SPAR: A Sequential Primacy and Attribution Ranking Framework for Skill Determination

Hui Yu[1,2], Xiao Ke[1,2✉], Zhihong Zeng[3✉], Huangbiao Xu[1,2✉], Huanqi Wu[1,2], and Yaru Su[1,2]

[1] Fujian Provincial Key Laboratory of Networking Computing and Intelligent Information Processing, College of Computer and Data Science, Fuzhou University, Fuzhou 350108, China
[2] Engineering Research Center of Big Data Intelligence, Ministry of Education, Fuzhou 350108, China
[3] College of Mathematics and Information Engineering, Longyan University, Longyan 364012, China
kex@fzu.edu.cn, {hui.yu.xiamen.work, huangbiaoxu.chn, wuhuanqi135}@gmail.com, 82009054@lyun.edu.cn

Abstract. Skill determination in video understanding remains challenging due to the intricate temporal dynamics of skill performance and the complex cognitive processes inherently underlying human evaluation. Existing approaches often overlook the progressive evolution of evaluators' cognitive states and the influence of key action stages on overall performance. To address this limitation, we propose **S**equential **P**rimacy and **A**ttribution **R**anking Modeling (**SPAR**). SPAR captures the temporal evolution of cognitive states throughout the evaluation process, enabling iterative refinement of evaluative understanding and producing progressively enhanced representations of overall performance quality. The framework integrates adaptive weighting based on the global temporal context, emphasizing critical segments through localized interaction modeling and multi-scale temporal aggregation. To correct inconsistencies in intermediate recognition signals, SPAR employs a bias-mitigation loss to ensure a faithful and consistent reflection of intrinsic action quality. Extensive experiments on five public datasets validate the effectiveness and generalization capability of the proposed framework.

Keywords: Cognitive Modeling · Cognition Adjust · Temporal Focus

1 Introduction

Skill determination aims to quantitatively evaluate the performance quality of individuals in executing specific tasks through detailed analysis of their action execution processes and dynamics. This capability has enabled a wide range of real-world applications, including the recommendation of expert-level instructional videos for novice learners on online platforms [7,8], rigorous and consistent

✉ Corresponding Author.

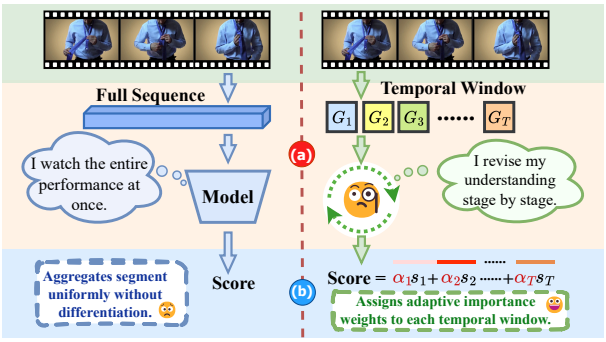

Fig. 1: (a) Previous approaches do not model the progressive formation and evolution of cognitive states during performance evaluation. Stage-aware Cognitive Memory (SaCM) captures this temporal evolution, representing iterative changes in the assessment of performance as the sequence unfolds. (b) Building on this, Temporal Window Weighting (TWW) assigns adaptive importance weights to video segments, emphasizing salient stages that facilitate a more accurate and comprehensive evaluation.

performance evaluation in competitive sports and Olympic events using rigorous and objective performance evaluation metrics [12,30,49,62], and the monitoring of motor function recovery in healthcare rehabilitation through systematic analysis and evaluation frameworks [3,38]. Contemporary methods employ pairwise ranking to construct hierarchical skill structures for task-specific proficiency [24].

However, skill determination introduces additional challenges. Evaluators must consider not only individual actions but also how earlier performance shapes perception of subsequent stages, reflecting the evolution of cognitive states during assessment [22, 37]. For example, in a tie-tying task, the precision of initial folds directly affects the complexity and feasibility of later steps. Participants may adjust subsequent maneuvers to maintain a correct knot, showing that evaluators interpret each step in the context of preceding actions [11]. Perception of later stages is thus influenced by earlier actions, forming a dynamic process characterized by temporal biases in human judgment [52], with early-stage representations persisting across extended sequences [34].

Despite this, most existing methods rely on global feature aggregation [7,25] or static temporal modeling [8, 27], which combine information from all stages or treat each stage independently with fixed importance. Although recent approaches identify salient skill-relevant segments [8], transfer nuanced assessment knowledge across stages or tasks [59], adopt hierarchical grading strategies [61], or enhance representations via semantic alignment [10], multimodal fusion [56], probabilistic embeddings [28], subaction localization [17], and vision-language semantic learning [43, 44, 46], they focus on segment importance, knowledge sharing, or feature refinement. None explicitly model the dynamic evolution of evaluators' cognitive states across the action sequence. Consequently, they fail

to capture stage-wise variations in skill perception or temporal dependencies induced by earlier actions, limiting their ability to reflect human-like evaluation.

While existing methods emphasize feature enrichment, uncertainty modeling, or substage detection, they largely neglect the progressive evolution of evaluators' cognitive states throughout performance observation [1, 36]. As a result, these approaches are limited in capturing temporal variations in perceived skill quality and representing nuanced quantitative differences in proficiency across segments, both essential for modeling skill variations across sequential stages [5].

To address the challenge of modeling the progressive formation and shaping of evaluators' cognitive states in long-term skill determination, we introduce the Stage-aware Cognitive Memory (SaCM) module within the **S**equential **P**rimacy and **A**ttribution **R**anking Modeling (**SPAR**) framework. SaCM partitions each video into a series of temporal segments and constructs a stage-specific cognitive state for each segment that captures the perception of skill quality. As illustrated in Figure 1, each segment is interpreted in the context of the cumulative cognitive state accumulated from preceding segments and prior domain knowledge, reflecting the process by which an evaluator integrates information over time to form an informed perception of current actions. Observed action features are projected into a latent space where prior cognitive context is incorporated and potential perceptual biases are filtered, yielding a refined stage-specific cognitive state. These states are iteratively updated across the sequence, with each stage-specific cognitive state conditioned on the cumulative cognitive state and prior observations, enabling gradual shaping and dynamic adjustment of evaluators' cognitive understanding. Through this iterative refinement, SaCM captures how skill evaluation emerges and evolves across the temporal sequence, providing structured stage-specific and cumulative cognitive states that reflect the progressive formation and adjustment of human evaluative cognition over time.

Nevertheless, accurately quantifying the contribution of each temporal segment remains a core challenge in long-term skill determination, as intricate temporal dependencies and inter-segment interactions can lead to inconsistent attribution and unstable scoring. To address this, we introduce a Temporal Window Weighting component that models the integration of information within a sliding temporal window, enabling evaluation of segment importance from an observer's perspective. Each video is partitioned into multiple temporal windows, within which local interactions between actions are analyzed and weighted according to their relative significance in the global temporal structure, capturing context-sensitive prioritization of critical action stages. Complementing this, a Judgment Bias Mitigation Loss component explicitly quantifies and reduces potential evaluator bias by comparing feature representations of paired sequences, using a bias mask to detect tendencies contrary to expected rankings and ensure faithful reflection of intrinsic action quality. Together, these components provide robust temporal attribution, stabilize rank-based evaluation, and enable fine-grained modeling of segment-level skill variations in long-term video assessments, enhancing both interpretability and methodological rigor of skill determination.

Our main contributions can be summarized as (1) the Stage-aware Cognitive Memory module, which models the progressive formation and refinement of evaluator cognition in long-term skill determination, (2) Temporal Window Weighting, which represents the process of evaluator assessment of segment importance after observing each temporal stage, (3) the Judgment Bias Mitigation Loss, which reduces cognitive tendencies contradicting expected rankings, ensuring reflection of intrinsic action quality, and (4) achieving state-of-the-art on pairwise ranking datasets and showing effectiveness on score prediction datasets.

2 Related Work

Skill Determination. Evaluating the proficiency of action execution is fundamental across multiple application domains. Daily tasks involve subtler skill variations that differ across individuals [9], while competitive sports demand precise coordination and strategic maneuvers [58], and group dance neatness assessment [45]. In surgical contexts, accurate motor control and timing are essential for patient safety [13]. Most existing approaches rely on supervised learning with datasets annotated by experts. One focuses on predicting absolute scores for an individual video using regression or deep learning models [25], whereas another assesses pairwise differences to quantify relative skill disparity between video pairs [2]. To address the scarcity of consistent annotations, ranking-based datasets such as EPIC-Skills [7] and BEST [8] have been widely used to evaluate comparative skill assessment methods. Building on these resources, recent studies have introduced adaptive pairwise comparison frameworks [31,59] that adjust comparison strategies according to observed skill differences, temporal modeling techniques [27] for capturing fine-grained motion dynamics and improving ranking stability, as well as multimodal AQA via mixture-of-experts [48] and LLM-driven reasoning [47]. More recently, large-scale long-term video datasets have emerged to support detailed skill estimation, including the BASKET dataset [32] for fine-grained basketball actions, the LOGO dataset [60] with multi-person videos and rich annotations on formations and action procedures, and CoffeeCraft [53] with diverse skill levels for skill determination.

Temporal Stage-aware Modeling. Long-term skill determination requires modeling fine-grained performance variations across temporal stages. However, many existing methods still rely on global feature aggregation or static temporal modeling, either merging all stages or treating them in isolation [23]. A method employs an attention mechanism to emphasize critical temporal segments [2], whereas another leverages procedure segmentation with contrastive regression to enhance interpretability [50]. Causal reasoning has been introduced to disentangle stage-wise dependencies and yield interpretable scores [14], and a recent study attempts to align stage-level attention or weight-score patterns with human judgment [6]. However, these approaches generally neglect the progressive integration of prior observations. In practice, human evaluators form an evolving understanding of overall skill by anchoring initial impressions [21,35] and continuously integrating previous observations with domain knowledge [11]. Existing

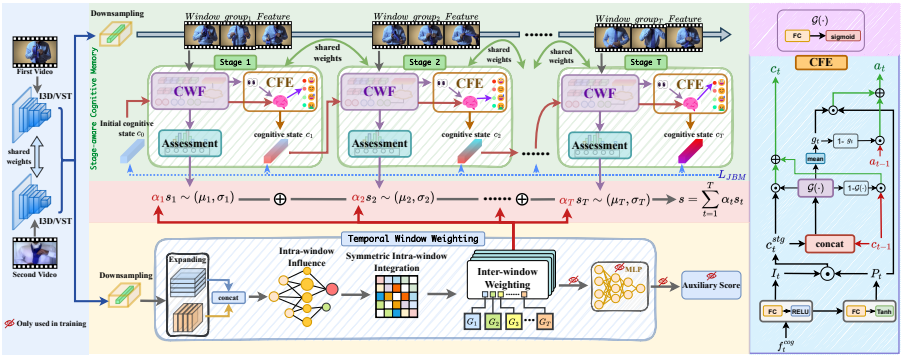

Fig. 2: Overview of the proposed **S**equential **P**rimacy and **A**ttribution **R**anking Modeling (**SPAR**) framework. Given a pair of video features, SPAR first employs the Stage-aware Cognitive Memory (SaCM) to emulate human evaluators who progressively integrate prior observations and domain knowledge when interpreting new segments, while generating stage-specific cognitive states that encode evolving skill understanding. Next, Temporal Window Weighting (TWW) evaluates intra-window feature relationships and calibrates window contributions via global context, highlighting the most decisive stages for overall skill determination. Finally, the Judgment Bias Mitigation (JBM) loss leverages cognition-derived masks to selectively penalize inconsistent stage judgments, mitigating bias accumulation during sequential comparison.

methods rarely capture this evolving cognition. Our SaCM, combined with temporal window weighting, models this progressive reasoning process, highlighting critical segments and enabling fine-grained skill determination.

3 Approach

In this section, we introduce our SPAR framework, as illustrated in Figure 2, and describe its main components and design in detail.

3.1 Problem Formulation

Primary Task: Pairwise Skill Ranking. The core objective of SPAR is to formulate skill determination as a relative ranking problem. Given a pair of videos $(V_i, V_j) \in \mathcal{V}$, where V_i represents a higher skill level than V_j, the goal is to predict a relative quality margin $\Delta\hat{s}$ that quantifies the skill difference:

$$\Delta\hat{s} = \mathcal{M}(V_i, V_j|\Theta), \tag{1}$$

where $\mathcal{M} = \{\mathcal{F}, \mathcal{T}\}$ denotes the overall framework, comprising the feature extractor $\mathcal{F}$ and the SPAR framework $\mathcal{T}$ that processes these features in detail, with Θ representing the learnable parameters. A positive $\Delta\hat{s} > 0$ indicates that V_i has a higher skill quality than V_j, and vice versa.

Extension to Absolute Score Prediction. While our framework is primarily optimized for relative comparisons, it seamlessly generalizes to absolute score estimation through a reference-based inference mechanism [54]. Specifically, for a query video V_q with an unknown score $\hat{s}_q$, we introduce an exemplar video V_e with a known ground-truth label s_e. The absolute score of the query is then derived by anchoring the predicted relative shift against this reference:

$$\Delta\hat{s} = \mathcal{M}(V_q, V_e \mid \Theta), \quad \hat{s}_q = s_e + \Delta\hat{s}. \tag{2}$$

3.2 Stage-aware Cognitive Memory

In long-term skill determination, a central challenge is modeling fine-grained variations across temporal stages while preserving stage-wise dependencies. While stage-specific features capture localized cues, independent modeling or naive aggregation obscures performance progression. We therefore model the progressive reasoning of human evaluators, who infer overall skill by integrating earlier-stage observations with domain knowledge. Based on this principle, we propose the **S**tage-**a**ware **C**ognitive **M**emory (**SaCM**), which maintains a cumulative cognitive state c_t that evolves across stages. At each stage, a stage-specific cognitive state c_t^{stg} is computed from the current segment and used to recursively update the cumulative state, enabling nuanced long-term skill assessment.

SaCM partitions the input video feature sequence $\mathbf{X} \in \mathbb{R}^{L\times D}$ into T non-overlapping temporal windows for each stage, with window size reflecting the temporal granularity at which human evaluators form judgments, and maps each window to a d_{mod}-dimensional latent space via a convolutional transformation:

$$\{\mathcal{W}_t\}_{t=1}^{T} = \text{Conv1D}\big(\{\mathbf{X}_t\}_{t=1}^{T}\big), \quad \mathcal{W}_t \in \mathbb{R}^{\frac{L}{T}\times d_{\text{mod}}}, \tag{3}$$

where L and D denote the total sequence length and original feature dimension, respectively. To incorporate domain knowledge, we introduce the initial cognitive state vector $\mathbf{c}_0 \in \mathbb{R}^{d_{\text{mod}}}$, a learnable vector iteratively refined as each temporal segment is processed. During training, $\mathbf{c}_0$ progressively encodes domain-specific expertise, forming the foundation for stage-wise cognitive states.

SaCM further employs a **C**ognitive **W**indow **F**iltering (**CWF**) module to iteratively refine cumulative cognitive state. A temporal embedding operator $\mathcal{E}_t(\cdot)$, implemented as a depth-wise 1D convolution along the temporal axis, encodes local temporal variations. To capture temporal dependencies and propagate contextual cues from preceding stages, we apply a two-layer convolution operator $\mathcal{R}_{\text{t}\to\text{k}}(\cdot)$ with GELU activation. Within each temporal window, the latent feature $\mathcal{W}_t$ is projected into latent multi-dimensional features $\{\mathcal{W}_{t,k}^{'}\}_{k=1}^{K}$, each combined with its local temporal context to produce the refined window representation:

$$f_{t,k}^{stg} = \text{DyT}\Big(\mathcal{W}_{t,k}^{'} + \mathcal{R}_{\text{t}\to\text{k}}\big(\mathcal{E}_t(\mathcal{W}_t)\big)\Big), \tag{4}$$

where $K = 4$ and DyT [64] stabilizes training in the recursive framework.

The cumulative cognitive state $\mathbf{c}_{t-1}$ encodes information accumulated from all preceding stages, while $\mathbf{c}_0$ provides prior domain knowledge for the first stage

in the absence of observations. Motivated by the human cognitive habit of evaluating while observing, this design models how evaluators interpret the current action segment by integrating prior knowledge $\mathbf{c}_0$ with previous assessments. As subsequent action segments are processed, earlier judgments are iteratively refined, with such refinement captured in the cumulative cognitive state $\mathbf{c}_{t-1}$. Based on this updated state, adaptive modulation parameters are computed to adjust the refined window representation f_t^{stg}:

$$[r_t^{\text{left}}, r_t^{\text{right}}] = \text{MLP}(\mathbf{c}_{t-1}), \quad r_t^{\text{scale}} = \text{sigmoid}(r_t^{\text{left}}), \quad r_t^{\text{shift}} = \tanh(r_t^{\text{right}}). \tag{5}$$

The cognitively modulated feature f_t^{mod} is obtained by applying a learnable linear projection $\mathcal{F}_{\text{mod}}(\cdot)$ to the modulated window representation, thereby integrating prior observations with domain knowledge for current action evaluation:

$$f_t^{\text{mod}} = \mathcal{F}_{\text{mod}}\big(r_t^{\text{scale}} \odot f_t^{\text{stg}} + r_t^{\text{shift}}\big). \tag{6}$$

To adaptively fuse information across the latent multi-dimensional features within each stage, attention weights are computed by combining each dimension with a context-enhanced mapping of the cognitively modulated feature. The weights are scaled by a temperature parameter τ and normalized via softmax:

$$\Omega_{t,k} = \text{Softmax}\left(\frac{\phi_k\Big(W_{t,k}^{'} + \psi_k(f_{t,k}^{mod})\Big) \cdot (f_{t,k}^{mod})^{\top}}{\tau}\right), \tag{7}$$

where $\phi_k(\cdot)$ is a multi-layer perceptron that maps the combined representation to raw attention scores, and $\psi_k(\cdot)$ is a lightweight MLP generating a context-aware mapping. The final stage-wise cognitive representation is obtained by summing the cognitively modulated features weighted by $\Omega_{t,k}$ across all dimensions:

$$f_t^{\text{cog}} = \sum_{k=1}^{K} \Omega_{t,k} \cdot f_{t,k}^{\text{mod}}. \tag{8}$$

Since human observations inherently contain noise, the stage-wise cognitive representation $\mathbf{f}_t^{\text{cog}}$ is probabilistically encoded by estimating its mean $\mu_t(\mathbf{f}_t^{\text{cog}})$ and log-standard deviation via a fully connected layer [57], with the standard deviation $\sigma_t(\mathbf{f}_t^{\text{cog}})$ obtained by exponentiating the log to ensure positivity. The stage-specific output $\hat{s}_t$ is then computed by sampling a variable $\epsilon \sim \mathcal{N}(0, 1)$:

$$\hat{s}_t = \mu_t(\mathbf{f}_t^{\text{cog}}) + \epsilon \cdot \sigma_t(\mathbf{f}_t^{\text{cog}}). \tag{9}$$

To model stage-specific cognition, as illustrated in Figure 2, the **C**ognitive **F**eature **E**ncoder (**CFE**) derives the stage-specific cognitive state c_t^{stg} from f_t^{cog}. Specifically, f_t^{cog} is first transformed through a ReLU-activated linear layer to produce a neutral stage impression I_t, and then processed by a linear layer with Tanh activation to generate a stage-specific approval signal $P_t \in [-1, 1]$, where the sign indicates the direction of evaluation (positive for approval, negative for disapproval) and the magnitude reflects the intensity of the impression. The element-wise product $c_t^{\text{stg}} = P_t \odot I_t$ encodes the signed assessment of the current

stage, with values near zero representing neutral impressions and large positive or negative values indicating strong favorable or unfavorable evaluations. The stage-specific cognitive state is subsequently integrated with the cumulative cognitive state from the previous stage, $\mathbf{c}_{t-1}$, to update the aggregated understanding across stages, forming the updated cumulative cognitive state $\mathbf{c}_t$:

$$c_t = \mathcal{G}[c_t^{stg}, c_{t-1}] \odot c_t^{stg} + (1 - \mathcal{G}[c_t^{stg}, c_{t-1}]) \odot c_{t-1}, \tag{10}$$

where $\mathcal{G}(\cdot)$ is a learnable gate that modulates the relative contributions of the current stage-specific cognitive state c_t^{stg} and the cumulative cognitive state $\mathbf{c}_{t-1}$, implemented as a linear layer with sigmoid activation. The scalar g_t represents the mean activation of $\mathcal{G}(\cdot)$, indicating the overall influence of the current stage. The cumulative approval intensity a_t, which captures the aggregated recognition aligned with the cumulative cognitive state $\mathbf{c}_t$, is updated by combining the current stage-specific approval with the previous cumulative approval. The initial approval a_0 is set to zero, representing a neutral starting point:

$$a_t = g_t \cdot \text{mean}(P_t) + (1 - g_t) \cdot a_{t-1}, \quad a_t \in [-1, 1]. \tag{11}$$

3.3 Temporal Window Weighting

In long videos, not all moments are equally informative. **T**emporal **W**indow **W**eighting (**TWW**) evaluates the importance of feature representations within and across temporal windows $\{\mathcal{W}_t\}_{t=1}^{T}$, highlighting decisive moments most relevant to skill assessment. For intra-window evaluation, stage features are expanded along the second and third dimensions, denoted $\mathcal{W}_t^{\mapsto 2}$ and $\mathcal{W}_t^{\mapsto 3}$, and concatenated to form pairwise inputs for a learnable similarity function:

$$e = \mathcal{F}_{\text{sim}}(\mathcal{W}_t^{\mapsto 3}, \mathcal{W}_t^{\mapsto 2}), \quad e_{\text{sym}} = \frac{1}{2}\left(e + e^{\top}\right), \tag{12}$$

where $\mathcal{F}_{\text{sim}}$ is implemented as two successive MLPs. To estimate the relative significance of each stage within a temporal window, the symmetrized similarity scores e_{sym} are normalized via a softmax function to obtain attention weights, which are then applied to the linearly projected features:

$$\mathcal{W}_t^{\text{intra}} = \sum \text{softmax}(e_{\text{sym}}) \cdot \text{FC}(\mathcal{W}_t)^{\top}, \tag{13}$$

where $\text{FC}(\cdot)$ denotes a linear projection, and summation along the window dimension aggregates the contributions within the temporal window.

For inter-window assessment, aggregated intra-window features $\{\mathcal{W}_t^{\text{intra}}\}_{t=1}^{T}$ are processed with positional encoding and contextual transformations to capture local and cross-window interactions. The relative contribution of each temporal window α_t is then obtained via a softmax over the transformed features:

$$\alpha_t = \text{softmax}\left(\mathcal{F}_{cxt}(\mathcal{R}(\mathcal{W}_t^{\text{intra}}) + \mathcal{W}_t^{\text{intra}})\right), \tag{14}$$

where $\mathcal{R}$ and $\mathcal{F}_{cxt}$ are convolutional and MLP blocks capturing intra-window interactions and contextual dependencies. The inter-window representation is

computed as the weighted sum of window features using these attention weights:

$$\mathcal{W}^{\text{inter}} = \sum_{t=1}^{T} \alpha_t \cdot \mathcal{W}_t^{\text{intra}}. \tag{15}$$

During training, $\mathcal{W}^{\text{inter}}$ is processed by an MLP to produce an auxiliary score $\hat{s}_{\text{aux}}$, providing additional supervision to facilitate learning of window-level discriminative cues. Independently, the importance weights α_t are used to aggregate the stage-specific predictions $\hat{s}_t$ into the overall video score:

$$\hat{s} = \sum_{t=1}^{T} \alpha_t \cdot \hat{s}_t, \quad \hat{s} \sim \mathcal{N}(\sum_{t=1}^{T} \alpha_t \mu_t, \sum_{t=1}^{T} \alpha_t^2 \sigma_t^2), \tag{16}$$

where $\hat{s}_t$ is defined in Eq. (9) with stage-specific mean μ_t and variance σ_t^2, and $\mathcal{N}(\cdot, \cdot)$ denotes a Gaussian distribution.

3.4 Judgment Bias Mitigation Loss

Human evaluative cognition resists abrupt updates to established beliefs in response to new information, as prior observations anchor subsequent judgments [20]. To mitigate systematic deviations in cumulative judgments arising from inconsistencies between stage-specific and cumulative cognitive states, we propose the **J**udgment **B**ias **M**itigation (**JBM**) loss. For a video pair (V_i, V_j) with ground-truth ordering $V_i \succ V_j$ and ranking direction ν, each stage-specific cognitive state c_t^{stg} and cumulative cognitive state c_t is associated with approval signals P_t and a_t, whose signs indicate preference direction and magnitudes indicate confidence. The JBM loss penalizes deviations in cumulative judgments and contrasts features with counterfactuals to maintain consistency between stage-specific evaluations and integration, promoting alignment with desired ranking.

Deviation Loss. To enforce discriminative and consistent cognitive states, we define a deviation loss that combines a discrimination term and a counterfactual deviation term. For a video pair with cognitive states x_i and x_j and their signed approval difference Δ_{sign}, a discrimination penalty is imposed based on their intrinsic similarity to prevent degenerate, indistinguishable states:

$$\mathcal{P}_{\text{disc}} = \tanh(1 + \cos(x_i, x_j)), \tag{17}$$

where x_i and x_j are the cognitive states of two videos, and $\cos(\cdot, \cdot)$ denotes cosine similarity. This term encourages distinguishable cognitive states across videos.

We identify stages where the cognitive approval contradicts the desired ranking ν using an indicator function $\mathbf{1}$, and define a flip indicator $\mathrm{p} = \mathbf{1}\big(\operatorname{sign}(\Delta_{\text{sign}}) \neq \nu\big)$. To correct these inconsistencies, counterfactual references are constructed by flipping the vectors toward the idealized direction. This approach encourages stronger adjustments for larger discrepancies while avoiding overly strict or ineffective interventions that arise from simply zeroing or saturating the vectors.

$$r_i^{\text{cf}} = \begin{cases} |x_i| \cdot \nu, & \text{if } \mathrm{p} = 1, \\ x_i, & \text{otherwise,} \end{cases} \quad r_j^{\text{cf}} = \begin{cases} -|x_j| \cdot \nu, & \text{if } \mathrm{p} = 1, \\ x_j, & \text{otherwise.} \end{cases} \tag{18}$$

The discrepancy between original and counterfactual representations is quantified via cosine similarity, forming the per-stage deviation penalty:

$$\text{sim}_i = \cos(x_i, r_i^{\text{cf}}), \;\; \text{sim}_j = \cos(x_j, r_j^{\text{cf}}), \;\; \mathcal{P}_{\text{dev}} = \tanh(1-\text{sim}_i) + \tanh(1-\text{sim}_j). \quad (19)$$

This term penalizes components that cause deviations in cognitive states, correcting inconsistent evaluations and ensuring higher-skill videos are properly represented relative to lower-skill ones. The overall per-stage deviation loss combines the discrimination and counterfactual deviation penalties:

$$\mathcal{L}_{\text{dev}}(x_i, x_j, \Delta_{\text{sign}}) = \mathcal{P}_{\text{dev}} + \mathcal{P}_{\text{disc}}. \quad (20)$$

Mask Construction. When the cumulative cognitive state at stage t deviates from the expected ranking ν, penalization is applied under two distinct conditions. The first condition corresponds to a simultaneous deviation of the current stage-specific cognitive state, which is indicated by the stage-wise mask $\text{mask}_t^{\text{stg}}$. The second condition corresponds to a concurrent deviation of the cumulative state from the preceding stage, as reflected in the cumulative mask $\text{mask}_t^{\text{cml}}$:

$$\text{mask}_t^{\text{stg}} = \mathbf{1}(\Delta P_t \neq \nu) \odot \mathbf{1}(\Delta a_t \neq \nu), \;\; \text{mask}_t^{\text{cml}} = \mathbf{1}(\Delta a_t \neq \nu) \odot \mathbf{1}(\Delta a_{t-1} \neq \nu), \quad (21)$$

where ΔP_t and Δa_t denote the stage-specific and cumulative cognitive differences at stage t, and Δa_{t-1} denotes the cumulative difference from the preceding stage.
Aggregation. Per-stage deviation penalties for stage-specific and cumulative cognitive states are computed using the loss $\mathcal{L}_{\text{dev}}$:

$$\mathcal{D}_t^{\text{stg}} = \mathcal{L}_{\text{dev}}\big(c_t^{\text{stg}}(V_i), c_t^{\text{stg}}(V_j), \Delta P_t\big), \;\; \mathcal{D}_t^{\text{cml}} = \mathcal{L}_{\text{dev}}\big(c_t(V_i), c_t(V_j), \Delta a_t\big). \quad (22)$$

The overall JBM loss is computed by weighting per-stage deviations with the stage-wise and cumulative masks:

$$\mathcal{L}_{\text{JBM}} = \sum_{t=1}^{T} \Big(\text{mask}_t^{\text{stg}} \odot \mathcal{D}_t^{\text{stg}} + \text{mask}_t^{\text{cml}} \odot \mathcal{D}_t^{\text{cml}}\Big). \quad (23)$$

3.5 Optimization

Pairwise Ranking. We replace the conventional margin ranking loss with a dynamic margin [53] to accommodate varying pairwise discrimination difficulty, where for a score difference Δx and a predefined margin m_0:

$$m = \beta_1 \tanh(\beta_2 \Delta x) + m_0, \quad \mathcal{L}_{\text{DyM}}(\Delta x, m_0) = \max(0,\, m - \Delta x). \quad (24)$$

We compute the ranking and auxiliary losses using $\mathcal{L}_{\text{DyM}}$, where $\Delta\hat{s}$ and $\Delta\hat{s}_{\text{aux}}$ represent the predicted and auxiliary differences for the same video pair:

$$\mathcal{L}_{\text{rank}} = \mathcal{L}_{\text{DyM}}(\Delta\hat{s}, m_1), \quad \mathcal{L}_{\text{aux}} = \mathcal{L}_{\text{DyM}}(\Delta\hat{s}_{\text{aux}}, m_2). \quad (25)$$

For a video pair (V_i, V_j), the predicted score difference $\Delta\hat{s} \sim \mathcal{N}(\Delta\mu, \Delta\sigma^2)$ is derived from Eq. 16 and used to formulate the Mean-Variance Probability (MVP) loss, which models ranking as probabilistic inference over score differences [53]:

$$P(\Delta\hat{s} > m_3) = \Phi\Big(\frac{\Delta\mu - m_3}{\sqrt{\Delta\sigma^2}}\Big), \quad (26)$$

where $\Delta\mu$ and $\Delta\sigma^2$ are the mean and variance of the Gaussian distribution of $\Delta\hat{s}$, Φ is the standard normal cumulative distribution function, and m_3 is a predefined threshold. The MVP loss is then computed directly from the $\Delta\hat{s}$ as:

$$\mathcal{L}_{\text{MVP}} = \lambda_1 \cdot \frac{-\log P(\Delta\hat{s} > m_3)}{\Delta\sigma^2} + \lambda_2 \cdot \log(\lambda_3 + \Delta\sigma^2), \tag{27}$$

where λ_1, λ_2, and λ_3 are learnable parameters. With a weighting factor η controlling the influence of $\mathcal{L}_{\text{JBM}}$, the overall objective function is then defined as:

$$\mathcal{L} = \mathcal{L}_{\text{rank}} + \mathcal{L}_{\text{aux}} + \mathcal{L}_{\text{MVP}} + \eta \cdot \mathcal{L}_{\text{JBM}}. \tag{28}$$

Score Prediction. To adapt our framework to datasets with absolute score annotations, we convert predicted relative score differences into absolute scores using a reference exemplar. Given a query video V_q and a exemplar video V_e with known score s_e, the absolute score is predicted as:

$$\hat{s}_q = s_e + \Delta\hat{s}, \quad \hat{s}_{\text{aux}} = s_e + \Delta\hat{s}_{\text{aux}}, \tag{29}$$

where $\Delta\hat{s}$ and $\Delta\hat{s}_{\text{aux}}$ are the main and auxiliary predicted score differences, and $\hat{s}_q$ and $\hat{s}_{\text{aux}}$ the corresponding predicted scores. The mean squared error (MSE) with the ground-truth score s_q of the query video defines the loss:

$$\mathcal{L}_{\text{s}} = \frac{1}{2}\Big[(\hat{s}_q - s_q)^2 + \zeta(\hat{s}_{\text{aux}} - s_q)^2\Big], \tag{30}$$

where ζ is a weighting factor for the auxiliary score loss. To further encourage discriminative skill ranking, we additionally employ $\mathcal{L}_{\text{DyM}}$. Combined with $\mathcal{L}_{\text{JBM}}$, the overall training objective for absolute score prediction is defined as:

$$\mathcal{L} = \mathcal{L}_{\text{s}} + \mathcal{L}_{\text{DyM}}(\Delta\hat{s}, \Delta\hat{s}_{\text{aux}}, m_4) + \eta \cdot \mathcal{L}_{\text{JBM}}. \tag{31}$$

4 Experiments

4.1 Experiment Settings

Datasets. We conduct experiments on five established long-term action quality assessment benchmarks. The first pairwise ranking dataset, **EPIC-Skills** [7], comprises tasks Surgery (SU), Dough Rolling (DR), Drawing (DW), and Chopsticks Using (CU). The second pairwise ranking dataset, **BEST** [8], includes tasks Apply Eyeliner (AE), Braid Hair (BH), Origami (OR), Scrambled Eggs (SE), and Tie Tie (TT). To extend our approach to score prediction, we evaluate it on **Fis-V** [42], **Rhythmic Gymnastics (RG)** [55], and **LOGO** [60].

Evaluation Metrics. In line with previous studies [7, 8, 31, 59], for pairwise ranking tasks we adopt accuracy (Acc) as the primary evaluation metric, with N_T and N_F denoting the numbers of correctly and incorrectly ranked pairs:

$$\text{Acc} = \frac{N_T}{N_T + N_F} \times 100\%. \tag{32}$$

Table 1: Performance on the EPIC-Skills dataset. **Bold** and underline indicate the best and second-best accuracies (%).

Method	Year	SU	DR	DW	CU	Avg. Acc
RankSVM [18]	2006	65.20	72.00	71.50	76.60	71.33
AlexNet+C3D [51]	2016	66.10	78.10	72.00	70.30	71.63
Who's Better [7]	2018	70.20	79.40	83.20	71.50	76.08
RAAN [8]	2019	68.50	86.90	82.30	84.70	80.60
AAS [31]	2022	71.89	88.47	88.15	87.73	84.06
ASAAS [59]	2024	**78.91**	89.64	91.96	92.78	88.32
T^2CR [19]	2024	75.82	87.82	89.58	87.56	85.20
Rhythmer [27]	2025	73.54	90.94	91.63	**97.68**	88.45
SPAR(Ours)	-	76.69	**91.42**	**94.19**	94.26	**89.14**

Table 2: Performance on the BEST dataset. **Bold** and underline indicate the best and second-best accuracies (%).

Method	Year	AE	BH	OR	SE	TT	Avg. Acc
STPN [29]	2018	-	-	-	-	-	70.00
Who's Better [7]	2018	-	-	-	-	-	75.80
RAAN [8]	2019	85.50	76.50	68.90	87.70	87.60	81.24
AAS [31]	2022	85.53	75.21	82.55	83.77	89.27	83.27
ASAAS [59]	2024	85.53	76.50	85.38	89.61	89.70	85.34
T^2CR [19]	2024	84.26	79.91	82.08	87.01	87.98	84.25
SPAR(Ours)	-	**86.38**	**82.91**	**87.74**	**92.21**	**93.56**	**88.56**

Consistent with prior studies [19, 25, 39], score prediction is evaluated using Spearman's Rank Correlation (ρ) and Relative L2-distance (R-ℓ_2), where ρ compares the predicted ranking series $\hat{q}$ with the ground truth q, and R-ℓ_2 measures the normalized difference between predicted scores $\hat{s}$ and ground-truth scores s:

$$\rho = \frac{\sum_i (q_i - \bar{q})(\hat{q}_i - \bar{\hat{q}})}{\sqrt{\sum_i (q_i - \bar{q})^2 \sum_i (\hat{q}_i - \bar{\hat{q}})^2}}, \quad R\text{-}\ell_2 = \frac{1}{N}\sum_{n=1}^{N} \left(\frac{|\hat{s}_n - s_n|}{s_{\max} - s_{\min}}\right)^2 \times 100. \tag{33}$$

Implementation Details. We implemented SPAR in PyTorch on a Tesla K80 GPU. To ensure fair comparison with prior works [25,31,59], we use the officially released pre-extracted features. Specifically, EPIC-Skills and BEST use I3D [4] features extracted by [7, 8], whereas LOGO, Rhythmic Gymnastics (RG), and Fis-V extract features using VST [26]. Each video includes 400 clips for pairwise ranking and 48 for score prediction. To emulate the progressive comprehension process of human evaluators, SPAR processes these clips in fixed-length temporal windows, with a window size of 16 for ranking and 2 for scoring, which ensures a comparable number of cognitive updates and maintains consistent reasoning across tasks. The model is trained with a batch size of 64 using the AdamW optimizer, with learning rates set to 1×10^{-4} for ranking and 5×10^{-4} for score prediction, and a weight decay of 1×10^{-5}. In all experiments, the loss parameters are initialized with the same values across datasets and are learnable during training: $\beta_1 = 1.0$, $\beta_2 = 0.5$, $m_1 = 1.55$, $m_2 = 1.5$ in Eq. 24; $\lambda_1 = \lambda_2 = 1.0$, $\lambda_3 = 0.5$, $m_3 = 1.05$ in Eq. 27; $\zeta = 0.6$ in Eq. 30; and $\eta = 0.2$, $m_4 = 1 \times 10^{-4}$ in Eq. 31. Further details and model complexity are provided in the **Appendix**.

4.2 Results and Analysis

Results on EPIC-Skills. Table 1 reports the empirical performance of SPAR on the EPIC-Skills dataset, evaluated using four-fold cross-validation. SPAR achieves the highest average accuracy of 89.14%, surpassing the previous best method by 0.7%. It attains the highest accuracy in most individual action tasks, demonstrating consistent and reliable performance across the entire dataset.

Results on BEST. Table 2 presents the performance of SPAR on the BEST dataset. The proposed method achieves an average accuracy of 88.56%, improving over the second-best approach by 3.2% and achieving notable gains in more

Table 3: Performance comparison on the RG and Fis-V datasets. **Bold** denotes the best, underline the second best.

Method	Year	RG (ρ ↑)					Fis-V (ρ ↑)		
		Ball	Clubs	Hoop	Ribbon	Avg.	TES	PCS	Avg.
C3D+SVR [33]	2017	0.375	0.551	0.495	0.516	0.483	0.400	0.590	0.501
MS-LSTM [42]	2019	0.621	0.661	0.670	0.695	0.663	0.660	0.809	0.744
ACTION-NET [55]	2020	0.684	0.737	0.733	0.754	0.728	0.694	0.809	0.757
GDLT [41]	2022	0.746	0.802	0.765	0.741	0.765	0.685	0.820	0.761
Skating-Mixer [40]	2023	–	–	–	–	–	0.680	0.820	0.759
LUSD-NET [17]	2023	–	–	–	–	–	0.679	0.823	0.760
SGN [10]	2023	–	–	–	–	–	0.700	0.830	0.773
PAMFN (only RGB) [56]	2024	0.636	0.720	0.769	0.708	0.711	0.665	0.823	0.755
ASGTN [25]	2025	0.792	0.825	0.784	0.793	0.799	0.703	0.845	0.784
MLAVL [44]	2025	0.826	**0.829**	**0.871**	**0.866**	**0.849**	**0.766**	0.863	**0.823**
PHI [63]	2025	0.818	0.803	0.812	0.805	0.810	0.726	**0.867**	0.804
CaFlow [15]	2025	**0.863**	0.822	0.843	0.833	0.841	0.729	0.865	0.805
SPAR (ours)	–	0.788	0.811	0.827	0.836	0.816	0.748	0.855	0.802

Table 4: Performance on the LOGO dataset.

Method	Year	ρ (↑)	R-ℓ_2 (↓)
MS-LSTM [42]	2019	0.542	5.763
USDL [39]	2020	0.473	5.076
CoRe [54]	2021	0.697	5.620
GDLT [41]	2022	0.647	4.148
TSA [49]	2022	0.475	4.778
TPT [2]	2022	0.589	5.228
HGCN [62]	2023	0.541	4.765
T^2CR [19]	2024	0.681	5.973
CoFInAl [61]	2024	0.661	5.754
QTD [6]	2024	0.698	4.948
ASGTN [25]	2025	0.704	5.695
MLAVL [44]	2025	0.804	2.269
PHI [63]	2025	0.835	2.752
CaFlow [15]	2025	**0.856**	**1.425**
SPAR (Ours)	-	0.749	3.743

Table 5: Ablation of key components on the BEST and LOGO datasets. The upper part sequentially adds components, while the lower part removes each component individually. Red / Blue indicates performance increase / decrease. * denotes LSTM [16], and † indicates that SaCM is replaced by the Attention Module [8].

Settings	BEST	LOGO	
	Avg. Acc (↑)	ρ (↑)	R-ℓ_2 (↓)
Reference*	72.25	0.474	21.146
+ SaCM	78.37↑8%	0.604↑27%	11.479↓46%
+ TWW (w/o JBM)	83.49↑7%	0.652↑8%	5.969↓48%
+ JBM (ours)	88.56↑6%	0.749↑15%	3.743↓37%
w/o SaCM† & w/o JBM	75.92↓14%	0.544↓27%	14.049↑275%
w/o TWW	80.38↓9%	0.637↓15%	9.668↑158%

challenging tasks, with improvements of 6.4% and 3.9% in specific cases. These results demonstrate that SPAR integrates information from prior segments and highlights critical stages, enabling a progressive and structured understanding of skill execution for accurate skill determination.

Results on absolute score prediction datasets. Table 3 and Table 4 present the overall performance of SPAR on three absolute score prediction benchmarks: RG, Fis-V, and LOGO. On LOGO, SPAR attains a Spearman's rank correlation ρ of 0.749 and an R-ℓ_2 of 3.743. Relative to QTD and ASGTN, SPAR improves $\rho(\times 100)$ by 5.1 and 4.5 points, respectively, while reducing R-ℓ_2 by 1.205 and 1.952. On RG and Fis-V, SPAR achieves average scores of 0.816 and 0.802, indicating consistently competitive performance across both benchmarks. Overall, these results suggest that SPAR generalizes effectively to absolute score prediction and represents a reliable and competitive approach within this domain.

4.3 Ablation Studies

A comprehensive ablation study is conducted to evaluate the contribution of each component within SPAR. As shown in Table 5, replacing the LSTM-based

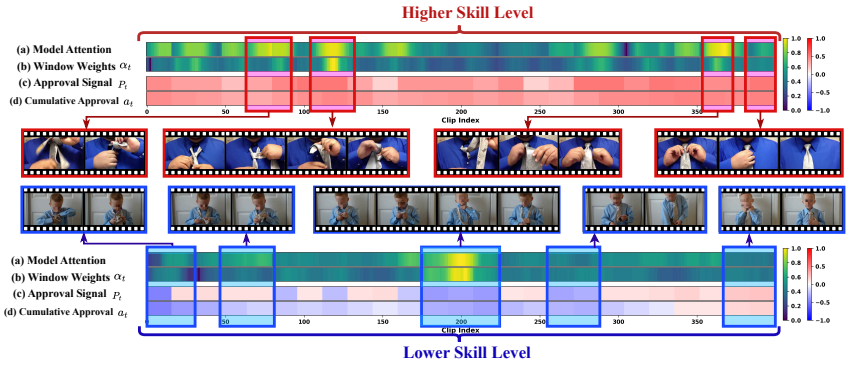

Fig. 3: Visualization of SPAR's evaluation process. The top row shows a high-skill video and the bottom row shows a low-skill video. (a) SPAR attention, (b) temporal window weights α_t, (c) stage-specific cognitive approval P_t, and (d) cumulative cognitive approval a_t. Red indicates positive approval and blue indicates negative approval.

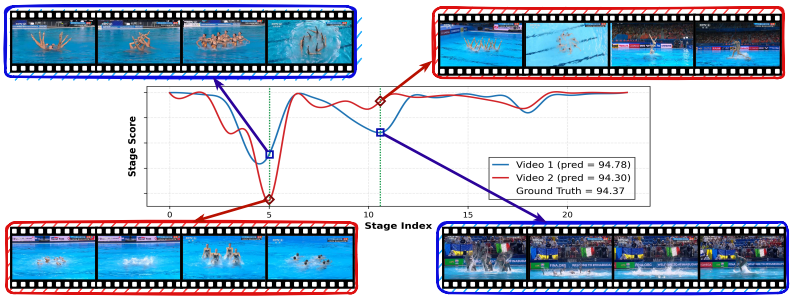

Fig. 4: Stage-wise skill evolution for two videos (V_1 and V_2) with identical ground-truth scores. The blue and red curves show their predicted stage-wise scores. Although both videos receive similar overall scores, V_1 performs better in early stages, while V_2 improves in later stages. Sampled frames illustrate these temporal variations.

reference [16] with SaCM yields substantial improvements across all metrics, demonstrating its effectiveness in encoding stage-specific information and integrating segment-level cues into evaluative representations. Incorporating TWW further enhances performance by adaptively emphasizing salient segments. Introducing JBM achieves the best performance by mitigating cumulative judgment drift and aligning local and global evaluations, reaching 88.56% accuracy on BEST and improving results on LOGO with ρ ($\times 100$) of 74.9 and R-ℓ_2 of 3.743. Ablating any component consistently degrades performance, confirming their complementary roles in long-term skill assessment.

Additionally, to further validate the effectiveness of the proposed SaCM design, an ASGTN-style [25] variant is constructed by replacing the original CWF filtering while retaining CFE, TWW, and JBM. On the BEST dataset, it achieves

accuracies of 80.85%, 72.65%, 75.94%, 85.71%, and 88.41% on AE, BH, OR, SE, and TT, respectively, with an average accuracy of 80.71%. These results are consistently lower than those of SPAR, indicating the effectiveness of the proposed SaCM design within the overall framework.

4.4 Visualization

Figure 3 reveals a consistent correspondence between temporal importance, stage-wise evaluation signals, and cumulative scoring dynamics. Regions corresponding to key actions exhibit consistently higher temporal window weights α_t, while less informative regions show lower responses, indicating a clear alignment between action relevance and temporal weighting. Meanwhile, stage-specific approval P_t exhibits both positive and negative variations over time, and these local evaluations are progressively integrated into a smoother cumulative trajectory a_t that captures the temporal evolution of the overall assessment process. Across videos, high-skill cases exhibit more concentrated positive evaluations on key temporal segments, whereas low-skill cases show weaker positive responses and more dispersed evaluation signals across time.

This capability is further illustrated by the stage-wise predictions in Figure 4, which show fine-grained temporal score trajectories rather than a single global estimate. Although videos with identical ground-truth scores yield similar overall predictions, their intermediate stage-wise scores differ across temporal segments. In particular, poorly performed segments correspond to lower predicted scores, while stronger execution stages receive higher scores, indicating that the model reflects variations in local skill quality over time. Overall, SPAR captures discriminative temporal scoring patterns beyond global score equivalence.

5 Conclusion

This work presents the Sequential Primacy and Attribution Ranking (SPAR) framework for long-term skill determination. SPAR models the progressive evolution of evaluative cognition, capturing how human assessors integrate prior observations and update judgments over time. By combining a stage-aware cognitive memory, temporal window weighting, and a judgment bias mitigation component, the framework enables coherent temporal reasoning, fine-grained segment attribution, and interpretable evaluation dynamics. Experiments on multiple benchmarks demonstrate its effectiveness and robustness in modeling skill progression and extending pairwise ranking to absolute score prediction.

Acknowledgments

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

References

1. Bae, G.Y., Chen, K.W.: Serial bias and response time: Prior stimulus not only biases but also modulates the speed of decision for a new stimulus. Journal of Vision **25**(2), 9–9 (2025)
2. Bai, Y., Zhou, D., Zhang, S., Wang, J., Ding, E., Guan, Y., Long, Y., Wang, J.: Action quality assessment with temporal parsing transformer. In: European conference on computer vision. pp. 422–438. Springer (2022)
3. Burggraaff, J., Dorn, J., D'Souza, M., Morrison, C., Kamm, C.P., Kontschieder, P., Tewarie, P., Steinheimer, S., Sellen, A., Dahlke, F., et al.: Video-based pairwise comparison: enabling the development of automated rating of motor dysfunction in multiple sclerosis. Archives of physical medicine and rehabilitation **101**(2), 234–241 (2020)
4. Carreira, J., Zisserman, A.: 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)
5. Dekel, R., Sagi, D.: Perceptual bias is reduced with longer reaction times during visual discrimination. Communications Biology **3**(1), 59 (2020)
6. Dong, X., Liu, X., Li, W., Adeyemi-Ejeye, A., Gilbert, A.: Interpretable long-term action quality assessment. In: 35th British Machine Vision Conference (2024)
7. Doughty, H., Damen, D., Mayol-Cuevas, W.: Who's better? who's best? pairwise deep ranking for skill determination. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 6057–6066 (2018)
8. Doughty, H., Mayol-Cuevas, W., Damen, D.: The pros and cons: Rank-aware temporal attention for skill determination in long videos. In: Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition. pp. 7862–7871 (2019)
9. Doughty, H.R.: Skill determination from long videos. Ph.D. thesis, University of Bristol (2021)
10. Du, Z., He, D., Wang, X., Wang, Q.: Learning semantics-guided representations for scoring figure skating. IEEE Transactions on Multimedia **26**, 4987–4997 (2023)
11. Feigin, H., Baror, S., Bar, M., Zaidel, A.: Perceptual decisions are biased toward relevant prior choices. Scientific reports **11**(1), 648 (2021)
12. Gao, J., Zheng, W.S., Pan, J.H., Gao, C., Wang, Y., Zeng, W., Lai, J.: An asymmetric modeling for action assessment. In: European Conference on Computer Vision. pp. 222–238. Springer (2020)
13. Gattupalli, S., Ebert, D., Papakostas, M., Makedon, F., Athitsos, V.: Cognilearn: A deep learning-based interface for cognitive behavior assessment. In: Proceedings of the 22nd International Conference on Intelligent User Interfaces. pp. 577–587 (2017)
14. Han, R., Zhou, K., Atapour-Abarghouei, A., Liang, X., Shum, H.P.: Finecausal: A causal-based framework for interpretable fine-grained action quality assessment. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 6018–6027 (2025)
15. Han, R., Zhou, K., Chen, S., Atapour-Abarghouei, A., Shum, H.P.: Caflow: Enhancing long-term action quality assessment with causal counterfactual flow. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 8231–8241 (2026)
16. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)

17. Ji, Y., Ye, L., Huang, H., Mao, L., Zhou, Y., Gao, L.: Localization-assisted uncertainty score disentanglement network for action quality assessment. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 8590–8597 (2023)
18. Joachims, T.: Training linear svms in linear time. In: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 217–226 (2006)
19. Ke, X., Xu, H., Lin, X., Guo, W.: Two-path target-aware contrastive regression for action quality assessment. Information Sciences **664**, 120347 (2024)
20. Kim, M., Park, B., Young, L.: The psychology of motivated versus rational impression updating. Trends in Cognitive Sciences **24**(2), 101–111 (2020)
21. Krieglstein, F., Beege, M., Wesenberg, L., Rey, G.D., Schneider, S.: The distorting influence of primacy effects on reporting cognitive load in learning materials of varying complexity. Educational Psychology Review **37**(1), 2 (2025)
22. Lange, R.D., Chattoraj, A., Beck, J.M., Yates, J.L., Haefner, R.M.: A confirmation bias in perceptual decision-making due to hierarchical approximate inference. PLOS Computational Biology **17**(11), e1009517 (2021)
23. Lei, Q., Zhang, H., Du, J.: Temporal attention learning for action quality assessment in sports video. Signal, Image and Video Processing **15**(7), 1575–1583 (2021)
24. Li, M., Zhang, H.B., Lei, Q., Fan, Z., Liu, J., Du, J.X.: Pairwise contrastive learning network for action quality assessment. In: European Conference on Computer Vision. pp. 457–473. Springer (2022)
25. Liu, J., Wang, H., Zhou, W., Stawarz, K., Corcoran, P., Chen, Y., Liu, H.: Adaptive spatiotemporal graph transformer network for action quality assessment. IEEE Transactions on Circuits and Systems for Video Technology (2025)
26. Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., Hu, H.: Video swin transformer. In: Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition. pp. 3202–3211 (2022)
27. Luo, Z., Xiao, Y., Yang, F., Zhou, J.T., Fang, Z.: Rhythmer: Ranking-based skill assessment with rhythm-aware transformer. IEEE Transactions on Circuits and Systems for Video Technology **35**(1), 259–272 (2025)
28. Majeedi, A., Gajjala, V.R., GNVV, S.S.S.N., Li, Y.: Rica^ 2: Rubric-informed, calibrated assessment of actions. In: European Conference on Computer Vision. pp. 143–161 (2024)
29. Nguyen, P., Liu, T., Prasad, G., Han, B.: Weakly supervised action localization by sparse temporal pooling network. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 6752–6761 (2018)
30. Pan, J.H., Gao, J., Zheng, W.S.: Action assessment by joint relation graphs. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6331–6340 (2019)
31. Pan, J., Gao, J., Zheng, W.: Adaptive action assessment. IEEE Trans. Pattern Anal. Mach. Intell. **44**(12), 8779–8795 (2022)
32. Pan, Y., Zhang, C., Bertasius, G.: Basket: A large-scale video dataset for fine-grained skill estimation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 28952–28962 (2025)
33. Parmar, P., Tran Morris, B.: Learning to score olympic events. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 20–28 (2017)
34. Rottman, B.M., Zhang, Y.: Learning about causal relations that change over time: primacy and recency over long timeframes in causal judgments and memory. Cognitive Research: Principles and Implications **10**(1), 9 (2025)

35. Rubínová, E., Price, H.L.: Primacy (and recency) effects in delayed recognition of items from instances of repeated events. Memory **32**(5), 627–645 (2024)
36. Shan, J., Hajonides, J.E., Myers, N.E.: Attractive serial dependence arises during decision-making. Plos Biology **23**(8), e3003333 (2025)
37. Steiner, D.D., Rain, J.S.: Immediate and delayed primacy and recency effects in performance evaluation. Journal of Applied Psychology **74**(1), 136 (1989)
38. Steinheimer, S., Dorn, J.F., Morrison, C., Sarkar, A., D'Souza, M., Boisvert, J., Bedi, R., Burggraaff, J., Kontschieder, P., Dahlke, F., et al.: Setwise comparison: efficient fine-grained rating of movement videos using algorithmic support–a proof of concept study. Disability and rehabilitation **42**(18), 2640–2646 (2020)
39. Tang, Y., Ni, Z., Zhou, J., Zhang, D., Lu, J., Wu, Y., Zhou, J.: 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)
40. Xia, J., Zhuge, M., Geng, T., Fan, S., Wei, Y., He, Z., Zheng, F.: Skating-mixer: Long-term sport audio-visual modeling with mlps. In: Proceedings of the AAAI conference on artificial intelligence. vol. 37, pp. 2901–2909 (2023)
41. Xu, A., Zeng, L.A., Zheng, W.S.: 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)
42. Xu, C., Fu, Y., Zhang, B., Chen, Z., Jiang, Y.G., Xue, X.: Learning to score figure skating sport videos. IEEE transactions on circuits and systems for video technology **30**(12), 4578–4590 (2019)
43. Xu, H., Ke, X., Li, Y., Xu, R., Wu, H., Lin, X., Guo, W.: Vision-language action knowledge learning for semantic-aware action quality assessment. In: European Conference on Computer Vision. pp. 423–440 (2024)
44. Xu, H., Ke, X., Wu, H., Xu, R., Li, Y., Guo, W.: Language-guided audio-visual learning for long-term sports assessment. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 23967–23977 (2025)
45. Xu, H., Ke, X., Wu, H., Xu, R., Li, Y., Xu, P., Guo, W.: Dancefix: An exploration in group dance neatness assessment through fixing abnormal challenges of human pose. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 8869–8877 (2025)
46. Xu, H., Wu, H., Ke, X., Li, Y., Xu, R., Guo, W.: Quality-guided vision-language learning for long-term action quality assessment. IEEE Transactions on Multimedia **27**, 7326–7339 (2025)
47. Xu, H., Wu, H., Ke, X., Peng, Y.: Limssr: Llm-driven sequence-to-score reasoning under training-time incomplete multimodal observations. arXiv preprint arXiv:2605.00434 (2026)
48. Xu, H., Wu, H., Ke, X., Wu, J., Xu, R., Xu, J.: Mcmoe: Completing missing modalities with mixture of experts for incomplete multimodal action quality assessment. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 11241–11249 (2026)
49. Xu, J., Rao, Y., Yu, X., Chen, G., Zhou, J., Lu, J.: 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)
50. Xu, J., Rao, Y., Zhou, J., Lu, J.: Procedure-aware action quality assessment: Datasets and performance evaluation. International Journal of Computer Vision **132**(12), 6069–6090 (2024)

51. Yao, T., Mei, T., Rui, Y.: Highlight detection with pairwise deep ranking for first-person video summarization. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 982–990 (2016)
52. Yoo, M., Bahg, G., Turner, B., Krajbich, I.: People display consistent recency and primacy effects in behavior and neural activity across perceptual and value-based judgments. Cognitive, Affective, & Behavioral Neuroscience pp. 1–18 (2025)
53. Yu, H., Ke, X., Zeng, Z., Xu, H., Wu, H.: Harp: Hierarchical adaptive ranking with probabilistic modeling for skill determination. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings. pp. 8337–8346 (2026)
54. Yu, X., Rao, Y., Zhao, W., Lu, J., Zhou, J.: Group-aware contrastive regression for action quality assessment. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 7919–7928 (2021)
55. Zeng, L.A., Hong, F.T., Zheng, W.S., Yu, Q.Z., Zeng, W., Wang, Y.W., Lai, J.H.: Hybrid dynamic-static context-aware attention network for action assessment in long videos. In: Proceedings of the 28th ACM international conference on multimedia. pp. 2526–2534 (2020)
56. Zeng, L.A., Zheng, W.S.: Multimodal action quality assessment. IEEE Transactions on Image Processing **33**, 1600–1613 (2024)
57. Zhang, B., Chen, J., Xu, Y., Zhang, H., Yang, X., Geng, X.: Auto-encoding score distribution regression for action quality assessment. Neural Computing and Applications **36**(2), 929–942 (2024)
58. Zhang, S.J., Pan, J.H., Gao, J., Zheng, W.S.: Semi-supervised action quality assessment with self-supervised segment feature recovery. IEEE Transactions on Circuits and Systems for Video Technology **32**(9), 6017–6028 (2022)
59. Zhang, S.J., Pan, J.H., Gao, J., Zheng, W.S.: Adaptive stage-aware assessment skill transfer for skill determination. IEEE Transactions on Multimedia **26**, 4061–4072 (2024)
60. Zhang, S., Dai, W., Wang, S., Shen, X., Lu, J., Zhou, J., Tang, Y.: Logo: A long-form video dataset for group action quality assessment. In: Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition. pp. 2405–2414 (2023)
61. Zhou, K., Li, J., Cai, R., Wang, L., Zhang, X., Liang, X.: Cofinal: Enhancing action quality assessment with coarse-to-fine instruction alignment. In: Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence. pp. 1771–1779 (2024)
62. Zhou, K., Ma, Y., Shum, H.P., Liang, X.: Hierarchical graph convolutional networks for action quality assessment. IEEE Transactions on Circuits and Systems for Video Technology **33**(12), 7749–7763 (2023)
63. Zhou, K., Shum, H.P., Li, F.W., Zhang, X., Liang, X.: Phi: Bridging domain shift in long-term action quality assessment via progressive hierarchical instruction. IEEE transactions on image processing **34**, 3718–3732 (2025)
64. Zhu, J., Chen, X., He, K., LeCun, Y., Liu, Z.: Transformers without normalization. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 14901–14911 (2025)