This WACV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

CaFlow: Enhancing Long-Term Action Quality Assessment with Causal Counterfactual Flow

Ruisheng Han[1], Kanglei Zhou[2], Shuang Chen[1], Amir Atapour-Abarghouei[1], Hubert P. H. Shum[1*]
[1]Durham University [2]Tsinghua University
{ruisheng.han, shuang.chen, amir.atapour-abarghouei, hubert.shum}@durham.ac.uk,
{zhoukanglei}@tsinghua.edu.cn

Abstract

*Action Quality Assessment (AQA) predicts fine-grained execution scores from action videos and is widely applied in sports, rehabilitation, and skill evaluation. Long-term AQA, as in figure skating or rhythmic gymnastics, is especially challenging since it requires modeling extended temporal dynamics while remaining robust to contextual confounders. Existing approaches either depend on costly annotations or rely on unidirectional temporal modeling, making them vulnerable to spurious correlations and unstable long-term representations. To this end, we propose **CaFlow**, a unified framework that integrates counterfactual de-confounding with bidirectional time-conditioned flow. The Causal Counterfactual Regularization (CCR) module disentangles causal and confounding features in a self-supervised manner and enforces causal robustness through counterfactual interventions, while the BiT-Flow module models forward and backward dynamics with a cycle-consistency constraint to produce smoother and more coherent representations. Extensive experiments on multiple long-term AQA benchmarks demonstrate that CaFlow achieves state-of-the-art performance. Code is available at* https://github.com/Harrison21/CaFlow

1. Introduction

Action Quality Assessment (AQA) [11, 41, 46, 51, 52] seeks to evaluate how well an action is performed in video sequences, going beyond action recognition to measure fine-grained execution quality [48, 49]. Accurate AQA is crucial for applications such as sports analytics [13, 42, 43, 48], medical rehabilitation [4, 47], and skill assessment [5, 10, 32]. Long-term AQA [38, 45], which evaluates extended video sequences instead of short clips, provides a more comprehensive and practical measure of performance but also introduces greater modeling challenges.

Recent approaches have attempted to address long-term AQA through representation learning, temporal modeling, or causal reasoning. Likert scoring [38] and hierarchical graph methods [48] capture temporal dependencies but struggle with robustness under domain shift. Causal AQA methods such as FineCausal [8] reduce spurious correlations by balancing foreground and background features but rely on costly human-annotated masks, limiting scalability. Meanwhile, advanced refinement approaches like PHI [53] model temporal dynamics but remain restricted to unidirectional flows, which accumulate errors over time and lead to unstable long-term representations.

Despite recent progress, we identify two fundamental yet overlooked challenges in existing methods: (i) **Confounding bias**, where irrelevant contextual factors such as background, environment, or camera viewpoint are entangled with quality scores, leading to spurious correlations. (ii) **Temporal instability**, where unidirectional refinement produces noisy or inconsistent features over long sequences, preventing accurate modeling of execution dynamics. Addressing both challenges simultaneously is essential for robust, interpretable, and generalizable long-term AQA.

To this end, we propose **CaFlow**, a unified framework that integrates two innovations: a mask-free Causal Counterfactual Regularization (CCR), which leverages the transformer's "desired feature" to partition features and enforce causal robustness via counterfactual testing; and Bidirectional Time-conditioned Flow (BiT-Flow), which refines features by modeling forward-backward dynamics with cycle consistency to yield smoother temporal trajectories. Together, these designs ensure that CaFlow focuses on causal cues while producing stable, temporally coherent features for long-term AQA. We validate CaFlow on three long-term AQA benchmarks: RG (rhythmic gymnastics, four apparatus), FIS-V (figure skating, two scores), and LOGO (diving, single score), totaling over 4,000 videos. Across all datasets, CaFlow achieves state-of-the-art performance. Our **source code** can be found in the supplementary mate-

*Corresponding author. This research is supported in part by the EPSRC NorthHFutures project (ref: EP/X031012/1).

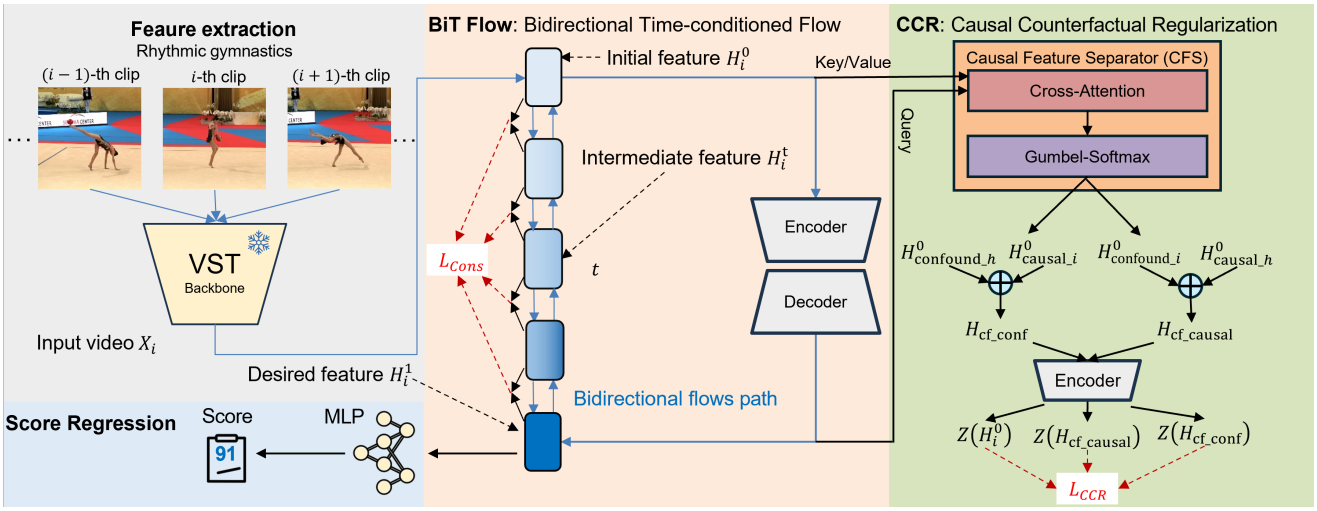


Figure 1. Framework of CaFlow. Our method tackles confounding and domain shift in AQA with two key contributions: (1) Causal Counterfactual Regularization (CCR), which uses a Causal Feature Separator and counterfactual mixing to separate causal from confounding clips and impose a triplet-style causal loss; (2) Bidirectional Time-conditioned Flow (BiT Flow), a time-conditioned bidirectional flow that progressively transforms H_i^0 to the AQA-specific representation H_i^1 with forward-backward consistency and optimal-transport regularization. The refined representation is finally regressed by an MLP to the quality score.

rial. Our main contributions are as follows:

- We introduce a self-supervised causal regularization module (CCR) that separates causal and confounding features without external annotations.
- We propose a bidirectional temporal refinement module (BiT-Flow) that enforces forward-backward consistency for stable long-term modeling.
- We present a unified causal-temporal framework (CaFlow) that achieves state-of-the-art performance on three long-term AQA benchmarks (RG, Fis-V, LOGO).

2. Related Work

2.1. Action Quality Assessment

Research on Action Quality Assessment (AQA) has progressed significantly over the past decade. Early efforts, such as Pirsiavash *et al.* [24], formulated AQA as a direct regression problem, mapping action representations to performance scores. Parisi *et al.* [22] instead evaluated action quality by measuring the correctness of action matches.

More recently, research has shifted towards *long-term AQA*, which requires handling extended durations, complex temporal dependencies, and subjective scoring. To address these challenges, Xu *et al.* [38] proposed Likert scoring with grade decoupling for more nuanced evaluation, while Zhou *et al.* [48] introduced Hierarchical GCNs to capture structural and temporal hierarchies. CoFInAl [50] further improved interpretability by aligning action segments with coarse-to-fine instructions, and PHI [53] tackled domain shift through progressive hierarchical instruction, enhancing robustness across datasets. In parallel, causal inference has been explored to improve robustness and interpretability. FineCausal [8], for instance, introduced a causal framework for fine-grained AQA. However, such approaches may sacrifice generalizability by overfitting to dataset-specific causal pathways, underscoring the need for methods that balance causal robustness with broad applicability.

2.2. Deep Causality Learning

The integration of causal inference into deep learning, often referred to as Deep Causality Learning, has emerged as a promising direction for improving robustness, interpretability, and generalization. A central idea is the use of *counterfactuals*, which reason about "what if" scenarios by altering specific features. For example, Xiao *et al.* [37] introduced masked images as counterfactual samples to enhance fine-tuning robustness, while Rao *et al.* [25] proposed counterfactual attention learning to highlight causally relevant cues in fine-grained categorization and person re-identification.

Causal reasoning has also been extended to video understanding. Wei *et al.* [35] developed visual causal scene refinement for Video Question Answering (VQA), aiming to disentangle object-event relations, and Liu *et al.* [17] modeled cross-modal causal links for video-text reasoning. In perceptual tasks, Shen *et al.* [26] investigated causal perceptual effects in Image Quality Assessment through abductive counterfactual inference, while Liang *et al.* [14] proposed de-confounded gaze estimation by mitigating spurious correlations. Collectively, these works demonstrate the potential of deep causality learning to explicitly model causal structures and counterfactuals, yielding more robust and interpretable visual models.

2.3. Flow Matching

Flow Matching (FM) models represent a cutting-edge advancement in generative modeling. This technique leverages neural Ordinary Differential Equations (ODEs) to implicitly learn a smooth transformation, or "flow," that maps samples from a simpler base distribution to a more complex target data distribution [31]. Building on the principles of normalizing flows [1, 18] and continuous normalizing flows [20], FM models excel at generating high-fidelity samples without the computational overhead of complex approximate inference methods [12]. A key innovation in recent FM approaches is the development of training algorithms that circumvent the computational difficulties of backpropagating through ODEs, requiring explicit ODE solving only during the inference phase [1, 15, 16, 19, 31]. Consequently, FM presents a highly promising and relatively underexplored direction for generative modeling, offering an efficient and effective means to learn and sample from intricate data distributions. In contrast to diffusion models [2, 3, 9, 33, 34], which rely on Stochastic Differential Equations (SDEs) and typically assume a Gaussian base distribution [28], FM provides enhanced flexibility. It allows for a broader selection of base distributions and employs ODEs for training rather than SDEs, resulting in smoother generative trajectories and often superior performance [27].

3. Methodology

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

3.1. Motivation and Framework Overview

Motivation Assessing long-term human actions requires models to capture subtle execution details while remaining robust to confounding factors such as environment, background, or recording conditions. However, existing AQA methods face two major limitations. First, they are vulnerable to spurious correlations, where irrelevant contextual cues (e.g., venue or camera angle) become entangled with performance scores, leading to biased predictions. Second, they typically adopt unidirectional temporal modeling, which produces unstable and poorly aligned representations over long sequences, making it difficult to capture the full temporal dynamics of an action. Together, these issues undermine robustness, interpretability, and generalization, limiting the reliability of current AQA approaches. Motivated by these challenges, we introduce **CaFlow**, a unified framework that integrates counterfactual de-confounding with bidirectional temporal flow, explicitly addressing both the causal and temporal limitations of prior methods.

Framework Overview CaFlow is designed for Action Quality Assessment (AQA). Given an input action video $\mathbf{X}_i \in \mathbb{R}^{T\times W\times H\times 3}$ with T frames of resolution $W \times H$ and 3 color channels, the video is first divided into M non-overlapping clips and passed through a pre-trained backbone to obtain an initial feature sequence $H_i^0 = \{h_m^0\}_{m=1}^M$, where $h_m^0 \in \mathbb{R}^D$ denotes the feature vector for the m-th clip. CaFlow then processes these features to predict a scalar action quality score $S \in \mathbb{R}$. As illustrated in Fig. 1, CaFlow consists of two key components, optimized jointly to achieve robust and generalized AQA:

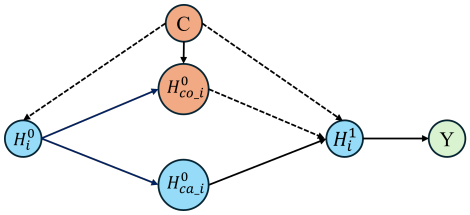


Figure 2. The causal graph of our AQA framework. Nodes represent variables: H_i^0 for initial video features, H_i^1 for desired features, C for confounder, $H_{co_i}^0$ for confound features, $H_{ca_i}^0$ for causal features, and Y for the final action score. Solid arrows ($\rightarrow$) indicate true causal relationships, whereas dashed arrows ($\dashrightarrow$) represent spurious causal relationships.

1. **Causal Counterfactual Regularization (CCR)**: This module introduces a Causal Feature Separator to disentangle causal from confounding clips within the input feature sequence H^0. It then applies counterfactual mixing with a triplet-style loss to enforce causally robust representations, ensuring the model focuses on true causal cues for action quality.
2. **Bidirectional Time-conditioned Flow (BiT-Flow)**: This time-conditioned bidirectional flow module progressively refines H^0 into stable AQA-specific representations $H^1 = \{h_m^1\}_{m=1}^M$ under forward-backward consistency. Here, $h_m^0, h_m^1 \in \mathbb{R}^{D'}$ represent the feature vectors before and after refinement at clip m. This refinement ensures temporal consistency and enhances the stability of the learned features.

Finally, the refined representation H^1 is aggregated (e.g., via pooling) and regressed by an MLP to predict the final action quality score S. The overall optimization objective for CaFlow is a high-level combination of these components:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{regression}} + \lambda_1 \mathcal{L}_{\text{CCR}} + \lambda_2 \mathcal{L}_{\text{BiT}},$$

where $\mathcal{L}_{\text{regression}}$ is the primary loss for predicting the action quality score, $\mathcal{L}_{\text{CCR}}$ (see Eq. (4)) is the loss associated with causal counterfactual regularization, and $\mathcal{L}_{\text{BiT-Flow}}$ (see Eq. (11)) enforces temporal consistency within the flow module. λ_1 and λ_2 are hyperparameters balancing these objectives. This comprehensive design compels the model to focus on truly causal cues while ensuring temporally consistent refinement, thereby enhancing robustness and generalization in long-term AQA. Each component will be detailed in the subsequent sections.

3.2. Causal Counterfactual Regularization

Design Idea Existing causal AQA methods [8] reduce spurious correlations by balancing foreground and background features, but their reliance on costly human-annotated masks limits generalizability. Our key insight is to achieve causal separation in a mask-free, self-supervised manner by exploiting the "desired feature" representation from the transformer as an internal guidance signal. This allows our Causal Feature Separator to disentangle causal and confounding features without external supervision. Moreover, instead of implicitly re-weighting features, we explicitly partition them and apply counterfactual testing, forcing the model to reject spurious dependencies. In the following, we present the causal graph, explain the separation mechanism, and introduce the counterfactual intervention strategy.

Structural Causal Model To model the feature deconfounding process in AQA, we formulate a Structural Causal Model (SCM) as illustrated in Fig. 2. The variables are defined as follows: H_i^0 denotes the initial sequence of clip features extracted from the backbone for video i; $H_{ca_i}^0$ and $H_{co_i}^0$ represent the decomposition of H_i^0 into score-causal features and score-confounding features, respectively; C is an unobserved confounder corresponding to environmental factors such as lighting or venue conditions, which may induce spurious correlations; H_i^1 is the desired feature representation generated by the transformer module, serving as a proxy for the scoring intent; and Y is the final ground-truth action quality score. The true causal pathway for robust AQA should follow $H_i^0 \rightarrow H_{ca_i}^0 \rightarrow H_i^1 \rightarrow Y$, where the extracted causal features influence the desired representation and ultimately determine the score. However, the confounder C introduces biased backdoor routes by simultaneously affecting both the raw stream and the desired representation, i.e., $H_i^0 \leftarrow C \rightarrow H_i^1 \rightarrow Y$ which create spurious correlations between appearance/context (e.g., venue or lighting) and the final score. In practice, this means that high scores recorded under certain environments can make the model falsely treat those environments as predictive, even when execution quality is unchanged.

To focus on the true causal effect $H_i^0 \rightarrow H_{ca_i}^0 \rightarrow H_i^1 \rightarrow Y$ and eliminate the non-causal impact of the backdoor paths opened by C, we adopt a front-door intervention that treats $H_{ca_i}^0$ as the mediator. Since $H_{ca_i}^0$ intercepts all directed paths from H_i^0 to Y and is not directly influenced by C, the causal effect of H^0 on Y can be estimated via:

$$P(Y \mid do(H^0 = h)) = \sum_{h_{\text{ca}}} P(h_{\text{ca}} \mid h) \times \sum_{h^1} P(h^1 \mid h_{\text{ca}}) \times \sum_{h'} P(Y \mid h^1, h')\, P(h'), \tag{1}$$

where h' indexes raw features from the observational distribution. This formulation blocks the spurious influences of $C \rightarrow H^0$ and $C \rightarrow H^1$, ensuring that the model learns only from the true causal pathway.

Causal Feature Separator (CFS) Inspired by recent advances in causal discovery through attention mechanisms [35], we design a Causal Feature Separator (CFS) to dynamically divide the initial feature sequence H_i^0 into a score-causal subset H_{causal} and a score-confounding subset H_{confound} in a self-supervised fashion. The separation is guided by the "desired feature" H_i^1 obtained from the temporal transformer, which acts as a high-level proxy for scoring intent. To realize this, we compute attention scores $\mathbf{a}_i \in \mathbb{R}^M$ through a cross-attention operation, where H_i^1 serves as the *Query* and the individual clip features $\{h_{i,t}^0\}_{t=1}^M$ from H_i^0 act as the *Keys* and *Values*. These attention scores measure the contribution of each clip feature to the desired representation, enabling the model to separate causal cues from confounding context.

To separate causal and confounding features, we apply the Gumbel-Softmax function to the continuous attention scores. This technique injects stochasticity through Gumbel noise and employs a temperature-controlled softmax to produce near-binary probability weights. Formally, we obtain a soft mask $\mathbf{m}_i \in [0,1]^M$, which is:

$$\mathbf{m}_i = \text{Gumbel-Softmax}(\mathbf{a}_i, \tau), \tag{2}$$

where τ is the temperature parameter annealed during training. Each element of $\mathbf{m}_i$ reflects the probability of the corresponding clip feature being causal. Using this mask, the initial feature sequence H_i^0 is partitioned into two complementary components:

$$H_{\text{causal}} = \mathbf{m}_i \odot H_i^0, H_{\text{confound}} = (1 - \mathbf{m}_i) \odot H_i^0, \tag{3}$$

where $\odot$ denotes element-wise multiplication. This procedure enables the model to softly separate causal and confounding features in a differentiable manner, without requiring external supervision.

Counterfactual Regularization via Causal Distillation. Drawing from the principles of counterfactual sample generation, we create counterfactuals in the feature space to explicitly break spurious correlations and regularize the model [37]. For a video i and another randomly selected video h from the same training batch, we first generate two types of counterfactual feature sequences by swapping their respective causal and confounding components:

1. **Confounder-Swapped Counterfactual ($H_{\text{cf_conf}}$):** We combine the causal features from video i with the confounding features from video h: $H_{\text{cf_conf}} = H_{\text{causal_}i} \cup H_{\text{confound_}h}$. This sample retains the original causal information but introduces a new, potentially conflicting confounding context.
2. **Causal-Swapped Counterfactual ($H_{\text{cf_causal}}$):** We combine confounding features from video i with causal

features from video h: $H_{\text{cf_causal}} = H_{\text{confound_}i} \cup H_{\text{causal_}h}$. This sample retains original confounding context but replaces core causal information.

With these generated counterfactuals, we introduce a causal distillation objective, $\mathcal{L}_{\text{CCR}}$, designed to make the model invariant to confounding information while remaining sensitive to causal information. Let $Z(H)$ denote the refined feature output by the model's primary feature encoder. This encoder is identical to the one used in the temporal transformer module, ensuring consistency in the feature space across both causal separation and temporal modeling. We formulate $\mathcal{L}_{\text{CCR}}$ using a triplet-style objective:

$$\begin{aligned} \mathcal{L}_{\text{CCR}} &= \max(0, D_{pos} - D_{neg} + m), \\ \text{where } D_{pos} &= D(Z(H_i^0), Z(H_{\text{cf_conf}})), \\ \text{and } D_{neg} &= D(Z(H_i^0), Z(H_{\text{cf_causal}})). \end{aligned} \tag{4}$$

where D is a distance metric like Mean Squared Error (MSE), and m is a margin hyperparameter. This loss encourages the distance between the representations of the original and the confounder-swapped features to be small, as their causal core is unchanged. Simultaneously, it pushes the distance between the original and the causal-swapped features to be large, as their causal core has been replaced. This directly produces representations that rely on the identified causal features and ignore the confounding ones.

3.3. Time-Conditioned Bidirectional Flow

Design Idea Standard flow-matching techniques in AQA often suffer from unstable representation refinement and temporal misalignment, as they model the process in a purely unidirectional manner [53]. To address this limitation, we draw inspiration from the Schrödinger Bridge problem [30], which provides a bidirectional formulation that simultaneously satisfies boundary conditions at both the start and end points. Building on this principle, BiT-Flow explicitly models both forward and backward temporal dynamics and enforces them to act as approximate inverses through a cycle-consistency constraint. Unlike unidirectional flow models that refine representations by repeatedly pushing them forward, BiT-Flow produces smoother and more coherent representation trajectories by coupling time-aware refinement with bidirectional symmetry. This design stabilizes training, prevents divergence, and yields temporally aligned features that are better suited for capturing fine-grained execution details in AQA.

Implementation The BiT-Flow module is implemented with three components. First, for time conditioning, a normalized time step $t \in [0, 1]$ is embedded using an MLP,

$$\mathbf{e}(t) = \text{MLP}(t), \tag{5}$$

and injected into both the encoder inputs $\mathbf{X}$ and the decoder queries $\mathbf{Q}$ to condition the refinement process on its temporal stage, which can be represented as:

$$\tilde{\mathbf{X}} = \mathbf{X} + \mathbf{e}(t), \tilde{\mathbf{Q}} = \mathbf{Q} + \text{mean}_n\big(\mathbf{e}(t)\big). \tag{6}$$

Second, for bidirectional flow, two direction-specific predictors generate forward and backward flows, which are blended by a schedule $\alpha(t)$:

$$\Delta\mathbf{H}_t^{\text{fwd}} = F_{\text{fwd}}(\tilde{\mathbf{X}}, \tilde{\mathbf{Q}}, t), \Delta\mathbf{H}_t^{\text{bwd}} = F_{\text{bwd}}(\tilde{\mathbf{X}}, \tilde{\mathbf{Q}}, 1-t), \tag{7}$$

$$\Delta\mathbf{H}_t = \alpha(t)\,\Delta\mathbf{H}_t^{\text{fwd}} + (1-\alpha(t))\,\Delta\mathbf{H}_t^{\text{bwd}}. \tag{8}$$

Finally, to ensure forward-backward consistency, we introduce losses that regularize the flow updates. With a prototype anchor $\mathbf{P}$, the flow loss aligns the blended update with the target change:

$$\mathcal{L}_{\text{flow}} = \|\Delta\mathbf{H}_t - (\mathbf{H}_{t+1} - \mathbf{P})\|_2^2, \tag{9}$$

while the consistency loss enforces a cycle:

$$\begin{aligned} \mathcal{L}_{\text{cons}} =& \|\mathbf{P} + \Delta\mathbf{H}_t - \mathbf{H}_{t+1}\|_2^2 \\ &+ \|\mathbf{H}_{t+1} + \Delta\mathbf{H}_t^{\text{bwd}} - \mathbf{P}\|_2^2. \end{aligned} \tag{10}$$

The final BiT-Flow objective combines both terms:

$$\mathcal{L}_{\text{BiT}} = \mathcal{L}_{\text{flow}} + \lambda_{\text{cons}}\,\mathcal{L}_{\text{cons}}. \tag{11}$$

The forward-backward consistency loss $\mathcal{L}_{\text{cons}}$ is the critical innovation of BiT-Flow, acting as a powerful regularizer that forces the two flows to be approximate inverses. This stabilizes training, prevents erratic updates, and produces smooth, temporally coherent representation trajectories. For AQA, where subtle variations in execution determine quality, such enforced smoothness is essential. By coupling time-aware refinement with bidirectional symmetry, BiT-Flow delivers more stable supervision, improved temporal alignment, and more accurate quality assessment.

4. Experiments

This section first introduces the experimental setup and then presents and analyzes the results.

4.1. Experimental Setting

Datasets Our evaluation covers three widely used benchmarks for long-term AQA. The first is the **Rhythmic Gymnastics (RG)** dataset [45], which contains 1,000 clips of athletes performing routines with four apparatus types: Ball, Clubs, Hoop, and Ribbon. Each sequence lasts roughly 1.6 minutes at 25 fps, and the official protocol allocates 200 training and 50 testing samples per discipline. The second benchmark, the **Figure Skating Video (Fis-V)** dataset [23, 24], consists of 500 ladies' singles short programs averaging 2.9 minutes in length. Following the standard split, 400 samples are reserved for training and 100 for

Table 1. Results of SRCC (↑) and R-ℓ_2 (↓) on the RG dataset. Best results are in **bold**, second best are underlined. "*" indicates our reimplementation based on the official code. Average SRCC is computed using Fisher-z. "–": not reported, "+": extra features/modalities.

Method	Publisher	Backbone	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
MS-LSTM [39]	TCSVT'19	VST	RGB	0.621	–	0.661	–	0.670	–	0.695	–	0.663	–
ACTION-NET [45]	ACM MM'20	VST$^+$	RGB	0.684	–	0.737	–	0.733	–	0.754	–	0.728	–
GDLT [38]	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* [48]	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 [44]	TIP'24	VST	RGB	0.636	–	0.720	–	0.769	–	0.708	–	0.711	–
VATP-Net [7]	TCSVT'24	VST	RGB+	0.800	–	0.810	–	0.780	–	0.769	–	0.800	–
CoFInAl [50]	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 [53]	TIP'25	VST	RGB	0.818	2.187	0.803	2.149	0.812	2.119	0.805	2.744	0.810	**2.300**
CaFlow (Ours)	–	VST	RGB	**0.863**	2.101	**0.822**	**2.048**	**0.843**	2.985	**0.833**	2.663	**0.841**	2.449

Table 2. Results of SRCC (↑) and R-ℓ_2 (↓) on the FIS-V dataset. Best results are in **bold** and second best are underlined. "*" indicates our reimplementation based on the official code. Average SRCC uses Fisher-z. "–": not reported, "+": extra features/modalities.

Method	Publisher	Backbone	Modality	TES		PCS		Average	
				SRCC	R-ℓ_2	SRCC	R-ℓ_2	SRCC	R-ℓ_2
MS-LSTM [39]	TCSVT'19	VST	RGB	0.660	–	0.809	–	0.744	–
ACTION-NET [45]	ACM MM'20	VST$^+$	RGB	0.694	–	0.809	–	0.757	–
GDLT [38]	CVPR'22	VST	RGB	0.685	3.717	0.820	2.072	0.761	2.895
HGCN* [48]	TCSVT'23	VST	RGB	0.246	12.628	0.221	20.531	0.234	16.580
MLP-Mixer [36]	AAAI'23	VST	RGB	0.680	–	0.820	–	0.750	–
SGN [6]	TMM'24	VST	RGB	0.700	–	0.830	–	0.765	–
PAMFN [44]	TIP'24	VST	RGB	0.665	–	0.823	–	0.755	–
VATP-Net [7]	TCSVT'24	VST	RGB+	0.702	–	0.863	–	0.796	–
CoFInAl [50]	IJCAI'24	VST	RGB	0.716	2.875	0.843	1.752	0.788	2.314
PHI [53]	TIP'25	VST	RGB	0.726	2.543	**0.867**	1.656	0.804	2.178
CaFlow (Ours)	–	VST	RGB	**0.729**	**2.480**	0.865	**1.619**	**0.805**	**2.050**

Table 3. Results of SRCC (↑) and R-ℓ_2 (↓) on the LOGO dataset. Best results are in **bold**, second best are underlined.

Method	Publisher	Backbone	Modality	SRCC	R-ℓ_2
USDL [29]	CVPR'20	VST	RGB	0.530	4.997
CoRe [43]	ICCV'21	VST	RGB	0.503	5.596
TSA [40]	CVPR'22	VST	RGB	0.570	4.536
CoRe-GOAT [46]	CVPR'23	VST	RGB	0.574	4.437
HGCN [48]	TCSVT'23	VST	RGB	0.475	4.640
CoFInAl [50]	IJCAI'24	VST	RGB	0.698	4.019
PHI [53]	TIP'25	VST	RGB	0.835	2.752
CaFlow (Ours)	–	VST	RGB	**0.856**	**1.425**

evaluation. Each program is annotated with two types of scores, namely the Total Element Score (TES) and the Program Component Score (PCS). In line with previous work [39], we train separate predictors for the two score categories. The third dataset, **LOng-form GrOup (LOGO)** [46], comprises 150 training and 50 test videos of synchronized swimming. With an average duration of 3.5 minutes per sequence, LOGO currently provides the longest video samples among AQA datasets and represents a particularly demanding benchmark for long-term assessment.

Evaluation Metrics To assess the effectiveness of the proposed method, we adopt two kinds of metrics.

In line with prior studies on long-term AQA [38, 45], we employ Spearman's Rank Correlation Coefficient (SRCC), denoted as ρ, which evaluates the monotonic agreement between predictions and ground-truth scores. SRCC is defined as the Pearson correlation between the ranking of predictions $r(\hat{s}_i)$ and the ranking of true labels $r(s_i)$:

$$\rho = \frac{\sum_{i=1}^{N} \big(r(s_i) - \bar{r}\big)\big(r(\hat{s}_i) - \bar{r}\big)}{\sqrt{\sum_{i=1}^{N} \big(r(s_i) - \bar{r}\big)^2} \sqrt{\sum_{i=1}^{N} \big(r(\hat{s}_i) - \bar{r}\big)^2}}, \quad (12)$$

where $\bar{r}$ is the mean rank. A larger ρ implies stronger consistency between predicted and true ranking orders. Following [21], we compute the average SRCC across different types in RG and across TES/PCS scores in Fis-V using Fisher's z-transformation to combine individual results.

Beyond correlation-based evaluation, we also report a stricter error measure, the relative ℓ_2 distance (R-ℓ_2) [43, 48, 50]. This metric captures the normalized discrepancy between predicted and ground-truth scores, which makes the comparison invariant to the absolute score range. Given the

maximum and minimum reference scores $s_{\max}$ and $s_{\min}$, the R-ℓ_2 distance is calculated as

$$\text{R-}\ell_2 = \frac{1}{N}\sum_{n=1}^{N}\left(\frac{|s_n - \hat{s}_n|}{s_{\max} - s_{\min}}\right)^2 \times 100, \quad (13)$$

where s_n and $\hat{s}_n$ denote the ground-truth and predicted scores of the n-th sample, respectively. For datasets with multiple action or score categories, the final performance is aggregated using Fisher's z-value.

Implementation Details All experiments are implemented in PyTorch and conducted on an RTX 4080 GPU. Video frames are sampled at 25 fps and uniformly divided into non-overlapping clips of 32 frames. A Vision Swin Transformer (VST) pre-trained on Kinetics-400 is adopted as the feature backbone, producing 1024-dimensional clip embeddings. During training, the start segment is randomly selected, with the number of clips set to M=68 for RG, M=124 for Fis-V, and M=48 for LOGO, respectively. We use the Adam optimizer with an initial learning rate of 1×10^{-2}, weight decay of 1×10^{-4}, and a batch size of 32. The learning rate is decayed by a factor of 0.1 after 50% and 75% of total epochs. To further optimize the networks, we apply a dropout of 0.3. For hyperparameters, the loss balancing factors λ_1 and λ_2 are set to 0.02 and 0.5, respectively, while the Gumbel-Softmax temperature is annealed from 1.0 during training. At inference time, clip features are aggregated through average pooling, and the final quality score is obtained from the regression head without any test-time augmentation. As shown in Table 1 of the supplementary material, CaFlow introduces additional parameters only in the offline stage while keeping the online stage lightweight and computationally efficient, yet still surpasses all prior long-term AQA methods with a significant improvement in average SRCC.

4.2. Comparisons with State-of-the-Arts

We compare CaFlow against a wide range of state-of-the-art AQA methods on the RG, FIS-V, and LOGO datasets, with results summarized in Tables 1–3. Across all three benchmarks, CaFlow consistently achieves the best or second-best results, demonstrating its effectiveness and robustness.

On the RG dataset (Table 1), CaFlow achieves the highest overall SRCC of 0.838, outperforming the previous best PHI [53] (0.810) by +3.5%. For error, CaFlow obtains an average R-ℓ_2 of 2.455, which is slightly higher than PHI (2.300), but still represents a 39.9% reduction compared to earlier methods such as HGCN (4.341). Looking into the four apparatuses, CaFlow demonstrates consistent advantages. For *Ball*, our method achieves an SRCC of 0.863, which improves upon PHI (0.818) by +5.5%, and reduces the R-ℓ_2 from 3.030 (HGCN) to 2.101, a 30.6% error reduction. For *Clubs*, CaFlow reaches 0.822 in SRCC, surpassing the previous best (0.810 by VATP-Net) by +1.5%, while also delivering the lowest R-ℓ_2 (2.048), reducing error by 40.6% compared to GDLT (2.179). For *Hoop*, CaFlow attains the best SRCC of 0.843, a +3.8% gain over PHI (0.812), though its R-ℓ_2 of 2.985 lags behind GDLT (2.012) and PHI (2.119). This suggests our method is particularly strong in correlation capture, even if some error variance remains. For *Ribbon*, CaFlow achieves 0.833 in SRCC, outperforming PHI (0.805) by +3.5%, while its R-ℓ_2 of 2.663 is close to PHI (2.744).

On the FIS-V dataset (Table 2), CaFlow achieves the best overall SRCC of 0.805, marginally surpassing PHI (0.804). While the correlation gain appears small (+0.1%), CaFlow reduces the average R-ℓ_2 to 2.050, a 5.9% improvement over PHI (2.178) and a 29.1% reduction relative to GDLT (2.895). In terms of event-level performance, CaFlow achieves the best TES score (0.729/2.480), improving R-ℓ_2 by 33.3% over PHI (2.543), while remaining highly competitive on PCS (0.865 vs. 0.867). This demonstrates that our bidirectional flow enhances stability even when correlation scores converge at the top end.

On the LOGO dataset (Table 3), CaFlow sets a new state of the art, achieving an SRCC of 0.856 and R-ℓ_2 of 1.425. Compared to PHI (0.835/2.752), this corresponds to a +2.5% improvement in correlation and a substantial 48.2% reduction in error. These large margins highlight the strength of explicitly combining counterfactual deconfounding with bidirectional temporal refinement, especially in handling long and complex sequences.

Overall, CaFlow delivers consistent and significant improvements across datasets. Even when SRCC margins over the strongest baselines are modest, CaFlow consistently achieves substantial reductions in R-ℓ_2, underscoring its robustness in mitigating spurious correlations and stabilizing temporal dynamics for fine-grained AQA.

4.3. Ablation Study

We evaluate the individual contributions of CCR and BiT-Flow on the RG and LOGO datasets. On RG (Table 4), both modules outperform the baseline. CCR consistently boosts correlations, especially on *Ball* (+3.1%) and *Ribbon* (+4.7%), validating its ability to disentangle causal cues. BiT-Flow alone stabilizes features but yields mixed results, with limited SRCC gains and higher R-ℓ_2. When combined, CaFlow achieves the best average SRCC (0.841) with competitive error (2.449), showing the two modules are complementary. On LOGO (Table 5), CCR reduces error by 30.4% and raises SRCC by +1.7%. BiT-Flow also improves over baseline, while their combination sets a new state-of-the-art with SRCC of 0.856 and R-ℓ_2 of 1.425. We omit ablations on FIS-V since improvements over the strong PHI baseline are small, with gains mainly emerging when both modules are integrated, consistent with RG and LOGO.

Table 4. Ablation on RG: SRCC (↑) and R-ℓ_2 (↓) per apparatus. Average SRCC uses Fisher-z. Best results in **bold**, second best underlined.

Variant	CCR	BiT	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
Backbone + Regressor (Baseline)	✗	✗	0.818	2.187	0.803	2.149	0.812	**2.119**	0.805	2.744	0.810	**2.300**
+ CCR	✓	✗	0.843	2.855	0.821	**1.950**	0.836	3.791	**0.843**	**2.109**	0.836	2.676
+ BiT-Flow	✗	✓	0.833	3.052	0.806	2.271	0.829	5.619	0.827	5.245	0.824	4.047
CCR + BiT-Flow (CaFlow, Ours)	✓	✓	**0.863**	**2.101**	**0.822**	2.048	**0.843**	2.985	0.833	2.663	**0.841**	2.449

Table 5. Ablation on LOGO: SRCC (↑) and R-ℓ_2 (↓). Best results in **bold**, second best underlined.

Variant	CCR	BiT	SRCC	R-ℓ_2
Backbone + Regressor (Baseline)	✗	✗	0.835	2.752
+ CCR	✓	✗	0.849	1.916
+ BiT–Flow	✗	✓	0.845	2.179
CCR + BiT–Flow (CaFlow, Ours)	✓	✓	**0.856**	**1.425**

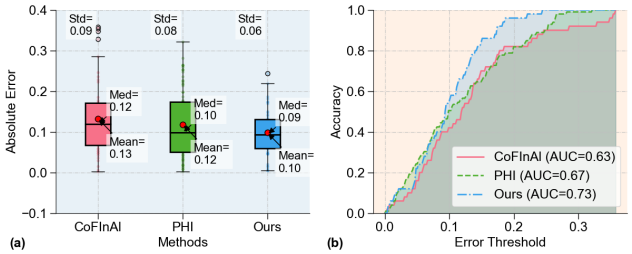


Figure 3. Error analysis on RG. (a) Boxplots of absolute errors with annotated statistics (mean/median/std). (b) Cumulative error-accuracy curves with area under the curve (AUC).

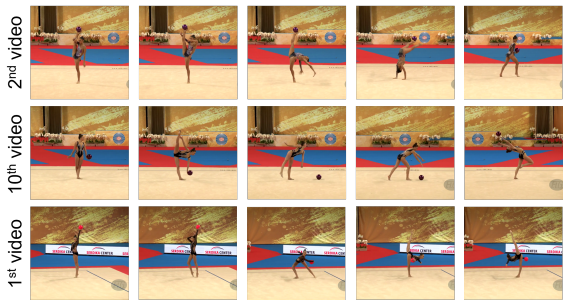


Figure 4. Three representative routines with key frames per case.

4.4. Qualitative Analysis

Fig. 3(a) shows that our method yields the lowest dispersion and bias among the three systems: the mean absolute error drops to 2.44 versus 2.95 for PHI and 3.31 for CoFInAI; the median decreases to 2.34 (PHI: 2.47, CoFInAI: 2.98); and the standard deviation narrows to 1.38 (PHI: 2.01, CoFInAI: 2.23). The cumulative accuracy curves in Figure 3(b) further confirm this trend: CaFlow attains the largest AUC of 0.73, compared to 0.67 for PHI and 0.63 for CoFInAI, with a clear advantage at small error thresholds—precisely the regime needed for reliable judging.

Fig. 4 presents three representative videos highlighting the comparative strengths and weaknesses of the evaluated methods. In the first case (Video 2, GT 16.70), CoFInAI underestimates by 14.90, PHI also underestimates at 15.79, while CaFlow predicts 17.57, closest to the ground truth, showing its ability to capture subtle execution details. In the second case (Video 10, GT 13.65), CoFInAI gives 11.80 and PHI 12.53, both lower than the truth, whereas CaFlow outputs 13.53, nearly identical to the ground truth, demonstrating robustness to confounding context. In the third case (Video 12, GT 16.70), all methods fail: CoFInAI predicts 15.10, PHI 18.07, and CaFlow 19.53, suggesting that highly complex or ambiguous executions remain challenging. Overall, these examples show that CaFlow generally delivers more reliable predictions than prior methods, though extreme cases still leave room for improvement.

5. Conclusion and Discussion

We presented **CaFlow**, a unified framework for Action Quality Assessment that integrates counterfactual deconfounding and bidirectional temporal refinement. Specifically, the *Causal Counterfactual Regularization (CCR)* module disentangles causal from confounding features in a self-supervised manner and enforces robustness through counterfactual interventions, while the *BiT-Flow* module models forward and backward temporal dynamics with cycle consistency, yielding smooth and coherent representation trajectories. Together, these components enable CaFlow to achieve state-of-the-art performance across multiple long-term AQA benchmarks.

Despite these advances, limitations remain. CCR removes spurious correlations without external supervision, yet separation still depends on internal representations, which may bias under extreme distribution shifts. BiT-Flow stabilizes refinement but adds computational overhead compared to lightweight unidirectional methods. Future work should explore more efficient flow architectures, lightweight regularization, and counterfactual reasoning in low-annotation or semi-supervised settings. Overall, CaFlow offers a robust, interpretable foundation for long-term AQA, showing the benefits of combining causal inference with bidirectional temporal modeling while paving the way for more generalizable, efficient frameworks.

References

[1] Michael S Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. *arXiv preprint arXiv:2209.15571*, 2022. 3

[2] Ziyi Chang, Edmund JC Findlay, Haozheng Zhang, and Hubert PH Shum. Unifying human motion synthesis and style transfer with denoising diffusion probabilistic models. *arXiv preprint arXiv:2212.08526*, 2022. 3

[3] Ziyi Chang, George A Koulieris, and Hubert PH Shum. On the design fundamentals of diffusion models: A survey. *arXiv preprint arXiv:2306.04542*, 2023. 3

[4] Swakshar Deb, Md Fokhrul Islam, Shafin Rahman, and Sejuti Rahman. Graph convolutional networks for assessment of physical rehabilitation exercises. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 30:410–419, 2022. 1

[5] Xinpeng Ding, Xiaowei Xu, and Xiaomeng Li. Sedskill: Surgical events driven method for skill assessment from thoracoscopic surgical videos. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 35–45. Springer, 2023. 1

[6] Zexing Du, Di He, Xue Wang, and Qing Wang. Learning semantics-guided representations for scoring figure skating. *IEEE Transactions on Multimedia*, 26:4987–4997, 2024. 6

[7] Kumie Gedamu, Yanli Ji, Yang Yang, Jie Shao, and Heng Tao Shen. Visual-semantic alignment temporal parsing for action quality assessment. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024. 6

[8] Ruisheng Han, Kanglei Zhou, Amir Atapour-Abarghouei, Xiaohui Liang, and Hubert PH Shum. Finecausal: A causal-based framework for interpretable fine-grained action quality assessment. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 6018–6027, 2025. 1, 2, 4

[9] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 3

[10] Christian Keilstrup Ingwersen, Artur Xarles, Albert Clapés, Meysam Madadi, Janus Nørtoft Jensen, Morten Rieger Hannemose, Anders Bjorholm Dahl, and Sergio Escalera. Video-based skill assessment for golf: Estimating golf handicap. In *Proceedings of the 6th International Workshop on Multimedia Content Analysis in Sports*, pages 31–39, 2023. 1

[11] Yanli Ji, Lingfeng Ye, Huili Huang, Lijing Mao, Yang Zhou, and Lingling Gao. Localization-assisted uncertainty score disentanglement network for action quality assessment. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 8590–8597, 2023. 1

[12] Ivan Kobyzev, Simon JD Prince, and Marcus A Brubaker. Normalizing flows: An introduction and review of current methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11):3964–3979, 2020. 3

[13] Mingzhe Li, Hong-Bo Zhang, Qing Lei, Zongwen Fan, Jinghua Liu, and Ji-Xiang Du. Pairwise contrastive learning network for action quality assessment. In *European Conference on Computer Vision*, pages 457–473. Springer, 2022. 1

[14] Ziyang Liang, Yiwei Bao, and Feng Lu. De-confounded gaze estimation. In *European Conference on Computer Vision*, pages 219–235. Springer, 2024. 2

[15] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022. 3

[16] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022. 3

[17] Yang Liu, Guanbin Li, and Liang Lin. Cross-modal causal relational reasoning for event-level visual question answering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):11624–11641, 2023. 2

[18] Yue Ma, Kanglei Zhou, Fuyang Yu, Frederick WB Li, and Xiaohui Liang. Uncertainty-aware probabilistic 3d human motion forecasting via invertible networks. In *IEEE International Conference on Robotics and Automation*, 2025. 3

[19] Kirill Neklyudov, Daniel Severo, and Alireza Makhzani. Action matching: A variational method for learning stochastic dynamics from samples. *arXiv preprint arXiv:2210.06662*, 2022. 3

[20] Derek Onken, Samy Wu Fung, Xingjian Li, and Lars Ruthotto. Ot-flow: Fast and accurate continuous normalizing flows via optimal transport. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9223–9232, 2021. 3

[21] Jia-Hui Pan, Jibin Gao, and Wei-Shi Zheng. Action assessment by joint relation graphs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6331–6340, 2019. 6

[22] German I Parisi, Sven Magg, and Stefan Wermter. Human motion assessment in real time using recurrent self-organization. In *2016 25th IEEE international symposium on robot and human interactive communication (RO-MAN)*, pages 71–76. IEEE, 2016. 2

[23] Paritosh Parmar and Brendan Tran Morris. Learning to score olympic events. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 20–28, 2017. 5

[24] Hamed Pirsiavash, Carl Vondrick, and Antonio Torralba. Assessing the quality of actions. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI 13*, pages 556–571. Springer, 2014. 2, 5

[25] Yongming Rao, Guangyi Chen, Jiwen Lu, and Jie Zhou. Counterfactual attention learning for fine-grained visual categorization and re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1025–1034, 2021. 2

[26] Wenhao Shen, Mingliang Zhou, Yu Chen, Xuekai Wei, Yong Feng, Huayan Pu, and Weijia Jia. Image quality assessment: Investigating causal perceptual effects with abductive counterfactual inference. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 17990–17999, 2025. 2

[27] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 3

[28] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 3

[29] Yansong Tang, Zanlin Ni, Jiahuan Zhou, Danyang Zhang, Jiwen Lu, Ying Wu, and Jie Zhou. Uncertainty-aware score distribution learning for action quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9839–9848, 2020. 6

[30] Alexander Tong, Nikolay Malkin, Kilian Fatras, Lazar Atanackovic, Yanlei Zhang, Guillaume Huguet, Guy Wolf, and Yoshua Bengio. Simulation-free schr\" odinger bridges via score and flow matching. *arXiv preprint arXiv:2307.03672*, 2023. 5

[31] Alexander Tong, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Kilian Fatras, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *arXiv preprint arXiv:2302.00482*, 2023. 3

[32] Loc Trinh, Tim Chu, Zijun Cui, Anand Malpani, Cherine Yang, Istabraq Dalieh, Alvin Hui, Oscar Gomez, Yan Liu, and Andrew 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*, pages 708–717. Springer, 2023. 1

[33] Yin Wang, Zhiying Leng, Frederick WB Li, Shun-Cheng Wu, and Xiaohui Liang. Fg-t2m: Fine-grained text-driven human motion generation via diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22035–22044, 2023. 3

[34] Yin Wang, Mu Li, Jiapeng Liu, Zhiying Leng, Frederick WB Li, Ziyao Zhang, and Xiaohui Liang. Fg-t2m++: Llms-augmented fine-grained text driven human motion generation. *International Journal of Computer Vision*, pages 1–17, 2025. 3

[35] Yushen Wei, Yang Liu, Hong Yan, Guanbin Li, and Liang Lin. Visual causal scene refinement for video question answering. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 377–386, 2023. 2, 4

[36] Jingfei Xia, Mingchen Zhuge, Tiantian Geng, Shun Fan, Yuantai Wei, Zhenyu He, and Feng Zheng. Skating-mixer: Long-term sport audio-visual modeling with mlps. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2901–2909, 2023. 6

[37] Yao Xiao, Ziyi Tang, Pengxu Wei, Cong Liu, and Liang Lin. Masked images are counterfactual samples for robust fine-tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20301–20310, 2023. 2, 4

[38] Angchi Xu, Ling-An Zeng, and Wei-Shi Zheng. Likert scoring with grade decoupling for long-term action assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3232–3241, 2022. 1, 2, 6

[39] Chengming Xu, Yanwei Fu, Bing Zhang, Zitian Chen, Yu-Gang Jiang, and Xiangyang Xue. Learning to score figure skating sport videos. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(12):4578–4590, 2019. 6

[40] Jinglin Xu, Yongming Rao, Xumin Yu, Guangyi Chen, Jie Zhou, and Jiwen 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*, pages 2949–2958, 2022. 6

[41] Jinglin Xu, Sibo Yin, Guohao Zhao, Zishuo Wang, and Yuxin 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*, pages 14628–14637, 2024. 1

[42] Long Yao, Qing Lei, Hongbo Zhang, Jixiang Du, and Shangce Gao. A contrastive learning network for performance metric and assessment of physical rehabilitation exercises. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2023. 1

[43] Xumin Yu, Yongming Rao, Wenliang Zhao, Jiwen Lu, and Jie Zhou. Group-aware contrastive regression for action quality assessment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7919–7928, 2021. 1, 6

[44] Ling-An Zeng and Wei-Shi Zheng. Multimodal action quality assessment. *IEEE Transactions on Image Processing*, 33: 1600–1613, 2024. 6

[45] Ling-An Zeng, Fa-Ting Hong, Wei-Shi Zheng, Qi-Zhi Yu, Wei Zeng, Yao-Wei Wang, and Jian-Huang Lai. Hybrid dynamic-static context-aware attention network for action assessment in long videos. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2526–2534, 2020. 1, 5, 6

[46] Shiyi Zhang, Wenxun Dai, Sujia Wang, Xiangwei Shen, Jiwen Lu, Jie Zhou, and Yansong 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*, pages 2405–2414, 2023. 1, 6

[47] Kanglei Zhou, Ruizhi Cai, Yue Ma, Qingqing Tan, Xinning Wang, Jianguo Li, Hubert PH Shum, Frederick WB Li, Song Jin, and Xiaohui Liang. A video-based augmented reality system for human-in-the-loop muscle strength assessment of juvenile dermatomyositis. *IEEE Transactions on Visualization and Computer Graphics*, 29(5):2456–2466, 2023. 1

[48] Kanglei Zhou, Yue Ma, Hubert PH Shum, and Xiaohui Liang. Hierarchical graph convolutional networks for action quality assessment. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(12):7749–7763, 2023. 1, 2, 6

[49] Kanglei Zhou, Ruizhi Cai, Liyuan Wang, Hubert PH Shum, and Xiaohui Liang. A comprehensive survey of action quality assessment: Method and benchmark. *arXiv preprint arXiv:2412.11149*, 2024. 1

[50] Kanglei Zhou, Junlin Li, Ruizhi Cai, Liyuan Wang, Xingxing Zhang, and Xiaohui Liang. Cofinal: Enhancing action quality assessment with coarse-to-fine instruction alignment. In *Proceedings of the 33rd International Joint Conference on Artificial Intelligence*, 2024. 2, 6

[51] Kanglei Zhou, Liyuan Wang, Xingxing Zhang, Hubert PH Shum, Frederick WB Li, Jianguo Li, and Xiaohui Liang.

Magr: Manifold-aligned graph regularization for continual action quality assessment. In *European Conference on Computer Vision*, pages 375–392, 2024. 1

[52] Kanglei Zhou, Qingyi Pan, Xingxing Zhang, Hubert PH Shum, Frederick WB Li, Xiaohui Liang, and Liyuan Wang. Continual action quality assessment via adaptive manifold-aligned graph regularization. *arXiv preprint arXiv:2510.06842*, 2025. 1

[53] Kanglei Zhou, Hubert PH Shum, Frederick WB Li, Xinxing Zhang, and Xiaohui Liang. Phi: Bridging domain shift in long-term action quality assessment via progressive hierarchical instruction. *IEEE Transactions on Image Processing*, 2025. 1, 2, 5, 6, 7