantial performance improvements compared to shift-unaware methods, as validated in Tables I to III, demonstrating the effectivenes and necessity of explicit methods in mitigating domain shifts in long-term AQA. In our experiments, we used the optimal parameters for I3D from the well-tuned parameters of VST to better validate the generalizability of the proposed method on different backbones. This means that we did not fine-tune the parameters of the I3D backbone. As shown in Tables I and II, PHI is still superior to CoFInAI that has been well-tuned on the I3D backbone. As a result, it still has significant room for improvement through further fine-tuning, which could potentially lead to even better performance. This explains why the VST backbone shows greater improvement over I3D when compared to CoFInAI.

Although PHI is weaker than CoFInAl in some categories, PHI exhibits a 2.03% performance gain on the challenging Fis-V dataset, while achieving a modest 0.37% gain on the simpler RG dataset. The smaller improvement observed on the RG dataset can be attributed to two key factors: its limited sample size and the relatively mild domain shift between the pre-training task and AQA. Specifically, RG comprises a smaller dataset compared to Fis-V (refer to the dataset details in Section V-A), which restricts the model's capacity to learn complex domain adaptation patterns. Furthermore, as illustrated in Figure 7(a), RG exhibits less pronounced feature misalignment, reducing the upper limit of performance improvement for extensive domain adaptation. In contrast, as can be seen in Figure 8(a), Fis-V demonstrates significant domain discrepancies, where PHI's domain shift mitigation mechanisms prove more impactful, resulting in a more substantial performance gain. Notably, PHI achieves a significant 42.9% reduction in R-ℓ_2 error on RG, indicating PHI's robustness in effectively managing subtle feature discrepancies. Furthermore, the performance gap in CoFInAl can be attributed to the discretization of the continuous score space for classification. Our novel contribution focuses on addressing feature-level discrepancies in AQA, which we have identified as crucial for achieving additional performance improvements, compared to the task-level one. By innovatively tackling these discrepancies, we have effectively enhanced the robustness and accuracy of AQA systems.

We further compare and analyze both solutions. While CoFInAl and PHI addresses the domain shift in complementary ways, they are fundamentally incompatible due to their opposing alignment directions. As can be seen from the different arrows in Figure 1(b), these two approaches operate the alignment process in opposite directions, making their direct combination infeasible. If pre-trained features are already well-optimized and generalizable, task-level alignment (CoFInAI) may be sufficient. However, our experimental results in Tables I to III demonstrate that additional performance

TABLE V
ABLATION RESULTS ON THE RG DATASET.

Setting	Ball		Clubs		Hoop		Ribbon		Average	
	SRCC	R-ℓ_2	SRCC	R-ℓ_2	SRCC	R-ℓ_2	SRCC	R-ℓ_2	SRCC	R-ℓ_2
PHI	0.818	2.187	0.803	2.149	0.812	2.199	0.805	2.744	0.810	2.300
PHI-half	0.808 $\downarrow$1.22%	2.027 $\downarrow$7.32%	0.790 $\downarrow$1.62%	2.460 $\uparrow$14.47%	0.788 $\downarrow$2.96%	5.776 $\uparrow$162.80%	0.752 $\downarrow$6.58%	2.900 $\uparrow$5.68%	0.785 $\downarrow$3.09%	3.291 $\uparrow$43.00%
w/o GMF	0.802 $\downarrow$1.96%	1.942 $\downarrow$11.20%	0.784 $\downarrow$2.36%	2.316 $\uparrow$7.77%	0.789 $\downarrow$2.83%	6.733 $\uparrow$206.27%	0.756 $\downarrow$6.09%	3.217 $\uparrow$17.24%	0.783 $\downarrow$3.33%	3.552 $\uparrow$54.43%
w/o TESA	0.661 $\downarrow$19.20%	3.347 $\uparrow$53.09%	0.637 $\downarrow$20.66%	3.904 $\uparrow$81.63%	0.371 $\downarrow$54.32%	6.490 $\uparrow$195.13%	0.550 $\downarrow$31.68%	5.016 $\uparrow$82.91%	0.564 $\downarrow$30.37%	4.689 $\uparrow$103.87%
w/o LCR	0.775 $\downarrow$5.26%	3.105 $\uparrow$42.00%	0.773 $\downarrow$3.74%	2.588 $\uparrow$20.40%	0.789 $\downarrow$2.83%	7.367 $\uparrow$235.09%	0.719 $\downarrow$10.68%	3.671 $\uparrow$33.81%	0.765 $\downarrow$5.56%	4.183 $\uparrow$81.87%
w/o KL	0.806 $\downarrow$1.47%	1.991 $\downarrow$8.96%	0.799 $\downarrow$0.50%	2.231 $\uparrow$3.81%	0.777 $\downarrow$4.31%	7.005 $\uparrow$218.56%	0.778 $\downarrow$3.35%	3.174 $\uparrow$15.68%	0.790 $\downarrow$2.47%	3.600 $\uparrow$56.52%

TABLE VI
ABLATION RESULTS ON THE FIS-V DATASET.

Setting	TES		PCS		Average	
	SRCC	R-ℓ_2	SRCC	R-ℓ_2	SRCC	R-ℓ_2
PHI	0.726	2.543	0.867	1.656	0.804	2.178
PHI-half	0.661 $\downarrow$9.07%	3.050 $\uparrow$19.88%	0.861 $\downarrow$1.15%	1.183 $\uparrow$28.58%	0.780 $\downarrow$3.48%	2.117 $\uparrow$2.79%
w/o GMF	0.668 $\downarrow$8.06%	3.413 $\uparrow$34.33%	0.857 $\downarrow$0.58%	1.769 $\uparrow$6.72%	0.780 $\downarrow$3.48%	2.591 $\uparrow$18.97%
w/o TESA	0.627 $\downarrow$13.87%	4.360 $\uparrow$166.33%	0.776 $\downarrow$10.97%	3.714 $\uparrow$123.67%	0.709 $\downarrow$11.84%	4.037 $\uparrow$84.81%
w/o LCR	0.280 $\downarrow$61.40%	5.576 $\uparrow$238.17%	0.294 $\downarrow$66.16%	4.984 $\uparrow$200.91%	0.282 $\downarrow$65.00%	5.280 $\uparrow$142.05%
w/o KL	0.576 $\downarrow$20.93%	2.437 $\downarrow$3.88%	0.841 $\downarrow$2.56%	2.564 $\uparrow$54.94%	0.780 $\downarrow$3.48%	2.501 $\uparrow$14.74%

gains are achievable through PHI, indicating that pre-trained features still require adaptation to better suit AQA. In the future, we aim to explore the potential of combining both approaches to fully leverage their respective strengths and achieve even higher performance. Moreover, PHI is designed as a modular enhancement and can be integrated into other baseline methods that do not explicitly consider domain shift. By refining feature alignment, PHI enhances model robustness and generalization, making it a versatile tool for improving AQA performance across various architectures.

d) Computational Efficiency and Model Complexity: To provide a more comprehensive comparison, we evaluated PHI and existing long-term AQA methods under identical conditions, with a focus on computational efficiency. The results are summarized in Table IV. PHI consists of online and offline components, strategically designed to balance computational efficiency with assessment accuracy. The online module (flow path) handles real-time inference with minimal latency, while the offline module (TETE) refines feature representations, thus reducing the computational burden during online inference. Combining the results from Tables I to III, the distillation design enables it to achieve a competitive balance between performance and efficiency compared to state-of-the-art models.

Notably, CoFInAl achieves the lowest FLOPs (0.1178G) and inference time (3.8834ms) among shift-aware methods, underscoring its computational efficiency. However, PHI compensates for a slightly higher computational cost by leveraging a progressive instruction strategy. This approach improves assessment accuracy with only a 0.1470G increase in FLOPs and a 2.4621ms additional inference delay compared to GDLT. Despite this, PHI outperforms in terms of online parameters and quality assessment performance, while maintaining a reasonable computational footprint. PHI introduces only a slight increase in FLOPs and parameters, while significantly improving the assessment accuracy. PHI reduces FLOPs (by 99.2%), parameters (by 89.3%), and inference time (by 98.1%) compared to ACTION-NET, and also achieves lower FLOPs and competitive inference time relative to HGCN. Additionally, PHI benefits from an offline distillation mechanism, allowing the TETE module to transfer knowledge to a lightweight flow model, further reducing online parameters and computational cost. In summary, while CoFInAl excels in computational efficiency, PHI strikes an optimal balance between computational cost and assessment accuracy, making it a practical solution for practical AQA applications.

2) Ablation Studies: This study aims to investigate the individual contributions of the core components of PHI, particularly focusing on GMF (see Hypothesis 1), TESA (see Section IV-A2), LCR (see Hypothesis 2), and the use of KL divergence loss (see Equation (15)). Tables V and VI and Figure 5 present the results on RG and Fis-V.

a) Validation of Hypothesis 1: To validate the effectiveness of Hypothesis 1, we compare removing GMF (*w/o GMF*) and TETE (*w/o TETE*), respectively. On the one hand, when removing GMF while retaining only TETE (*w/o GMF*), we observe a decrease in performance across all categories, with an average SRCC decrease of 3.33% and an average R-ℓ_2 increase of 54.43% on the RG dataset (see Table V) and an average SRCC decrease of 3.48% and an average R-ℓ_2 increase of 18.97% on the Fis-V dataset (see Table VI). These results indicate that the desired features estimated by TETE alone contain inaccuracies due to the inherent domain gap. The GMF module plays a crucial role in refining these estimations by progressively reducing the domain shift, thereby enhancing the reliability of feature representations and improving action assessment performance. On the one hand, by replacing TESA with vanilla attention (*w/o TESA*), we observe a more substantial decrease in performance, with an average SRCC decrease of 30.37% on the RG dataset (see Table V) and an average SRCC decrease of 11.84% on the Fis-V dataset (see Table VI). This emphasizes the critical role of TESA in capturing long-range dependencies efficiently, which is essential for accurate AQA. These results collectively validate the effectiveness of

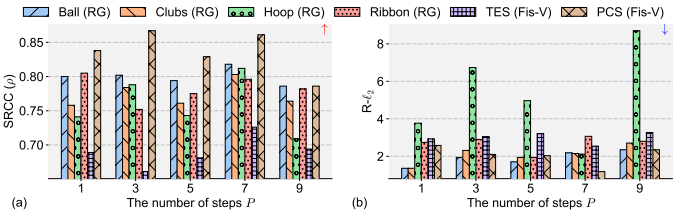


Fig. 5: Results of (a) SRCC and (b) R-ℓ_2 on the impact of different steps. The symbol "↑" indicates higher is better, while the symbol "↓" indicates lower is better.

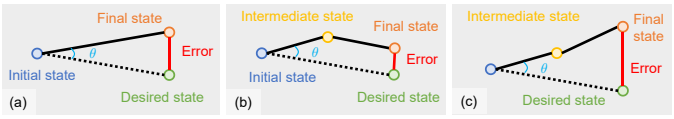


Fig. 6: Conceptual illustrations of (a) one-step and (b,c) multi-step flows. In (a), the one-step flow allows for a fast and direct reduction of domain gaps, making it computationally efficient. However, it may lead to larger errors. Suppose the second state deviates from the desired direction with the same degree θ. In contrast, the multi-step flow (b) provides a more robust alignment through gradual refinements, reducing the risk of large errors. However, in certain cases (c), multi-step alignment can accumulate errors, potentially leading to worse performance than the one-step approach.

Hypothesis 1

b) Validation of Hypothesis 2: To validate the effectiveness of Hypothesis 2, we compare removing LCR (*w/o LCR*) and KL (*w/o KL*), respectively. On the one hand, removing LCR (*w/o LCR*) results in a notable decrease in performance, with an average SRCC decrease of 5.56% and an average R-ℓ_2 increase of 81.87% on the RG dataset (see Table V) and an average SRCC decrease of 65.00% and an average R-ℓ_2 increase of 142.15% on the Fis-V dataset (see Table VI). This demonstrates that LCR plays a crucial role in learning representations focusing on fine-grained cues, which are vital for mitigating domain shift and improving AQA performance. On the one hand, replacing the KL divergence loss with MSE (*w/o KL*) leads to a significant decrease in performance, with an average R-ℓ_2 increase of 56.52% on the RG dataset (see Table V) and an average R-ℓ_2 increase of 14.74% on the Fis-V dataset (see Table VI). This highlights the effectiveness of the KL divergence loss in guiding the model to learn more robust representations aligned with AQA. These results collectively validate the effectiveness of Hypothesis 2.

c) Impact of the Model Size: As shown in Tables V and VI, reducing the parameter size of the flow network ϕ by half (PHI-half) results in a noticeable decrease in performance across all categories. Specifically, we observe an average decrease of 3.09% in SRCC and an average increase of 43.00% in R-ℓ_2 on the RG dataset and an average decrease of 3.48% in SRCC and an average increase of 2.79% in R-ℓ_2 on the RG dataset. Combined with the reported results in Tables I and II, we observe that PHI-half still outperforms some strong baselines [16], [6]. This finding shows the importance of model size in maintaining performance levels, suggesting that a larger parameter space with simple MLPs contributes to better overall performance, indicating the effectiveness of PHI's network design.

d) Impact of Different Steps: The number of flow steps plays a crucial role in refining pre-trained features for AQA. Adjusting this parameter influences the model's ability to reduce the domain gap between pre-trained features and AQA tasks (see Figure 5). Firstly, while increasing the number of steps yields slight improvements, the gains are not always substantial. To provide an intuitive understanding, Figure 6 illustrates the conceptual differences between one-step and multi-step flows. The one-step approach (see Figure 6(a)) achieves competitive performance due to its direct alignment but is more susceptible to deviations. In contrast, the multi-step flow (see Figure 6(b)) provides more controlled and gradual refinement, potentially reducing large alignment errors. However, multi-step alignment can also introduce accumulative errors (see Figure 6(c)), especially when the initial step already aligns features effectively. This explains why, in some cases, multi-step flows offer only marginal improvements or even perform worse than the one-step approach (in Figure 5). Nonetheless, multi-step alignment enhances stability and robustness, particularly in scenarios where direct alignment might lead to suboptimal feature adaptation. Additionally, increasing the number of steps allows for a more gradual transformation of initial features into task-specific representations, potentially improving accuracy and reliability. However, this comes at the cost of higher computational complexity and longer training time. Conversely, reducing the number of steps accelerates training but may limit the model's ability to effectively align with AQA tasks. Importantly, our method is designed to support both one-step and multi-step alignment, offering flexibility depending on the computational constraints and accuracy requirements. The absence of a clear performance trend across different step settings further highlights the robustness and adaptability of our approach.

TABLE VII
RESULTS OF DIFFERENT TRAINING STRATEGIES.

Strategy	RG	Fis-V
Two-stage	0.810	0.804
One-stage	0.807 $\downarrow 0.37\%$	0.800 $\downarrow 0.50\%$

e) Impact of Different Training Strategies: We adopt a two-stage training process in our approach. Initially, we train the TETE to refine the estimation of desired features, focusing on obtaining accurate representations essential for the flow model. Subsequently, we integrate these refined features into our overall framework, enabling joint training with components like GMF. To demonstrate the efficacy of the two-stage training approach, we conducted experiments comparing it to a one-stage joint training process. The results, shown in Table VII, reveal that the two-stage process outperforms the one-stage process, with improvements of 0.37% on the RG dataset and 0.50% on the Fis-V dataset. This highlights the clear advantages of the two-stage training strategy.

3) Qualitative and Quantitative Results: We present a diverse range of visualizations to demonstrate both the qualitative and quantitative performance of our method.

a) Visualization of the Domain Shift: We visually demonstrate the effectiveness of our PHI method in mitigating domain shifts through feature distribution visualizations in the latent space and correlation analyses between predicted and ground truth scores. Specifically, we employ the t-SNE toolbox [54] to generate feature distribution plots, which are presented in the first rows of Figures 7 and 8. Additionally, we illustrate the correlation between predicted and actual scores through scatter plots, as shown in the second rows of Figures 7 and 8. To assess the model's generalization capability, we visualize two action categories, Hoop and PCS, selected from the RG and PCS datasets, respectively. In the feature distribution plots, different score ranges (or grades) are delineated using an SVC classifier. Due to variations in dataset scales, we categorize the test samples into four score ranges for Hoop (RG), assigning labels from 0 to 3, and six score ranges for PCS (Fis-V), assigning labels from 0 to 5. Improved feature clustering, where samples of the same grade (represented by similar color shading) occupy the same region, indicates a more structured feature space for action assessment. Furthermore, we provide comparisons with GDLT [16] and CoFInAl [23] to highlight the advantages of our approach. Since the Fis-V dataset contains a larger number of test samples compared to RG, it provides a more robust and reliable evaluation of model performance and generalizability. Therefore, we primarily focus on comparisons conducted on the PCS (Fis-V) dataset for a more comprehensive assessment of our method in the following.

Specifically, the features extracted by the VST backbone, as shown in Figure 8(a), exhibit a mixed distribution, indicating difficulties in distinguishing between different score ranges. This suggests that the pre-trained backbone may not be well-suited for the AQA task, leading to significant domain shift issues. In Figure 8(b), the feature distribution of GDLT appears confused, reflecting ineffective feature learning. In contrast, Figures 8(c) and 8(d) illustrate the feature distributions of CoFInAl and PHI, respectively, which display more distinct clustering. This clearer separation facilitates the identification of samples within each score range, indicating that both CoFInAl and PHI effectively mitigate domain shift. While the advantage of PHI can be observed by comparing the feature plots in Figures 7(c) and 7(d), the t-SNE visualizations in Figures 8(c) and 8(d) do not provide a definitive comparison between the domain shift mitigation capabilities of PHI and CoFInAl. Notably, CoFInAl suffers from a loss of precision in score prediction, which affects its fine-grained assessment capability. In contrast, PHI maintains higher precision, leading to improved reliability in action assessment. To further clarify this, we demonstrate that PHI outperforms CoFInAl in the following analysis.

Figures 8(e), 8(f) and 8(h) compare correlation plots between GDLT, CoFInAl, and PHI. In Figure 8(e), the correlation line ($\hat{s}_i = 0.28s_i + 18.72$) of GDLT shows a deviation from the ideal correlation line ($\hat{s}_i = s_i$), indicating a weak correlation with ground truth scores. In Figure 8(f), the correlation line ($\hat{s}_i = 0.34s_i + 16.95$) of CoFInAl shows a smaller deviation from the ideal line, suggesting a stronger correlation with ground truth scores compared to GDLT. Notably, Figure 8(h)

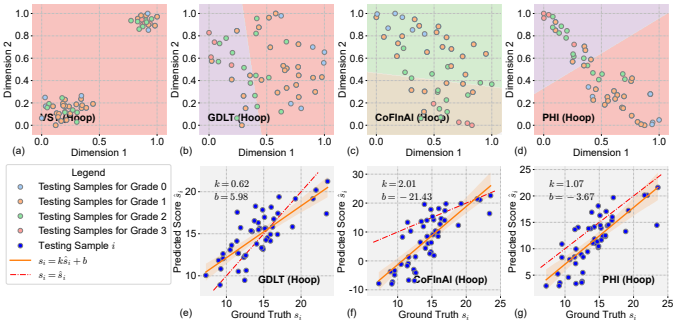


Fig. 7: Visualization depicting the mitigation of domain shift on the Hoop (RG) dataset: t-SNE feature distribution plots (a, b, c, d) and correlation comparison plots (e, f, g) of GDLT [16], CoFInAl [23], and our PHI method. The dataset is split into four grades.

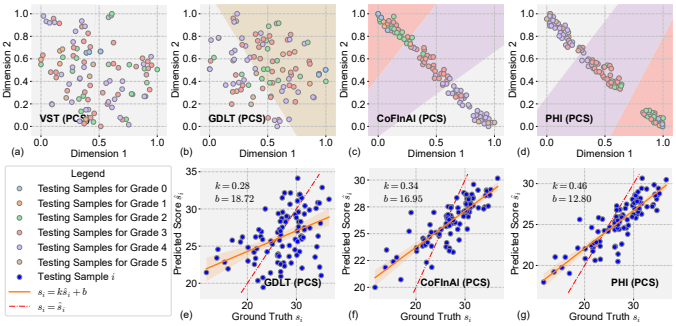
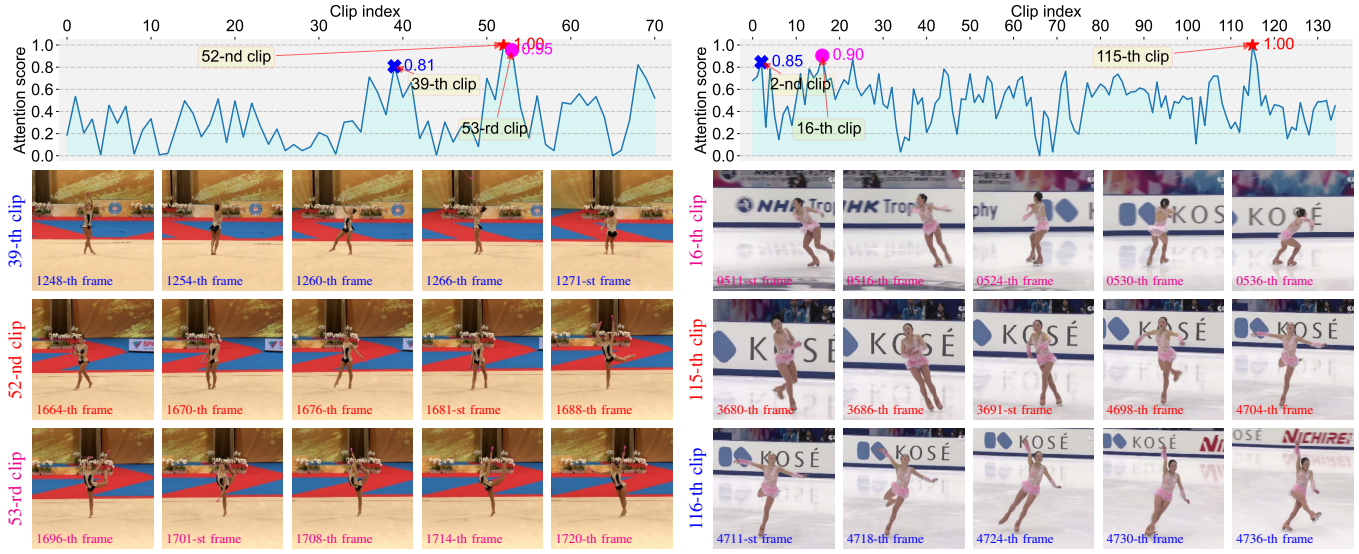


Fig. 8: Visualization depicting the mitigation of domain shift on the PCS (Fis-V) dataset: t-SNE feature distribution plots (a, b, c, d) and correlation comparison plots (e, f, g) of GDLT [16], CoFInAl [23], and our PHI method. The dataset is split into six grades.

shows that the correlation line of PHI ($\hat{s}_i = 0.46s_i + 12.80$) is the closest to the ideal line, demonstrating the highest correlation with ground truth scores among the three methods. This suggests that PHI achieves a higher correlation with ground truth scores compared to GDLT and CoFInAl, further highlighting its superiority in mitigating domain shift and improving AQA performance. Overall, the visualizations in Figure 8 provide compelling evidence of PHI's effectiveness in mitigating domain shift and improving long-range AQA performance compared to existing methods.

b) Visualization of Attention Weights: To gain deeper insights into the attention mechanism within the TETE module, we visualize the attention weights and the highlighted clips for two representative samples from the RG and Fis-V datasets. Figures 9(a) and 9(b) illustrate the attention weight distributions for each clip in the Club (RG) and PCS (Fis-V) models, respectively. The first row in each subfigure presents the normalized attention weights assigned to all action clips, revealing the varying levels of importance attributed to different segments of the action. The following rows highlight the top three clips that received the highest attention scores, which are crucial for the model's decision-making. In Figure 9(a), the model predicts a score of 14.30, while the ground-truth score is 14.52, resulting in a minimal error of 0.22. Similarly, in Figure 9(b), the model predicts a score of 27.78, compared to the ground-truth score of 26.68, yielding an error of 1.10. For example, in Figure 9(b), the three most attended clips are 39,



(a) Sample #056 from the RG (Club) dataset

(b) Sample #056 from the Fis-V (PCS) dataset

Fig. 9: Visualization of attention weights and highlighted clips for samples on (a) the RG (Club) dataset and (b) the Fis-V (PCS) dataset.

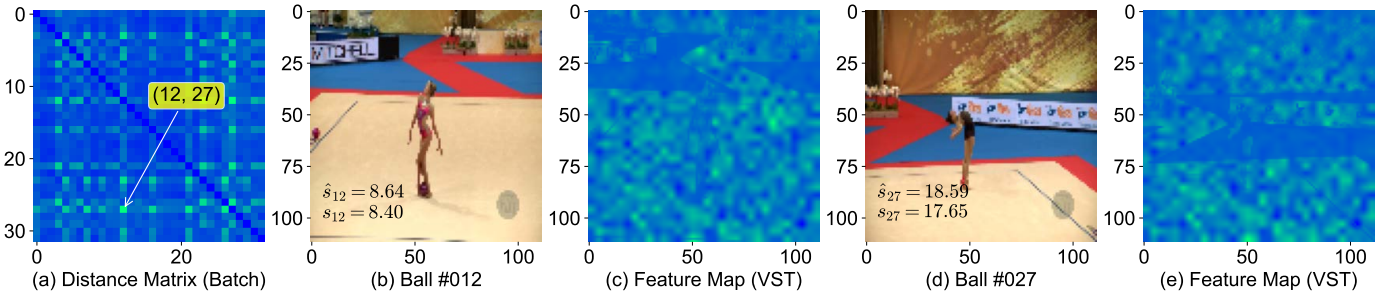


Fig. 10: Visualization of the distance matrix on the Ball (RG) dataset.

52, and 53, corresponding to key subactions where the player executes precise hand-leg coordination. These visualizations provide valuable insights into how the AQA model prioritizes different temporal segments within an action sequence. By identifying the most critical moments, we enhance our understanding of the model's interpretability and its evaluation process for action quality assessment.

c) Visualization of the Distance Matrix: In Figure 10, the heatmap of the distance matrix is depicted. The highest value in the matrix, such as $(12, 27)$, signifies the greatest distance between the 12-th and 27-th actions. To further assess the efficacy, the two actions and their respective heatmaps are visualized. These heatmaps in Figures 10(c) and 10(d) provide insights into the domain shift issue. The minimal difference between the predicted and ground-truth scores indicates the effectiveness of our method. For instance, with a ground-truth score of $s_{12} = 8.40$ and the predicted score is $\hat{s}_{12} = 8.64$. Additionally, observing the score distance between the two videos reveals a wide range, consistent with their feature distance, demonstrating the effectiveness of our PHI method.

VI. Conclusions and Discussions

In this work, we identify and analyze task-level and feature-level domain shifts in long-term AQA and propose PHI as a hierarchical adaptation framework to address feature-level discrepancies. Rather than a mere extension of existing methods, PHI introduces a novel integration of shallow-to-deep adaptation and coarse-to-fine alignment strategies to enhance AQA performance. The shallow-to-deep adaptation strategy, enabled by GMF, effectively reduces domain gaps while maintaining computational efficiency. Simultaneously, the coarse-to-fine alignment mechanism, facilitated by LCR, refines coarse features extracted from pre-trained models, aligning them with fine-grained representations crucial for AQA. Experimental results on three representative long-term AQA datasets demonstrate the significant effectiveness of PHI, underscoring the importance of mitigating feature-level discrepancies in improving AQA performance. Notably, compared to the task-level adaptation method CoFInAI, PHI exhibits superior performance in mitigating domain shifts and enhancing AQA accuracy, emphasizing the critical role of feature-level alignment in long-term AQA tasks. Furthermore, the hierarchical adaptation framework of PHI is highly generalizable beyond AQA, with potential applications in rehabilitation analysis, sports motion scoring, and movement disorder diagnosis. The principles of shallow-to-deep adaptation and coarse-to-fine alignment also extend to domains such as multi-modal alignment. By addressing domain shifts through hierarchical feature refinement, PHI provides a theoretically grounded and versatile framework,

paving the way for future research across multiple fields.

Despite its strong performance, PHI has several limitations, each suggesting directions for future work. First, the auto-regressive nature of GMF introduces cumulative errors, which may progressively degrade prediction accuracy over multiple steps. Future research will focus on advanced optimization strategies and novel regularization techniques to mitigate these challenges. Second, while our clip alignment method effectively minimizes discrepancies between actions, it may lead to information loss in cases of significant temporal variation. Future work will explore adaptive alignment strategies to enhance temporal correlation capture while maintaining computational efficiency. Third, although PHI and CoFInAI address domain shifts from complementary perspectives, feature-level and task-level alignment, respectively, their integration into a unified framework remains an open challenge. Future research will investigate synergistic strategies to combine these approaches for a more comprehensive domain adaptation solution. Additionally, PHI is currently designed for unimodal inputs, limiting its applicability in multi-modal scenarios. Extending the framework to support multi-modal integration will be a key avenue for future work, broadening its utility across diverse applications. Finally, PHI is tailored for mitigating the domain shift issue in long-term AQA tasks, which may not be directly applicable to short-term AQA scenarios. Future research will explore adaptations to extend its applicability to short-term AQA, further increasing its versatility. Addressing these limitations will enhance the robustness, generalizability, and applicability of PHI, advancing the field of action quality assessment and its related domains.

REFERENCES

[1] K. Zhou, L. Wang, X. Zhang, H. P. Shum, F. W. Li, J. Li, and X. Liang, "Magr: Manifold-aligned graph regularization for continual action quality assessment," in *Proceedings of the 18th European Conference on Computer Vision*, 2024.

[2] Y. Ji, L. Ye, H. Huang, L. Mao, Y. Zhou, and L. Gao, "Localization-assisted uncertainty score disentanglement network for action quality assessment," in *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 8590–8597, 2023.

[3] S. Zhang, W. Dai, S. Wang, X. Shen, J. Lu, J. Zhou, and Y. Tang, "Logo: A long-form video dataset for group action quality assessment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2405–2414, 2023.

[4] J. Xu, S. Yin, G. Zhao, Z. Wang, and Y. Peng, "Fineparser: A fine-grained spatio-temporal action parser for human-centric action quality assessment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14628–14637, 2024.

[5] Y.-M. Li, L.-A. Zeng, J.-K. Meng, and W.-S. Zheng, "Continual action assessment via task-consistent score-discriminative feature distribution modeling," *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.

[6] K. Zhou, Y. Ma, H. P. Shum, and X. Liang, "Hierarchical graph convolutional networks for action quality assessment," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 12, pp. 7749–7763, 2023.

[7] M. Li, H.-B. Zhang, Q. Lei, Z. Fan, J. Liu, and J.-X. Du, "Pairwise contrastive learning network for action quality assessment," in *European Conference on Computer Vision*, pp. 457–473, Springer, 2022.

[8] X. Yu, Y. Rao, W. Zhao, J. Lu, and J. Zhou, "Group-aware contrastive regression for action quality assessment," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 7919–7928, 2021.

[9] L. Yao, Q. Lei, H. Zhang, J. Du, and S. Gao, "A contrastive learning network for performance metric and assessment of physical rehabilitation exercises," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2023.

[10] K. Zhou, R. Cai, Y. Ma, Q. Tan, X. Wang, J. Li, H. P. Shum, F. W. Li, S. Jin, and X. Liang, "A video-based augmented reality system for human-in-the-loop muscle strength assessment of juvenile dermatomyositis," *IEEE Transactions on Visualization and Computer Graphics*, vol. 29, no. 5, pp. 2456–2466, 2023.

[11] S. Deb, M. F. Islam, S. Rahman, and S. Rahman, "Graph convolutional networks for assessment of physical rehabilitation exercises," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 30, pp. 410–419, 2022.

[12] C. K. Ingwersen, A. Xarles, A. Clapés, M. Madadi, J. N. Jensen, M. R. Hannemose, A. B. Dahl, and S. Escalera, "Video-based skill assessment for golf: Estimating golf handicap," in *Proceedings of the 6th International Workshop on Multimedia Content Analysis in Sports*, pp. 31–39, 2023.

[13] L. Trinh, T. Chu, Z. Cui, A. Malpani, C. Yang, I. Dalieh, A. Hui, O. Gomez, Y. Liu, and A. Hung, "Self-supervised sim-to-real kinematics reconstruction for video-based assessment of intraoperative suturing skills," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 708–717, Springer, 2023.

[14] X. Ding, X. Xu, and X. Li, "Sedskill: Surgical events driven method for skill assessment from thoracoscopic surgical videos," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 35–45, Springer, 2023.

[15] L.-A. Zeng, F.-T. Hong, W.-S. Zheng, Q.-Z. Yu, W. Zeng, Y.-W. Wang, and J.-H. Lai, "Hybrid dynamic-static context-aware attention network for action assessment in long videos," in *Proceedings of the 28th ACM international conference on multimedia*, pp. 2526–2534, 2020.

[16] A. Xu, L.-A. Zeng, and W.-S. Zheng, "Likert scoring with grade decoupling for long-term action assessment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3232–3241, 2022.

[17] A. Dadashzadeh, S. Duan, A. Whone, and M. Mirmehdi, "Pecop: Parameter efficient continual pretraining for action quality assessment," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 42–52, 2024.

[18] P. Parmar and B. Morris, "Action quality assessment across multiple actions," in *2019 IEEE winter conference on applications of computer vision (WACV)*, pp. 1468–1476, IEEE, 2019.

[19] Y. Bai, D. Zhou, S. Zhang, J. Wang, E. Ding, Y. Guan, Y. Long, and J. Wang, "Action quality assessment with temporal parsing transformer," in *European conference on computer vision*, pp. 422–438, Springer, 2022.

[20] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6299–6308, 2017.

[21] Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, and H. Hu, "Video swin transformer," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3202–3211, 2022.

[22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[23] K. Zhou, J. Li, R. Cai, L. Wang, X. Zhang, and X. Liang, "Cofinal: Enhancing action quality assessment with coarse-to-fine instruction alignment," in *Proceedings of the 33rd International Joint Conference on Artificial Intelligence*, 2024.

[24] X. Liu, C. Gong, and Q. Liu, "Flow straight and fast: Learning to generate and transfer data with rectified flow," *arXiv preprint arXiv:2209.03003*, 2022.

[25] Y. Lipman, R. T. Chen, H. Ben-Hamu, M. Nickel, and M. Le, "Flow matching for generative modeling," *arXiv preprint arXiv:2210.02747*, 2022.

[26] P. Parmar and B. Tran Morris, "Learning to score olympic events," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 20–28, 2017.

[27] H. Pirsiavash, C. Vondrick, and A. Torralba, "Assessing the quality of actions," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI 13*, pp. 556–571, Springer, 2014.

[28] S. Wang, D. Yang, P. Zhai, C. Chen, and L. Zhang, "Tsa-net: Tube self-attention network for action quality assessment," in *Proceedings of the 29th ACM international conference on multimedia*, pp. 4902–4910, 2021.

[29] C. Xu, Y. Fu, B. Zhang, Z. Chen, Y.-G. Jiang, and X. Xue, "Learning to score figure skating sport videos," *IEEE transactions on circuits and systems for video technology*, vol. 30, no. 12, pp. 4578–4590, 2019.

[30] X. Wang, J. Li, and H. Hu, "Skeleton-based action quality assessment via partially connected lstm with triplet losses," in *Chinese Conference*

on Pattern Recognition and Computer Vision (PRCV), pp. 220–232, Springer, 2022.
[31] L.-A. Zeng and W.-S. Zheng, "Multimodal action quality assessment," *IEEE Transactions on Image Processing*, vol. 33, pp. 1600–1613, 2024.
[32] S. Wang, D. Yang, P. Zhai, Q. Yu, T. Suo, Z. Sun, K. Li, and L. Zhang, "A survey of video-based action quality assessment," in *2021 International Conference on Networking Systems of AI (INSAI)*, pp. 1–9, IEEE, 2021.
[33] P. Parmar and B. T. Morris, "What and how well you performed? a multitask learning approach to action quality assessment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 304–313, 2019.
[34] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, pp. 4489–4497, 2015.
[35] J.-H. Pan, J. Gao, and W.-S. Zheng, "Action assessment by joint relation graphs," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6331–6340, 2019.
[36] Z. Du, D. He, X. Wang, and Q. Wang, "Learning semantics-guided representations for scoring figure skating," *IEEE Transactions on Multimedia*, vol. 26, pp. 4987–4997, 2024.
[37] J. Xia, M. Zhuge, T. Geng, S. Fan, Y. Wei, Z. He, and F. Zheng, "Skating-mixer: Long-term sport audio-visual modeling with mlps," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, pp. 2901–2909, 2023.
[38] A. Tong, N. Malkin, G. Huguet, Y. Zhang, J. Rector-Brooks, K. Fatras, G. Wolf, and Y. Bengio, "Improving and generalizing flow-based generative models with minibatch optimal transport," *arXiv preprint arXiv:2302.00482*, 2023.
[39] M. S. Albergo and E. Vanden-Eijnden, "Building normalizing flows with stochastic interpolants," *arXiv preprint arXiv:2209.15571*, 2022.
[40] D. Onken, S. W. Fung, X. Li, and L. Ruthotto, "Ot-flow: Fast and accurate continuous normalizing flows via optimal transport," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, pp. 9223–9232, 2021.
[41] I. Kobyzev, S. J. Prince, and M. A. Brubaker, "Normalizing flows: An introduction and review of current methods," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 11, pp. 3964–3979, 2020.
[42] K. Neklyudov, D. Severo, and A. Makhzani, "Action matching: A variational method for learning stochastic dynamics from samples," *arXiv preprint arXiv:2210.06662*, 2022.
[43] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.
[44] Z. Chang, G. A. Koulieris, and H. P. Shum, "On the design fundamentals of diffusion models: A survey," *arXiv preprint arXiv:2306.04542*, 2023.
[45] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," *arXiv preprint arXiv:2011.13456*, 2020.
[46] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," *arXiv preprint arXiv:2010.02502*, 2020.
[47] K. Gedamu, Y. Ji, Y. Yang, J. Shao, and H. T. Shen, "Fine-grained spatio-temporal parsing network for action quality assessment," *IEEE Transactions on Image Processing*, vol. 32, pp. 6386–6400, 2023.
[48] J. Xu, Y. Rao, X. Yu, G. Chen, J. Zhou, and J. Lu, "Finediving: A fine-grained dataset for procedure-aware action quality assessment," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2949–2958, 2022.
[49] F. Raiber and O. Kurland, "Kullback-leibler divergence revisited," in *Proceedings of the ACM SIGIR international conference on theory of information retrieval*, pp. 117–124, 2017.
[50] X. Ke, H. Xu, X. Lin, and W. Guo, "Two-path target-aware contrastive regression for action quality assessment," *Information Sciences*, vol. 664, p. 120347, 2024.
[51] K. Gedamu, Y. Ji, Y. Yang, J. Shao, and H. T. Shen, "Visual-semantic alignment temporal parsing for action quality assessment," *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.
[52] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, pp. 770–778, 2016.
[53] Y. Tang, Z. Ni, J. Zhou, D. Zhang, J. Lu, Y. Wu, and J. Zhou, "Uncertainty-aware score distribution learning for action quality assessment," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9839–9848, 2020.
[54] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne.," *Journal of machine learning research*, vol. 9, no. 11, 2008.

Kanglei Zhou is a Ph.D. candidate in the School of Computer Science and Engineering at Beihang University, specializing in action quality assessment and augmented reality. From February to August 2024, he was a visiting student in the Department of Computer Science at Durham University. He received his Bachelor's degree in the College of Computer and Information Engineering from Henan Normal University in 2020.

Hubert P. H. Shum (Senior Member, IEEE) is a Professor of Visual Computing and the Director of Research of the Department of Computer Science at Durham University, specialising in modelling spatio-temporal information with responsible AI. He is also a Co-Founder and the Co-Director of Durham University Space Research Centre. Before this, he was an Associate Professor/Senior Lecturer at Northumbria University and a Postdoctoral Researcher at RIKEN Japan. He received his PhD degree from the University of Edinburgh. He chaired conferences such as Pacific Graphics, BMVC and SCA, and has authored over 180 research publications.

Frederick W. B. Li received a B.A. and an M.Phil. degree from Hong Kong Polytechnic University, and a Ph.D. degree from the City University of Hong Kong. He is currently an Associate Professor at Durham University, researching computer graphics, deep learning, collaborative virtual environments, and educational technologies. He is also an Associate Editor of Frontiers in Education and an Editorial Board Member of Virtual Reality & Intelligent Hardware. He chaired conferences such as ISVC and ICWL.

Xingxing Zhang received the BE and PhD degrees from the Institute of Information Science, Beijing Jiaotong University, in 2015 and 2020, respectively. She was also a visiting student with the Department of Computer Science, University of Rochester, from 2018 to 2019. She was a postdoc with the Department of Computer Science and Technology, Tsinghua University, from 2020 to 2022. Her research interests include continual learning and zero/few-shot learning. She has received the excellent PhD thesis award from the Chinese Institute of Electronics, in 2020.

Xiaohui Liang received his Ph.D. degree in computer science and engineering from Beihang University, China. He is currently a Professor, working in the School of Computer Science and Engineering at Beihang University. His main research interests include computer graphics and animation, visualization, and virtual reality.